

# On the Information in Deep Biometric Templates: from Vulnerability of Unprotected Templates to Leakage in Protected Templates

Présentée le 30 septembre 2024

Faculté des sciences et techniques de l'ingénieur  
Laboratoire de traitement des signaux 5  
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

**Hatef OTROSHI SHAHREZA**

Acceptée sur proposition du jury

Prof. P. Frossard, président du jury  
Prof. J.-Ph. Thiran, Dr S. Marcel, directeurs de thèse  
Prof. A. Ross, rapporteur  
Prof. V. M. Patel, rapporteur  
Prof. F. Tramèr, rapporteur





*A journey of a thousand miles  
begins with a single step.*  
— Lao Tzu, Chinese Sage

To my wife...



# Acknowledgements

Before delving into the contents of the thesis, I would like to take a moment to express my heartfelt appreciation to everyone who has been a part of my academic and research journey and has contributed to the completion of this thesis.

First of all, I am profoundly grateful to **Sébastien Marcel** for the opportunity to pursue my PhD with him and his unwavering trust and belief in me from the beginning of this journey. Thank you for all the guidance, encouragement, and endless support throughout the years. I am also grateful to you **Jean-Philippe Thiran** for your generous unconditional support and encouragement throughout my PhD. I would like to also thank my thesis committee, **Arun Ross**, **Vishal Patel**, and **Florian Tramèr** for kindly accepting to review the output of my four years of research and for their insightful comments. I would like to also thank **Christophe Busch** for serving in my candidacy exam and for his invaluable comments in the beginning stage of my PhD. I extend my gratitude to **Pascal Frossard** for generously serving in both, my candidacy exam and thesis defence, as the president of the jury. I would like to also thank **Jean-Michel Sallese**, my EPFL mentor, for his guidance and support throughout my PhD.

Looking back on my PhD journey, I am proud of the chances that I had to meet and collaborate with fantastic people from different groups. A special thank you goes to **Yanina Shkel** for her unconditional and generous time to discuss different topics from the Information Theory point of view. Thank you for all the fruitful discussions and collaborations as well as all your support. I would like to also thank **Christophe Busch** and **Christian Rathgeb** for warmly opening their arms to me for visiting their team at Hochschule Darmstadt, and for the collaborations during my visit and afterward. I also thank all their team members, especially **Daniel**, **Janier**, **Jascha**, **Robert**, and **Jannis** for their help to make my stay in Darmstadt a great experience. I would like to thank **Hugues Salamin** for warmly welcoming my internship in ams OSRAM AG. I extend my gratitude to all team members in ams OSRAM for being such fantastic hosts during my internship, specially **Alex** and **Antoine** for their help and support during my internship. I would like to thank all PIs and ESRs in the TReSPAsS-ETN<sup>1</sup> and PriMa projects, with a special mention to **Mathias**, **Pietro**, **Dailé**, **Amina**, **Ahmad**, **Mahdi**, and **Giuseppe**, for all the discussions, idea-sharing, collaborations, and furthermore for the friendship and the good moments.

I would like to also thank the (current and former) members of the Biometrics group at Idiap for all their help in the course of my PhD as well as fruitful discussions and collaborations:

---

<sup>1</sup>My PhD was mainly funded by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network, and I would like to thank the European Commission for this generous grant.

## Acknowledgements

---

Alex, Vedrana, Tiago, Anjith, Alain, Karine, David, Christophe, Laurent, Amir, Sushil, Pavel, Parsa, Ketan, Behrooz, Eklavya, Junduan, Luis, and Vedit. Also, I would like to thank amazing people at IT and development teams at Idiap, who were always supportive with any technical issue, special mention to **Louis-Marie, Frank, Samuel, Bastien, Mattéo, Vincent, Laurent, Yannick, William, Olivier, and Philip**. I extend my gratitude to the HR team at Idiap who were always supportive and made my work at Idiap smooth: **Laura, Sylvie, Christophe, Elisa, Qëndresa, and Valentin**. A special thank you to **Andrea Cavallaro**, the director of Idiap, for his generous and unwavering support.

I would like to thank all my friends at Idiap: **Sina, Sepehr, Alireza, Rabeeh, Ansul, Roberto, Amirreza, Mahdi, and Arya**. I would like to also thank my friends at EPFL: **Alireza, Mohammadhossein, Seyed, and Saeed**. Thank you for your friendship, your support, and the fun moments. I am also very thankful to my friends from Iran who, apart from the distance, have been very close in several instances of my life. Specially, I would like to thank **Peyman, Mohsen, Hadi, Hamid, Ali, and Farshad**.

I am both grateful and indebted to you, **Arash Amini** and **Hamid Behrooz**, the advisors of my MSc at Sharif University of Technology, who opened the doors of research to me and supported me in every stage. I would like to extend my gratitude to **Farokh Marvasti** and **Mahdieh Soleymani** for all the lessons I learned from them and for their support over the years.

My sincerest gratitude goes to **my parents**. You are the true reason that I am here. Your love and encouragement have and will always take me wherever I go, and I feel extremely grateful for having you all by my side. I extend similar gratitude to **my sister** for her unlimited support and for having played an immense role in shaping my life. I would also like to extend my gratitude to **my parents-in-law** for their unwavering support and love through these years. Thank you indefinitely!

Finally, there is one person who deserves my biggest acknowledgment, and I cannot thank her enough for her support, patience, brilliance or her love: **my wife, Sahar**. You have been beside me from the very beginning of my PhD and supporting me in every single moment of this journey. Words cannot express the level of my appreciation, and so I wholeheartedly dedicate this thesis to you!

*Lausanne, July 24, 2024*

Hatef Otroshi Shahreza

# Abstract

Biometric recognition systems tend toward ubiquity and are widely being used in different applications for authentication purposes. Compared to conventional authentication tools, such as PIN or password, which are always in danger of being forgotten or stolen, biometric authentication offers excellent convenience for the user. In contrast, in addition to security threats, biometric systems are also in danger of privacy issues. This is because biometric data include privacy-sensitive information of enrolled subjects, which causes privacy concerns in the application of biometric systems. Generally, in biometric systems, some features (also known as templates) are often extracted from biometric data, and are stored in the database of the system. Then, during recognition, similar templates are extracted and compared to the ones stored in the database. In this thesis, we focus on templates stored in biometric systems and investigate the vulnerability of systems to different attacks based on templates stored in the database of a biometric system.

In the first part of the thesis, we consider the face recognition system as one of the popular biometric systems and show that if an adversary gains access to the database of a face recognition system, they may be able to reconstruct face images of underlying leaked facial templates. The reconstructed face images not only reveal privacy-sensitive information but also can be used to impersonate the systems that the user is enrolled in. We evaluate the adversary's successful attack rate in entering the system based on an injection attack by bypassing the camera. In addition, we consider the real-world scenario where the adversary may perform a practical presentation attack to impersonate and evaluate the attack rate.

In the second part of the thesis, we propose new methods to protect biometric templates. We present MLP-Hash, a new cancelable biometric scheme that works based on random multi-layer perceptrons (MLP). We also discuss a hybrid template protection mechanism that leverages cancelable biometric and homomorphic encryption. Using cancelable biometric and homomorphic encryption not only boosts for a higher security of the protected templates but also reduces the required computation compared to applying homomorphic encryption only. The proposed template protection schemes can be used in systems of different biometric modalities (face, voice, finger vein, etc.). Finally, we present a new method to protect and enhance vascular biometric recognition methods using BioHashing and an auto-encoder network.

The last part of this thesis is focused on the evaluation of template protection schemes. We first benchmark different template protection schemes based on the ISO/IEC 24745 standard requirements. We discuss the metrics to evaluate the leakage of information in the protected

## Abstract

---

biometric templates. In particular, we investigate the invertibility of protected biometric templates and also propose a new measure to evaluate the linkability of protected templates. The proposed linkability metric is based on maximal leakage, which is a well-studied measure in information-theoretic literature. We show that the resulting linkability measure has a number of important theoretical properties and an operational interpretation in terms of statistical hypothesis testing. We further explore the application of our proposed method for the case that the adversary gains access to multiple protected templates, which has not been investigated in the literature.

**Keywords:** Biometrics, Biometric Template Protection, Face Recognition, Face Reconstruction, Finger Vein Recognition, Information Leakage, Linkability, Maximal Leakage, Multi-modal, Presentation Attack, Speaker Recognition, Template, Template Inversion, Vascular Biometrics.

# Résumé

Les systèmes de reconnaissance biométrique sont omniprésents et largement utilisés dans différentes applications d'authentification. Ces systèmes biométriques offrent une commodité exceptionnelle pour l'utilisateur, comparés aux outils d'authentification conventionnels, tels que les codes PIN ou les mots de passe, qui comportent le risque d'être oubliés ou volés. En revanche, en plus des risques de sécurité, les systèmes biométriques sont également exposés à des problèmes de confidentialité. En effet, les données biométriques comprennent des informations sensibles sur la vie privée des sujets enregistrés, ce qui suscite des préoccupations en matière de confidentialité dans l'application des systèmes biométriques. Généralement, dans les systèmes biométriques, certaines caractéristiques (également appelées gabarits (*templates*)) sont souvent extraites des données biométriques et stockées dans la base de données du système. Ensuite, lors de la reconnaissance, des gabarits similaires sont extraits et comparés à ceux stockés dans la base de données. Dans cette thèse, nous nous concentrons sur les gabarits stockés dans les systèmes biométriques. Plus précisément, nous examinons la vulnérabilité des systèmes face à différentes attaques qui exploitent les gabarits stockés dans la base de données d'un système biométrique.

Dans la première partie de la thèse, nous considérons le système de reconnaissance faciale comme l'un des systèmes biométriques populaires et montrons que si un adversaire accède à la base de données d'un système de reconnaissance faciale, il peut être capable de reconstruire des images faciales à partir des gabarits faciaux divulgués. Les images faciales reconstruites révèlent non seulement des informations sensibles sur la vie privée, mais peuvent également être utilisées par un adversaire pour accéder au système biométrique. Nous évaluons le taux de réussite des attaques de l'adversaire pour pénétrer dans le système en utilisant une attaque par injection, contournant ainsi la caméra. De plus, nous considérons le scénario réel où l'adversaire peut effectuer une attaque de présentation dans un but d'usurpation et évaluons le taux de réussite de cette attaque.

Dans la deuxième partie de la thèse, nous proposons de nouvelles méthodes pour protéger les gabarits biométriques. Nous présentons MLP-Hash, un nouveau schéma de biométrie révocable basé sur des perceptrons multicouches (MLP) aléatoires. Nous discutons également un mécanisme hybride de protection des gabarits qui exploite la biométrie révocable et le chiffrement homomorphe. L'utilisation de la biométrie révocable et du chiffrement homomorphe augmente non seulement la sécurité des gabarits protégés, mais réduit également le calcul requis pour l'application du chiffrement homomorphe. Les schémas de protection des modèles proposés peuvent être utilisés dans des systèmes de différentes modalités biomé-

## Résumé

---

triques (visage, voix, veine des doigts, etc.). Enfin, nous présentons une nouvelle méthode pour protéger et améliorer les méthodes de reconnaissance biométrique vasculaire en utilisant le BioHashing et un réseau auto-encodeur.

La dernière partie de cette thèse est consacrée à l'évaluation des schémas de protection des modèles. Nous comparons d'abord différents schémas de protection des gabarits selon les exigences de la norme ISO/IEC 24745. Nous discutons des métriques pour évaluer la fuite d'informations dans les modèles biométriques protégés. En particulier, nous étudions l'inversibilité des gabarits biométriques protégés et proposons également une nouvelle mesure pour évaluer la liabilité (*linkability*) les gabarits protégés. La métrique de *linkability* proposée est basée sur la fuite maximale (*maximal leakage*), une mesure bien étudiée dans la littérature de la théorie de l'information. Nous montrons que la mesure de *linkability* résultante présente un certain nombre de propriétés théoriques importantes et une interprétation opérationnelle en termes de test d'hypothèse statistique. Nous explorons également l'application de notre méthode proposée dans le cas où l'adversaire accède à plusieurs gabarits protégés, ce qui n'a pas été étudié dans la littérature.

**Mots-clés :** Biométrie, Protection des gabarits biométriques, Reconnaissance faciale, Reconstruction faciale, Reconnaissance des veines des doigts, Fuite d'information, Liabilité, Fuite maximale, Multimodal, Attaque de présentation, Reconnaissance vocale, Modèles, Inversion de gabarits, Biométrie vasculaire.



# Zusammenfassung

Biometrische Erkennungssysteme sind allgegenwärtig und werden in verschiedensten Anwendungen zur Authentifizierung genutzt. Im Vergleich zu herkömmlichen Authentifizierungsmethoden wie PINs oder Passwörtern, die immer Gefahr laufen, vergessen oder gestohlen zu werden, bietet die biometrische Authentifizierung dem Benutzer einen hervorragenden Komfort. Im Gegensatz dazu sind biometrische Systeme neben Sicherheitsbedrohungen auch von Datenschutzproblemen betroffen. Dies liegt daran, dass biometrische Daten sensible private Informationen der registrierten Personen enthalten, was bei der Anwendung von biometrischen Systemen Bedenken hinsichtlich des Datenschutzes aufwirft. Im Allgemeinen werden in biometrischen Systemen Merkmale (auch als Templates bekannt) aus biometrischen Daten extrahiert und in einer Datenbank des Systems gespeichert. Während des Erkennungsvorganges werden dann vergleichbare Templates extrahiert und mit den zuvor in der Datenbank gespeicherten verglichen. In dieser Dissertation konzentrieren wir uns auf die in den Datenbanken der biometrischen Systeme gespeicherten Templates und untersuchen deren Verwundbarkeit gegenüber verschiedenen Angriffen.

Im ersten Teil der Dissertation betrachten wir Gesichtserkennungssysteme, das sie eines der beliebtesten biometrischen Systeme darstellen. Wir zeigen, dass ein Angreifer, der Zugriff auf die Datenbank eines solchen Gesichtserkennungssystems erhält, möglicherweise in der Lage ist, Gesichtsbilder aus den kompromitierten biometrischen Templates zu rekonstruieren. Die rekonstruierten Gesichtsbilder können nicht nur sensible Informationen über die Anwender offenbaren und so ihre Privatsphäre verletzen, sondern können auch dazu verwendet werden, die Identität von Personen zu kopieren und sich so in Systemen, in denen der Benutzer registriert ist, als dieser zu authentisieren. Wir bewerten die Erfolgsrate des Angreifers beim Eindringen in das System anhand eines Injektionsangriffs durch Umgehung der Kamera. Darüber hinaus betrachten wir das realistische Szenario, in dem der Angreifer einen Präsentationsangriff durchführt, um sich als andere Person auszugeben, und bewerten die Erfolgsrate dieses Angriffs.

Im zweiten Teil der Dissertation schlagen wir neue Methoden zum Schutz der biometrischen Templates vor. Wir präsentieren MLP-Hash, ein neues annullierbares biometrisches Schema, das auf einem zufälligen mehrschichtigen Perzeptron (MLP) basiert. Weiter diskutieren auch einen hybriden Schutzmechanismus für biometrische Templates, der ein annullierbares biometrisches Schema mit homomorpher Verschlüsselung kombiniert. Die Kombination der beiden Ansätze erhöht nicht nur die Sicherheit der gespeicherten Templates, sondern reduziert auch den erforderlichen Rechenaufwand im Vergleich zur homomorphen Verschlüsselung

## Zusammenfassung

---

allein. Die vorgeschlagenen Schutzmechanismen können in Systemen mit unterschiedlichen biometrischen Modalitäten (Gesicht, Stimme, Fingeradern, etc.) verwendet werden. Abschließend präsentieren wir eine neue Methode zum Schutz und zur Verbesserung der vaskulären biometrischen Erkennungsmethoden unter Verwendung von BioHashing und einem Autoencoder-Netzwerk.

Der letzte Teil dieser Dissertation konzentriert sich auf die Bewertung von Template-Schutzmechanismen. Zunächst bewerten wir verschiedene Template-Schutzmechanismen anhand der Anforderungen der ISO/IEC 24745-Norm. Wir diskutieren die Metriken zur Bewertung des Durchsickerns von Informationen aus den geschützten biometrischen Templates. Insbesondere untersuchen wir die Umkehrbarkeit der geschützten biometrischen Templates und präsentieren eine neue Messmethode zur Bewertung der Verknüpfbarkeit geschützter Templates. Die vorgeschlagene Metrik basiert auf dem maximalem Informationsabfluss, einem in der informationstheoretischen Literatur gut untersuchten Maß. Wir zeigen, dass die resultierende Metrik eine Reihe wichtiger theoretischer Eigenschaften und eine operationelle Interpretation Sinne statistischer Hypothesentests aufweist. In einer neuartigen Untersuchung prüfen wir die Anwendung unserer vorgeschlagenen Methode für den Fall, dass ein Angreifer Zugriff auf mehrere geschützte Templates erhält.

**Schlüsselwörter:** Biometrie, Schutz biometrischer Templates, Gesichtserkennung, Gesichtsrekonstruktion, Fingeradererkennung, Informationsleck, Verknüpfbarkeit, Maximales Durchsickern, Multimodal, Präsentationsangriff, Spracherkennung, Template, Template-Umkehrung, Vaskuläre Biometrie.

# Contents

<b>Acknowledgements</b>	<b>i</b>
<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xix</b>
<b>Acronyms</b>	<b>xxiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivations . . . . .	2
1.2 Objectives and Contributions . . . . .	4
1.3 Thesis Outline . . . . .	8
<b>2 Related Work</b>	<b>11</b>
2.1 Template Inversion in Biometric Systems . . . . .	11
2.2 Biometric Template Protection Schemes . . . . .	15
2.2.1 Cancelable Biometrics . . . . .	16
2.2.2 Biometric Cryptosystems . . . . .	17
2.2.3 Homomorphic Encryption . . . . .	17
2.3 Evaluation of Biometric Template Protection . . . . .	18
2.3.1 Irreversibility . . . . .	18
2.3.2 Unlinkability . . . . .	19
2.3.3 Recognition Performance . . . . .	21
2.4 Biometric Feature Extractors and Datasets . . . . .	22
2.4.1 Face Recognition . . . . .	22
2.4.2 Voice Recognition . . . . .	23
2.4.3 Vascular Recognition . . . . .	23
2.4.4 Iris Recognition . . . . .	23
<b>3 Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates</b>	<b>25</b>
3.1 Low-resolution Face Reconstruction . . . . .	26
3.1.1 Proposed Method . . . . .	27
3.1.1.1 Threat Model . . . . .	27
3.1.1.2 Face Reconstruction Method . . . . .	28

## Contents

---

3.1.1.3	Vulnerability Evaluation to TI Attacks . . . . .	32
3.1.2	Experiments . . . . .	32
3.1.2.1	Experimental Setup . . . . .	33
3.1.2.2	Primary Experiment . . . . .	35
3.1.2.3	Comparison with Previous Methods . . . . .	35
3.1.2.4	Ablation Study . . . . .	37
3.1.2.5	TI Vulnerability Analysis of SOTA FR Models . . . . .	39
3.1.2.6	Discussion . . . . .	43
3.2	High-resolution Face Reconstruction using Real Data . . . . .	48
3.2.1	Proposed Method . . . . .	50
3.2.1.1	Threat Model . . . . .	50
3.2.1.2	Face Reconstruction Method . . . . .	50
3.2.2	Experiments . . . . .	52
3.2.2.1	Experimental Setup . . . . .	52
3.2.2.2	Comparison with Previous Methods . . . . .	54
3.2.2.3	Ablation Study . . . . .	56
3.3	High-resolution Face Reconstruction using Synthetic Data . . . . .	57
3.3.1	Proposed Method . . . . .	58
3.3.1.1	Threat Model . . . . .	58
3.3.1.2	Face Reconstruction Method . . . . .	59
3.3.2	Experiments . . . . .	61
3.3.2.1	Experimental Setup . . . . .	61
3.3.2.2	Comparison with Previous Methods . . . . .	62
3.3.2.3	Ablation Study . . . . .	63
3.4	3D Face Reconstruction . . . . .	64
3.4.1	Proposed Method . . . . .	66
3.4.1.1	Threat Model . . . . .	66
3.4.1.2	3D Face Reconstruction Method . . . . .	68
3.4.1.3	Camera Parameters Optimization . . . . .	71
3.4.2	Experiments . . . . .	73
3.4.2.1	Experimental Setup . . . . .	73
3.4.2.2	TI Attack by Injecting Reconstructed Face Images . . . . .	75
3.4.2.3	Practical Presentation Attack using Reconstructed Face Images . . . . .	80
3.4.2.4	Discussion . . . . .	84
3.5	Conclusion . . . . .	91
<b>4</b>	<b>Protection of Biometric Templates</b>	<b>93</b>
4.1	MLP-Hash: Protecting Biometric Templates via Randomized MLP . . . . .	93
4.1.1	Proposed method . . . . .	94
4.1.1.1	MLP-Hash Algorithm . . . . .	94
4.1.1.2	Comparing MLP-Hash Templates . . . . .	95
4.1.2	Experiments . . . . .	95

4.1.2.1	Experimental Setup and Baselines . . . . .	96
4.1.2.2	Unlinkability Evaluation . . . . .	96
4.1.2.3	Irreversibility Evaluation . . . . .	97
4.1.2.4	Recognition Accuracy Evaluation . . . . .	98
4.1.2.5	Discussion . . . . .	99
4.2	Hybrid Protection of Deep Templates using Cancelable Biometrics and Homomorphic Encryption . . . . .	99
4.2.1	Hybrid template protection method . . . . .	101
4.2.1.1	Notations and Formulation . . . . .	101
4.2.1.2	Combinations of different CB methods with HE . . . . .	102
4.2.2	Experiments . . . . .	103
4.2.2.1	Experimental Setup . . . . .	103
4.2.2.2	Analysis . . . . .	104
4.2.2.3	Discussion . . . . .	107
4.3	Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition . . . . .	107
4.3.1	Proposed Framework . . . . .	108
4.3.1.1	Auto-encoder . . . . .	109
4.3.1.2	Protecting Deep Templates . . . . .	110
4.3.2	Experiments . . . . .	110
4.3.2.1	Experimental Setup . . . . .	110
4.3.2.2	Performance Evaluation for Previous FVR Methods . . . . .	112
4.3.2.3	Ablation Study . . . . .	114
4.3.2.4	Using our Framework as a FVR Method . . . . .	116
4.3.2.5	Palm and Wrist Vein Recognition . . . . .	116
4.3.2.6	Discussion . . . . .	116
4.4	Conclusion . . . . .	120
<b>5</b>	<b>Evaluation of Biometric Template Protection Schemes</b>	<b>123</b>
5.1	Benchmarking Cancelable Biometric Protection Schemes for Deep Templates .	123
5.1.1	Evaluation metrics . . . . .	124
5.1.1.1	Recognition Performance . . . . .	124
5.1.1.2	Unlinkability . . . . .	125
5.1.1.3	Irreversibility . . . . .	125
5.1.2	Experimental Results . . . . .	126
5.1.2.1	Recognition Performance Evaluation . . . . .	126
5.1.2.2	Unlinkability Evaluation . . . . .	129
5.1.2.3	Irreversibility Evaluation . . . . .	129
5.2	Inversion of Protected Biometric Templates . . . . .	131
5.2.1	Face Reconstruction from Protected Templates . . . . .	131
5.2.1.1	Threat Model . . . . .	131
5.2.1.2	Face Reconstruction . . . . .	132

## Contents

---

5.2.2	Experimental Results . . . . .	133
5.2.2.1	Experimental Setup . . . . .	134
5.2.2.2	Face Reconstruction from Protected Templates . . . . .	135
5.2.2.3	Discussion . . . . .	136
5.3	Measuring Linkability of Protected Templates using Maximal Leakage . . . . .	138
5.3.1	Maximal Linkability . . . . .	139
5.3.1.1	Notations and the Definition of Maximal Leakage . . . . .	139
5.3.1.2	Maximal Linkability of Biometric Templates . . . . .	141
5.3.1.3	On Estimating System Linkability . . . . .	142
5.3.1.4	Maximal Linkability and Hypothesis Testing . . . . .	143
5.3.2	Comparison With Other Measures . . . . .	145
5.3.2.1	On Linkability via Differential Privacy . . . . .	145
5.3.2.2	Comparison with Gomez-Barrero <i>et al.</i> Measure . . . . .	148
5.3.3	Experiments . . . . .	151
5.3.3.1	Experimental Setup . . . . .	151
5.3.3.2	Analyze . . . . .	152
5.3.3.3	Discussion . . . . .	156
5.4	Measuring Linkability of Multiple Protected Templates . . . . .	157
5.4.1	Problem Definition and Formulation . . . . .	158
5.4.1.1	Notations . . . . .	158
5.4.1.2	Different Scenarios with Multiple Similarity Scores . . . . .	159
5.4.2	Measuring linkability using multiple similarity scores . . . . .	160
5.4.2.1	Data Processing Inequality . . . . .	161
5.4.2.2	Composition Theorems . . . . .	163
5.4.2.3	Maximal Linkability for Multiple Similarity Scores . . . . .	164
5.4.3	Experiments . . . . .	166
5.4.3.1	Experimental Setup . . . . .	166
5.4.3.2	Analysis of different scenarios in biometric systems . . . . .	166
5.4.3.3	Extending studied scenarios to three similarity scores . . . . .	170
5.4.3.4	Discussion . . . . .	170
5.5	Conclusion . . . . .	171
<b>6</b>	<b>Conclusion and Future Directions</b>	<b>175</b>
<b>A</b>	<b>Recognition Performance of Face Recognition Models</b>	<b>179</b>
<b>B</b>	<b>Face Reconstruction from Partially Leaked Templates</b>	<b>181</b>
B.1	Methodology . . . . .	181
B.1.1	Threat Model . . . . .	182
B.1.2	Face Reconstruction . . . . .	182
B.2	Experiments . . . . .	184
B.2.1	Experimental Setup . . . . .	184

B.2.2 Analysis . . . . .	184
<b>C Vulnerability of an Anti-spoofing Face Recognition System to Reconstructed Face Images from TI Attacks</b>	<b>187</b>
<b>D Proofs for Chapter 5</b>	<b>189</b>
D.1 Proof for Section 5.3.1 . . . . .	189
D.2 Proof for Section 5.3.2 . . . . .	189
<b>E Reproducibility</b>	<b>191</b>
<b>Bibliography</b>	<b>210</b>
<b>Curriculum Vitae</b>	<b>211</b>





# List of Figures

1.1	General block diagram of a biometric recognition system (unprotected)	2
3.1	Block diagram of the proposed framework for evaluating vulnerability of FR system to TI attacks in Section 3.1	26
3.2	Block diagram of the $m$ th DSCasConv block	30
3.3	Structure of the proposed face reconstruction network in Section 3.1	30
3.4	Histogram of scores for the TI attack in Section 3.1	34
3.5	Effect of loss function on the performance of the TI attack in Section 3.1	37
3.6	Effect of weights in our loss function performance of the TI attack in Section 3.1	37
3.7	Effect of network structure on the TI attack performance in Section 3.1	39
3.8	Sample faces and their reconstructed versions by different TI attack network blocks in Section 3.1	41
3.9	Sample faces and their reconstructed versions by the TI attack in Section 3.1 for different backbones	42
3.10	Sample faces and their reconstructed versions by the TI attack in Section 3.1 for different heads	44
3.11	Sample faces and their reconstructed versions by the TI attack in Section 3.1 trained w/wo data augmentation	47
3.12	Sample faces and their reconstructed versions by the TI attack in Section 3.2	49
3.13	General block diagram of the proposed method in Section 3.2	49
3.14	Block diagram of the proposed face reconstruction network in Section 3.2	51
3.15	Sample faces and their reconstructed versions by the blackbox TI attack in Section 3.2	55
3.16	Sample faces and their reconstructed versions by the TI attack in Section 3.3	57
3.17	Block diagram of the proposed method in Section 3.3	59
3.18	Sample faces and their reconstructed versions by the TI attack in Section 3.3	63
3.19	Sample faces as well as their 3D and frontal 2D reconstructions by whitebox TI attack in Section 3.4	64
3.20	General block diagram of the proposed method in Section 3.4	64
3.21	Block diagram of the threat model in Section 3.4	66
3.22	Block diagram of our proposed TI attack in Section 3.4	69
3.23	Block diagram of a FR system and data flows in normal usage, TI attack by injecting, and presentation attack using the reconstructed face images in Section 3.4	73

## List of Figures

---

3.24 Sample faces, their frontal reconstructs, and reconstructed versions within the camera parameters sub-grid by the whitebox TI attack in Section 3.4 . . . . .	76
3.25 Sample faces, their frontal reconstructs, and reconstructed versions within the camera parameters sub-grid by the blackbox TI attack in Section 3.4 . . . . .	79
3.26 The evaluation setup for presentation attack in Section 3.4 . . . . .	80
3.27 A sample image, its reconstructed versions, the digital replay attacks, and presentation attacks in Section 3.4 . . . . .	82
3.28 A sample image, its frontal reconstruction, its 3D face reconstruction, and reconstructed versions camera parameters grid by whitebox TI attack in Section 3.4 .	85
3.29 Ablation study on the effect of different hyperparameters in grid search for camera parameters optimization in attack 1 in Section 3.4 . . . . .	86
3.30 Ablation study on the effect of different hyperparameters in continuous optimization for camera parameters in attack 1 in Section 3.4 . . . . .	86
3.31 Histogram of pitch and yaw in original and reconstructed images by attack 1 in Section 3.4 . . . . .	88
3.32 Reconstruction of sample images in whitebox and blackbox TI attacks in Section 3.4 . . . . .	89
4.1 Block diagram of MLP-Hash protected FR system in Section 4.1 . . . . .	94
4.2 Unlinkability evaluation of unprotected and MLP-Hash protected templates in Section 4.1 . . . . .	96
4.3 General scheme of the proposed hybrid protection method in Section 4.2 . . .	100
4.4 Block diagram of the proposed hybrid protection method in Section 4.2 . . . .	101
4.5 ROC curves of the unprotected, HE-protected, CB-protected, and hybrid-protected templates in Section 4.2 . . . . .	104
4.6 Block diagram of the proposed framework in Section 4.3 . . . . .	108
4.7 Sample finger vein images and their WLD, RLT, and MC features in Section 4.3 .	111
4.8 ROC curves of previous FVR methods with BioHash protected templates, BioHash protected of their PCA transformation, and their protected version via the proposed framework in Section 4.3 . . . . .	113
4.9 Evaluating the effect of $L_{embedding}$ in the proposed framework in Section 4.3 .	114
4.10 Evaluating the effect of $L_{BioHash}$ in the proposed framework in Section 4.3 . . .	115
4.11 Evaluating the effect of $\alpha$ in the proposed framework in Section 4.3 . . . . .	115
4.12 ROC curves of the framework in Section 4.3 in two modes: given the raw finger vein images, or given the features extracted from WLD, RLT, and MC . . . . .	117
4.13 2D representation of extracted features for 5 different identities with different methods in Section 4.3 . . . . .	117
4.14 Evaluating the performance of the proposed framework with different values of $L_{embedding}$ and $L_{BioHash}$ in Section 4.3 . . . . .	119
5.1 System performance evaluation for different physiological and behavioral biometric traits in Section 5.1 . . . . .	127

5.2	Sample faces and their reconstructed versions from protected templates with the TI attack in Section 5.2 . . . . .	137
5.3	Bounds on adversary's ability to perform hypothesis testing for different maximal linkability scores in Section 5.3 . . . . .	145
5.4	Synthetic distributions of mated and non-mated scores and their ROC plots in Section 5.3 . . . . .	146
5.5	Distributions of scores for mated and non-mated templates for two systems that are ranked differently by our linkability measure proposed in Section 5.3 and Gomez <i>et al.</i> . . . . .	149
5.6	Histogram of mated and non-mated scores for linkable protected templates in Section 5.3 . . . . .	155
5.7	Histogram of mated and non-mated scores for unprotected templates in Section 5.3 . . . . .	156
5.8	General block diagram of a biometric recognition system in Section 5.4 . . . . .	158
5.9	Different scenarios with two scores from different leaked templates in Section 5.4 . . . . .	160
5.10	Histograms of synthetic distributions of mated and non-mated scores and their ROC plots in Section 5.4 . . . . .	162
5.11	Linkability of protected templates using multiple scoring functions in Section 5.4 . . . . .	170
B.1	Vulnerability of FR systems to inversion of partially leaked template in Appendix B . . . . .	185
B.2	Sample faces and their reconstructed versions from different percentages of leaked facial templates by TI attack in Appendix B . . . . .	185



## List of Tables

2.1	Comparison with related works on TI attacks against FR systems . . . . .	12
2.2	Summary of methods for evaluating the linkability of protected biometric templates. . . . .	19
3.1	Recognition performance and vulnerability to the TI attack in Section 3.1 . . . .	35
3.2	Comparison of different face reconstruction methods in Section 3.1 . . . . .	36
3.3	Comparison of FR models with different SOTA backbones and the same head in Section 3.1 . . . . .	40
3.4	Comparison of FR models with different SOTA heads and the same backbone in Section 3.1 . . . . .	43
3.5	Comparison of SAR against FR systems using different similarity score functions in Section 3.1 . . . . .	45
3.6	Complexity comparison of different network structures in Section 3.1 . . . . .	45
3.7	Comparison of performance of our face reconstruction network when trained on different datasets in Section 3.1 . . . . .	46
3.8	Comparison of reconstruction quality of TI attack in Section 3.1 trained w/wo data augmentation . . . . .	47
3.9	Vulnerability evaluation of SOTA pretrained FR models in Section 3.1 . . . . .	48
3.10	Comparison with different methods for TI attack in Section 3.2 . . . . .	54
3.11	Evaluating the effect of mapping space and each loss term in the whitebox TI attack in Section 3.2 . . . . .	57
3.12	Comparison with different TI attacks in Section 3.3 . . . . .	62
3.13	The effect of each loss term in the whitebox TI attack in Section 3.3 . . . . .	63
3.14	Different TI attacks against FR systems in our threat model in Section 3.4 . . . .	68
3.15	Evaluation of whitebox attacks when injecting reconstructed face images in Section 3.4 . . . . .	77
3.16	Evaluation of blackbox attacks when injecting reconstructed face images in Section 3.4 . . . . .	78
3.17	Vulnerability evaluation of the simulation and practical blackbox TI attacks in Section 3.4 . . . . .	83
3.18	Comparison of the proposed method with previous <i>blackbox</i> TI methods in practical presentation attacks in Section 3.4 . . . . .	84

## List of Tables

---

3.19 Ablation study on the semi-supervised learning approach and the effect of loss terms in attack 1 in Section 3.4 . . . . .	86
3.20 SAR in whitebox and blackbox TI attacks against different target FR systems in Section 3.4 . . . . .	90
4.1 Unlinkability evaluation of MLP-Hash, BioHash, IoM-GRP, and IoM-URP protected templates in Section 4.1 . . . . .	97
4.2 Irreversibility evaluation of MLP-Hash, BioHash, IoM-GRP, and IoM-URP protected templates in Section 4.1 . . . . .	97
4.3 Comparison of MLP-Hash-protected, BioHash-protected, IoM-GRP-protected, IoM-URP-protected, and unprotected SOTA FR models in Section 4.1 . . . . .	98
4.4 Complexity comparison of template protection methods in Section 4.1 . . . . .	99
4.5 Protection of CB, HE, and hybrid (CB+HE) protection against different threat models in the ISO/IEC 30136 standard in Section 4.2 . . . . .	100
4.6 The average execution time and recognition performance of HE and the hybrid method when applying BioHashing in Section 4.2 . . . . .	105
4.7 The average execution time and recognition performance of HE and the hybrid method when applying MLP-Hashing in Section 4.2 . . . . .	106
4.8 The average execution time and recognition performance of HE and the hybrid method when applying IoM Hashing in Section 4.2 . . . . .	106
4.9 Size of the features extracted by WLD, RLT, and MC methods and their execution times in Section 4.3 . . . . .	111
4.10 Performance of previous FVR methods with BioHash protected templates and their enhanced version via the proposed framework in Section 4.3 . . . . .	114
4.11 Performance of the proposed framework in Section 4.3 . . . . .	118
4.12 The average execution time to get embedding features from the encoder and the required memory for encoder network in the proposed framework in Section 4.3	120
5.1 Summary of feature extraction models, datasets, numbers of mated and non-mated comparisons, and biometric performance in the experiments in Section 5.1	126
5.2 Recognition performance evaluation in benchmarking CB schemes in Section 5.1	128
5.3 Unlinkability evaluation in benchmarking CB schemes in Section 5.1 . . . . .	129
5.4 Irreversibility evaluation in benchmarking CB schemes in Section 5.1 . . . . .	130
5.5 Performance of reconstructed face images from protected templates in the TI attack in Section 5.2 with FMR of $10^{-2}$ . . . . .	134
5.6 Performance of reconstructed face images from protected templates in the TI attack in Section 5.2 with FMR of $10^{-3}$ . . . . .	135
5.7 Performance of reconstructed face images from HE-protected templates with different models as $F_{adv}$ in the TI attack in Section 5.2 . . . . .	136
5.8 Linkability of synthetic distributions of scores for mated and non-mated templates in Section 5.3 . . . . .	150
5.9 Summary of BTP schemes used in the linkability experiments in Section 5.3. . .	152

5.10 Summary of biometric recognition systems used in the linkability experiments in Section 5.3 . . . . .	152
5.11 Linkability of different BTP schemes in Section 5.3 . . . . .	153
5.12 Linkability of BioHash-protected templates with different scoring functions in Section 5.3 . . . . .	154
5.13 Linkability of BioHash-protected templates across different biometric characteristics in Section 5.3 . . . . .	154
5.14 Linkability of BioHash-protected templates for different feature extractors in Section 5.3 . . . . .	155
5.15 Linkability of BioHash-protected templates for different biometric modalities in Section 5.4 . . . . .	167
5.16 Linkability of BioHash-protected templates for different face feature extractors in Section 5.4 . . . . .	167
5.17 Linkability of different BTP schemes in Section 5.4 . . . . .	168
5.18 Linkability of protected templates with different keys in Section 5.4 . . . . .	169
5.19 Linkability of BioHash-protected templates with different scoring functions in Section 5.4 . . . . .	169
A.1 Recognition performance of FR models used in the experiments . . . . .	179
A.2 PLCC of comparison scores for FR models used in the experiments . . . . .	180
B.1 Recognition performance of FR models in Appendix B . . . . .	184
B.2 Vulnerability of FR systems to TI attack in Appendix B from different percentage of template leakage . . . . .	185
C.1 Evaluation of a PAD FR system for presentation attacks using the reconstructed face images from the proposed TI attacks . . . . .	187





# Acronyms

**ACER** Average Classification Error Rate

**AD** Auxiliary Data

**AE** Auto-Encoder

**APCER** Attack Presentation Classification Error Rate

**BFV** Brakerski/Fan-Vercauteren

**BIPA** Illinois Biometric Information Privacy Act

**BPCER** Bona fide Presentation Classification Error Rate

**BTP** Biometric Template Protection

**CB** Cancelable Biometrics

**CMC** Cumulative Match Curve

**CNN** Convolutional Neural Network

**CO** Continuous Optimization

**COTS** Commercial-Off-The-Shelf

**DNN** Deep Neural Network

**DSCasConv** Enhanced Deconvolution using Cascaded Convolution and Skip connections

**EER** Equal Error Rate

**EU-GDPR** European Union General Data Protection Regulation

**FADP** Federal Act on Data Protection

**FFHQ** Flickr-Faces-HQ

**FHE** Fully Homomorphic Encryption

## Acronyms

---

**FID** Fréchet Inception Distance

**FMR** False Match Rate

**FNMR** False Non-Match Rate

**FR** Face Recognition

**FVR** Finger Vein Recognition

**GaFaR** Geometry-aware Face Reconstruction

**GAN** Generative Adversarial Network

**GNeRF** Generative Neural Radiance Fields

**GS** Grid Search

**HE** Homomorphic Encryption

**HiResT2F** High-resolution Template to Face

**IAPMR** Impostor Attack Presentation Match Rate

**IoM** Index-of-Maximum

**IoM-GRP** Gaussian Random Projection Index-of-Maximum Hashing

**IoM-URP** Uniformly Random Permutation Index-of-Maximum Hashing

**LFW** Labeled Faced in the Wild

**LQP** Local Quantized Patterns

**MC** Maximum Curvature

**MI** Mutual Information

**NeRF** Neural Radiance Fields

**PAD** Presentation Attack Detection

**PCA** Principal Component Analysis

**PHE** Partially Homomorphic Encryption

**PI** Pseudonymous Identifiers

**PLCC** Pearson Linear Correlation Coefficient

<b>RBF</b>	Radial Basis Function
<b>RBR</b>	Renewable Biometric Reference
<b>RLT</b>	Repeated Line Tracking
<b>ROC</b>	Receiver Operating Characteristic
<b>RTMR</b>	Renewable Template Matching Rate
<b>SAR</b>	Success Attack Rate
<b>SOTA</b>	state-of-the-art
<b>SWHE</b>	Somewhat Homomorphic Encryption
<b>TI</b>	Template Inversion
<b>TMR</b>	True Match Rate
<b>WGAN</b>	Wasserstein Generative Adversarial Network
<b>WLD</b>	Wide Line Detector



# 1 Introduction

Biometric recognition, or biometrics, refers to the process of automated establishment of identity based on the biological or behavioral characteristics<sup>1</sup>, such as face, voice fingerprint, iris, palm/finger vein, etc. [1], [2]. With recent advancements in deep neural networks, biometric recognition systems have achieved significant performance, and become a reliable solution in applications which requires automatic recognition. Furthermore, compared to traditional authentication mechanisms, such as passwords or tokens (which are in danger of being lost or forgotten), biometric recognition systems are more convenient for users. For these reasons, biometric systems tend towards ubiquity and are increasingly deployed in numerous applications, from personal (e.g., smartphone unlocking with face<sup>2</sup> or fingerprint<sup>3</sup> recognition, etc.) to large-scale applications (e.g., face<sup>4</sup>, fingerprint<sup>5</sup>, and iris<sup>6</sup> recognition in national identity system, or face recognition for passport control at borders and airports<sup>7</sup>, etc.).

Biometric recognition systems (or shortly biometric systems) operates by acquiring biometric data from each user<sup>8</sup>, extracting a set of representative features (called templates) from the acquired data, and comparing this feature set against the templates in the database [2]. A biometric template is a compact representation of the biometric characteristic, containing the discriminatory information that is used for recognizing the user [3]. Biometric templates (also known as features) are often extracted from biometric data and are stored in the system's database during the enrolment stage. Later, during the recognition stage, a new biometric template is extracted and compared with the templates in the database (Figure 1.1). Based on the application, a biometric system may operate in *verification* mode or *identification* mode. In the verification mode, an individual claims an identity (using a user name, a smart card,

---

<sup>1</sup>also known as traits or identifiers

<sup>2</sup><https://apple.co/3mLGCV>

<sup>3</sup><https://bit.ly/3cTJ7Gp>

<sup>4</sup><https://bbc.in/3QeIsO2>

<sup>5</sup><https://bit.ly/3SkvAbi>

<sup>6</sup><https://uidai.gov.in>

<sup>7</sup><https://cnet.co/3sG8qSd>

<sup>8</sup>also referred to as subject

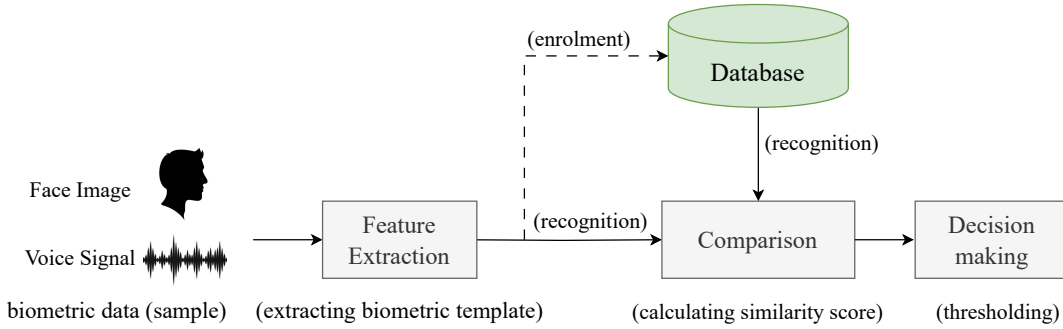


Figure 1.1: General block diagram of a biometric recognition system (unprotected)

etc.) and the system verifies the person's identity by comparing the captured biometric data with the biometric template(s), corresponding to the claimed identity, stored in the system database. In the identification mode, the system recognizes an individual by searching the templates of all the users enrolled in the system database and conducting a one-to-many comparison to establish an individual's identity (or fails if the subject is not enrolled in the system) without the subject having to claim an identity [2].

### 1.1 Background and Motivations

Because of the strong and permanent link between individuals and their biometric characteristics, leakage of biometric information, such as biometric templates, can seriously jeopardize the security and privacy of users. In addition, unlike the traditional authentication (such as password), it is not possible to discard the leaked biometric template and enroll the corresponding user with a new template. In fact, a leaked biometric template can be used by an adversary to retrieve important information about the corresponding user, such as age, gender, ethnicity, etc. Moreover, an adversary can reconstruct the biometric data (such as face image) from biometric templates and use the reconstructed data to enter the biometric system, causing serious threats against biometric systems.

Along the same lines, since biometric templates contain important information about the user's identity, data protection regulations<sup>9</sup> consider biometric templates as sensitive data and impose legal obligations to protect biometric templates. To protect biometric templates and address privacy issues in biometric recognition systems, several schemes are proposed in the literature [3], [7], [8]. In addition, several ISO/IEC standards are published to establish requirements for biometric template protection schemes and their evaluation [9], [10].

In light of these concerns regarding sensitive information in biometric templates, which are stored in biometric systems, this thesis focuses on the information in biometric templates

---

<sup>9</sup>such as the Switzerland Federal Act on Data Protection (FADP) [4], the European Union General Data Protection Regulation (EU-GDPR) [5], the Illinois Biometric Information Privacy Act 740 ILCS 14 (BIPA) [6], etc.

and explores different privacy and security aspects of the leakage of information in biometric templates. The thesis starts with the vulnerability evaluation of unprotected biometric systems to different attacks based on templates stored in the database of a biometric system and demonstrates serious threats in unprotected systems. We consider the face recognition system, as one of the popular biometric systems, and show that an adversary can reconstruct face images of users using their facial templates in a template inversion attack. The reconstructed face images not only reveal privacy-sensitive information, but also can be used to impersonate different systems where the user is enrolled in. We consider an injection attack, where the adversary can bypass the camera and use the reconstructed face image to enter the system. In addition, we consider a real-world scenario where the adversary may perform a practical presentation attack to impersonate. We investigate template inversion attacks under different assumptions, and propose new methods for reconstructing face from facial templates.

Our study on the vulnerability of biometric systems to template inversion attacks motivates the necessity of protecting biometric templates. Therefore, in the second part of the thesis, we delve into privacy-preserving techniques and propose new template protection mechanisms. We propose different methods based on different technologies to protect biometric templates. The proposed methods are not limited to face recognition systems, and cover different types of biometric characteristics.

After proposing different template protection methods, we focus on the evaluation of template protection schemes in the last part of the thesis. We first benchmark different template protection schemes using existing metrics in the literature and find limitations in the evaluation of template protection schemes. Then, we focus on evaluating the invertibility and linkability of protected templates. For evaluating the invertibility property, we propose a learning-based method to investigate the invertibility of protected biometric templates, which has not been explored in the literature. For evaluating the linkability of protected templates, we propose a new metric based on maximal leakage, which is a well-studied measure in information-theoretic literature. We show that the resulting linkability measure has a number of important theoretical properties and an operational interpretation in terms of statistical hypothesis testing. We further explore the application of our proposed method for the case that the adversary gains access to multiple protected templates, which has not been investigated in the literature.

In summary, this thesis investigates information in biometric templates<sup>10</sup>, starting from vulnerability evaluation of unprotected templates, which motivates template protection. With this motivation, new template protection schemes are presented. Finally, evaluation of biometric template protection schemes is explored and new measures are proposed.

---

<sup>10</sup>especially, focusing on biometric templates extracted by deep neural networks (or shortly deep biometric templates.)

### 1.2 Objectives and Contributions

The focus of this thesis is on the information in biometric templates and the contributions can be categorized into three groups<sup>11</sup>:

- **Group 1: Face Reconstruction from Facial Templates**

- [J1] **H. Otroshi Shahreza** and S. Marcel, “Comprehensive Vulnerability Evaluation of Face Recognition Systems to Template Inversion Attacks Via 3D Face Reconstruction”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023. (Reference [11])
- [J2] **H. Otroshi Shahreza**, V. Krivokuća Hahn, S. Marcel, “Vulnerability of State-of-the-Art Face Recognition Models to Template Inversion Attack”, IEEE Transactions on Information Forensics and Security, 2024. (Reference [12])
- [J3] **H. Otroshi Shahreza** and S. Marcel, “Template Inversion Attack using Synthetic Face Images against Real Face Recognition Systems”, IEEE Transactions on Biometrics, Behavior, and Identity Science, 2024. (Reference [13])
- [C1] **H. Otroshi Shahreza**, V. Krivokuća Hahn, and S. Marcel, “Face Reconstruction from Deep Facial Embeddings using a Convolutional Neural Network”, In Proceedings of the IEEE International Conference on Image Processing (ICIP), 2022. (Reference [14])
- [C2] **H. Otroshi Shahreza** and S. Marcel, “Face Reconstruction from Facial Templates by Learning Latent Space of a Generator Network”, In Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS), 2023. (Reference [15])
- [C3] **H. Otroshi Shahreza** and S. Marcel, “Template Inversion Attack against Face Recognition Systems using 3D Face Reconstruction”, IEEE/CVF International Conference on Computer Vision (ICCV), 2023. (Reference [16])
- [C4] **H. Otroshi Shahreza** and S. Marcel, “Blackbox Face Reconstruction from Deep Facial Embeddings Using A Different Face Recognition Model”, In Proceedings of the IEEE International Conference on Image Processing (ICIP), 2023. (Reference [17])
- [C5] **H. Otroshi Shahreza** and S. Marcel, “Inversion of Deep Facial Templates using Synthetic Data”, In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), 2023. (Reference [18])
- [C6] **H. Otroshi Shahreza** and S. Marcel, “Face reconstruction from partially leaked facial embeddings”, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. (Reference [19])

---

<sup>11</sup>J: Journal, C: Conference, A:ArXiv (under-review)



- **Group 2: Protection of Biometric Templates**

- [J4] **H. Otroshi Shahreza** and S. Marcel, “Towards Protecting and Enhancing Vascular Biometric Recognition Methods via Biohashing and Deep Neural Networks”, IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021. (Reference [20])
- [C7] **H. Otroshi Shahreza** and Sébastien Marcel, “Deep Auto-Encoding and Biohashing for Secure Finger Vein Recognition”, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021. (Reference [21])
- [C8] **H. Otroshi Shahreza**, C. Rathgeb, D. Osorio-Roig, V. Krivokuća Hahn, S. Marcel, and C. Busch, “Hybrid Protection of Biometric Templates by Combining Homomorphic Encryption and Cancelable Biometrics”, In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), 2022. (Reference [22])
- [C9] **H. Otroshi Shahreza**, V. Krivokuća Hahn, S. Marcel, “MLP-Hash: Protecting Face Templates via Hashing of Randomized Multi-Layer Perceptron”, In Proceedings of the European Signal Processing Conference (EUSIPCO), 2023. (Reference [23])

- **Group 3: Evaluation of Biometric Template Protection Schemes**

- [J5] **H. Otroshi Shahreza**, Y. Y Shkel, and S. Marcel, “Measuring Linkability of Protected Biometric Templates Using Maximal Leakage”, IEEE Transactions on Information Forensics and Security, 2023. (Reference [24])
- [J6] **H. Otroshi Shahreza**, Y. Y Shkel, and S. Marcel, “On Measuring Linkability of Multiple Protected Biometric Templates using Maximal Leakage”, IEEE Access, 2024. (Reference [25])
- [C10] **H. Otroshi Shahreza**, V. Krivokuća Hahn, and S. Marcel, “On the Recognition Performance of BioHashing on state-of-the-art Face Recognition models”, In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), 2021. (Reference [26])
- [C11] **H. Otroshi Shahreza** and S. Marcel, “Breaking Template Protection: Reconstruction of Face Images from Protected Facial Templates”, In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2024. (Reference [27])
- [A1] **H. Otroshi Shahreza**, P. Melzi, D. Osorio-Roig, C. Rathgeb, C. Busch, S. Marcel, R. Tolosana, and R. Vera-Rodriguez, “Benchmarking of cancelable biometrics for deep templates”, arXiv preprint arXiv:2302.13286, 2023. (Reference [28])

In summary, the contributions of this thesis are published in **6 peer-reviewed journal** papers and **11 peer-reviewed conference** papers. In addition to these publications, my research

## Chapter 1. Introduction

---

during the course of my Ph.D. is extended to other topics in the field of biometrics which are not discussed in this thesis, composing a total of **10 peer-reviewed journal** papers and **22 peer-reviewed conference** papers. The additional papers can also be categorized as follows:

- **Additional Group 1: Lensless Imaging**

- [J7] **H. Otroshi Shahreza**, A. Veuthey, and S. Marcel, “Towards High-Resolution Face Image Generation From Coded Aperture Camera”, IEEE Sensors Letters, 2023. (Reference [29])
- [C12] **H. Otroshi Shahreza**, A. Veuthey, and S. Marcel, “Face recognition using lensless camera”, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. (Reference [30])

- **Additional Group 2: Protection of Biometric Templates**

- [C13] D. Osorio-Roig, C. Rathgeb, **H. Otroshi Shahreza**, C. Busch, and S. Marcel, “Indexing Protected Deep Face Templates by Frequent Binary Patterns”, In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), 2022. (Reference [31])
- [C14] P. Melzi, **H. Otroshi Shahreza**, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, J. Fierrez, S. Marcel, and C. Busch, “Multi-IVE: Privacy Enhancement of Multiple Soft-Biometrics in Face Embeddings”, In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). 2023. (Reference [32])
- [C15] **H. Otroshi Shahreza**, A. Bassit, S. Marcel, and Raymond Veldhuis “Remote Cancelable Biometric System for Verification and Identification Applications”, In Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG), 2023. (Reference [33])

- **Additional Group 3: Synthetic Data for Face Recognition**

- [J8] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, W. Zhao, X. Zhu, Z. Yan, X. Zhang, J. Wu, Z. Lei, S. Tripathi, M. Kothari, M. Zama, D. Deb, B. Biesseck, P. Vidal, R. Granada, G. Fickel, G. Führ, D. Menotti, A. Unnervik, A. George, C. Ecabert, **H. Otroshi Shahreza**, P. Rahimi, S. Marcel, I. Sarridis, C. Koutlis, G. Baltso, S. Papadopoulos, C. Diou, N. Domenico, G. Borghi, L. Pellegrini, E. Mas-Candela, Á. Sánchez-Pérez, A. Atzori, F. Boutros, N. Damer, G. Fenu, M. Marras, “FRCSyn-onGoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems”, Information Fusion, 2024. (Reference [34])
- [C16] **H. Otroshi Shahreza**, A. George and S. Marcel, “SynthDistill: Face Recognition with Knowledge Distillation from Synthetic Data”, In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), 2023. (Reference [35])

- [C17] **H. Otroshi Shahreza**, C. Ecabert, A. George, A. Unnervik, S. Marcel, N. Domenico, G. Borghi, D. Maltoni, F. Boutros, J. Vogel, N. Damer, Á. Sánchez-Pérez, E. Mas-Candela, J. Calvo-Zaragoza, B. Biesseck, P. Vidal, R. Granada, D. Menotti, I. DeAndres-Tame, S. Cava, S. Concas, P. Melzi, R. Tolosana, R. Vera-Rodriguez, G. Perelli, G. Orrù, G. Marcialis, J. Fierrez, “SDFR: Synthetic Data for Face Recognition Competition”, In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2024. (Reference [36])
- [C18] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, W. Zhao, X. Zhu, Z. Yan, X. Zhang, J. Wu, Z. Lei, S. Tripathi, M. Kothari, M. Zama, D. Deb, B. Biesseck, P. Vidal, R. Granada, G. Fickel, G. Führ, D. Menotti, A. Unnervik, A. George, C. Ecabert, **H. Otroshi Shahreza**, P. Rahimi, S. Marcel, I. Sarridis, C. Koutlis, G. Baltso, S. Papadopoulos, C. Diou, N. Domenico, G. Borghi, L. Pellegrini, E. Mas-Candela, Á. Sánchez-Pérez, A. Atzori, F. Boutros, N. Damer, G. Fenu, M. Marras, “FRCSyn Challenge at WACV 2024: Face Recognition Challenge in the Era of Synthetic Data.” In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), 2024. (Reference [37])
- [C19] I. DeAndres-Tame, R. Tolosana, P. Melzi, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, A. Morales, J. Fierrez, J. Ortega-Garcia, Z. Zhong, Y. Huang, Y. Mi, S. Ding, S. Zhou, S. He, L. Fu, H. Cong, R. Zhang, Z. Xiao, E. Smirnov, A. Pimenov, A. Grigorev, D. Timoshenko, K. Asfaw, C. Low, H. Liu, C. Wang, Q. Zuo, Z. He, **H. Otroshi Shahreza**, A. George, A. Unnervik, P. Rahimi, S. Marcel, P. Neto, M. Huber, J. Kolf, N. Damer, F. Boutros, J. Cardoso, A. Sequeira, A. Atzori, G. Fenu, M. Marras, V. Štruc, J. Yu, Z. Li, J. Li, W. Zhao, Z. Lei, X. Zhu, X. Zhang, B. Biesseck, Pedro V. Coelho, R. Granada, D. Menotti, “Second Edition FRCSyn Challenge at CVPR 2024: Face Recognition Challenge in the Era of Synthetic Data”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024. (Reference [38])
- [A2] D. Geissbühler, **H. Otroshi Shahreza**, and S. Marcel, “Synthetic Face Datasets Generation via Latent Space Exploration from Brownian Identity Diffusion”, arXiv preprint arXiv:2405.00228, 2024. (Reference [39])

### • Additional Group 4: Efficient Face Recognition

- [J9] A. George, C. Ecabert, **H. Otroshi Shahreza**, K. Kotwal and S. Marcel, “Edge-Face: Efficient Face Recognition Model for Edge Devices”, IEEE Transactions on Biometrics, Behavior, and Identity Science, 2024. (Reference [40])
- [C20] J. N. Kolf, F. Boutros, J. Elliesen, M. Theuerkauf, N. Damer, M. Alansari, O. A. Hay, S. Alansari, S. Javed, N. Werghi, K. Grm, V. Štruc, F. Alonso-Fernandez, K. H. Diaz, J. Bigun, A. George, C. Ecabert, **H. Otroshi Shahreza**, K. Kotwal, S. Marcel, I. Medvedev, B. Jin, D. Nunes, A. Hassanpour, P. Khatiwada, A. A. Toor, and B. Yang,

## Chapter 1. Introduction

---

“EFAr 2023: Efficient face recognition competition”, In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), 2023. (Reference [41])

- **Additional Group 5: Different Attacks against Face Recognition Systems**

- [C21] L. Colbois, **H. Otroshi Shahreza**, S. Marcel, “Approximating Optimal Morphing Attacks using Template Inversion”, IEEE International Joint Conference on Biometrics (IJCB), 2023. (Reference [42])
- [A3] A. Unnervik, **H. Otroshi Shahreza**, A. George, S. Marcel, “Model Pairing Using Embedding Translation for Backdoor Attack Detection on Open-Set Classification Tasks”, arXiv preprint arXiv:2402.18718, 2024. (Reference [43])

- **Additional Group 6: Image Forensics**

- [J10] A. Daryani, M. Mirmahdi, A. Hassanpour, **H. Otroshi Shahreza**, B. Yang, J. Fierrez, “IRL-Net: Inpainted Region Localization Network via Spatial Attention”, IEEE Access, 2023. (Reference [44])

- **Additional Group 7: Large Language Models and Biometrics**

- [C22] A. Hassanpour, Y. Kowsari, **H. Otroshi Shahreza**, B. Yang, and S. Marcel, “ChatGPT and biometrics: an assessment of face recognition, gender detection, and age estimation capabilities”, In Proceedings of the IEEE International Conference on Image Processing (ICIP), 2024. (Reference [45])

It is noteworthy that this thesis and all the experimental results reported in this thesis are geared towards reproducible research. Details about reproducing the results and link to source code as well as published materials are available in Appendix E.

## 1.3 Thesis Outline

This thesis is composed of 6 chapters, which are organized as follows:

In this chapter (chapter 1), the motivations, objectives, and contributions of this thesis are briefly summarized.

Chapter 2 gives an overview of related work in the literature on different topics covered in this thesis, including template inversion, biometric template protection schemes, and evaluation of biometric template protection. In addition, this chapter introduces all the databases used in this work, which are used in the experimental sections of other chapters.

Chapter 3 presents vulnerability evaluation of face recognition systems to template inversion attacks. The chapter proposes different methods to reconstruct face from facial templates. It includes methods for low-resolution face reconstruction, high-resolution face reconstruction,

and geometry-aware 3D face reconstruction. We also show the vulnerability of face recognition systems to presentation attacks using reconstructed face images.

Chapter 4 presents new template protection mechanisms. This chapter introduces MLP-Hash, a new cancelable biometric scheme that works based on random multi-layer perceptrons (MLP). It also discusses a new hybrid template protection mechanism that leverages cancelable biometric and homomorphic encryption. The proposed template protection schemes can be used in systems of different biometric modalities (face, voice, finger vein, etc). Finally, this chapter presents a new method to protect and enhance vascular biometric recognition methods using Biohashing and an auto-encoder network.

Chapter 5 delves into the evaluation of biometric template protection schemes. The chapter starts with benchmarking cancelable biometric methods, and uses existing methods in the literature to evaluate different schemes. The limitations of existing measures are discussed and motivates the necessity of new measures for the evaluation of protected templates. The chapter introduces a learning-based approach to invert protected templates and evaluate the invertibility of protected templates. In addition, a new measure for evaluating the linkability of protected templates is proposed. The proposed linkability measure is extended for the case of multiple leaked protected templates, across different systems or within a single uni-modal or multi-modal system.

Chapter 6 concludes this thesis by providing a summary of the major contributions and findings. Potential directions for future work are also discussed.



## 2 Related Work

In this chapter, we review some of the works from the literature that are linked to the problems studied in this thesis. Section 2.1 review previous works in the literature on template inversion in biometric systems. Section 2.2 focuses on template protection in biometric systems and describe different biometric template protection schemes. Section 2.3 summarizes the methods for evaluating template protection in the literature. Finally, Section 2.4 describes datasets and feature extractors for different biometric modalities that are used in our experiments in the next chapters.

### 2.1 Template Inversion in Biometric Systems

In this section, we review previous work in the literature on template inversion (TI) attack against biometric systems. While there are different works on template inversion for different biometric characteristics, such as fingerprint images [46] and vascular images [47], the majority of publications on this topic are related to reconstruction of face from face images which is discussed in this section.

Among different types of potential attacks against face recognition (FR) systems that have been studied in the literature [48]–[50], TI attack can jeopardize both the privacy and security of the enrolled users. In the TI attack, the adversary gains access to the database of a face recognition system and tries to invert the stored templates to reconstruct the underlying face images. The reconstructed face images can reveal privacy-sensitive information about the users, including age, gender, ethnicity, etc. In addition, the adversary can use the reconstructed face image to impersonate the enrolled users and enter the system.

Methods in the literature for face reconstruction in TI attacks against FR systems can be generally categorized from different aspects, including the basis of the method (optimization/learning-based), the type of attack (*whitebox/blackbox* attack), and the resolution of reconstructed face images (high/low resolution). However, all previous methods generate 2D images in TI attacks against FR systems.

## Chapter 2. Related Work

Table 2.1: Comparison with related works on template inversion attacks against face recognition systems.

Reference	Method Basis	2D/3D	Resolution	Whitebox/ Blackbox	Transferability Evaluation	Practical Presentation Attack Evaluation	Available Source Code
[51]	1) optimization 2) learning	2D	low	whitebox	✗	✗	✗
[52]	learning	2D	low	both*	✗	✗	✗
[53]	learning	2D	low	blackbox	✗	✗	✓
[54]	learning	2D	low	both <sup>†</sup>	✗	✗	✗
[55]	learning	2D	low	both <sup>†</sup>	✗	✗	✗
[56]	learning + optimization	2D	low	blackbox	✗	✗	✗
[57]	learning	2D	high	blackbox	✗	✗	✓
[58]	1) learning 2) optimization	2D	high	blackbox	✗	✗	✗
[59]	learning	2D	low + high <sup>§</sup>	blackbox	✗	✗	✗
[60]	optimization	2D	high	blackbox	✗	✗	✓
[61]	optimization	2D	high	blackbox	✗	✗	✓
[Ours] (Section 3.1)	learning	2D	low	both <sup>‡</sup>	✗	✗	✓
[Ours] (Section 3.2)	learning	2D	high	both <sup>‡</sup>	✗	✗	✓
[Ours] (Section 3.3)	learning	2D	high	both <sup>‡</sup>	✗	✗	✓
[Ours] (Section 3.4)	learning	3D	high	both <sup>‡</sup>	✓	✓	✓

\*The method is based on the *whitebox* attack, and is also applied in the *blackbox* scenario by removing a loss term that required the FR model.

<sup>†</sup>The method is based on the *whitebox* attack, and is extended to the *blackbox* with knowledge distillation of the FR model.

<sup>§</sup>They first reconstruct low-resolution face images and then apply a super-resolution model to generate high resolution face images.

<sup>‡</sup>The method is based on the *whitebox* attack, and is extended to *blackbox* using a different FR model.

Several methods have been proposed for reconstructing low-resolution 2D face images from facial templates [51]–[56]. In [51], authors proposed two *whitebox* methods to reconstruct 2D low-resolution face images from facial templates. In the first method (optimization-based), they used a gradient-descent-based approach on a guiding image or random (noise) image to find an image that minimizes the distance between the template of the reconstructed face image and the target template. In addition, they used several regularization terms to generate a smooth image, including the total variation and Laplacian pyramid gradient normalization [62] of the reconstructed face image. In their learning-based method, they trained a deconvolutional neural network with the same loss function as in their optimization-based method, to generate reconstructed face images. For the evaluation of their method, they only discussed the visual reconstruction quality and did not provide any security evaluation on a FR system.

In [52], authors trained a multi-layer perceptron (MLP), to find the facial landmark coordinates, and a convolutional neural network (CNN), to generate face texture from the given facial template. Next, they used a differentiable warping to combine the estimated landmarks (from MLP) with the generated textures (from CNN) and reconstruct low-resolution 2D face images. They used their method for *whitebox* and *blackbox* attacks. In the *whitebox* attack, they trained their MLP and CNN by minimizing the distance between templates of the original



and reconstructed face images. However, for their *blackbox* attack, they trained MLP and CNN separately, and used the warping in the inference only. For the security evaluation, they only reported the histogram of scores between the templates extracted from the original and reconstructed face images and compared it with the histogram of genuine scores.

In [53], authors proposed a learning-based method to generate low-resolution 2D face images in the *blackbox* attacks against FR systems. They proposed two new deconvolutional networks, called NbBlock-A and NbBlock-B, and trained them with either pixel loss ( $\ell_1$  norm of pixel-level reconstruction error) or perceptual loss (distance of middle layers of VGG-19 [63] when given the original and reconstructed face images). For the security evaluation, they considered two types of attacks and evaluated vulnerability of FR systems. In their first type of attack, they compared the templates extracted from the original and reconstructed face images, and in their second type of attack, they compared the templates extracted from reconstructed images with templates of a different face image of the same user.

In [54] and [55], a same method based on bijection learning is used to train GAN networks with PO-GAN [64] and TransGAN [65] structures, respectively. In the *whitebox* attack, authors minimized the distance between target templates and templates extracted from the reconstructed face images using the FR model. To extend their method to the *blackbox* attack, they proposed to use the distillation of knowledge to train a student network that mimics the target FR model. However, they did not report any detail about the training of the student network (e.g., network structure, etc.) nor published their source code. For the security evaluation, they reported the matching accuracy between the reconstructed image and another original image in each positive pair in their TI attacks. However, they did not evaluate the vulnerability of FR systems at different threshold configurations.

In [56], authors proposed a 3-step method to reconstruct low-resolution 2D face images in the *blackbox* attack. In the first step, they trained a general face generator network based on GAN. In the second step, they trained a MLP to map the templates to the templates of a known (i.e., *whitebox* knowledge) FR model. In the third step, they used an optimization on the latent space of their face generator to find a latent code that can generate a face image that maximizes two terms; the cosine similarity between the templates (mapped templates and the templates extracted by the known FR model) and the discriminator score (for being a real face image). For their security evaluation, they reported the adversary's success attack rate (SAR), but they did not specify the system's operation configuration, such as the system's recognition false match rate (FMR).

In contrast to the most works in the literature that generate low-resolution 2D face images, recently few methods are proposed for high-resolution 2D face reconstruction. In [57], authors proposed a learning-based method to reconstruct high-resolution 2D face images in the *blackbox* attack. They used a pretrained StyleGAN2 [66] to generate some face images and extracted the templates using the FR model. Then, they trained a MLP to map facial templates to the input latent codes of StyleGAN2 [66]. For the security analysis, they considered two types

## Chapter 2. Related Work

---

of attacks as defined in [53] and evaluated the vulnerability of FR systems. They also evaluated their reconstructed face images with a commercial-off-the-shelf (COTS) presentation attack detection (PAD) system, also known as face liveness detection in their paper. However, the authors did not perform a *practical* presentation attack scenario, in which the images should have been recaptured by camera prior to be fed to the COTS PAD. Similarly, in [58], authors proposed a learning-based method for high-resolution 2D face reconstruction in the *blackbox* attack. They learned three mapping networks from the facial templates to three separate parts in the intermediate latent space of StyleGAN. Each of these mapping networks is composed of a MLP and is used to reconstruct coarse to fine information of face image. They also proposed to find this mapping with optimization instead of learning the mapping networks. For the security analysis, they did not report success attack rate (percentage) for any configuration. They only reported the histogram of the distance between templates of reconstructed and original face images and compared it with the histogram of templates for random pair of images (i.e., zero-effort impostor).

In [59], authors used a learning-based method based on a conditional denoising diffusion probabilistic model to reconstruct 2D face images in *blackbox* attack. They used the conditional diffusion model in [67] and iteratively denoise an input Gaussian noise conditioned with facial templates to generate low resolution (i.e.,  $64 \times 64$ ) face images from facial templates. Then, they used a super-resolution network to generate face images with a higher resolution (i.e.,  $256 \times 256$ ). Compared to other learning-based methods, their proposed method is relatively very slow<sup>1</sup>, because of iterative reconstruction in the inference stage. In addition, compared to other methods, that directly generate high-resolution face images, the method in [59] first reconstructs low-resolution face images and then uses a super-resolution to generate high-resolution face images. For security analysis, similar to [54], [55], they reported the matching accuracy between the reconstructed and a different original image in each positive pair, and did not evaluate the vulnerability of FR systems at different threshold configurations.

In [60], authors proposed a optimization on the latent vector (i.e., input noise) of StyleGAN2 [66] to find latent codes which generates face images with templates similar to the target templates. They solved this optimization with a grid-search and simulated annealing [68] approach for the *blackbox* scenario. However, since their method is computationally expensive<sup>2</sup>, they evaluated their method on only 20 face images and reported distance between the original templates and templates of the reconstructed face images. Along the same lines, in [61] authors considered a similar optimization to [60] on the latent vector of StyleGAN2 [66], but instead of grid-search, they solved the optimization using the standard genetic algorithm [69] for the *blackbox* attack. For the security analysis, they also considered two types of attacks as defined in [53] and evaluated the vulnerability of FR systems. Moreover, they evaluated their reconstructed face images using three COTS PAD systems (called liveness detection in their

---

<sup>1</sup>They reported four minutes to reconstruct  $64 \times 64$  face images and the super-resolution to  $256 \times 256$  on a NVIDIA RTX 3090 GPU.

<sup>2</sup>They reported 5 minutes execution time to reconstruct each single image on a system equipped with graphic card.

paper). However, similar to [57], they did not perform a *practical* presentation attack scenario by recapturing the reconstructed face images.

In this thesis, three different methods to reconstruct face images from facial templates are proposed in Chapter 3. Table 2.1 compares our methods with the previous works in the literature. To our knowledge, we proposed the first method on 3D face reconstruction from facial templates (which are extracted from 2D face recognition models). Moreover, in contrast to most works in the literature, two of our method generates high-resolution face images. We also propose our methods for both *whitebox* and *blackbox* attacks against FR systems and evaluate the *transferability* of our reconstructed face images (which has not been reported before for TI attacks). Furthermore, we perform practical presentation attacks against FR systems using the reconstructed face images. Last but not least, the source code of all the experiments of our methods are publicly available in the corresponding paper package to facilitate the reproducibility of our work.

We should note that in addition to the aforementioned methods, which work with DNN-based FR systems, there are a few methods on classic FR systems (before the era of deep learning), such as [70] and [71]. In 2007, Mohanty *et al.* [71] considered a blackbox scenario and assumed that the template comparison function is also known. They proposed to use a set of face images and calculated the templates of these images using the FR model. Next, they calculated the similarity of these templates and found an affine space where face images can approximate the original similarity matrix. Then, they compared these templates with the target template and used the similarity values to embed the target subject in the affine space, which can lead to reconstruction of the face image using the inverse of the affine transformation. In 2013, Mignon and Jurie [70] proposed a method to invert face templates in a blackbox scenario by training a radial basis function (RBF) least squares regression to map facial templates to the face representation in eigenspace (composed of whitened eigenfaces), and then convert them to the face image. They reported the classification error rate of inverting features from Local Quantized Patterns (LQP) [72]. Later, Mai *et al.* [53] re-implemented the method of Mignon and Jurie [70] and evaluated it on the inversion of FaceNet [73] features. However, the inversion performance on FaceNet features was not promising [53].

## 2.2 Biometric Template Protection Schemes

To protect biometric data, several biometric template protection (BTP) schemes have been proposed in the literature [74], [75]. In general for a BTP scheme, the ISO/IEC 24745 standard [9] establishes four main requirements:

- *Renewability*: We should be able to generate a new protected template for a subject whose template is compromised.
- *Unlinkability*: There should be no leakage of information between different protected templates of the same (unprotected) biometric template.

## Chapter 2. Related Work

---

- *Irreversibility*: It should be computationally infeasible to reconstruct the original biometric templates or eventually the biometric data from the protected templates.
- *Recognition Performance*: The protected templates should be discriminative enough to be used for accurate recognition.

BTP methods are commonly categorized into *cancelable biometrics* and *biometric cryptosystems*, which are described in Section 2.2.1 and Section 2.2.2, respectively. Some work also use homomorphic encryption for biometric template protection which is described in Section 2.2.3.

### 2.2.1 Cancelable Biometrics

In cancelable biometrics (CB) protection methods (such as BioHashing [76], Multi-Layer Perceptron (MLP) Hashing [23], Index-of-Maximum (IoM) Hashing [77], etc.) a transformation function (which is dependent on a *key*) is often utilized to generate protected templates, and then the recognition is carried out by comparing the transformed templates [3], [8], [75].

In this thesis, different cancelable biometrics schemes are used, which are described in the following:

#### BioHashing

This CB scheme generates a user-specific orthogonal matrix and multiply it to the unprotected embeddings. The result is then binarized to generate binary-valued protected templates [76]. Then, for comparing probe and reference templates, BioHashing generates binary-valued templates and use Hamming distance for calculating the comparison scores during recognition [76].

#### Bloom Filters

Bloom Filters were initially proposed for 2D iris-codes [78], where each binary iris-code is divided into multiple blocks. Then, the XOR operation is applied to each column of the block with a user-specific secret, and the resulting vector is mapped to a decimal number. Finally, the index corresponding to the decimal number is turned to 1 in the protected template. To apply Bloom Filters on the features extracted by DNNs, we first quantise the features and divide each quantised feature to multiple blocks and XOR each block with the user-specific secret. For comparing probe and reference templates, the Hamming distance is normalized by the number of 1's in protected templates.

### Index-of-Maximum Hashing

BTP schemes based on Index-of-Maximum (IoM) apply several user-specific transformations on the unprotected embeddings, and then considers index of maximum values for each transformation to build the protected template. We consider two variants of IoM-Hashing, including uniformly random permutation-based hashing (shortly IoM-URP) and Gaussian random projection-based hashing (IoM-GRP). In IoM-GRP, several user-specific Gaussian projections are applied on the unprotected template and index of maximum in each projection is used to build the protected template. In IoM-URP, multiple uniform random permutations are applied and their Hadamard product is calculated. Then index of maximum in different windows are used to generate the protected template. The IoM-Hashing-based schemes generate integer-valued protected templates and use the average number of collisions for calculating the comparison scores during recognition [77].

### 2.2.2 Biometric Cryptosystems

in biometric cryptosystems (such as fuzzy vault [79], fuzzy commitment [80], etc.), a key is either generated from a biometric template or bound with a biometric template. Then, recognition is performed based on the correct generation or retrieval of the key [7], [81], [82]. Both fuzzy commitment [80] and fuzzy vault [79] schemes enable an error-tolerant protection of biometric templates. While fuzzy commitment was used for protection of biometric templates in different work [83], [84], it has been shown that the fuzzy commitment scheme can not effectively protect against to a certain kind of linkage attack, i.e. the correlation attack [85]. This conflicts with the unlinkability as well as irreversibility requirements for a template protection scheme defined by the ISO/IEC 24745 standard [9]. Similarly, different security analyses have shown that the original fuzzy vault scheme is also vulnerable to correlation attack [86], [87]. In [88], an improved version of the fuzzy vault scheme was proposed which generates smaller records, and was later shown that can be effectively robust to correlation attack [89]. The fuzzy vault scheme has been applied to different biometric characteristics [90], [91]. Recently, [82] proposed a method based on fuzzy vault to protect face features extracted by deep neural networks.

### 2.2.3 Homomorphic Encryption

As an alternative to cancelable biometrics and biometric cryptosystems, we could use *Homomorphic Encryption HE*, which allows computations to be carried out on the ciphertexts and generates encrypted results. Then, the results (which are in the encrypted domain) can be decrypted to plaintexts. The decrypted results will exactly correspond to the results of the operations performed in the plaintext domain (i.e., on the original features). Based on the allowed types and numbers of operations on the ciphertexts, the HE-based systems can also be categorized into three classes:

## Chapter 2. Related Work

---

- *Partially Homomorphic Encryption (PHE)* systems that support only one type of arithmetic operation (i.e., either addition or multiplication) in the encrypted domain with no limit on the number of operations (e.g., [92], [93]).
- *Somewhat Homomorphic Encryption (SWHE)* systems that support both addition and multiplication but with a limited number of operations (e.g., [94], [95]).
- *Fully Homomorphic Encryption (FHE)* systems that support additions and multiplications in the encrypted domain with no limit on the number of operations (e.g., [96]–[98]).

Several works in the literature have used FHE for template protection in biometric recognition systems. In [99], a secure face verification system based on HE was proposed. The application of HE for face identification and face verification were also investigated in [100] and [101], respectively. In [102], HE was used for iris verification and identification. In [103], a multimodal biometric verification system using HE was proposed, and different fusion strategies were studied. There are also some works in the literature that focus on reducing computation and enhancing the efficiency of applying HE in biometric recognition systems. For example, [99] and [104] performed dimensionality reduction on the biometric features prior to HE. In [105] and [106], indexing and searching in the system's database was enhanced (for the identification application).

## 2.3 Evaluation of Biometric Template Protection

Following the requirements of the ISO/IEC 24745 standard [9] (mentioned in Section 2.2), different properties should be evaluated for each biometric template protection scheme. The ISO/IEC ISO30136 [10] also outline evaluation protocols for evaluating each requirement of biometric template protection scheme, which are explained in the following sections.

### 2.3.1 Irreversibility

Irreversibility of protected templates is one of the most important requirements of each template protection mechanisms. Since the inversion process of protected templates may differ in different BTP schemes, each of the proposed template protection methods in the literature has used an ad-hoc approach to investigate the inversion of protected templates. Therefore, the irreversibility of protected templates is often measured empirically according to the computational complexity of inversion attacks and the number of guesses involved in recovering the original template [107], [108]. These methods have been proposed for specific applications and hardly generalize to all BTP schemes. The ISO/IEC ISO30136 [10] defines different threat models and suggests different threat models to evaluate the security and invertibility of protected templates, including:

## 2.3 Evaluation of Biometric Template Protection

Table 2.2: Summary of generic methods in the literature for evaluating the linkability of protected biometric templates.

Ref.	Measure Basis	Quantify Linkability Degree of system	Assumptions
[110]	Accuracy (EER)	✗	-
[111]	Accuracy (ROC)	✗	-
[112]	Accuracy (ROC)	✗	-
[113]	Accuracy (combined)	✗	-
[114]	Accuracy (CMC)	✗	-
[115]	Score distribution	✗	-
[116]	Score distribution	✗	closed-set
[117]	Score distribution	✓	prior probabilities
[Ours] (Section 5.3)	Score distribution	✓	-

- *Naive threat model*: This is the case where the adversary has a blackbox knowledge about the protection method, with no further information about the underlying algorithm and any associated secrets. We can also assume that the adversary has access to a small set of protected templates (not a large biometric database).
- *Standard threat model*: This is the case where the adversary has full knowledge of the protection algorithm, but does not know the secrets and, therefore, cannot execute submodules that require the secrets.
- *Full-disclosure threat model*: This threat model refers to the case where the adversary knows everything about the system, including all the submodules and secrets.

### 2.3.2 Unlinkability

According to the ISO/IEC 30136 standard [10], the more precise definition of unlinkability is: “**unlinkability** is the difficulty of **distinguishing** between Auxiliary Data (AD)s and/or Pseudonymous Identifiers (PIs) of two Renewable Biometric References (RBRs) generated from **the same subject's** characteristic and ADs and/or PIs of two RBRs generated from **different subjects' characteristics**” [emphasis added]. In the context of BTP, we can extend the definition of *mated* and *non-mated* pairs in the ISO/IEC 2382-37 standard [109] as:

- *mated*: two protected templates are mated if they correspond to *the same subject* (they can be either from the same sample or different samples) and with different keys.
- *non-mated*: two protected templates are non-mated if they correspond to *different subjects* with different keys.

Therefore, to gain the unlinkability criterion, the protected templates should be such that an adversary would not be able to distinguish mated and non-mated protected pairs.

## Chapter 2. Related Work

---

Table 2.2 summarizes the previous works in the literature which have used a generic method to evaluate the linkability of protected biometric templates. Buhan *et al.* [110] considered a biometric cryptosystem and compared the recognition accuracy of the system in terms of Equal Error Rate (EER) in two scenarios: i) templates protected with a single key (i.e., regular recognition accuracy analysis), ii) templates protected with different keys (i.e., unlinkability analysis). While the increase of EER implies some degree of unlinkability, the unlinkability is not quantified in their work. Kelkboom *et al.* [111] considered similar scenarios and compared the recognition performance of the system in terms of the Receiving Operating Characteristic (ROC). Then, if the recognition accuracy shown by the ROC curve decreases, the system is considered to be unlinkable. However, the unlinkability can neither be quantified in this approach. Similarly, Nagar *et al.* [112] found the ROC curve of matching templates with different keys to evaluate the unlinkability of the system.

Piciuccio *et al.* [113] used a similar approach to [110]–[112], but combined the results of regular analysis and unlinkability analysis. They plotted the True Match Rate (TMR) in the unlinkability analysis<sup>3</sup> versus the system's False Non-Match Rate (FNMR) in the regular analysis. Their method does not evaluate the True Match Rate (TMR) in the unlinkability analysis, and the degree of general unlinkability is also not quantified in their method. Along the same lines, Rua *et al.* [114] found the probability that the adversary can determine the correct identity in a top-N list and plotted this probability similar to Cumulative Match Curves (CMC). Then, as an evaluation of the unlinkability of the system, they compared this plot with the curve corresponding to the probability of random guesses being correct (i.e., full unlinkability). However, their method does not provide a single number to quantify the general unlinkability of the system.

In contrast to [110]–[114] which have evaluated unlinkability based on accuracy metrics, [115]–[117] considered score distributions in their unlinkability evaluations. In [115], Ferrara *et al.* calculated three distributions of scores, including scores of templates with different keys from: 1) the same sample, 2) different samples of the same subject, and 3) samples of different subjects. Then, according to visual comparisons of these distributions, they evaluate the unlinkability of templates. Wang and Hu [116] used the latter two score distributions only and evaluated unlinkability by visual comparison of these distributions. Gomez-Barrero *et al.* [117] proposed two quantitative measures (local and global) based on score distributions. Similar to [116], they considered two distributions of scores for mated and non-mated pairs. Then, as their local measure for each score, they consider the difference in conditional probabilities of the hypothesis of being mated and the hypothesis of being non-mated. To calculate their local measure, they use the likelihood ratio of mated and non-mated hypotheses and the ratio of prior probabilities. For their global measure, they considered the conditional expectation of their local measure over score values. The global measure ( $D_{\rightarrow}^{sys}$ ) proposed in [117] was the first quantitative evaluation that measures the degree of unlinkability of the biometric systems. It is also properly defined and bounded in the  $[0, 1]$  interval. However, it has several drawbacks

---

<sup>3</sup>referred as Renewable Template Matching Rate (RTMR) in their work.



that we discuss in Section 5.3.2.2.

In addition to prior work on linkability, there is ample work on general privacy measures in information theory and computer science communities [118], [119]. The most prominent notions of privacy are  $\epsilon$ -differential privacy and  $(\epsilon, \delta)$ -differential privacy which were developed for the database release problem [120]–[122]. The main idea behind this approach is to control the influence of a single database entry on the output of differentially private queries. BTP schemes have been studied from the differential privacy perspective in [123] where a differentially private distributed face-recognition system is proposed. A hypothesis testing perspective on differential privacy has been introduced in [124] and extended in [125]. In particular, [125] show that  $(\epsilon, \delta)$ -differential privacy guarantees could be interpreted as bounds on the ROC curves of appropriately defined hypothesis tests.

Another recent measure of interest is maximal leakage which seeks to control the adversary's ability to refine his or her estimate of any function of data [126], [127]. Maximal leakage has been recently discussed in the context of hypothesis testing: Privacy-utility trade-offs using maximal leakage as a privacy metric and the type II (false alarm) error exponent as the utility metric have been studied in [128]; In [129] the so-called “noiseless privacy” is related to hypothesis testing and to maximal leakage; And, maximal leakage is used to bound generalization errors of learning algorithms in [130].

In Section 5.3 of this thesis, we propose a new measure for evaluating the linkability of protected biometric templates. Our proposed measure combines the work on maximal leakage from information-theoretic literature [126], [127] with the perspective on global linkability introduced in [117]. Since our proposed measure is based on a well-studied information measure, it inherits many of the theoretic properties of this measure. In addition, we show that the proposed linkability measure has an appealing operational interpretation in terms of hypothesis testing that the adversary could perform on a pair of protected templates. This hypothesis testing interpretation of our proposed measure makes it consistent with the definition of linkability in the ISO/IEC 30136 standard [10]. We further compare our proposed measure to a similar measure based on differential privacy [122] and show that the differential privacy-based measure is too strict for the linkability application. Finally, the experimental implementation of our proposed measure shows that it gives intuitively correct linkability scores across different BTP schemes, biometric characteristics, and scoring functions.

### 2.3.3 Recognition Performance

To report the recognition performance of biometric systems, we can use different metrics which evaluate the recognition error rates of the system, such as the Equal Error Rate (EER) of the system or the False Non-Match Rate (FNMR) at the decision thresholds corresponding to different False Match Rates (FMRs) of the system, e.g., FMRs of 1% and 0.1%. We can also

plot the Detection Error Trade-off (DET)<sup>4</sup> or Receiver Operating Characteristic (ROC) curves to compare the recognition performance of the system at different FMR values. Similarly, to report the recognition performance of protected biometric systems, we can use similar recognition error rates and also plot ROC or DET curves. In addition, we can consider two scenarios and evaluate the recognition performance of protected biometric systems:

- *unknown-key* or *normal* scenario: it is the expected case in practice, where we generate protected templates with user-specific keys.
- *known-key* or *stolen-token* scenario: we assume that keys are disclosed, hence we evaluate the recognition performance considering the same key for each user.

## 2.4 Biometric Feature Extractors and Datasets

In this thesis, different biometric characteristics (including face recognition, voice recognition, vascular recognition, and iris recognition) are used. For each biometric characteristic, the state-of-the-art feature extractor and dataset are used, which are described in this section.

### 2.4.1 Face Recognition

For face recognition, we use different models in our experiments, including iResNet100-ArcFace [132], ElasticFace [133], FaceNet [73], and InceptionResnetV2-CenterLoss [134] models as different feature extractors. We also use face recognition models with different backbones, including MobileFaceNet [135], ResNet [136], SE-ResNet [137], HRNet [138], EfficientNet [139], GhostNet [140], AttentionNet [141], TF-NAS [142], ResNeSt [143], ReXNet [144], RepVGG [145], LightCNN [146], and Swin [147]. Moreover, we use face recognition models which are trained with different loss functions (different heads), including AM-Softmax [148], ArcFace [132], AdaCos [149], AdaM-Softmax [150], CircleLoss [151], CurricularFace [152], MV-Softmax [153], NPCFace [154], and MagFace [155]. We should note that when we perform experiments for different backbones, we use the same head (i.e., MV-Softmax [153]) for all the models, and when we perform experiments for different heads, we use the same backbone (i.e., MobileFaceNet [135]) for all the models. We use the FaceX-Zoo<sup>5</sup> [156] toolbox for models for experiments with different heads and different backbones. All these models are trained on the MS-Celeb-1M dataset [157].

For face recognition datasets, we use the MOBIO [158], Labeled Faces in the Wild (LFW) [159], and AgeDB [160] datasets. The MOBIO [158] dataset is a bimodal dataset including face and voice data taken with mobile and laptop devices from 150 individuals, captured in 12 sessions (6-11 samples in each session) for each subject. The LFW database includes 13,233 face images

---

<sup>4</sup>The ISO/IEC 19795-1 standard [131] suggests using DET curves for reporting and comparing recognition performance of biometric systems.

<sup>5</sup>Available at <https://github.com/JDAI-CV/FaceX-Zoo>

of 5,749 people, among which 1,680 people have two or more images. The AgeDB datasets consists of 16,488 images of 568 famous people, with 29 images in average per each subject and the average age range of 50.3 years.

In this thesis, we also use the Flickr-Faces-HQ (FFHQ) dataset [161] for training face reconstruction network. The FFHQ dataset consists of 70,000 face images (with no identity labels) and includes variations in terms of age, ethnicity, accessories, and image background.

### 2.4.2 Voice Recognition

For voice (speaker) recognition, we extract deep features using the ECAPA-TDNN [162] model. We also use the voice data of the MOBIO dataset (the same dataset described for face) for voice recognition experiments. Since MOBIO dataset is a bimodal dataset, we use this dataset for unimodal biometric systems (face recognition and voice recognition separately) and also for multi-modal experiments.

### 2.4.3 Vascular Recognition

In this thesis, we consider different vascular images, including finger vein, palm vein, and wrist vein images. To extract features from the vascular images, we use the Wide Line Detector (WLD) [163], Repeated Line Tracking (RLT) [164], and Maximum Curvature (MC) [165] algorithms to extract biometric features from vascular images. In addition, as a deep-learning-based feature extractor for finger vein images, we use the modified version of DenseNet-161 [166] based on the approach proposed in [167].

For finger vein recognition, we use the SDUMLA [168] dataset which consists of finger vein images of 106 individuals. For each individual, 6 instances (index, middle and ring fingers of both hands) are considered, and for each instance 6 samples are captured. We assume each instance as a different subject. We also use the finger vein UTFVP dataset [169] in our experiments. This dataset contains in total 1440 finger vein images which have been collected from 60 subjects. We used the training (subjects 1-10, 240 images), development (subjects 11-28, 432 images) and evaluation (subjects 29-60, 768 images) subsets of this dataset. For other vascular biometric modalities, we use PUT Vein dataset [170] which includes palm vein and wrist vein images. This dataset consists of 2400 images, where half of images contains palm vein images (1200 images) and another half contains wrist vein images (another 1200 images) which were acquired from both hands of 50 individuals.

### 2.4.4 Iris Recognition

We consider iris images obtained with near-infrared (NIR), and use the CASIA Thousand database [171], composed of 1000 individuals, each one represented with 10 images from both their right and left eyes. To extract deep features from iris images, we use the DenseNet-201

## Chapter 2. Related Work

---

network proposed in [172], specifically fine-tuned for iris recognition with the samples of the first 750 individuals of CASIA Thousand. We pre-process the samples of the remaining 250 individuals according to the procedure proposed in [172]. Compared to the other biometric characteristics, iris data requires additional steps for data selection. Samples that contain glasses are identified according to the method proposed in [173] and filtered out. After a manual check of pre-processed images, we also discard three samples that provide an incorrect segmentation. Then, we assign the samples obtained from the right and left eyes of the same original individual to different subjects in the experimental part. We rule out subjects with less than six samples, and limit to the first six the set of samples considered for the remaining subjects.

### 3 Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

In this chapter, we focus on face recognition systems (as one of the most popular biometric systems), and explore their vulnerability to template inversion attacks. We consider a face recognition system, where several users are enrolled and investigate some possible attacks based on breaching face templates from the database of the system. We start from simple scenario in section 3.1, where the adversary tries to reconstruct low-resolution face images and inject the reconstructed images to the system. We investigate the vulnerability of face recognition systems to our attacks in both whitebox and blackbox scenarios. We extend our approach for the case where the adversary gains access to a portion of face templates and tries to perform the template inversion attack in Appendix B. In section 3.2 and section 3.3, we assume that the adversary aims to reconstruct high-resolution face images from leaked facial templates. We consider two approaches, based on *real* and *synthetic* data to generate high-resolution face images in section 3.2 and 3.3, respectively. In section 3.4, we consider a more complex approach where the adversary aims to reconstruct 3D face from facial templates. The 3D reconstruction enables adversary to optimize reconstruction pose, and therefore increase the success attack rate. We also consider the scenario where the adversary uses the reconstructed face from templates of one system to attack another system and evaluate the transferability of the attack. The adversary can use the reconstructed images to inject to the system (injection attack), or in a real-world attack may perform presentation attack using reconstructed face images. Therefore, we extend our evaluation in this section to presentation attack and evaluate the vulnerability of the systems in practical scenarios. We further evaluate the performance of a presentation attack detection (PAD) system to our presentation attacks using the reconstructed face images from our TI attacks in Appendix C.

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

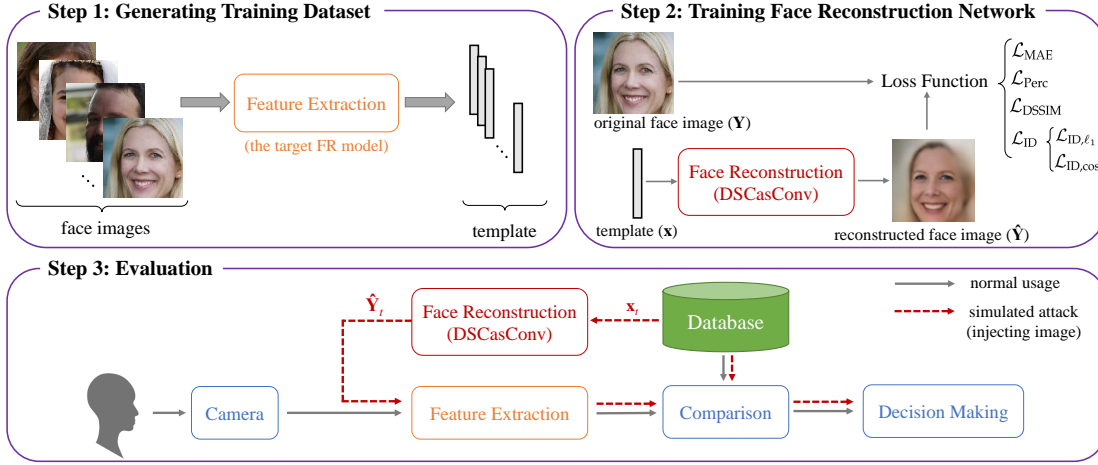


Figure 3.1: Block diagram of the proposed framework for evaluating vulnerability of FR system to TI attacks with low-resolution face reconstruction

#### 3.1 Low-resolution Face Reconstruction

In this section, we propose a new neural network to reconstruct face images from face templates. Our network is based on a new block, called “enhanced deconvolution using cascaded convolution and skip connections” (shortly, *DSCasConv*). *DSCasConv* generates outputs with the same size as the deconvolution layer, while enhancing the deconvolution output through several cascaded convolutional layers with skip connections. Each convolutional layer in *DSCasConv* enhances the output of previous layer, and in total, the residual cascaded convolutional layers recover and improve local dependencies of deconvolution output. We train our face reconstruction network with a multi-term loss function. In particular, to enhance the face reconstruction during training, we use a loss term that minimizes the mean absolute error and cosine distance between the templates extracted from the original and reconstructed face images. For vulnerability evaluation, we consider the scenario where the adversary gains access to the system’s database and wants to enter the system by inverting the enrolled face templates and impersonating the enrolled users by employing the reconstructed face images. We assume the case where the adversary can bypass the camera and inject the reconstructed face image into the system. Our experiments show that the face images reconstructed by our template inversion (TI) network can be recognized by FR systems, and they also provide crucial identity information about the users enrolled in the system database. In our experiments, we evaluate the vulnerability of SOTA FR methods to our TI attack. We consider different SOTA FR methods with various network structures (different “backbones”) as well as with different loss functions (different “heads”).

### 3.1.1 Proposed Method

In this section, we introduce our proposed framework (as depicted in Fig. 3.1), to evaluate the vulnerability of FR systems to a TI attack. First, we describe the threat model that we consider in this study, in section 3.1.1.1. Then, we propose our face reconstruction network in section 3.1.1.2. Finally, we describe our vulnerability evaluation protocol in section 3.1.1.3.

#### 3.1.1.1 Threat Model

To evaluate the vulnerability of a given FR system, we need to first define the threat model that characterises the adversary on which we wish to base our vulnerability analysis [10], [48], [53]. Considering a real-world attack scenario<sup>1</sup>, we define the following properties for the adversary:

- *Adversary's goal:* The adversary aims to impersonate a user enrolled in the target FR system.
- *Adversary's knowledge:* The adversary is assumed to have only the following information:
  1. The target face templates  $\mathbf{t}_i$  of a user enrolled in the system's database.
  2. The whitebox or blackbox knowledge of the feature extraction model  $F_{\text{template}}(\cdot)$ , which can be used to generate a face template  $\mathbf{t} = F_{\text{template}}(\mathbf{I})$  from a face image  $\mathbf{I}$ . In the case of blackbox attack, the adversary is assumed to have whitebox knowledge of another FR model  $F_{\text{proxy}}(\cdot)$ .

However, the adversary is assumed not to have any other information, neither from the target system nor from the target template. In particular, the adversary is assumed not to have the following information:

- Any additional information or prior knowledge about the identity of the target template, including age, gender, etc.
  - Any information about the training set of the feature extraction model. Therefore, the adversary is assumed not to be able to use the same (or similar) dataset to learn TI.
  - Any knowledge about the comparison and decision making submodules of the target system, including the similarity score function and the system's decision threshold.
- *Adversary's capability:* The adversary is assumed to have the following capabilities:
    1. The adversary can inject the reconstructed face images directly into the feature extractor of the target system and bypass the sensor (i.e., camera).

<sup>1</sup>We should note that our threat model is aligned with the *full-disclosure* scenario defined in the ISO/IEC 30136 standard [10] for evaluating invertibility of biometric templates.

2. For each target template, the adversary is allowed only one attempt to enter the system.

- *Adversary's strategy:* Under the above assumptions, the adversary can reconstruct face image  $\hat{\mathbf{I}} = G_W(\mathbf{t}_{\text{leaked}})$  from the target template  $\mathbf{t}_{\text{leaked}}$  using a reconstruction model  $G_W(\cdot)$ . Then, the adversary can use the reconstructed face image  $\hat{\mathbf{I}}$  as a query to enter the target FR system. The weights  $W$  of the reconstruction model  $G_W(\cdot)$  can be learned using a dataset of face images and their corresponding face templates extracted by the feature extractor model  $F_{\text{template}}(\cdot)$ .

### 3.1.1.2 Face Reconstruction Method

In this section, we introduce our neural network  $G_W(\cdot)$  to invert a face template  $\mathbf{t}$  and reconstruct a face image  $\hat{\mathbf{I}} = G_W(\mathbf{t})$ . To train our network, we first need to generate training data, including pairs of face template  $\mathbf{t}$  and face image  $\mathbf{I}$ , which is described in section 3.1.1.2. We train our network with a multi-term loss function as described in section 3.1.1.2. Our network structure, which includes multiple DSCasConv blocks, is described in section 3.1.1.2.

**Generating Training Data** To generate our training dataset, let us assume that we have a dataset of face images  $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$ , where  $\mathbf{I}_i$  and  $N$  indicate the  $i$ th image and the total number of images, respectively. Also, let us assume that we have the coordinates of the facial landmarks (e.g., eyes),  $\mathbf{L}_i$ , for image  $\mathbf{I}_i$  in the dataset  $\mathcal{I}$ . In addition, let  $A(\mathbf{I}_i, \mathbf{L}_i)$  denote the function used prior to feature extraction, which accepts  $\mathbf{I}_i$  and  $\mathbf{L}_i$  as inputs and returns an aligned and cropped face image. We can generate our training dataset of aligned images and associated templates,  $\mathcal{D}$ , by extracting facial features from all face images in the face dataset  $\mathcal{I}$  after alignment:

$$\mathcal{D} = \{([F_{\text{template}} \circ A](\mathbf{I}_i, \mathbf{L}_i), A(\mathbf{I}_i, \mathbf{L}_i))\}_{i=1}^N, \quad (3.1)$$

where  $F_{\text{template}}(\cdot)$  indicates the feature extraction model of the FR system.

However, our experiments show that an augmented dataset can improve the generalization and performance of our TI network. So, we augment the dataset  $\mathcal{I}$  and generate a new dataset  $\mathcal{I}_a = \{\mathbf{I}_{a,j}\}_{j=1}^M$  using a random transformation function  $T(\cdot)$ , where  $M$  is the number of images in the augmented dataset  $\mathcal{I}_a$ , and  $\mathbf{I}_{a,j}$  is an image augmented by  $T(\cdot)$ , i.e.,  $\mathbf{I}_{a,j} = T(\mathbf{I}_k)$ ,  $0 \leq k < N$ . To increase the robustness of the inversion network, we also add random noise to the coordinates of landmarks before feature extraction. However, in our augmented training dataset,  $\mathcal{D}_a$ , we pair up each extracted feature with its corresponding aligned face using the original values of landmark coordinates (i.e., without noise). Hence, we generate the augmented training dataset  $\mathcal{D}_a$  as follows:

$$\mathcal{D}_a = \{([F \circ A](\mathbf{I}_{a,j}, \mathbf{L}_j + \mathbf{N}_j), A(\mathbf{I}_{a,j}, \mathbf{L}_j))\}_{j=1}^M, \quad (3.2)$$

where  $\mathbf{N}_j$  is random noise with a uniform distribution in  $(-\delta, \delta)$ . We consider  $\delta = 4$  in our ex-



periments. It is worth mentioning that using the original facial landmark coordinates in feature extraction helps our inversion network to generate all images with the same alignment and thus eliminates the additional work of finding the landmark coordinates and reconstructing face images in different locations. Adding noise to the coordinates of landmarks before feature extraction increases the variation in the template space, thereby enhancing the robustness of our inversion network to the alignment.

It is worth mentioning that for the random transformation function  $T(\cdot)$ , we use a random combination of the following transformations: random PCA color augmentation [174], randomly adjusting contrast, randomly adjusting brightness, Gaussian blurring (random standard deviation), and JPEG compression (random compression rate).

For simplicity, we denote the augmented training dataset in Eq. 3.2 with  $\mathcal{D}_a = \{(\mathbf{t}_j, \mathbf{I}_j)\}_{j=1}^M$ , where  $\mathbf{t}_j = [F \circ A](\mathbf{I}_{a,j}, \mathbf{L}_j + \mathbf{N}_j)$  and  $\mathbf{I}_j = A(\mathbf{I}_{a,j}, \mathbf{L}_j)$ .

**Loss Function** To train the reconstruction network,  $G_W(\cdot)$ , we optimize its weights  $W$  using loss function  $\mathcal{L}(\cdot, \cdot)$  on the augmented training dataset  $\mathcal{D}_a$  such that:

$$W^* = \underset{W}{\operatorname{argmin}} \mathbb{E}_{(\mathbf{t}, \mathbf{I}) \in \mathcal{D}_a} \mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}), \quad (3.3)$$

where  $(\mathbf{t}, \mathbf{I})$  denotes a pair of face template  $\mathbf{t}$  and face image  $\mathbf{I}$  in our augmented training dataset  $\mathcal{D}_a$ , and  $\hat{\mathbf{I}} = G_W(\mathbf{t})$  is the reconstructed face image. To this end, we define a multi-term loss function including:

- *Mean Absolute Error (MAE)*: To help the network to generate a face image that is similar to the original image, we use the Mean Absolute Error (MAE) loss term, which includes the  $\ell_1$ -norm of the reconstruction error:

$$\mathcal{L}_{\text{MAE}}(\hat{\mathbf{I}}, \mathbf{I}) = \|\hat{\mathbf{I}} - \mathbf{I}\|_1 \quad (3.4)$$

- *Dissimilarity Structural Index Metric (DSSIM)*: In addition to MAE of the reconstructed face, we maximize the objective quality of the reconstructed image. To this end, we use the Similarity Structural Index Metric (SSIM) [175] of the reconstructed image and optimize the DSSIM loss term as follows:

$$\mathcal{L}_{\text{DSSIM}}(\hat{\mathbf{I}}, \mathbf{I}) = \frac{1 - \text{SSIM}(\hat{\mathbf{I}}, \mathbf{I})}{2} \quad (3.5)$$

- *Perceptual Loss*: In addition to DSSIM, we use a perceptual loss by minimizing the  $\ell_1$ -norm of the difference between the features extracted from  $\mathbf{I}$  and  $\hat{\mathbf{I}}$  by a convolutional neural network trained on ImageNet [176]. This helps the model to generate images with a similar representation of the original image (i.e., face). We use a pre-trained VGG-16 [177] model and consider its middle feature maps to calculate the perceptual

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

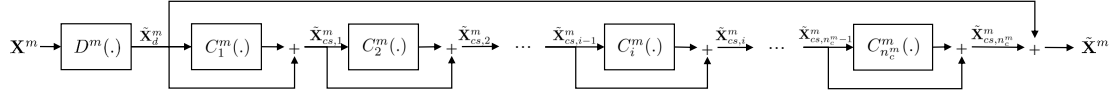


Figure 3.2: Block diagram of the  $m$ th DSCasConv block

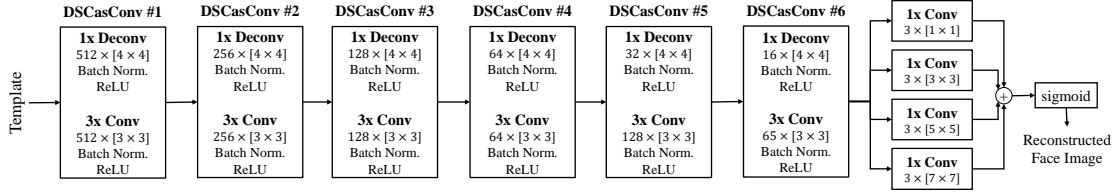


Figure 3.3: Structure of the proposed face reconstruction network

loss. Let us denote the feature mapping of VGG-16 as  $P(\cdot)$ . Then the perceptual loss can be expressed as:

$$\mathcal{L}_{\text{Perc}}(\hat{\mathbf{I}}, \mathbf{I}) = \|P(\hat{\mathbf{I}}) - P(\mathbf{I})\|_1 \quad (3.6)$$

- *ID loss*: In addition to the aforementioned loss terms, we would like the templates extracted from the reconstructed face image to be close to the templates of the original face image, to increase the chances of a successful TI attack. So, we minimize the distance between the templates extracted from the reconstructed face  $\hat{\mathbf{I}}$  and original face  $\mathbf{I}$ . To achieve this, we minimize the  $\ell_1$ -norm of the difference between the extracted features and also maximize their cosine similarity. Thereby, we define ID loss with two terms as follows:

$$\begin{aligned} \mathcal{L}_{\text{ID}}(\hat{\mathbf{I}}, \mathbf{I}) &= \mathcal{L}_{\text{ID}, \ell_1}(\hat{\mathbf{I}}, \mathbf{I}) + \mathcal{L}_{\text{ID}, \cos}(\hat{\mathbf{I}}, \mathbf{I}) \\ &= \underbrace{\|F_{\text{proxy}}(\hat{\mathbf{I}}) - F_{\text{proxy}}(\mathbf{I})\|_1}_{\text{minimizing } \ell_1\text{-norm}} + \underbrace{\frac{-F_{\text{proxy}}(\hat{\mathbf{I}}) \cdot F_{\text{proxy}}(\mathbf{I})}{\|F_{\text{proxy}}(\hat{\mathbf{I}})\|_2 \cdot \|F_{\text{proxy}}(\mathbf{I})\|_2}}_{\text{maximizing cosine similarity}}, \end{aligned} \quad (3.7)$$

where  $F_{\text{proxy}}(\cdot)$  is the FR model which the adversary has access to. In whitebox attack  $F_{\text{proxy}}(\cdot)$  is the same as  $F_{\text{template}}(\cdot)$  (i.e.,  $F_{\text{proxy}} = F_{\text{template}}$ ), but in blackbox attack  $F_{\text{proxy}}(\cdot)$  is a different model.

We use a linear combination of the above loss terms as the total loss:

$$\mathcal{L} = \mathcal{L}_{\text{MAE}} + \alpha_1 \mathcal{L}_{\text{DSSIM}} + \alpha_2 \mathcal{L}_{\text{Perc}} + \alpha_3 \mathcal{L}_{\text{ID}}, \quad (3.8)$$

where  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  are weights of each loss term. To train the proposed face reconstruction network with this multi-term loss function, we use the Adam [178] optimizer with the initial learning rate of  $10^{-3}$  and decrease the learning rate by a factor of 0.5 every 10 epochs.

**Network Structure** To reconstruct face images from their corresponding templates, we can use deconvolution layers to build our face reconstruction network (e.g., [51], [53]). However, since deconvolution acts as upsampling, the deconvolution output suffers from insufficient detail [53]. In particular, similar to upsampling, in deconvolution, local dependencies are weakened, which leads to a blurry output. To address these shortcomings, we propose a new block, called “enhanced deconvolution using cascaded convolutions and skip connections” (shortly, *DSCasConv*), which generates outputs with the same size as the deconvolution layer. In the proposed block, we apply multiple cascaded convolutional layers on the deconvolution output. Considering the significant effect of residual learning [136], we also use skip connections to further enhance the output by forcing the convolutional layers to learn residuals. In addition, skip connections can enhance the gradient flow and prevent gradient vanishing problem [136] in our deep *DSCasConv* block. Hence, we use a skip connection for each of the convolution layers as well as a skip connection over all cascaded convolution layers in our *DSCasConv* block. Using cascaded convolutional layers with skip connections after deconvolution can recover and improve local dependencies, and therefore result in sharper and more detailed outputs. Indeed, each residual convolutional layer enhances the result of previous layers, and in total, the residual cascaded convolutional layers improves the deconvolution output.

To formulate the proposed block, let  $\mathbf{x}^m$  and  $\tilde{\mathbf{x}}^m$  denote the input and output of the  $m$ th *DSCasConv* block, respectively. Assume that the  $m$ th *DSCasConv* block consists of the deconvolution operator,  $D^m(\cdot)$ , and a set of convolution operations,  $\mathcal{C}_m = \{C_i^m(\cdot) | i = 1, \dots, n_c^m\}$ , where  $n_c^m$  is the number of convolution operations at the  $m$ th *DSCasConv* block. Let us define  $\tilde{\mathbf{x}}_d^m = D^m(\mathbf{x}^m)$  as the output of the deconvolution operator and  $\tilde{\mathbf{x}}_{cs,i}^m$  as the summation of the  $i$ th convolution operation and its corresponding skip connection as follows:

$$\tilde{\mathbf{x}}_{cs,i}^m = \begin{cases} \tilde{\mathbf{x}}_{cs,i-1}^m + C_i^m(\tilde{\mathbf{x}}_{cs,i-1}^m) & \text{if } i > 0 \\ \tilde{\mathbf{x}}_d^m & \text{if } i = 0 \end{cases}. \quad (3.9)$$

Then, we define the output of the  $m$ th *DSCasConv* block as below:

$$\tilde{\mathbf{x}}^m = \tilde{\mathbf{x}}_d^m + \tilde{\mathbf{x}}_{cs,n_c^m}^m \quad (3.10)$$

Fig.3.2 illustrates the block diagram of the  $m$ th *DSCasConv* block.

We build our network with 6 *DSCasConv* blocks (each includes 1 deconvolution and 3 convolution operations) with 512, 256, 128, 64, 32, 16 filters, respectively. For deconvolution and convolution layers in our *DSCasConv* blocks, we use kernels of sizes 4 and 3, respectively. In addition, we use Batch Normalization [179] and a rectified linear unit (ReLU) after each deconvolution and convolution operation in our *DSCasConv* blocks. Finally, we pass the output of the last *DSCasConv* block to 4 parallel convolutional layers with different kernel sizes (including sizes of 1, 3, 5, and 7), which are added and passed through a sigmoid function

## Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

---

to generate the final reconstructed face. Fig. 3.3 depicts the general structure of our face reconstruction neural network.

### 3.1.1.3 Vulnerability Evaluation to TI Attacks

To evaluate the vulnerability of a FR model to a TI attack, we consider a real-world scenario based on the assumptions described in section 3.1.1.1. To this end, we consider a FR system with several enrolled users. Based on our threat model, we assume that the adversary can access the system's database and aims to invert the enrolled templates to reconstruct the underlying face images. The images are then injected into the feature extractor to impersonate the enrolled users, and therefore, enter the system. We should note that there might be several templates for each user stored in the system's database, but according to our threat model, the adversary does not have any knowledge of this.

To train the face reconstruction network, based on our threat model we assume that the adversary does not have any information about the training set of the feature extractor. Therefore, the adversary uses a different training dataset for the inversion model. So, we have three different datasets in our evaluation: 1) dataset used for training the feature extractor (by the system designer), 2) dataset enrolled in the FR system, and 3) dataset used for training the TI model (by the adversary).

We assume that the adversary trains the inversion model (as in section 3.1.1.2), then uses the inversion model to invert the enrolled templates, and injects the reconstructed face images into the system. Again according to our threat model, the adversary is allowed only one attempt to enter the system for each inverted target template. Hence, in our evaluation, for each template stored in the system's database, we invert the template and reconstruct the face image. Then, we extract the template from the reconstructed face image and find the system's comparison score between this template and the corresponding reference templates. If the score is greater than the system's threshold, the attack is considered successful, meaning that the adversary can enter the system. Hence, similar to the Receiver Operating Characteristic (ROC) plot, we use the comparison scores of inverted templates to plot the Success Attack Rate (SAR) versus the system's False Match Rate (FMR) by changing the system's decision threshold in the impostors' score range. This plot can be used to compare the vulnerability of FR models at different FMRs. Fig. 3.1 illustrates the general block diagram of the proposed TI evaluation framework.

### 3.1.2 Experiments

In this section, we describe the experiments used to evaluate our framework and analyze the vulnerability of SOTA FR models using this framework. First, in section 3.1.2.1, we describe the experimental setup and the FR models used in our experiments. Next, as a primary experiment, we evaluate the vulnerability of the iResNet100-ArcFace [132] model, which is a well-known

SOTA FR model, in section 3.1.2.2. Next, we compare our proposed face reconstruction method with previous methods in the literature in section 3.1.2.3 against the iResNet100-ArcFace [132] model. Then, in section 3.1.2.4, we provide an ablation study on the effect of our network structure (section 3.1.2.4) and our loss function (section 3.1.2.4) on the performance of the primary experiment. After evaluating our proposed framework, in section 3.1.2.5 we evaluate the vulnerability of different SOTA FR models, with different backbones and different heads, using our framework. Finally, we discuss the experimental findings in section 3.1.2.6.

#### 3.1.2.1 Experimental Setup

In our experiments, we evaluate the vulnerability of SOTA FR models to our TI attack. For the primary experiment, we use the iResNet100-ArcFace<sup>2</sup> [132] model to study the performance of our face reconstruction network and compare the proposed face reconstruction network with previous methods in the literature. Furthermore, we evaluate the vulnerability of different FR model backbones, including MobileFaceNet [135], ResNet [136], SE-ResNet [137], HR-Net [138], EfficientNet [139], GhostNet [140], AttentionNet [141], TF-NAS [142], ResNeSt [180], ReXNet [144], RepVGG [145], LightCNN [146], and Swin [147]. Moreover, we evaluate the vulnerability of FR models which are trained with different loss functions (different heads), including AM-Softmax [148], ArcFace [132], AdaCos [149], AdaM-Softmax [150], CircleLoss [151], CurricularFace [152], MV-Softmax [153], NPCFace [154], and MagFace [155]. We should note that when we compare the vulnerability of different backbones, we use the same head (i.e., MV-Softmax [153]) for all the models, and when we compare the vulnerability of different heads, we use the same backbone (i.e., MobileFaceNet [135]) for all the models. In addition to iResNet100-ArcFace and also SOTA backbones and heads, we also evaluate the vulnerability of four other pretrained SOTA FR models in the literature, including ElasticFace [133], AdaFace [181], EdgeFace [40], and PocketNet [182]. All the aforementioned FR models, except EdgeFace, are trained on the MS-Celeb-1M dataset [157], and EdgeFace is trained on the WebFace260M dataset [183].

To train our face reconstruction network for each FR model, we use the FFHQ dataset [161], which consists of 70,000 face images (with no identity labels) and includes variations in terms of age, ethnicity, accessories, and image background. For a fair comparison, we train each of the face reconstruction networks with 90 epochs. We use a random 90% portion of the FFHQ dataset to generate the training set as explained in section 3.1.1.2, by generating 10 augmented images for each original image. The remaining 10% portion is used for validation. After training our face reconstruction network, we use the MOBIO [158], Labeled Faces in the Wild (LFW) [159], and AgeDB [160] datasets to build the FR systems and evaluate their vulnerability in our framework. The MOBIO dataset is a bimodal dataset including audio and face data acquired using mobile devices from 150 people. We use the *development* subset of the *mobio-all* protocol<sup>3</sup> in our experiments. The LFW database includes 13,233 images of

<sup>2</sup>iResNet100 backbone trained with ArcFace loss.

<sup>3</sup>The implementation of the *mobio-all* protocol for the MOBIO dataset is available at <https://gitlab.idiap.ch/>

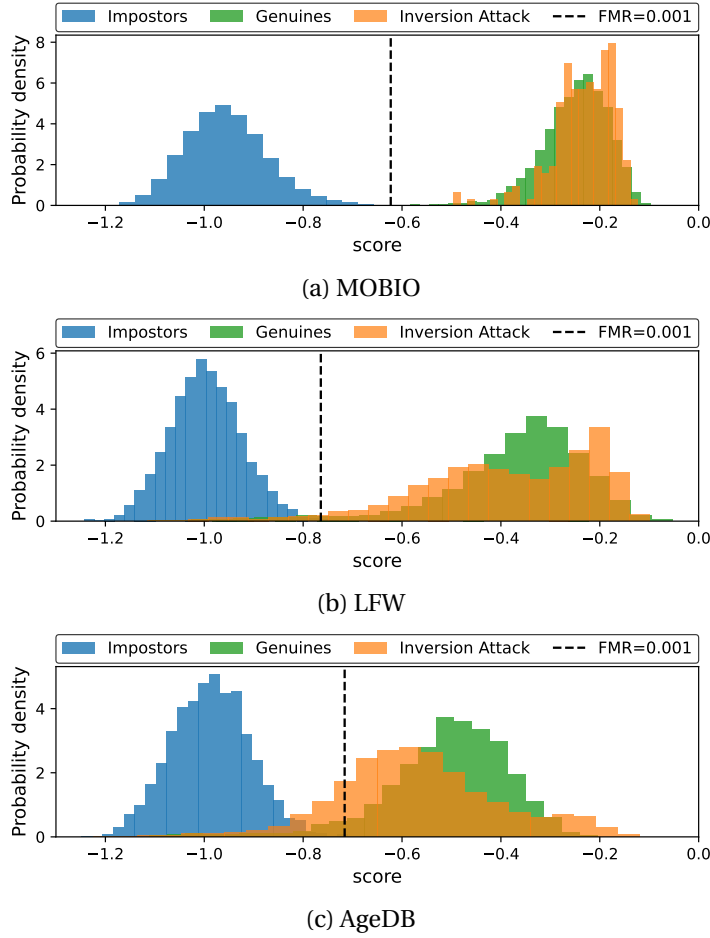


Figure 3.4: Histogram of scores (negative cosine distance) for genuine templates and impostor templates, as well as scores for the templates extracted from reconstructed face images when injecting the reconstructed face images into the system using (a) MOBIO, (b) LFW, and (c) AgeDB datasets.

5,749 people, where 1,680 people have two or more images. We use the *View 2* protocol<sup>4</sup> to evaluate the models. The AgeDB [160] dataset contains 16,488 images of 568 famous people. The minimum and maximum age in this dataset are 1 and 101, respectively, and the average age range for each subject is 50.3 years. We use the *30-year* protocol (i.e., the age difference of each pair's faces is equal to 30) in our experiments.

In our experiments, we consider  $\alpha_1 = 0.75$ ,  $\alpha_2 = 0.02$ , and  $\alpha_3 = 0.025$  as the weights of our loss function in Eq. 3.8. We also provide an ablation study on the effect of these weights in section 3.1.2.4.

bob/bob.db.mobio

<sup>4</sup>The implementation of the *View 2* protocol for the LFW dataset is available at <https://gitlab.idiap.ch/bob/bob.db.lfw>

Table 3.1: Recognition performance (TMR) and vulnerability to template inversion attack (SAR) of the iResNet100-ArcFace model, at  $\text{FMR} = 10^{-2}$  and  $\text{FMR} = 10^{-3}$  on the MOBIO, LFW, and AgeDB datasets.

Dataset	FMR = $10^{-2}$		FMR = $10^{-3}$	
	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
MOBIO	100.00	100.00	100.00	100.00
LFW	97.70	97.44	96.63	96.48
AgeDB	96.97	94.03	93.77	84.60

### 3.1.2.2 Primary Experiment

As a primary experiment, we evaluate the vulnerability of the iResNet100-ArcFace [132] model in whitebox TI attacks using our framework, on the MOBIO, LFW, and AgeDB datasets. To this end, as discussed in section 3.1.1.3, we train our face reconstruction network using the FFHQ dataset. Next, we use our face reconstruction network to invert the facial templates stored in the FR system and inject the reconstructed face images into the system. Fig. 3.4 illustrates the histogram of scores for genuine and impostor templates, as well as scores for the templates extracted from reconstructed face images when injecting the reconstructed face images into the FR system. As this figure shows, the scores between the templates extracted from the reconstructed face images and the reference templates enrolled in the system’s database are close to the genuine scores and, therefore, are likely to break the system. Table 3.1 reports the recognition performance and vulnerability of the iResNet100-ArcFace model to a TI attack in terms of True Match Rate (TMR) and SAR, respectively, at  $\text{FMR} = 10^{-2}$  and  $\text{FMR} = 10^{-3}$ , on the MOBIO, LFW, and AgeDB datasets. As this table shows, while the iResNet100-ArcFace model achieves high recognition performance on the MOBIO, LFW, and AgeDB datasets, it is seriously vulnerable to our whitebox TI attack.

### 3.1.2.3 Comparison with Previous Methods

We compare the performance of our face reconstruction network with the methods proposed<sup>5</sup> in [11], [13], [15], [51], [53], [57], [60], [61] in whitebox and blackbox attacks against iResNet100-ArcFace model. As mentioned in Table 2.1, [53], [57], [60] are based on the blackbox scenario and [51] is based on the whitebox scenario. Methods in [11], [13], [15] can be used for both whitebox and blackbox scenarios, and we use their both whitebox and blackbox implementations in our experiment. For each method, we train a separate face reconstruction networks using FFHQ dataset, and use the trained network to invert facial templates stored in the FR system. We use the reconstructed face images to inject into the system and evaluate vulnerability of iResNet100-ArcFace on the MOBIO, LFW, and AgeDB datasets using our framework. Table 3.2 compares the performance of these different methods in terms of SAR in attacks

<sup>5</sup>The source codes of other methods in Table 2.1 are not publicly available and we could not reproduce their results.

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.2: Comparison of different face reconstruction methods against the iResNet100-ArcFace model in terms of SAR at  $\text{FMR}=10^{-2}$  and  $\text{FMR}=10^{-3}$  on the MOBIO, LFW, and AgeDB datasets.

Attack Type	Method	MOBIO		LFW		AgeDB	
		$\text{FMR}=10^{-2}$	$\text{FMR}=10^{-3}$	$\text{FMR}=10^{-2}$	$\text{FMR}=10^{-3}$	$\text{FMR}=10^{-2}$	$\text{FMR}=10^{-3}$
whitebox	Zhmoginov and Sandler [51]	100.00	85.71	93.01	85.87	82.40	54.18
	Otroshi Shahreza and Marcel [15]	100.00	92.38	93.64	86.82	75.87	62.08
	Otroshi Shahreza and Marcel [13]	96.67	84.76	92.31	85.01	73.78	60.03
	GaFaR [11]	95.71	82.86	89.27	79.84	63.30	48.94
	GaFaR+GS [11]	97.62	85.23	90.77	82.52	67.86	53.10
	GaFaR+CO [11]	97.62	89.05	91.87	84.25	71.95	58.00
	[Ours]	<b>100.00</b>	<b>100.00</b>	<b>97.44</b>	<b>96.48</b>	<b>94.03</b>	<b>84.60</b>
blackbox	NBNetA-M [53]	2.86	0.48	16.06	5.35	3.75	0.42
	NBNetA-P [53]	15.24	1.43	29.61	12.16	8.26	1.14
	NBNetB-M [53]	19.52	0.48	26.10	10.79	6.06	0.49
	NBNetB-P [53]	51.90	21.9	60.33	39.49	21.56	5.18
	Dong <i>et al.</i> [57]	24.85	3.33	28.21	13.21	9.56	1.80
	Vendrow and Vendrow [60]	69.52	29.05	77.00	57.70	40.94	16.56
	Dong <i>et al.</i> [61]	85.71	58.57	87.25	75.31	58.79	43.22
	Otroshi Shahreza and Marcel [15]	89.52	81.90	87.93	77.13	59.46	43.87
	Otroshi Shahreza and Marcel [13]	88.57	80.00	84.69	71.31	57.94	42.35
	GaFaR [11]	77.14	47.62	71.24	51.78	36.00	21.67
	GaFaR+GS [11]	85.71	64.76	78.12	61.56	44.37	28.95
	[Ours]	<b>99.52</b>	<b>97.14</b>	<b>94.98</b>	<b>91.08</b>	<b>82.78</b>	<b>70.38</b>

against the iResNet100-ArcFace model on the MOBIO, LFW, and AgeDB datasets. For blackbox attacks in methods which use another model as proxy (such as ours and [11], [13], [15]), we use ElasticFace as the proxy model. As this table shows, our proposed method outperforms previous methods in [11], [13], [15], [51], [53], [57], [60], [61] in both whitebox and blackbox attacks. In particular, our method achieves better performance compared to low-resolution face reconstruction methods (i.e., [51], [53]). This is achieved as the result of our network structure and loss function which are further studied in section 3.1.2.4. Comparing other methods which generate high-resolution face images (i.e., [13], [15], [57], [60], [61]) or 3D face (i.e., [11]), the reconstructed face images by our method still achieve superior performance, which elaborates on a trade-off between resolution of reconstructed face images and performance in terms of SAR in our method and these methods in the literature. Among different face reconstruction methods in the literature, [11], [15] achieve the best performance after our proposed method in attack against FR systems and generate high-resolution and 3D face, respectively. We present and further discuss our methods proposed in [11], [13], [15] in sections 3.2, 3.3, and 3.4, respectively. In the remainder of our experiments in section 3.1.2, we focus on whitebox TI attacks against FR systems.



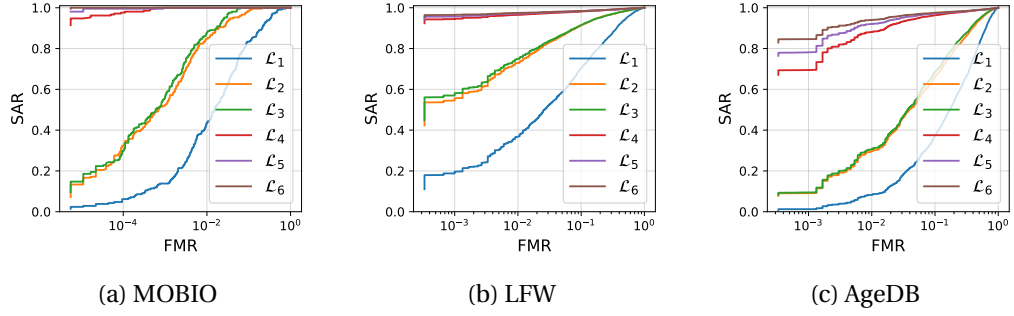


Figure 3.5: Effect of loss function (as in Eq.3.11) on the performance of our face reconstruction network in a template inversion attack against a face recognition system based on the iResNet100-ArcFace model evaluated using (a) MOBIO, (b) LFW, and (c) AgeDB datasets.

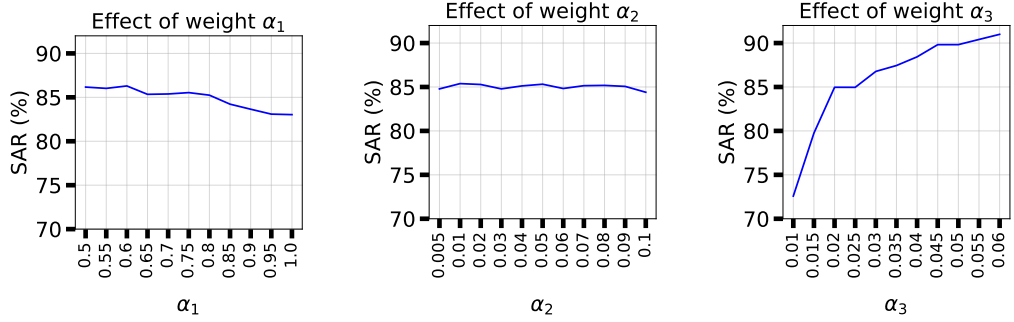


Figure 3.6: Effect of weights ( $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$ ) in our loss function (as in Eq. 3.8) on the performance of our face reconstruction network in a template inversion attack against a face recognition system based on the iResNet100-ArcFace model evaluated on the AgeDB dataset.

#### 3.1.2.4 Ablation Study

In this section, we describe our ablation study on the effect of network structure (section 3.1.2.4) and loss function (section 3.1.2.4) on the face reconstruction performance. In our ablation studies, we consider a FR system based on the iResNet100-ArcFace model, and we evaluate the SAR over different values of the system's FMR using the MOBIO, LFW, and AgeDB datasets.

**Ablation Study on the Loss Function** To evaluate the effect of each loss term, we train different face reconstruction networks with different loss functions. Considering our multi-term loss function in Eq. 3.8, let us denote linear combinations of different loss terms as

follows:

$$\begin{aligned}
\mathcal{L}_1 &= \mathcal{L}_{\text{MAE}}, \\
\mathcal{L}_2 &= \mathcal{L}_{\text{MAE}} + \alpha_1 \mathcal{L}_{\text{DSSIM}}, \\
\mathcal{L}_3 &= \mathcal{L}_{\text{MAE}} + \alpha_1 \mathcal{L}_{\text{DSSIM}} + \alpha_2 \mathcal{L}_{\text{Perc}}, \\
\mathcal{L}_4 &= \mathcal{L}_{\text{MAE}} + \alpha_1 \mathcal{L}_{\text{DSSIM}} + \alpha_2 \mathcal{L}_{\text{Perc}} + \alpha_3 \mathcal{L}_{\text{ID}, \ell_1}, \\
\mathcal{L}_5 &= \mathcal{L}_{\text{MAE}} + \alpha_1 \mathcal{L}_{\text{DSSIM}} + \alpha_2 \mathcal{L}_{\text{Perc}} + \alpha_3 \mathcal{L}_{\text{ID}, \cos}, \\
\mathcal{L}_6 &= \mathcal{L}_{\text{MAE}} + \alpha_1 \mathcal{L}_{\text{DSSIM}} + \alpha_2 \mathcal{L}_{\text{Perc}} + \alpha_3 \mathcal{L}_{\text{ID}},
\end{aligned} \tag{3.11}$$

where  $\mathcal{L}_{\text{ID}} = \mathcal{L}_{\text{ID}, \ell_1} + \mathcal{L}_{\text{ID}, \cos}$  as in Eq. 3.7. Fig. 3.5 compares the performance of face reconstruction networks trained with these different loss functions. As this figure shows, each of the terms enhances the performance of the face reconstruction network. In particular, using either of the terms in ID loss improves the performance, but using both results in the best performance. However, we should note that ID loss terms require full knowledge of the FR model (i.e., whitebox scenario), which is the assumption we make in the ablation study. In blackbox scenario, we assume that the adversary has access to a proxy FR model. To further investigate the effect of weights  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  in our loss function, we perform an ablation study where we change the value of each of these weights and keep the other ones unchanged. Fig. 3.6 shows the effect of these weight in the performance of our method on the AgeDB dataset. The results in this figure show that compared to other weights,  $\alpha_3$  is very sensitive and has a significant effect on the performance of our model. In other words, ID loss has the most contribution to the performance of our method, which is aligned with our ablation study in Fig. 3.5.

**Ablation Study on the Network Structure** For evaluating the efficacy of the proposed network, we train several face reconstruction networks with similar structures but built with different blocks<sup>6</sup>, including typical deconvolution block, NBNet-A block [53], NBNet-B block [53], and DSCasConv block. We train these networks with  $\mathcal{L}_3$ ,  $\mathcal{L}_4$ , and  $\mathcal{L}_6$  (our loss function) of Eq. 3.11. Fig. 3.7 compares the performance of these networks in terms of SAR over different values of FMR, evaluated on the MOBIO, LFW, and AgeDB datasets. As this figure shows, due to the dominant effect of our loss function, these blocks achieve competitive performance when trained with this loss function ( $\mathcal{L}_6$  of Eq. 3.11). However, when trained with  $\mathcal{L}_4$ , our proposed network achieves the best performance on the LFW dataset and competitive performance with NBNet-B on the MOBIO dataset. Finally, when using loss  $\mathcal{L}_3$ , our proposed network clearly outperforms other network structures on both the MOBIO, LFW, and AgeDB datasets.

Fig. 3.8 illustrates sample reconstructed face images using face reconstruction networks with deconvolution, NBNet-A, NBNet-B, and our proposed DSCasConv blocks trained with the same loss function (i.e., Eq. 3.8). As the results in this figure show, the reconstructed face

---

<sup>6</sup>We should note that the typical deconvolution block is used in [51]. Also, [15], [57], [60], [61] used StyleGAN which generates high-resolution images and is not comparable to our reconstructed face images. Similarly, [11] used a GNeRF model to generate 3D face which is neither directly comparable to our method.

### 3.1 Low-resolution Face Reconstruction

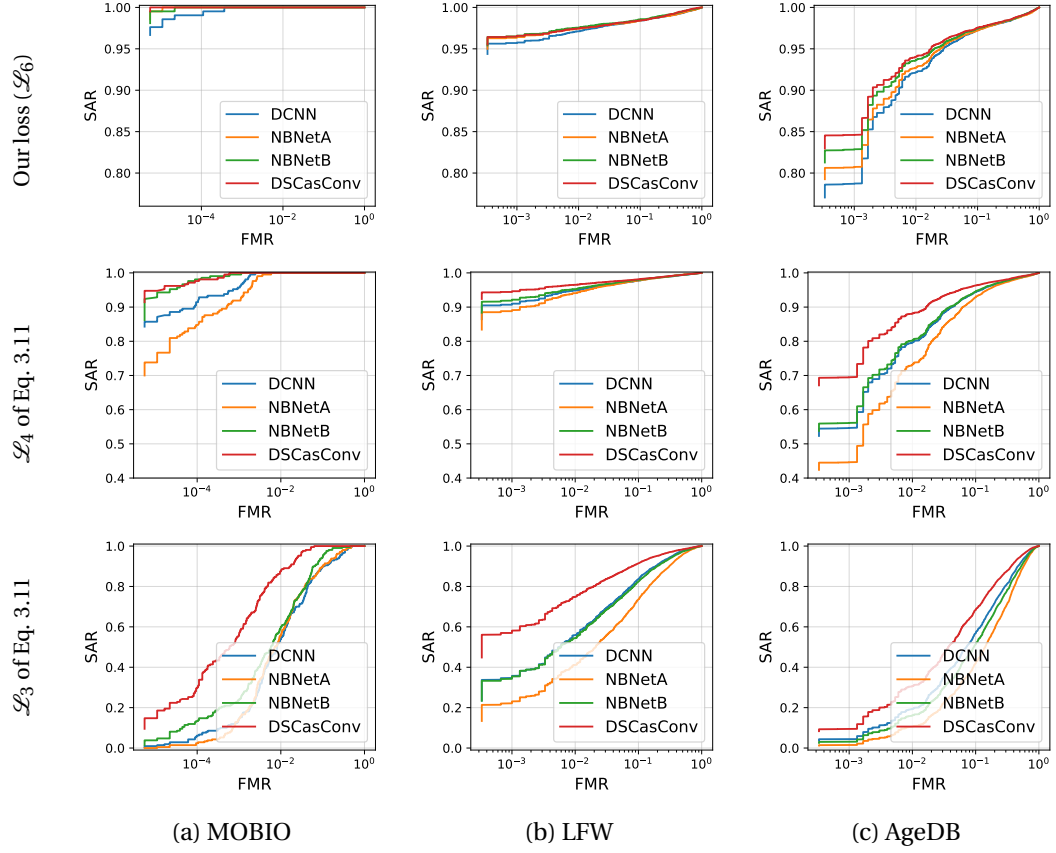


Figure 3.7: Effect of network structure on the face reconstruction performance, when trained with our loss function (first row) as well as  $\mathcal{L}_4$  (second row) and  $\mathcal{L}_3$  (third row) of Eq. 3.11, in a template inversion attack against a face recognition system based on the iResNet100-ArcFace model, evaluated using (a) MOBIO, (b) LFW, and (c) AgeDB datasets.

images using the network with DSCasConv blocks have better visual quality and fewer visual artifacts.

#### 3.1.2.5 TI Vulnerability Analysis of SOTA FR Models

In this section, we evaluate the vulnerability of SOTA FR models to a TI attack using our proposed framework, on the MOBIO, LFW, and AgeDB datasets. The vulnerability of FR models with different SOTA backbones and the same head is evaluated in section 3.1.2.5. We then evaluate the vulnerability of FR models with different SOTA heads and the same backbone in section 3.1.2.5.

**Different Backbones** Table 3.3 compares FR models with different SOTA backbones and the same head (i.e., MV-Softmax) in terms of the number of parameters and the number of

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.3: Comparison of face recognition models with different SOTA backbones and the same head in terms of the number of parameters, MACs, recognition performance (TMR), and vulnerability to template inversion (SAR) at FMR =  $10^{-3}$  on the MOBIO, LFW, and AgeDB datasets.

Model	Params	MACs	MOBIO		LFW		AgeDB	
			↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
MobileFaceNet	1.19M	227.57M	89.86	98.57	53.47	67.49	65.70	69.48
Resnet50	43.57M	6.31G	98.00	100.00	72.87	72.77	89.67	75.97
Resnet152	71.14M	12.33G	98.00	100.00	77.07	73.96	90.63	72.94
HRNet	70.63M	4.35G	98.34	99.52	79.77	79.88	88.50	68.70
EfficientNet-B0	33.44M	77.83M	94.13	99.05	61.53	70.44	69.37	66.49
TF-NAS-A	39.59M	534.41M	92.59	99.52	62.87	68.80	78.27	69.37
LightCNN-29	11.60M	2.84G	87.76	93.81	52.70	66.15	56.17	57.40
GhostNet	26.76M	194.49M	87.51	100.00	55.00	65.05	49.90	48.75
Attention-56	98.96M	6.34G	98.75	99.52	67.33	69.05	87.30	67.04
Attention-92	134.56M	10.62G	98.12	99.05	73.00	74.32	92.13	72.46
ResNeSt50	76.79M	5.55G	99.02	97.62	89.37	86.97	91.53	71.51
ReXNet	15.20M	429.64M	92.15	97.62	67.57	73.30	70.03	62.70
RepVGG-A0	39.94M	1.55G	89.77	99.05	45.43	57.87	72.37	66.01
RepVGG-B0	46.65M	3.44G	93.58	95.71	44.80	52.85	83.60	75.85
RepVGG-B1	106.75M	13.21G	96.87	98.10	62.70	62.66	85.17	65.91
Swin-T	46.74M	4.37G	96.78	100.00	79.97	83.58	90.30	85.11
Swin-S	68.01M	8.53G	99.02	100.00	88.07	89.81	91.23	83.94



Figure 3.8: Sample face images from the FFHQ dataset and their reconstructed images from iResNet100-ArcFace templates using face reconstruction networks based on different blocks: (a) original image, (b) deconvolution, (c) NBNet-A, (d) NBNet-B, and (e) DSCasConv.

multiply/accumulate operations (MACs)<sup>7</sup>, as well as recognition performance (i.e., TMR) and vulnerability to TI (i.e., SAR) at  $\text{FMR} = 10^{-3}$  on the MOBIO, LFW, and AgeDB datasets. Fig. 3.9 also compares the reconstructed face images of these models from the validation subset of the FFHQ dataset. The values below each image in this figure report the cosine similarity between the templates extracted by the corresponding FR model from the original image and the reconstructed image.

**Different Heads** Table 3.4 compares FR models with different SOTA heads and the same backbone (i.e., MobileFaceNet<sup>8</sup>) in terms of recognition performance (i.e., TMR) and vulnerability to TI (i.e., SAR) at  $\text{FMR} = 10^{-3}$  on the MOBIO, LFW, and AgeDB datasets. Fig. 3.10 also compares the reconstructed face images of these models from the validation subset of the FFHQ dataset. Similarly to Fig. 3.9, the values below each image in Fig. 3.10 report the cosine similarity between templates extracted from the original image and the reconstructed image

<sup>7</sup>The values for the number of parameters and the number of MACs are from [156].

<sup>8</sup>which has the least number of parameters in Table 3.3.

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

	Original						
							
MobFace							
	0.868	0.805	0.832	0.770	0.870	0.825	0.851
ResNet50							
	0.866	0.761	0.746	0.808	0.833	0.790	0.832
ResNet152							
	0.874	0.786	0.871	0.817	0.812	0.744	0.813
HRNet							
	0.863	0.838	0.806	0.740	0.795	0.833	0.810
EffNet-B0							
	0.888	0.763	0.828	0.789	0.866	0.776	0.830
TF-NAS-A							
	0.872	0.815	0.851	0.698	0.861	0.804	0.857
LightCNN							
	0.876	0.838	0.871	0.813	0.881	0.833	0.840
GhostNet							
	0.837	0.828	0.846	0.820	0.848	0.821	0.840
AttNet56							
	0.859	0.837	0.853	0.805	0.786	0.766	0.819
AttNet92							
	0.872	0.830	0.840	0.766	0.826	0.852	0.798
ResNeSt50							
	0.887	0.833	0.853	0.836	0.835	0.807	0.813
ResXNet							
	0.855	0.819	0.840	0.832	0.844	0.823	0.803
R.VGG-A0							
	0.822	0.800	0.861	0.772	0.819	0.788	0.835
R.VGG-B0							
	0.839	0.771	0.804	0.837	0.808	0.802	0.805
R.VGG-B1							
	0.843	0.803	0.799	0.773	0.807	0.817	0.763
Swin-T							
	0.877	0.866	0.829	0.837	0.853	0.847	0.858
Swin-S							
	0.877	0.831	0.828	0.822	0.877	0.870	0.817

Figure 3.9: Faces from the FFHQ dataset and their reconstructed versions using our TI method for different backbones. The values show the cosine similarity between the templates.

### 3.1 Low-resolution Face Reconstruction

Table 3.4: Comparison of face recognition models with different SOTA heads and the same backbone (MobileFaceNet) in terms of recognition performance (TMR) and vulnerability to template inversion (SAR) at  $\text{FMR} = 10^{-3}$  on the MOBIO, LFW, and AgeDB datasets.

Model	MOBIO		LFW		AgeDB	
	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
AM-Softmax	92.29	98.57	63.37	72.37	56.90	63.92
AdaM-Softmax	89.57	98.57	66.03	73.85	50.87	58.33
AdaCos	84.83	97.62	61.33	70.69	48.20	56.33
ArcFace	88.73	98.57	60.43	71.17	66.90	71.60
MV-Softmax	89.86	98.57	53.47	67.49	65.70	69.48
CurricularFace	87.51	97.62	48.87	66.28	49.20	59.38
CircleLoss	87.98	99.52	45.43	63.87	63.70	71.72
NPCFace	87.07	98.1	63.93	72.98	60.87	68.53
MagFace	88.00	97.62	59.23	70.74	63.20	70.21

by the corresponding FR model.

#### 3.1.2.6 Discussion

Our experiments in sections 3.1.2.2 to 3.1.2.5 show the privacy and security threat of a TI attack to FR systems. In particular, Fig. 3.9 and Fig. 3.10 suggest that the reconstructed face images reveal important information about the users, including race, gender, age, etc. In addition, as shown by the relatively high SAR values in Fig. 3.5, Fig. 3.7, Table 3.3, and Table 3.4, the reconstructed face images can be used to enter the system by impersonating the corresponding enrolled users, which threatens the security of the FR system. In many cases in Table 3.1, Table 3.3, and Table 3.4, the values of SAR are even higher than the values for system recognition performance in terms of TMR. This is due to the fact that in our evaluation framework, the templates of reconstructed face images are compared to the templates of original face images (i.e., reference templates stored in the system’s database). However, in the evaluation of recognition performance (i.e., TMR) other samples of the enrolled users are used to enter the system. Therefore, a good reconstructed face image may have a higher chance than another sample of the same subject to enter the system. Our experiments in section 3.1.2.3 show that the proposed face reconstruction achieves higher SAR values than previous methods in the literature [51], [53], [57], [60] in TI attacks against FR systems.

In our threat model in section 3.1.1.1, we consider the case where the adversary is assumed not to have any other information about the target FR system except the feature extractor. In particular, we assume that the adversary does not have any knowledge about the comparison and decision making submodules of the target system. However, in our experiments in sections 3.1.2.2-3.1.2.5, we used negative cosine distance as the similarity score between



### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates







































































Original							
AM-Soft.							
	0.896	0.857	0.866	0.833	0.841	0.865	0.867
AdaM-Soft.							
	0.870	0.872	0.894	0.823	0.889	0.841	0.772
AdaCos							
	0.952	0.921	0.933	0.838	0.933	0.922	0.925
ArcFace							
	0.907	0.886	0.884	0.870	0.904	0.869	0.880
MV-Soft.							
	0.868	0.805	0.832	0.770	0.870	0.825	0.851
Curricular							
	0.947	0.889	0.904	0.880	0.869	0.912	0.892
CircleLoss							
	0.852	0.835	0.863	0.813	0.862	0.860	0.822
NPCFace							
	0.902	0.865	0.837	0.866	0.890	0.805	0.908
MagFace							
	0.915	0.855	0.915	0.869	0.919	0.848	0.885

Figure 3.10: Faces from the FFHQ dataset and their reconstructed versions using our TI method for different heads. The values below each image show the cosine similarity between the corresponding templates.

reference and probe templates. To evaluate the effect of this assumption in the performance of our proposed method, as another experiment, we consider different functions<sup>9</sup> for the comparison and decision making submodules of the target system. Table 3.5 compares the recognition performance of the iResNet100-ArcFace model and its vulnerability to our attack on the MOBIO, LFW, and AgeDB datasets. As this table shows, regardless of scoring function, the values for SAR are considerably high for each case and comparable to the value of the system's recognition performance.

Our ablation study in section 3.1.2.4 shows that our loss function and proposed network structure are very effective at reconstructing the underlying face images from their enrolled

<sup>9</sup>Implementations of all these scoring functions are available in the SciPy package: <https://scipy.org>



### 3.1 Low-resolution Face Reconstruction

Table 3.5: Comparison of SAR against FR systems with iResNet100-ArcFace model and using different similarity score functions at FMR= $10^{-3}$  on the MOBIO, LFW, and AgeDB datasets. In each case, the value of distance for probe and reference comparison is multiplied by  $-1$  to get a similarity score.

Function	MOBIO		LFW		AgeDB	
	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
Cosine distance	100.00	100.00	96.63	96.48	93.77	84.60
Euclidean distance	99.98	100.00	87.27	81.61	86.1	77.74
Manhattan (L1) distance	99.98	100.00	86.13	80.02	85.93	77.34
Correlation distance	100.00	100.00	96.63	96.46	93.70	84.51
Canberra distance	100.00	100.00	95.63	95.22	91.80	77.42
Bray-Curtis distance	100.00	100.00	96.07	96.16	93.50	83.06

Table 3.6: Complexity comparison of different network structures.

Network	Params	Exe Time (ms)	
		CPU	GPU
DCNN	6.98M	4.87	0.10
NBNet-A	4.13M	3.59	0.24
NBNet-B	5.23M	4.43	0.33
DSCasConv	16.44M	13.69	0.57

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.7: Comparison of performance of our face reconstruction network when trained on different datasets (FFHQ and CASIA-WebFace) in template inversion attack (SAR) against FR with the iResNet100-ArcFace model, evaluated on the MOBIO and LFW datasets.

	<b>FMR = <math>10^{-2}</math></b>		<b>FMR = <math>10^{-3}</math></b>	
	<b>FFHQ</b>	<b>CASIA-WebFace</b>	<b>FFHQ</b>	<b>CASIA-WebFace</b>
<b>MOBIO</b>	100.00	100.00	100.00	100.00
<b>LFW</b>	97.44	97.84	96.48	97.03

face templates. In addition to the experiment in section 3.1.2.4, which shows the effectiveness of the proposed loss function, our experiments in section 3.1.2.4 also confirm that using  $\mathcal{L}_{ID,cos}$  plus  $\mathcal{L}_{ID,\ell_1}$  (as in Eq. 3.7) improves the reconstruction such that all the studied network structures achieve competitive performance. However, when using weaker loss functions such as  $\mathcal{L}_3$  and  $\mathcal{L}_4$  of Eq. 3.11, our network structure was generally found to outperform other network structures. Sample reconstructed face images in Fig. 3.8 show that DSCasConv can result in better perceptual reconstruction quality and fewer visual artifacts. We also compare the complexity and execution times of the different network structures studied in section 3.1.2.4. Table 3.6 compares the network complexity in terms of the number of parameters and the average inference execution time (milliseconds) in the reconstruction of  $112 \times 112$  face images from 512 dimensional templates, using a system equipped with an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and an NVIDIA GeForce GTX 1080 Ti GPU. As this table shows, with a similar number of blocks, our network has the highest number of parameters and the slowest execution time, because compared to other structures, DSCasConv block has additional convolutional operations (i.e., multiple cascaded convolutional layers with skip connections). Therefore, there is a trade-off between the complexity and the reconstruction performance in our proposed network. However, we should note that the number of parameters in our face reconstruction network is still much smaller than that of almost all SOTA FR models in Table 3.3, and therefore the TI network is inverting FR models using a lower network capacity (in terms of the number of parameters).

To investigate the effect of training data, as another experiment, we train our face reconstruction network using the CASIA-WebFace dataset [184]. Table 3.7 compares the performance of models trained with FFHQ and CASIA-WebFace. As the results in this table show, the performance of our method in terms of SAR remains comparable when trained with different datasets. There is, however, a slightly better performance when the model is trained with CASIA-WebFace. This may be due to the fact that CASIA-WebFace contains 494,414 face images while FFHQ contains 70,000 images. Therefore, when the model is trained with CASIA-WebFace, there are more variations in the images, and thus the model can be more generalizable in the test stage. In addition, while FFHQ has high-quality images, the quality of images in CASIA-WebFace is more similar to the test datasets, which can also contribute to improvement in the performance of the face reconstruction model on the test set when the model is trained on CASIA-WebFace.

Table 3.8: Comparison of reconstruction quality of the proposed network trained with ( $\mathcal{D}_a$ ) and without ( $\mathcal{D}$ ) data augmentation, on the validation set of FFHQ.

Training Data	$\uparrow$ SSIM	$\downarrow$ FID	$\downarrow \mathcal{L}_{\text{Perc}}$
w data aug. ( $\mathcal{D}_a$ )	0.37	114.66	2.69
wo data aug. ( $\mathcal{D}$ )	0.35	149.93	2.77

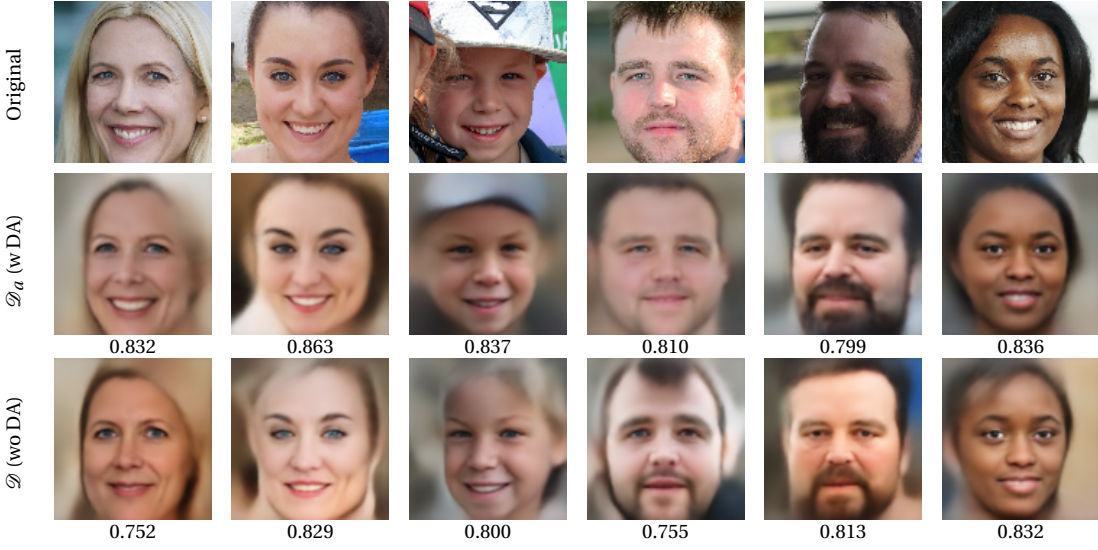


Figure 3.11: Sample face images from the FFHQ dataset (first row) and their corresponding reconstructed image using our face reconstruction network trained *with* (second row) and *without* (third row) data augmentation in template inversion of a face recognition system based on the iResNet100-ArcFace model. The values below each image show the cosine similarity between the corresponding templates.

It is noteworthy that in all our experiments, we used the data augmentation method described in section 3.1.1.2. Fig. 3.11 compares sample face images reconstructed from iResNet100-ArcFace templates using our network trained over training set  $\mathcal{D}$  as in Eq. 3.1 (*without* data augmentation) and also over the augmented training set  $\mathcal{D}_a$  as in Eq. 3.2. As depicted in Fig. 3.11, the reconstruction network trained *with* data augmentation generates face images with better visual reconstruction quality. Table 3.8 also compares the quality of the reconstructed images in terms of  $\mathcal{L}_{\text{Perc}}$  as in Eq.3.6, SSIM [175], and Fréchet Inception Distance (FID) [185] for the validation data of the FFHQ dataset. As this table shows, the network trained *with* data augmentation generates images of better quality.

Last but not least, the experiments in section 3.1.2.5 and section 3.1.2.5 show the vulnerability of SOTA FR models to our TI attack. Comparing Table 3.3 and Table 3.4, we can see that changing the FR model backbone seems to have more effect than changing the head, on the recognition performance (in terms of TMR) of the FR model and also its TI vulnerability (in terms of SAR). For example, in the case of the LFW dataset, changing the head of the FR

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.9: Vulnerability evaluation of state-of-the-art pretrained FR models in terms of SAR at  $\text{FMR}=10^{-3}$  on the MOBIO, LFW, and AgeDB datasets.

method	MOBIO		LFW		AgeDB	
	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
ArcFace [132]	100.00	100.00	97.44	96.48	94.03	84.60
ElasticFace [133]	100.00	99.52	94.70	94.33	92.30	87.26
AdaFace [181]	100.00	99.52	98.40	95.85	95.10	68.39
EdgeFace-S [40]	99.48	100.00	92.70	91.27	75.23	64.67
PocketNet-S [182]	99.34	100.00	90.27	91.09	76.43	78.09

model can change the TMR and SAR in the range of 45.43% – 66.03% and 63.87% – 73.85%, respectively, while changing the backbone of the FR model can change the TMR and SAR in the range of 44.80% – 89.37% and 52.85% – 89.81%, respectively. Therefore, not only are the ranges of change in recognition performance (TMR) and TI vulnerability (SAR) larger when the backbone is changed, but the maximum value of each range is also greater. Table 3.3 and Table 3.4 further suggest that models with higher recognition performance are more likely to be more vulnerable to this type of attack. As another experiment, we evaluate the vulnerability of four other pretrained SOTA FR models in the literature, including ElasticFace [133] and AdaFace [181] as two SOTA FR models as well as EdgeFace [40] and PocketNet [182] as two SOTA lightweight FR models. Table 3.9 reports the vulnerability of these pretrained state-of-the-art face recognition models in the literature (including iResNet100-ArcFace) on the MOBIO, LFW, and AgeDB datasets. As the results in this table show, all these models are highly vulnerable to TI attacks. Since these models also have high recognition performance, this table also supports the hypothesis that models with higher recognition performance are more likely to be more vulnerable to this type of attack.

## 3.2 High-resolution Face Reconstruction using Real Data

In this section, we propose a novel method to reconstruct high-resolution (i.e.,  $1024 \times 1024$ ) face images from deep facial templates (dubbed *HiResT2F*). Fig. 3.12 illustrates sample face images from the FFHQ dataset [186] and their reconstructed face images using our proposed method from the templates extracted from ArcFace-Insightface [132] model. In our method, we train a neural network to map face templates to the *intermediate* latent space of StyleGAN [187], and then we generate the reconstructed face image using StyleGAN’s pretrained synthesis network. StyleGAN’s synthesis network can generate a realistic and high-resolution face image from

### 3.2 High-resolution Face Reconstruction using Real Data



Figure 3.12: Sample face images from the FFHQ dataset and their corresponding reconstructed images using our template inversion method from ArcFace templates (used in a face recognition system). The values below each image show the cosine similarity between the corresponding templates of original and reconstructed face images.

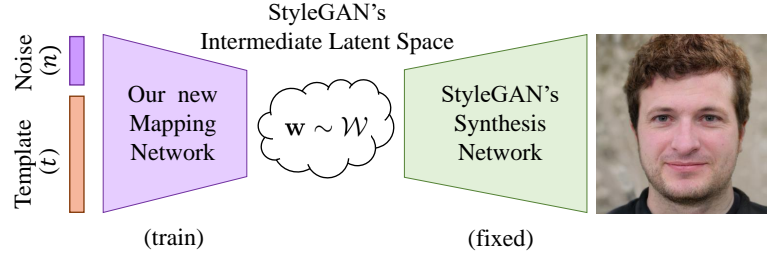


Figure 3.13: General block diagram of the proposed method

a given code in the *intermediate* latent space of StyleGAN. We train our mapping network with a Generative Adversarial Network (GAN)-based framework to learn the distribution of intermediate latent space of StyleGAN. We also apply a multi-term loss function to reconstruct face images, including a pixel loss to minimize pixel-level reconstruction error and an identity loss to preserve the identity of synthesized images. Fig. 3.13 depicts the general block diagram of the proposed method. We propose our method for whitebox and blackbox attacks against FR systems. In our experiments, we evaluate our proposed TI method to reconstruct face images from deep facial templates extracted by SOTA FR models on the MOBIO [158] and LFW [159] datasets.

In the following, we describe the threat model and our proposed method in Section 3.2.1. Next, we present the experiments in Section 3.2.2.

### 3.2.1 Proposed Method

In this section, we describe the proposed method to reconstruct face images from facial embeddings. In section 3.2.1.1, we describe our threat model and how the adversary can attack the system. Then, in section 3.2.1.2 we present our face reconstruction method. Fig. 3.14 depicts the block diagram of the proposed face reconstruction network.

#### 3.2.1.1 Threat Model

We consider the situation where the adversary gains access to the system database, and aims to invert the templates and impersonate. Therefore, we consider the following properties for the adversary:

- *Adversary's goal:* The adversary aims to impersonate a user enrolled in the FR system.
- *Adversary's knowledge:* The adversary has the following information:
  1. The leaked face template  $\mathbf{t}_{\text{leaked}}$ , which is enrolled in the system's database.
  2. The whitebox or blackbox knowledge of the feature extraction model  $F_{\text{template}(\cdot)}$ , that can generate a facial template  $\mathbf{t} = F_{\text{template}}(\mathbf{I})$  from a face image  $\mathbf{I}$ . In case of blackbox knowledge of the feature extraction model  $F_{\text{template}(\cdot)}$ , we assume that the adversary has also whitebox knowledge of another FR model,  $F_{\text{proxy}(\cdot)}$ .
- *Adversary's capability:* The adversary can use the reconstructed face image to perform a presentation attack (e.g., using a 2D printed face photograph or a 3D face mask) to the target FR system. We assume that the adversary can only perform one attempt for each target template, to enter the system.
- *Adversary's strategy:* Under the aforementioned assumptions, the adversary can train a template inversion model  $G(\cdot)$ , and then reconstruct the face image  $\hat{\mathbf{I}}_t = G(\mathbf{t}_t)$  from the target template  $\mathbf{t}_{\text{leaked}}$ . Then, the adversary can use the reconstructed face image  $\hat{\mathbf{I}}_t$  as a query to enter the target FR system.

#### 3.2.1.2 Face Reconstruction Method

We consider a dataset  $\mathcal{S} = \{\mathbf{I}_i\}_{i=1}^N$  containing  $N$  face images (no identity label is required). For each face image  $\mathbf{I}_i$ , we extract embedding  $\mathbf{t}_i = F_{\text{template}}(\mathbf{I}_i)$  using the model  $F_{\text{template}(\cdot)}$ , and consider  $\mathcal{D} = \{(\mathbf{t}_i, \mathbf{I}_i)\}_{i=1}^N$  as our training dataset. We would like to train a face reconstruction network which gets the embedding  $\mathbf{t}$  corresponding to image  $\mathbf{I}$ , where  $\mathbf{t} = F_{\text{template}}(\mathbf{I})$ , and generates the reconstructed face  $\hat{\mathbf{I}}$ . To generate the reconstructed face image  $\hat{\mathbf{I}}$ , we would like to use StyleGAN as a pretrained face generation network that generates high-resolution face images. The StyleGAN model consists of two networks, a mapping network which generates

### 3.2 High-resolution Face Reconstruction using Real Data

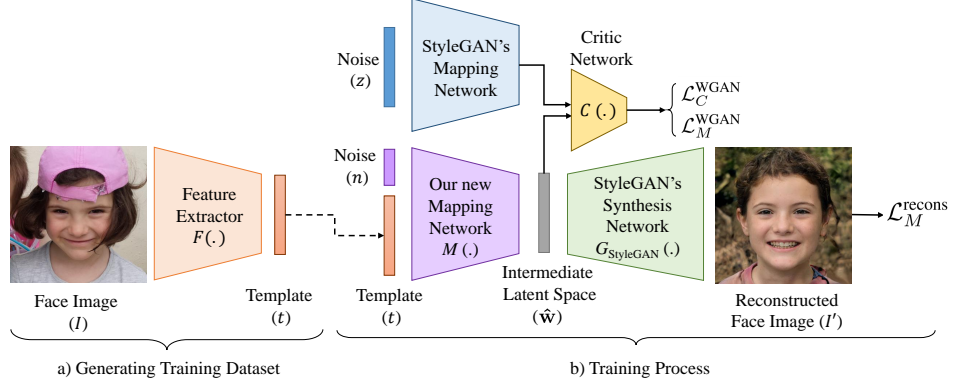


Figure 3.14: Block diagram of the proposed face reconstruction network

an *intermediate* latent code  $\mathbf{w} \sim \mathcal{W}_{\text{StyleGAN}}$  from a random input  $\mathbf{z} \in \mathcal{Z}$  from the Gaussian distribution, and a synthesis network  $G_{\text{StyleGAN}}(\cdot)$ , which generates a high-resolution image from the *intermediate* latent code  $\mathbf{w}$ . To use the pretrained StyleGAN for reconstructing face image from face embeddings, we train a new mapping network  $M(\cdot)$ , which maps the input face embedding  $\mathbf{t}$  to the *intermediate* latent space of StyleGAN  $\hat{\mathbf{w}}$ , and then we can use the mapped *intermediate* latent code to generate the reconstructed face image  $\hat{\mathbf{I}} = G_{\text{StyleGAN}}(\hat{\mathbf{w}})$  using the StyleGAN's synthesis network  $G_{\text{StyleGAN}}(\cdot)$ . However, it is important that the generated latent code  $\hat{\mathbf{w}}$  by our mapping network be on the distribution of the *intermediate* latent space of StyleGAN  $\mathcal{W}_{\text{StyleGAN}}$ . To this end and to help the mapping network learn the distribution of the *intermediate* latent space of StyleGAN  $\mathcal{W}_{\text{StyleGAN}}$ , we train our new mapping network in a GAN-based framework. We concatenate a random vector  $\mathbf{n} \sim \mathcal{N}$  with a normal distribution to the embedding vector  $\mathbf{t}$  as an input to our mapping network (i.e.,  $\hat{\mathbf{w}} = M([\mathbf{n}, \mathbf{t}])$ ). We use the Wasserstein GAN (WGAN) [188] framework and train a critic network  $C(\cdot)$  along our mapping network  $M(\cdot)$ . We generate *real* samples for intermediate latent codes by sampling random noise  $\mathbf{z} \in \mathcal{Z}$  and generating corresponding intermediate latent codes  $\mathbf{w} \in \mathcal{W}_{\text{StyleGAN}}$  using the mapping network of StyleGAN. Then, having samples for intermediate latent codes of StyleGAN  $\mathbf{w} \in \mathcal{W}_{\text{StyleGAN}}$  and the intermediate latent codes generated by our mapping network  $\hat{\mathbf{w}} \sim M([\mathbf{n}, \mathbf{t}])$ , we train the critic network with our mapping network by optimizing the following loss functions:

$$\mathcal{L}_C^{\text{WGAN}} = \mathbb{E}_{\mathbf{w} \sim \mathcal{W}_{\text{StyleGAN}}} [C(\mathbf{w})] - \mathbb{E}_{\hat{\mathbf{w}} \sim M([\mathbf{n}, \mathbf{t}])} [C(\hat{\mathbf{w}})] \quad (3.12)$$

$$\mathcal{L}_M^{\text{WGAN}} = \mathbb{E}_{\hat{\mathbf{w}} \sim M([\mathbf{n}, \mathbf{t}])} [C(\hat{\mathbf{w}})] \quad (3.13)$$

In the loss functions, on one hand, the parameters of the critic network are optimized to give lower scores when given  $\mathbf{w} \sim \mathcal{W}_{\text{StyleGAN}}$  and higher scores when given  $\hat{\mathbf{w}} \sim M([\mathbf{n}, \mathbf{t}])$ . On the other hand, the parameters of the mapping network are optimized so that the score of the critic network when given  $\hat{\mathbf{w}} \sim M([\mathbf{n}, \mathbf{t}])$  will return lower values. For our mapping network

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

---

$M$  and critic network  $C$ , we use fully connected networks with two hidden layer and Leaky ReLU activation function. In addition to the WGAN loss functions for our critic and mapping networks, we optimize our mapping network with the following reconstruction loss function:

$$\mathcal{L}_M^{\text{recons}} = \mathcal{L}_M^{\text{pixel}} + \mathcal{L}_M^{\text{ID}}, \quad (3.14)$$

where  $\mathcal{L}_M^{\text{pixel}}$  and  $\mathcal{L}_M^{\text{ID}}$  are pixel and identity losses in face reconstruction, respectively, and are defined as:

$$\mathcal{L}_M^{\text{pixel}} = \mathbb{E}_{\hat{\mathbf{w}} \sim M((\mathbf{n}, \mathbf{t}))} [\|\mathbf{I} - G_{\text{StyleGAN}}(\hat{\mathbf{w}})\|_2^2] \quad (3.15)$$

$$\mathcal{L}_M^{\text{ID}} = \mathbb{E}_{\hat{\mathbf{w}} \sim M((\mathbf{n}, \mathbf{t}))} [\|F_{\text{proxy}}(\mathbf{I}) - F_{\text{proxy}}(G_{\text{StyleGAN}}(\hat{\mathbf{w}}))\|_2^2] \quad (3.16)$$

where  $F_{\text{proxy}}(\cdot)$  is the FR model which the adversary has complete knowledge about its internal functioning and parameters. In the whitebox attack,  $F_{\text{proxy}}$  is the same as the FR model of the system. In the blackbox scenario,  $F_{\text{proxy}}$  is different from the FR model of the system. The pixel loss  $\mathcal{L}_M^{\text{pixel}}$  considers  $\ell_2$ -norm of pixel-level reconstruction error, which should be minimized so that the reconstructed face image will be similar to the original face image. The ID loss  $\mathcal{L}_M^{\text{ID}}$  also aims to minimize  $\ell_2$ -norm of difference between embeddings (extracted by  $F_{\text{proxy}}$ ) of original and reconstructed face images, to reconstruct identity in the generated face images.

To train our networks, we use Adam [178] optimizer and optimize the parameters of our new mapping network  $M(\cdot)$  for  $\mathcal{L}_M^{\text{recons}}$  loss in every iteration of our training process. In addition, in the WGAN framework, we update weights of our new mapping network  $M(\cdot)$  and critic network  $C(\cdot)$  every  $n_M$  (for minimizing  $\mathcal{L}_M^{\text{WGAN}}$ ) and every  $n_C$  (for minimizing  $\mathcal{L}_C^{\text{WGAN}}$ ) iterations, respectively. Choosing different numbers of iteration ( $n_M$  and  $n_C$ ) for training our new mapping network  $M(\cdot)$  and critic network  $C(\cdot)$ , helps us to find the balance and convergence in training two networks in our WGAN training. For optimizing our three loss functions in Eqs. 3.12-3.14, we used a separate optimizer (i.e., three optimizers in total). Algorithm 1 represents our training process.

### 3.2.2 Experiments

In this section, we describe our experiments and discuss our results. First, in Section 3.2.2.1, we describe our experimental setup. Next, in Section 3.2.2.2 we compare the performance of our method with previous methods in the literature. Then, in Section 3.2.2.3 we present an ablation study to investigate the effect of each part in our proposed method.

#### 3.2.2.1 Experimental Setup

We consider different SOTA FR models in our experiments, including ArcFace-Insightface [132], ElasticFace [133], HRNet [138], AttentionNet [141], RepVGG[145], and Swin [147]. Table A.1 of Appendix A reports the recognition performance of these models in terms of True Match Rate



### 3.2 High-resolution Face Reconstruction using Real Data

---

**Algorithm 1** Training process of our new mapping network.

---

**Require:**  $\theta_M$ , parameters of the new mapping network.  $\theta_C$ , parameters of the critic network.

**Require:**  $n_{\text{epoch}}$ , the number of epochs.  $n_{\text{iteration}}$ , the number of iterations in each epoch.  $n_M$ , the number of training iterations, after which to optimize  $\theta_M$  in WGAN.  $n_C$ , the number of training iterations, after which to optimize  $\theta_C$  in WGAN.  $\delta$ , the WGAN clipping parameter.

**Require:**  $\alpha_M^{\text{recons}}$ , the learning rate for optimizing  $\theta_M$  based on  $\mathcal{L}_M^{\text{recons}}$ .  $\alpha_M^{\text{WGAN}}$ , the learning rate for optimizing  $\theta_M$  in WGAN.  $\alpha_C^{\text{WGAN}}$ , the learning rate for optimizing  $\theta_C$  in WGAN.

**Require:**  $\mathcal{D}$ , a dataset of real face images and corresponding embeddings

```

1: procedure TRAINING
2:   Initialize  $\theta_C$  and  $\theta_M$ 
3:   for epoch = 1, ...,  $n_{\text{epoch}}$  do
4:     for itr = 1, ...,  $n_{\text{iteration}}$  do
5:       Sample a batch from  $\mathcal{D}$  and calculate:
6:        $g_{\theta_M} \leftarrow \nabla_{\theta_M} \mathcal{L}_M^{\text{recons}}$ 
7:        $\theta_M \leftarrow \theta_M - \alpha_M^{\text{recons}} \cdot \text{Adam}(\theta_M, g_{\theta_M})$ 
8:       if itr mod  $n_M = 0$  then
9:          $g_{\theta_M} \leftarrow \nabla_{\theta_M} \mathcal{L}_M^{\text{WGAN}}$ 
10:         $\theta_M \leftarrow \theta_M - \alpha_M^{\text{WGAN}} \cdot \text{Adam}(\theta_M, g_{\theta_M})$ 
11:      end if
12:      if itr mod  $n_C = 0$  then
13:        Sample a batch  $w \sim \mathcal{W}_{\text{StyleGAN}}$  and calculate:
14:         $g_{\theta_C} \leftarrow \nabla_{\theta_C} \mathcal{L}_C^{\text{WGAN}}$ 
15:         $\theta_C \leftarrow \theta_C - \alpha_C^{\text{WGAN}} \cdot \text{Adam}(\theta_C, g_{\theta_C})$ 
16:         $\theta_C \leftarrow \text{clip}(\theta_C, -\delta, \delta)$ 
17:      end if
18:    end for
19:  end for
20: end procedure

```

---

(TMR) at thresholds corresponding to False Match Rates (FMRs) of  $10^{-2}$  and  $10^{-3}$  evaluated on the MOBIO, LFW, and AgeDB datasets. We should note that all these models are trained on the MS-Celeb-1M dataset [157].

For face reconstruction, we use StyleGAN3 [187] as a pretrained face generation network, and train our new mapping network as described in Section 3.2.1. To train our face reconstruction network, we use Flickr-Faces-HQ (FFHQ) [186] dataset. This dataset includes 70,000 high-quality images at  $1024 \times 1024$  resolution crawled from Flickr<sup>10</sup> (without any identity label). The face images in the FFHQ dataset have variations in terms of age and ethnicity. In addition, there are different accessories in the pictures, such as eyeglasses, sunglasses, hats, etc.

To evaluate the performance of our face reconstruction network, we use the MOBIO [158] and Labeled Faces in the Wild (LFW) [159] datasets. The MOBIO dataset consists of face and voice data of 150 people captured using mobile devices in 12 sessions. We use the *development*

---

<sup>10</sup><https://www.flickr.com/>

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.10: Comparison with different methods for TI attack against SOTA FR models at systems’ FMR= $10^{-3}$  on the MOBIO and LFW datasets in terms of success attack rate (SAR). For attacks using our proposed method, we use ArcFace and ElasticFace as  $F_{\text{proxy}}$  to calculate the ID loss in Eq. 3.16 in the blackbox attacks. The values are in percentage. The best two values in attacks against each system are embolden and the best value is highlighted.

method	MOBIO						LFW					
	ArcFace	Els.Face	Att.Net	HRNet	RepVGG	Swin	ArcFace	Els.Face	Att.Net	HRNet	RepVGG	Swin
NBNetA-M [53]	0	2.38	0	0	0	0	4.32	10.90	1.24	1.60	1.13	3.82
NBNetA-P [53]	4.76	16.19	0.48	0	14.29	7.14	16.83	26.98	0.66	1.44	5.72	9.70
NBNetB-M [53]	1.90	3.80	3.33	7.14	3.33	8.57	10.98	21.44	3.22	4.47	3.21	11.23
NBNetB-P [53]	15.24	43.81	31.90	26.67	23.81	44.29	40.26	58.16	16.29	18.42	15.24	40.76
Dong <i>et al.</i> [57]	3.33	8.10	10.48	6.67	9.05	3.33	13.21	12.61	3.90	4.07	3.22	12.38
Vendrow and Vendrow [60]	29.05	43.81	27.14	26.67	20.95	45.24	57.70	53.03	21.12	18.85	9.62	46.84
Dong <i>et al.</i> [61]	61.43	76.67	42.86	49.05	20.00	65.71	74.48	73.67	32.07	31.73	10.89	53.59
[Ours] HiResT2F ( $F_{\text{proxy}} = \text{Els.Face}$ )	<b>81.90</b>	<b>88.09</b>	<b>76.67</b>	<b>82.38</b>	<b>63.33</b>	<b>85.71</b>	<b>77.13</b>	<b>82.78</b>	<b>38.19</b>	<b>43.71</b>	<b>24.63</b>	<b>64.57</b>
[Ours] HiResT2F ( $F_{\text{proxy}} = \text{ArcFace}$ )	<b>92.38</b>	<b>89.05</b>	<b>81.43</b>	<b>86.19</b>	<b>58.10</b>	<b>90.48</b>	<b>86.82</b>	<b>83.43</b>	<b>45.02</b>	<b>48.22</b>	<b>24.34</b>	<b>61.95</b>

subset of the *mobio-all* protocol<sup>11</sup> to evaluate the models. The LFW database includes 13,233 face images of 5,749 people, among which 1,680 people have two or more images. We use the *View 2* protocol<sup>12</sup> in our experiments.

We use the pretrained model of StyleGAN3 to generate  $1024 \times 1024$  high-resolution images. To train our face reconstruction networks, we consider  $n_C = 4$  and  $n_M = 2$  in Algorithm 1. The input noise vectors to the mapping network of StyleGAN’s pretrained network (i.e.,  $\mathbf{z} \in \mathcal{Z}$ ) and to our mapping network  $M(\cdot)$  (i.e.,  $\mathbf{n} \in \mathcal{N}$ ) are both from the standard normal distribution and with 512 and 8 dimensions, respectively. The embeddings extracted by the aforementioned FR models also have 512 dimensions. For a fair comparison with previous methods, we used the available source codes of each method<sup>13</sup> and evaluated them within the same evaluation protocol.

#### 3.2.2.2 Comparison with Previous Methods

We compare the performance of our proposed face reconstruction network with previous methods in the literature<sup>14</sup>, including NBNetA-M [53], NBNetA-P [53], NBNetB-M [53], NBNetB-P [53], Dong *et al.* [57], Vendrow and Vendrow [60], and Dong *et al.* [61]. Table 3.10 compares the performance of our face reconstruction method with previous works in the blackbox<sup>15</sup> TI attack against different SOTA FR models in terms of success attack rate (SAR) on the MOBIO and LFW datasets. For our method, we use ArcFace and ElasticFace as  $F_{\text{proxy}}$  to calculate the ID loss in Eq. 3.16. As the results in this table shows, our method outperforms previous works in the literature in blackbox attacks against SOTA FR models on four different datasets. Our differ-

<sup>11</sup>The implementation of the *mobio-all* protocol for the MOBIO dataset is available at <https://gitlab.idiap.ch/bob/bob.db.mobio>

<sup>12</sup>The implementation of the *View 2* protocol for the LFW dataset is available at <https://gitlab.idiap.ch/bob/bob.db.lfw>

<sup>13</sup>We compared our works with methods in Table 3.10 that have available source code.

<sup>14</sup>Other methods in Table 2.1 do not have available source code.

<sup>15</sup>Whitebox methods in Table 2.1 do not have available source code.

### 3.2 High-resolution Face Reconstruction using Real Data

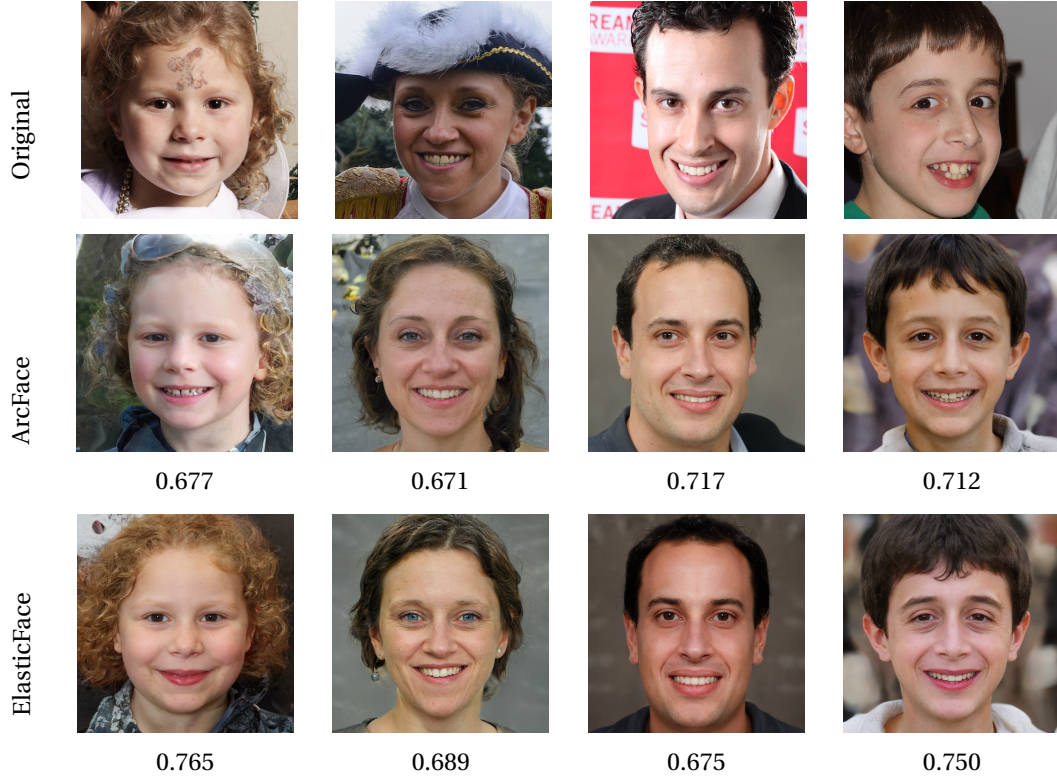


Figure 3.15: Sample face images from the FFHQ dataset and their corresponding reconstructed images using our template inversion in blackbox TI attack against ArcFace (second row) and ElasticFace (third row). The values below each image show the cosine similarity between the corresponding templates of original and reconstructed face images.

ent attacks when using ArcFace and ElasticFace as  $F_{\text{proxy}}$  for training our face reconstruction network result in competitive performance. However for the most cases, ArcFace as  $F_{\text{proxy}}$  has a better SAR, which might be due to the superior performance of ArcFace compared to ElasticFace as reported in Table A.1 of Appendix A. Comparing vulnerability of different FR models in Table 3.10 and considering the recognition performances in Table A.1 of Appendix A, models with a better recognition performance are generally more vulnerable to our attack. Fig. 3.15 shows sample face images from validation set of the FFHQ dataset and their reconstructed images using our method in the *blackbox* TI attacks against ArcFace (using ElasticFace as  $F_{\text{proxy}}$ ) and ElasticFace (using ArcFace as  $F_{\text{proxy}}$ ). As this figure shows, the reconstructed face images are realistic and reveal important privacy-sensitive information (such as age, gender, etc.) about the corresponding subjects. In addition, since the reconstructed face images are high-resolution (i.e.,  $1024 \times 1024$ ), they can be used to perform practical presentation attacks against FR systems.

We should note that SOTA template inversion methods in the literature, Vendrow and Vendrow [60] and Dong *et al.* [61], similarly use StyleGAN in their method, but they find the corresponding

latent code for given facial template based on iterative<sup>16</sup> optimization approaches. In contrast, we use a learning-based approach to find the intermediate latent code, which is much faster in the inference stage. For reconstructing a high-resolution (i.e.,  $1024 \times 1024$ ) face image for a single facial template, our method requires in average only 0.08 seconds execution time on a system equipped with an NVIDIA GeForce RTX<sup>TM</sup> 3090, while Vendrow and Vendrow [60] and Dong *et al.* [61] need 72.06 and 1355.20 seconds in average, respectively. In contrast, these methods do not need a training stage.

#### 3.2.2.3 Ablation Study

To evaluate the effect of each part in our proposed method, we perform an ablation study and train different face reconstruction networks. Table 3.11 reports our ablation study on the mapping space as well as the application of WGAN training (Eqs. 3.12 and 3.13) and the effect of each of our loss terms in our reconstruction loss function (i.e., Eq. 3.14). In our ablation study, we consider whitebox attacks against ArcFace and evaluate the performance in terms of SAR for a system with FMRs of  $10^{-2}$  and  $10^{-3}$  on the MOBIO and LFW datasets. As the results of our ablation study show, the WGAN training on  $\mathcal{W}_{\text{StyleGAN}}$  space in our proposed method has a crucial effect on the performance of our method. Indeed, our adversarial training helps our new mapping network to learn the distribution of the *intermediate* latent space  $\mathcal{W}_{\text{StyleGAN}}$  of StyleGAN and generate face-like images. However, if we do not use adversarial training, the training process of the new mapping network will diverge. It is particularly important since on one hand, we fix the synthesis part of StyleGAN and do not update its weights in the training process. On the other hand, optimizing the reconstruction loss without adversarial training cause the generated intermediate latent codes be out of the distribution of  $\mathcal{W}_{\text{StyleGAN}}$ . In particular, since  $\mathcal{W}_{\text{StyleGAN}}$  is a subset of  $\mathbb{R}^{16 \times 512}$ , optimizing only with the gradients of reconstruction loss can easily cause the network to generate intermediate latent codes out of the distribution of  $\mathcal{W}_{\text{StyleGAN}}$ , and thus divergence of the face reconstruction network. According to Table 3.11, each term in our reconstruction loss function (i.e., Eq. 3.14) also improves the performance of our face reconstruction. In particular, the ID loss has a significant effect on the performance of our method. It is because the face recognition model, which is used as  $F_{\text{proxy}}$  in our ID loss, extracts the identity features, and therefore it helps to preserve identity in the generated face image. In the case of pixel loss, while it achieves very poor performance when being used as our reconstruction loss, it improves the performance of our face reconstruction in combination with our ID loss.

In our ablation study, we also investigate the mapping space in our proposed method. As the results in Table 3.11 shows, mapping to the *intermediate* latent space  $\mathcal{W}_{\text{StyleGAN}}$  achieves a higher performance compared to mapping embeddings to the input latent space  $\mathcal{Z}$ . As a matter of fact and, the *intermediate* latent space  $\mathcal{W}_{\text{StyleGAN}}$ , compared to input latent space  $\mathcal{Z}$ , provides more control on the generated face images. Therefore, training mapping from

---

<sup>16</sup>In each iteration, they need to generate a batch of images through StyleGAN and query them into the face recognition model.

### 3.3 High-resolution Face Reconstruction using Synthetic Data

Table 3.11: Evaluating the effect of mapping space and also each loss term in our loss function in the whitebox attack against ArcFace in terms of SAR for a system with FMRs of  $10^{-2}$  and  $10^{-3}$  evaluated on the MOBIO and LFW datasets. The values are in percentage.

Mapping Space	WGAN training (Eqs. 3.12 and 3.13)	Reconstruction Loss Function	MOBIO		LFW	
			FMR= $10^{-2}$	FMR= $10^{-3}$	FMR= $10^{-2}$	FMR= $10^{-3}$
$\mathcal{W}_{\text{StyleGAN}}$	$\times$	$\mathcal{L}_M^{\text{recons}} = \mathcal{L}_M^{\text{ID}}$	0	0	0.14	0.02
		$\mathcal{L}_M^{\text{recons}} = \mathcal{L}_M^{\text{pixel}}$	0	0	0.44	0.09
		$\mathcal{L}_M^{\text{recons}} = \mathcal{L}_M^{\text{pixel}} + \mathcal{L}_M^{\text{ID}}$	0	0	0.32	0.02
	$\checkmark$	$\mathcal{L}_M^{\text{recons}} = \mathcal{L}_M^{\text{ID}}$	98.10	82.38	90.56	80.74
		$\mathcal{L}_M^{\text{recons}} = \mathcal{L}_M^{\text{pixel}}$	0	0	0.65	0.07
		$\mathcal{L}_M^{\text{recons}} = \mathcal{L}_M^{\text{pixel}} + \mathcal{L}_M^{\text{ID}}$	<b>100.00</b>	<b>92.38</b>	<b>93.64</b>	<b>86.82</b>
$\mathcal{I}$	$\checkmark$	$\mathcal{L}_M^{\text{recons}} = \mathcal{L}_M^{\text{pixel}} + \mathcal{L}_M^{\text{ID}}$	71.42	41.42	75.94	57.18

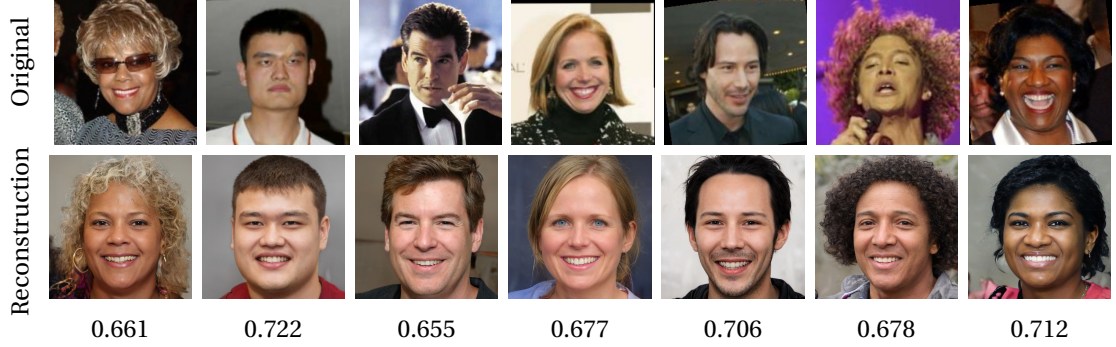


Figure 3.16: Sample face images from the LFW dataset and their reconstructed images using our template inversion method from facial templates extracted by ArcFace. The values below each image show the cosine similarity between the corresponding templates of original and reconstructed face images. It is noteworthy that while our proposed face reconstruction network is trained on synthetic data the reconstruction is generalizable on facial templates extracted from real face images. The decision threshold corresponding to FMR =  $10^{-3}$  is 0.24 on the LFW dataset, and thus all these reconstructed face images pass this threshold.

face embeddings to the *intermediate* latent space  $\mathcal{W}_{\text{StyleGAN}}$  of StyleGAN leads to a better reconstruction.

### 3.3 High-resolution Face Reconstruction using Synthetic Data

In this section, we propose a new method to reconstruct high-resolution face images from facial templates using a pre-trained face generator network without real data. We use StyleGAN [187] as a face generation network and generate synthetic face images. Then, we build our training set by extracting face templates from the synthesized face images. We also keep the intermediate latent codes in the face generator network when synthesizing each face image in our training set. We learn a mapping from facial templates to the intermediate latent space of the StyleGAN model using a multi-term loss function. In the inference stage, we use the trained

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

---

mapping to generate an intermediate latent code for StyleGAN and use the remaining part of the StyleGAN network to generate the reconstructed face image. We propose our method for whitebox (where the adversary knows the parameters and internal functioning of the feature extractor of the FR system) and blackbox (where the adversary does not have knowledge about the internal functioning of the feature extractor of the FR system) template inversion attacks against face recognition systems. We evaluate the vulnerability of state-of-the-art (SOTA) FR systems to our TI attack on two datasets of real face images, including Labeled Faces in the Wild (LFW) [159] and MOBIO [158] datasets. While our model is trained on the synthetic data, the experimental results show that on real data our model outperforms previous methods in the literature. Our experiments also show the vulnerability of SOTA FR systems to our TI attack. Fig. 3.16 illustrates sample face images from the LFW [159] dataset and their corresponding reconstructed face images.

We should highlight that using synthetic face images as training data in our proposed method has two merits: first, the adversary does not need to find a dataset of real face images to use for training. Second, we can have corresponding latent code for each face image and use it directly in our training.

In the following, we describe our proposed method in Sec. 3.3.1, and present our experimental results in Sec. 3.3.2.

#### 3.3.1 Proposed Method

We consider the threat model as described in Sec. 3.3.1.1 and use the proposed face reconstruction method in Sec. 3.3.1.2 to invert facial templates.

##### 3.3.1.1 Threat Model

We assume a TI attack against FR systems with the following threat model:

- *Adversary's Goal:* The adversary aims to invert face templates stored in the database of the FR systems and impersonate.
- *Adversary's Knowledge:* The adversary has access to the database of the face recognition system (complete or partial) and also has whitebox or blackbox knowledge of the feature extractor of the FR system.
- *Adversary's Capability:* The adversary can use the reconstructed face image to inject to the feature extractor of the system as a query.
- *Adversary's Strategy:* The adversary plans to train a face reconstruction network and invert the facial templates. The adversary then uses the reconstructed face image to impersonate by injecting the reconstructed face image into the FR system.

### 3.3 High-resolution Face Reconstruction using Synthetic Data

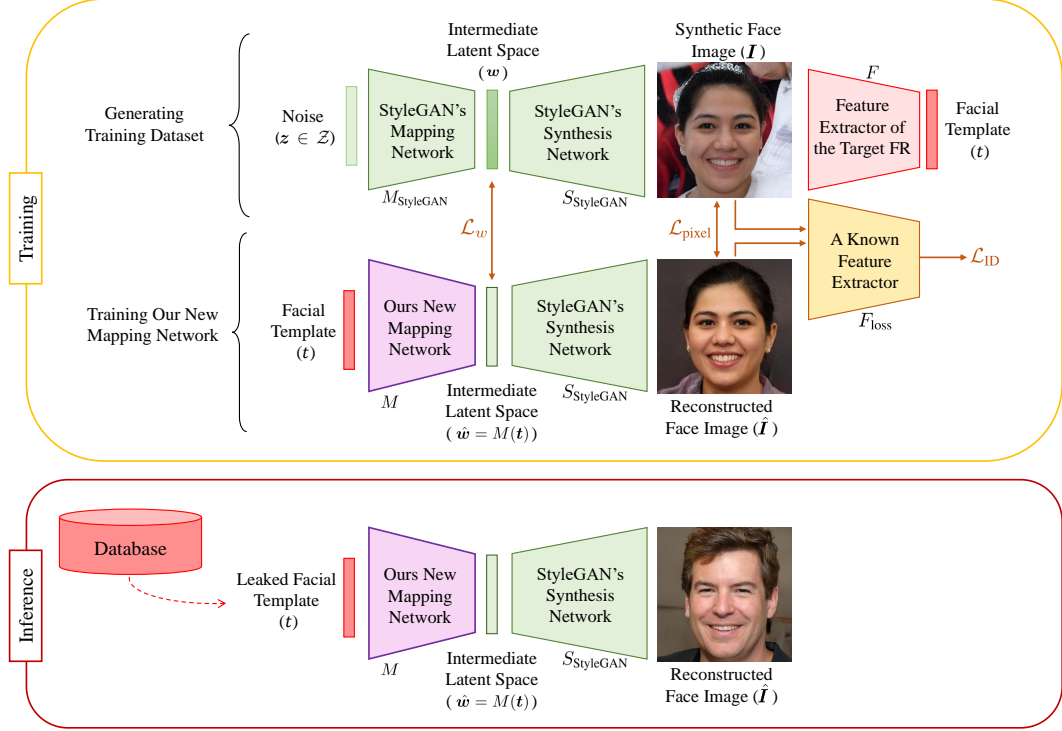


Figure 3.17: Block diagram of the proposed method

#### 3.3.1.2 Face Reconstruction Method

To reconstruct face images from facial templates, we consider the situation where the adversary has access to a pretrained face generator model such as StyleGAN [187]. The StyleGAN model is composed of two sub-networks, a mapping network and a synthesis network. Let us denote the mapping network with  $M_{\text{StyleGAN}}$  and the synthesis network with  $S_{\text{StyleGAN}}$ . The mapping network gets a random noise vector  $\mathbf{z} \in \mathcal{Z}$  as input and generates an *intermediate* latent code  $\mathbf{w} = M_{\text{StyleGAN}}(\mathbf{z}) \in \mathcal{W}_{\text{StyleGAN}}$ , which is then fed to the synthesis network to generate the face image  $\mathbf{I} = S_{\text{StyleGAN}}(\mathbf{w})$ .

To generate a training dataset for learning a face reconstruction network, we use the StyleGAN model to generate synthetic face images and extract facial templates from the synthesized face images. To this end, we sample  $K$  noise  $\{\mathbf{z}_i | \mathbf{z} \in \mathcal{Z} \sim \mathcal{N}(0, \mathbb{I}), i = 1, \dots, K\}$  from Gaussian distribution  $\mathcal{N}(0, \mathbb{I})$  for the input of StyleGAN. Next, we generate corresponding intermediate latent codes  $\mathbf{w}_i = M_{\text{StyleGAN}}(\mathbf{z}_i)$  and synthetic images  $\mathbf{I}_i = S_{\text{StyleGAN}}(\mathbf{w}_i)$ , and then use the feature extractor of the target FR system<sup>17</sup> to extract facial templates  $\mathbf{t}_i = F_{\text{template}}(\mathbf{I}_i)$  from our synthetic face images. Finally, we can have our training dataset  $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{t}_i, \mathbf{w}_i) | i = 1, \dots, K\}$  which has triples of synthetic face images  $\mathbf{I}_i$  as well as their corresponding facial templates  $\mathbf{t}_i$  and the StyleGAN intermediate latent codes  $\mathbf{w}_i$ .

<sup>17</sup>As mentioned in our threat model in Section 3.3.1.1, we only need the *blackbox* knowledge of target FR model, and it is not necessary to have the *whitebox* knowledge.

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

---

**Algorithm 2** Training process in our proposed method.

---

**Require:** :  $n_{\text{epoch}}$ : number of epochs,  $n_{\text{iteration}}$ : number of iterations in each epoch,  $\alpha$ : learning rate.

```

1: procedure TRAINING
2:   Initialize weights  $\theta_M$  of our new mapping network
3:   for epoch = 1, ...,  $n_{\text{epoch}}$  do
4:     for itr = 1, ...,  $n_{\text{iteration}}$  do
5:       Sample a batch of random noise vectors:
6:        $\mathbf{z} \in \mathcal{Z} \sim \mathcal{N}(0, \mathbb{I})$ 
7:       Generate training data:
8:        $\mathbf{w} = M_{\text{StyleGAN}}(\mathbf{z})$ 
9:        $\mathbf{I} = S_{\text{StyleGAN}}(\mathbf{w})$ 
10:       $\mathbf{t} = F_{\text{template}}(\mathbf{I})$ 
11:      Reconstruct image from template  $\mathbf{t}$ :
12:       $\hat{\mathbf{w}} = M(\mathbf{t})$ 
13:       $\hat{\mathbf{I}} = S_{\text{StyleGAN}}(\hat{\mathbf{w}})$ 
14:      Calculate loss  $\mathcal{L}_{\text{total}}$  and optimize  $\theta_M$ :
15:       $g_{\theta_M} \leftarrow \nabla_{\theta_M} \mathcal{L}_{\text{total}}$ 
16:       $\theta_M \leftarrow \theta_M - \alpha \cdot \text{Adam}(\theta_M, g_{\theta_M})$ 
17:    end for
18:  end for
19: end procedure

```

---

After generating our dataset  $\mathcal{D}$ , we can use this dataset to train a new mapping network  $M(\cdot)$  to project the facial template  $t$  to the intermediate latent code  $\hat{\mathbf{w}} = M(\mathbf{t})$  in  $\mathcal{W}_{\text{StyleGAN}}$  space of StyleGAN. Then, we use the the intermediate latent code  $\hat{\mathbf{w}} = M(\mathbf{t})$  as input to the synthesis network of StyleGAN  $S_{\text{StyleGAN}}$  to generate the reconstructed face image  $\hat{\mathbf{I}} = S_{\text{StyleGAN}}(\hat{\mathbf{w}})$ . We train our new mapping network  $M(\cdot)$  with parameters  $\theta_M$  using the following multi-term loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_w + \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{ID}}, \quad (3.17)$$

where  $\mathcal{L}_w$ ,  $\mathcal{L}_{\text{pixel}}$  and  $\mathcal{L}_{\text{ID}}$  are the intermediate latent space loss, pixel loss, and ID loss, respectively, and are defined as follows:

$$\mathcal{L}_w = \|\mathbf{w} - M(\mathbf{t})\|_2^2, \quad (3.18)$$

$$\mathcal{L}_{\text{pixel}} = \|\mathbf{I} - S_{\text{StyleGAN}}(M(\mathbf{t}))\|_2^2, \quad (3.19)$$

$$\mathcal{L}_{\text{ID}} = \|F_{\text{proxy}}(\mathbf{I}) - F_{\text{proxy}}(\hat{\mathbf{I}})\|_2^2. \quad (3.20)$$

The intermediate latent space loss ( $\mathcal{L}_w$ ) is used to minimize the error in the estimated intermediate latent code  $\hat{\mathbf{w}} = M(\mathbf{t})$  in the intermediate latent space  $\mathcal{W}_{\text{StyleGAN}}$  of StyleGAN. Since



### 3.3 High-resolution Face Reconstruction using Synthetic Data

we use synthetic face images, we have the correct values of intermediate latent codes  $w$  to calculate  $\mathcal{L}_w$ . The pixel loss ( $\mathcal{L}_{\text{pixel}}$ ) is also applied to minimize the pixel-level reconstruction error for the reconstructed face image  $\hat{I} = S_{\text{StyleGAN}}(M(t))$  compared to the original image  $I$ . Finally, the ID loss is used to optimize the similarity between the facial templates extracted from the reconstructed and original face images using a FR feature extractor  $F_{\text{proxy}}(\cdot)$ . In the whitebox TI attack, the adversary can use the same feature extractor as the one in the target FR system (i.e.,  $F_{\text{template}}$ ) as  $F_{\text{proxy}}(\cdot)$ ; however, in the blackbox scenario, the adversary needs to use a different feature extractor that has access to<sup>18</sup>. Therefore, in blackbox TI attacks,  $F_{\text{proxy}}(\cdot)$  is different from the target FR system<sup>19</sup>. Algorithm 2 summarizes our training process and Fig. 3.17 illustrates the block diagram of our proposed face reconstruction method. In our experiments, we generate 25,000 synthetic face images for our training set and use Adam optimizer [178] with the learning rate of  $10^{-4}$ . In the inference stage, we use our trained mapping network to project the facial template to the intermediate space of StyleGAN, and then use the synthesis network to generate the reconstructed face image.

#### 3.3.2 Experiments

In this section, we present our experiments and discuss our results. First, in Section 3.3.2.1 we describe our experimental setup. Then, in Section 3.3.2.2 we compare the performance of our method with previous methods in the literature in TI attacks against SOTA FR models. In Section 3.3.2.3, we report an ablation study for our proposed method.

##### 3.3.2.1 Experimental Setup

We consider SOTA FR systems as target systems and evaluate their vulnerability to our TI attack. We use ArcFace [132], ElasticFace [133], and also different FR models with SOTA backbones from FaceX-Zoo [156], including AttentionNet [141], HRNet [138], RepVGG [145], and Swin [147]. The recognition performances of these models are reported in Table A.1 of Appendix A.

To evaluate the vulnerability of these FR models, we use the MOBIO [158] and Labeled Faces in the Wild (LFW) [159] datasets. The MOBIO dataset includes face images of 150 subjects captured using mobile devices in 12 sessions (6-11 samples in each session). The LFW dataset contains 13,233 face images of 5,749 subjects collected from the internet, in which 1,680 subjects have two or more images. For each of the MOBIO or LFW datasets, we build a FR system and then invert the enrolled facial templates to reconstruct face images. Next, according to our threat model described in Sec. 3.3.1.1, we inject the reconstructed face

<sup>18</sup>Note that the adversary needs to have whitebox knowledge of feature extractor used in  $F_{\text{proxy}}$  to be able to calculate gradients in optimizing loss function for training face reconstruction model.

<sup>19</sup>Note that the alternate model is only used in the *blackbox* scenario and is only applied for  $F_{\text{proxy}}(\cdot)$  in the loss function  $\mathcal{L}_{\text{ID}}$  (not to extract the initial templates  $t$ ). In both whitebox and blackbox scenarios, feature extractor of the target FR system (i.e.,  $F_{\text{template}}$ ) is always used to extract the initial templates  $t$  in generating training dataset  $\mathcal{D}$ .

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.12: Comparison with different face reconstruction methods in TI attacks against SOTA FR models in terms of success attack rate (SAR) at systems'  $\mathbf{FMR} = 10^{-3}$  on the MOBIO and LFW datasets . For attacks using our proposed method, we use ArcFace and ElasticFace as  $F_{\text{proxy}}$  in our loss function. The best two values in attack against each system is embolden. The values are in percentage.

method	MOBIO						LFW					
	ArcFace	Els.Face	Att.Net	HRNet	RepVGG	Swin	ArcFace	Els.Face	Att.Net	HRNet	RepVGG	Swin
NBNetA-M [53]	0	2.38	0	0	0	0	4.32	10.90	1.24	1.60	1.13	3.82
NBNetA-P [53]	4.76	16.19	0.48	0	14.29	7.14	16.83	26.98	0.66	1.44	5.72	9.70
NBNetB-M [53]	1.90	3.80	3.33	7.14	3.33	8.57	10.98	21.44	3.22	4.47	3.21	11.23
NBNetB-P [53]	15.24	43.81	31.90	26.67	23.81	44.29	40.26	58.16	16.29	18.42	15.24	40.76
Dong <i>et al.</i> [57]	3.33	8.10	10.48	6.67	9.05	3.33	13.21	12.61	3.90	4.07	3.22	12.38
Vendrow and Vendrow [60]	29.05	43.81	27.14	26.67	20.95	45.24	57.70	53.03	21.12	18.85	9.62	46.84
Dong <i>et al.</i> [61]	61.43	76.67	42.86	49.05	20.00	65.71	<b>74.48</b>	73.67	32.07	31.73	10.89	53.59
[Ours] ( $F_{\text{proxy}} = \text{Els.Face}$ )	<b>80.00</b>	<b>87.62</b>	<b>78.10</b>	<b>78.10</b>	<b>68.57</b>	<b>79.05</b>	71.31	<b>80.41</b>	<b>36.92</b>	<b>43.13</b>	<b>29.33</b>	<b>61.63</b>
[Ours] ( $F_{\text{proxy}} = \text{ArcFace}$ )	<b>84.76</b>	<b>86.67</b>	<b>81.90</b>	<b>85.24</b>	<b>70.95</b>	<b>84.76</b>	<b>85.01</b>	<b>81.70</b>	<b>43.58</b>	<b>50.04</b>	<b>35.75</b>	<b>66.57</b>

images into the feature extractor of the FR system as a query and evaluate the adversary's success attack rate (SAR) at different false match rates (FMRs) of the FR system.

In our experiments, we use the pre-trained model of StyleGAN3 to generate  $1024 \times 1024$  high-resolution face images. We generated 25,000 synthetic face images for our training set in our experiments.

#### 3.3.2.2 Comparison with Previous Methods

We compare the performance of our proposed method with previous works in the literature with available source code, including NBNet-A-M [53], NBNet-A-P [53], NBNet-B-M [53], NBNet-B-P [53], Vendrow and Vendrow [60], Dong *et al.* [57], and Dong *et al.* [61]. Table 3.12 compares the performance of our method with these methods in terms of adversary's success attack rate (SAR) against different SOTA FR systems at  $\text{FMR } 10^{-3}$ , on the MOBIO and LFW datasets. For our method, we use ArcFace and ElasticFace as  $F_{\text{proxy}}$  in our loss function (Eq. 3.20) to reconstruct face images from facial templates extracted from different FR systems and train a separate model for each FR model. As the results in Table 3.12 show, our method outperforms previous methods in the literature. In particular, compared to [57], [60], [61] which used StyleGAN to generate high-resolution and realistic face images our method achieves superior performance. Comparing the results in this table with the recognition performances of FR systems reported in Table A.1 of Appendix A, we observe that a FR system with a higher recognition accuracy is more vulnerable to our attack. Comparing the results of ArcFace and ElasticFace in the loss function of our method, the results show that ArcFace leads to better SAR values, which may be due to the fact that ArcFace has a better recognition performance than ElasticFace as shown in Table A.1 of Appendix A. Fig. 3.18 illustrates sample face images from LFW dataset and their corresponding reconstructed face images in whitebox and blackbox TI attacks using ArcFace templates.

### 3.3 High-resolution Face Reconstruction using Synthetic Data

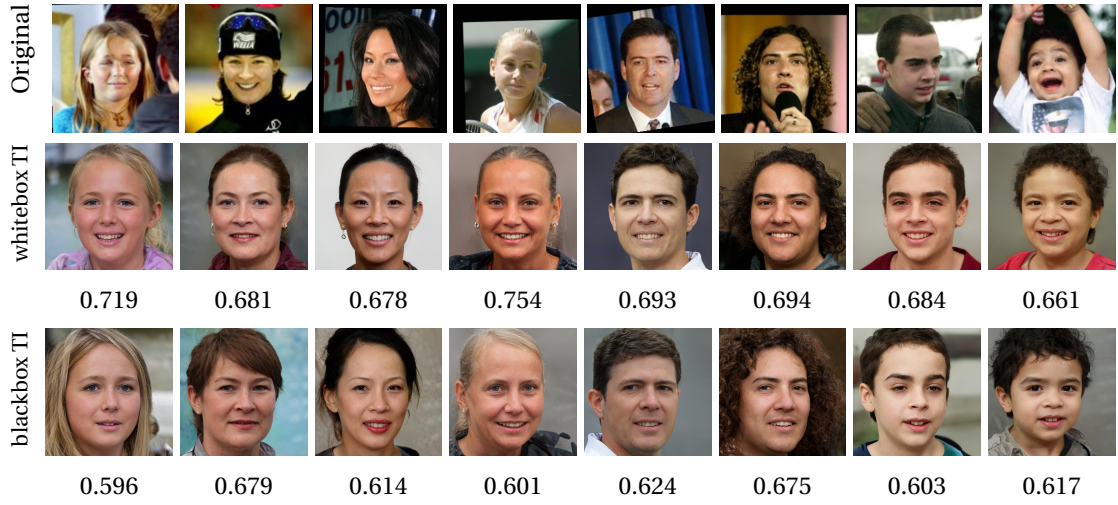


Figure 3.18: Sample real face images from the LFW dataset (first row) and their reconstructed images from ArcFace templates in whitebox (second row) and blackbox (third row). The values below each image show the cosine similarity between the corresponding templates of original and reconstructed face images. The decision threshold corresponding to  $\text{FMR} = 10^{-3}$  is 0.24 on the LFW dataset, and thus all these reconstructed images pass this threshold.

Table 3.13: Ablation study on the effect of each loss term in whitebox attack against ArcFace in terms of SAR for a system with FMRs of  $10^{-2}$  and  $10^{-3}$  on the MOBIO and LFW datasets.

Loss function	MOBIO		LFW	
	FMR= $10^{-2}$	FMR= $10^{-3}$	FMR= $10^{-2}$	FMR= $10^{-3}$
$\mathcal{L}_{\text{total}} = \mathcal{L}_w$	43.81	13.80	47.69	27.54
$\mathcal{L}_{\text{total}} = \mathcal{L}_w + \mathcal{L}_{\text{pixel}}$	40.00	13.81	45.61	25.98
$\mathcal{L}_{\text{total}} = \mathcal{L}_w + \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{ID}}$	97.62	89.05	92.89	85.84
$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{ID}}$	0	0	0.32	0.02

#### 3.3.2.3 Ablation Study

To evaluate the effect of each loss term in our proposed method, we perform an ablation study, where we train different mapping networks with different loss functions and evaluate the adversary's SAR in a TI attack against a FR system. To this end, we consider a whitebox TI attack against a FR system based on ArcFace on the MOBIO and LFW datasets. Table 3.13 reports the effect of each loss term in our proposed method. As the results in this table show, each term in our loss function improves the reconstructed face images in TI attacks against FR systems. In particular, we observe that using the latent code loss (i.e.,  $\mathcal{L}_w$ ) helps the training compared to using all other terms except the latent code loss term. This also highlights the use of synthetic data in our proposed method where we have the correct latent code for each single image in our synthetic training dataset. When using the latent code loss, our ID loss also significantly improves the reconstruction compared to other cases without ID loss. The pixel-level loss, however, slightly degrades the SAR values but reduces the pixel-level errors (e.g., hair color, etc.) in the reconstructed face images.

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

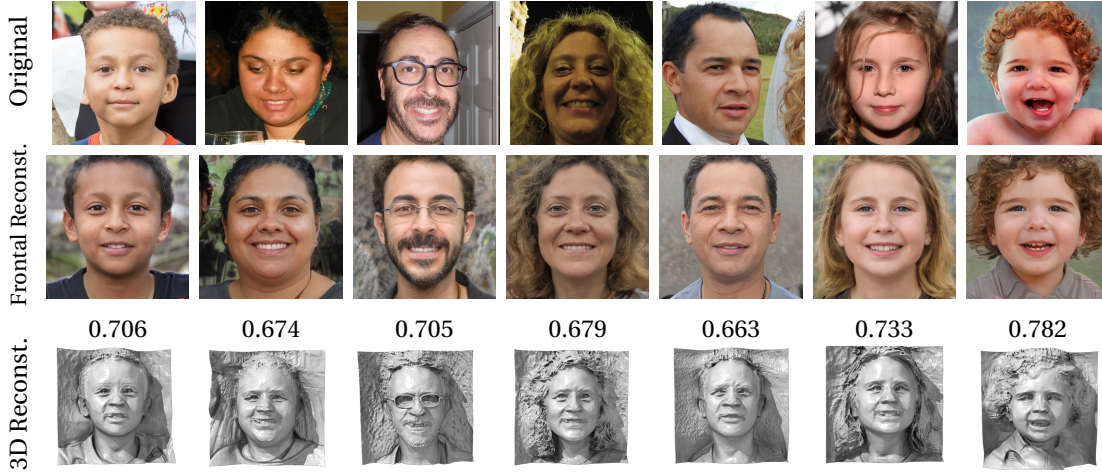


Figure 3.19: Sample face images from the FFHQ dataset (first row) as well as their corresponding 3D (third row) and frontal 2D reconstruction (second row) from facial templates in the whitebox *template inversion attack* against ArcFace. Values show the cosine similarity between the templates of the original and frontal reconstructed face images.

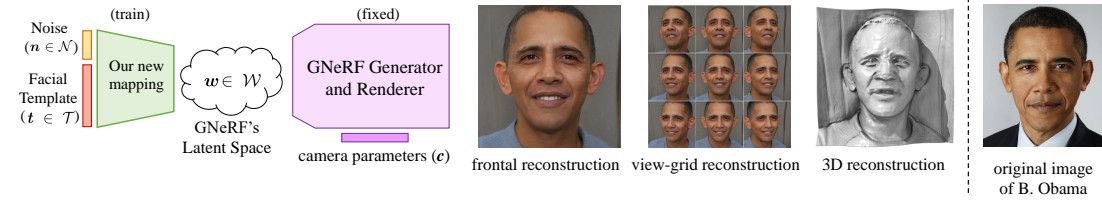


Figure 3.20: General block diagram of the proposed method: we train a mapping network from facial templates (input) to the intermediate latent space  $\mathcal{W}_{\text{GNeRF}}$  of GNeRF model. The mapped latent codes along with camera parameters are fed to the GNeRF generator and renderer network (fixed) to generate face image from desired view. Sample outputs of our model (frontal image, view-grid, and 3D face reconstruction) for face reconstruction from B. Obama's facial template are depicted.

### 3.4 3D Face Reconstruction

In this section, we present a comprehensive vulnerability evaluation of FR systems to TI attacks using 3D face reconstruction. We propose a new method (called geometry-aware face reconstruction, shortly *GaFaR*) to 3D reconstruct faces from facial templates using a geometry-aware face generator network. To our knowledge, this is the first work to reconstruct 3D faces from facial templates. Fig. 3.19 illustrates sample face images from the FFHQ [186] dataset and their corresponding 3D reconstruction from ArcFace [132] templates using our proposed method.

In recent years, the neural radiance fields (NeRF) [189] has attracted attentions in the computer vision community because of its impressive results in the novel-view generation problem. Generative NeRF (GNeRF) methods such as [190]–[202] combine conditional NeRF with generative models, such as a generative adversarial network (GAN), for geometry-aware image generation tasks. In GNeRF methods, a generative model is used to embed the appearance

and shape of an object into a latent space. Then, the camera parameters along with the latent code of the generative model are fed into a NeRF model for the rendering process. Among GNeRF methods, several works proposed geometry-aware 3D face generation models that can generate face images from different views [194]–[201].

In our proposed 3D face reconstruction method, we use a geometry-aware face generator network based on GNeRF, and learn a mapping from facial templates to the *intermediate* latent space of the GNeRF model. We train our model with a *semi-supervised* approach using real and synthetic face images. For real training face images, where we do not have the corresponding GNeRF latent codes, we train our mapping within a GAN-based framework to learn the distribution of GNeRF *intermediate* latent space (*unsupervised* learning). However, for the synthetic training face images, we have the corresponding GNeRF latent codes, and directly learn the mapping from facial templates to the GNeRF *intermediate* latent space (*supervised* learning). At the inference stage, we have the 3D reconstructed face and can generate a face image from any arbitrary pose. Thus, we apply optimization on the camera parameters to generate face images with a pose that can increase the success attack rate against the FR system. Fig. 3.20 illustrates the general block diagram of our proposed template inversion attack.

We introduce our face reconstruction method for *whitebox* and *blackbox* TI attacks against FR systems. In the *whitebox* scenario, the adversary knows the internal functioning and parameters of the feature extraction model. However, in the *blackbox* scenario, the adversary does not have any knowledge about the internal functioning of the feature extraction model and can only use it to extract features from an arbitrary image. We consider the scenario where the adversary uses another FR model, with known internal functioning and parameters (i.e., *whitebox* knowledge), and uses this FR model for training the face reconstruction network. We present a comprehensive vulnerability evaluation of state-of-the-art (SOTA) FR systems to our TI attacks in *whitebox* and *blackbox* scenarios. We evaluate the *transferability* of the reconstructed face images by considering the situation where the adversary tries to reconstruct face images of the templates leaked from a FR system and use the reconstructed face images to impersonate the same users in another FR system (with a different feature extraction model) that the users are enrolled. Indeed, the transferability of TI attacks reveals a critical threat to FR systems, since the reconstructed face images can be used to enter other FR systems that the victim is enrolled in. Considering the *whitebox/blackbox* scenario and the adversary's knowledge of the target FR system, we define five different TI attacks, and comprehensively evaluate the vulnerability of SOTA FR systems to TI attacks. Furthermore, we perform practical evaluations based on presentation attacks using the digital replay and printed photographs of the reconstructed face images, and evaluate the vulnerability of SOTA FR systems.

The remainder of this section is structured as follows. First, we describe the threat model, our five different defined attacks, and our proposed method in Section 3.4.1. Next, in Section 3.4.2, we present our experiments and discuss our results.

## Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

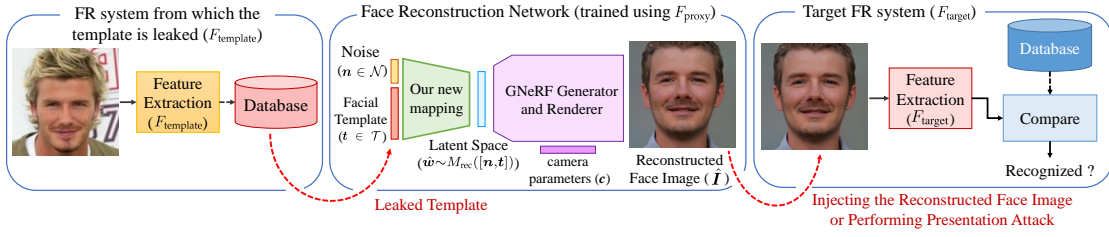


Figure 3.21: Block diagram of our threat model in Section 3.4

### 3.4.1 Proposed Method

We describe our threat model and define different TI attacks against FR systems in Section 3.4.1.1 (as depicted in Fig. 3.21). Then, we describe our proposed method to reconstruct 3D faces from facial templates in Section 3.4.1.2. In the inference stage, we optimization on the camera parameters to generate a face image that can improve the success attack rate, as described in Section 3.4.1.3. Fig. 3.22 illustrates the block diagram of the proposed TI attack, including our 3D face reconstruction method and our optimization on camera parameters during the inference stage.

#### 3.4.1.1 Threat Model

We consider the situation where the adversary gains access to the database of a FR system ( $F_{\text{template}}$ ), and aims to invert its templates. The adversary is also assumed to have access<sup>20</sup> to a feature extractor model  $F_{\text{proxy}}$  (which can be the same or different than  $F_{\text{template}}$ ). The adversary trains a face reconstruction model to reconstruct face images from templates extracted by  $F_{\text{template}}$ , and uses the reconstructed face images to impersonate into the same or a different FR system ( $F_{\text{target}}$ ). Therefore, we consider the following properties for the adversary:

- *Adversary's goal:* The adversary aims to reconstruct face images from templates stored in the database of a FR system ( $F_{\text{template}}$ ), and use the reconstructed face images to enter the same or a different FR system (we call it the target FR system,  $F_{\text{target}}$ ).
- *Adversary's knowledge:* The adversary has the following information:
  - The leaked face templates  $t_{\text{leaked}}$  of users, which are enrolled in the database of  $F_{\text{template}}$ .
  - The adversary also has the *whitebox* knowledge of a feature extractor model ( $F_{\text{proxy}}$ ). It is worth mentioning that  $F_{\text{proxy}}$  can be similar to or different from  $F_{\text{template}}$  and  $F_{\text{target}}$ .
- *Adversary's capability:* We consider two scenarios for the adversary's capability:

<sup>20</sup>The adversary can use  $F_{\text{proxy}}$  for training the face reconstruction network.

- The adversary can perform a presentation attack using the reconstructed face images to impersonate and enter the target FR system (e.g., using digital replay attacks or printed photographs).
- The adversary can inject the reconstructed face image as a query to the target FR system.
- **Adversary's strategy:** The adversary trains a face reconstruction model to invert the leaked facial templates  $t_{\text{leaked}}$ . Then, based on the adversary's capability, the adversary can use the reconstructed face images to either perform a presentation attack or inject the reconstructed face image as a query to the target FR system.

In our threat model, we consider three different feature extraction models, including  $F_{\text{template}}(\cdot)$ ,  $F_{\text{proxy}}(\cdot)$ , and  $F_{\text{target}}(\cdot)$ . Fig. 3.21 illustrates the block diagram of our threat model. Based on the target FR system and the adversary's knowledge, we can define five different attacks:

- **Attack 1:** The adversary has the *whitebox* knowledge of the feature extractor of the FR system from which the template is leaked and aims to impersonate to the same FR system (i.e.,  $F_{\text{template}} = F_{\text{proxy}} = F_{\text{target}}$ ).
- **Attack 2:** The adversary has the *whitebox* knowledge of the feature extractor of the FR system from which the template is leaked, but aims to impersonate to a different FR system (i.e.,  $F_{\text{template}} = F_{\text{proxy}} \neq F_{\text{target}}$ ).
- **Attack 3:** The adversary aims to impersonate to the same FR system from which the template is leaked, but has only the *blackbox* access to the feature extractor of the FR system. Instead, the adversary has the *whitebox* knowledge of another FR model to use for training the face reconstruction model (i.e.,  $F_{\text{template}} = F_{\text{target}} \neq F_{\text{proxy}}$ ).
- **Attack 4:** The adversary aims to impersonate to a different FR system than the one which from the template is leaked. In addition, the adversary has the *whitebox* knowledge of the feature extractor of the target FR system (i.e.,  $F_{\text{template}} \neq F_{\text{proxy}} = F_{\text{target}}$ ).
- **Attack 5:** The adversary aims to impersonate to a different FR system from which the template is leaked, and has only the *blackbox* knowledge of the both the FR systems. However, the adversary instead has the *whitebox* knowledge of another FR model to use for training the face reconstruction model (i.e.,  $F_{\text{template}} \neq F_{\text{proxy}} \neq F_{\text{target}}$ ).

Table 3.14 summarizes different TI attack types in our threat model as well as the adversary's knowledge of different FR models in each type of attack. In all types of attacks, the leaked facial templates to be reconstructed are from  $F_{\text{template}}$  and the reconstructed face image is used to attack target FR system  $F_{\text{target}}$ . In attack 1 and attack 3, the target FR system is the same as the FR system from which the template is leaked (i.e.,  $F_{\text{template}} = F_{\text{target}}$ ). However, in attacks 2, 4, and 5, the target FR system is different from the FR system from which the template is leaked



### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.14: Different TI attacks against FR systems in our threat model. FR models are indicated with symbols, where having the same (different) symbol means the same (different) FR models are used. Each symbol is also either filled with white or black, indicating whitebox or blackbox knowledge to the corresponding model, respectively.

Attack Type	$F_{\text{template}}^1$	$F_{\text{proxy}}^2$	$F_{\text{target}}^3$
Attack 1	□	□	□
Attack 2	□	□	◆
Attack 3	■	△	■
Attack 4	■	△	△
Attack 5	■	△	◆

<sup>1</sup> $F_{\text{template}}$ : the FR system from which the template is leaked.

<sup>2</sup> $F_{\text{proxy}}$ : the FR model which adversary has access to and use it for training the TI model (i.e., always whitebox).

<sup>3</sup> $F_{\text{target}}$ : the target FR system that the adversary aims to enter using the reconstructed face image from the TI attack.

(i.e.,  $F_{\text{template}} \neq F_{\text{target}}$ ), and therefore in attack 2, 4, and 5, the *transferability* of reconstructed face images in attacks against different FR systems is evaluated. Comparing different types of attacks, in attack 1 the adversary has knowledge of the FR system from which the template is leaked and aims to enter the same FR system, therefore it is expected that attack 1 may be the easiest attack. In contrast, in attack 5 the adversary does not have the whitebox knowledge of the FR system from which the template is leaked or the target FR system, and thus attack 5 may be the hardest attack for the adversary.

#### 3.4.1.2 3D Face Reconstruction Method

To reconstruct 3D faces from facial templates, we use a pretrained EG3D [199] model as a geometry-aware face generator network based on GNeRF. This model consists of two networks, a mapping network and a generator and renderer network. The mapping network  $M_{\text{GNeRF}}$  takes a random noise  $\mathbf{z} \in \mathcal{Z}$  in the input and generates an *intermediate* latent code  $\mathbf{w} = M_{\text{GNeRF}}(\mathbf{z}) \in \mathcal{W}_{\text{GNeRF}}$ . The *intermediate* latent code  $\mathbf{w}$  provides more control over the generated face images than input random noise  $\mathbf{z}$ . The generator and renderer network  $G(\cdot, \cdot)$  takes the *intermediate* latent code  $\mathbf{w}$  and camera parameters  $\mathbf{c}$ , to generate a face image  $\mathbf{I} = G(\mathbf{w}, \mathbf{c})$  from an arbitrary view. To reconstruct 3D faces from facial templates, we learn a new mapping  $M_{\text{rec}}: \mathcal{T} \rightarrow \mathcal{W}_{\text{GNeRF}}$  from the facial templates  $\mathbf{t} \in \mathcal{T}$  to the *intermediate* latent space  $\mathcal{W}_{\text{GNeRF}}$  of the GNeRF model. Then, we feed the mapped *intermediate* latent vector  $\hat{\mathbf{w}}$  along with camera parameters  $\mathbf{c}$  into the GNeRF model  $G(\cdot, \cdot)$  to generate a face image  $\hat{\mathbf{I}} = G(\hat{\mathbf{w}}, \mathbf{c})$  from an arbitrary view corresponds to the camera parameters  $\mathbf{c}$ . We train our mapping network  $M_{\text{rec}}$  simultaneously using real and synthetic training data with a *semi-supervised* approach as follows:

**Unsupervised learning using real training data** To train our mapping network  $M_{\text{rec}}(\cdot)$  with the real training data, we use a set of real face images  $\{\mathbf{I}_{\text{real}, i}\}_{i=0}^N$  and extract the facial template



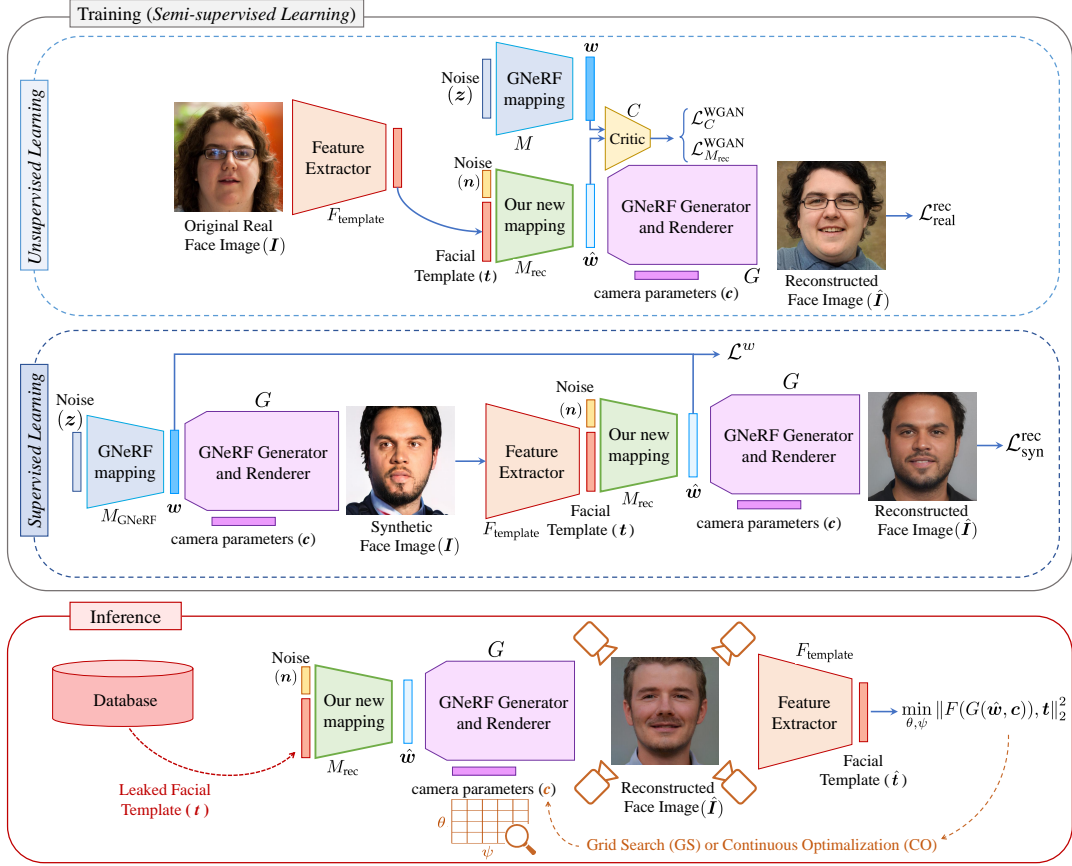


Figure 3.22: Block diagram of our proposed TI attack: During the training process, a *semi-supervised* approach is used to learn our mapping  $M_{rec}$  (illustrated as a green block) from the facial templates to the *intermediate* latent space of the GNeRF model. We use *real* training data (where we don't have the corresponding latent code) and *synthetic* training data (where we have the corresponding latent code  $w$ ) simultaneously for *unsupervised* and *supervised* learning in our method. In the inference stage, the leaked template  $t$  is fed into our mapping network to find corresponding vector  $\hat{w} = M_{rec}([n, t])$  in the *intermediate* latent space of the GNeRF. Then, camera parameters  $c$  along with  $\hat{w}$  are given to the generator and renderer of GNeRF  $G$  to generate a reconstructed face image  $\hat{I} = G(\hat{w}, c)$ . To enhance the attack, we propose an optimization (grid search or continuous optimization) on two of the camera parameters,  $\theta$  and  $\psi$ , from  $c$ , to find the best pose, which minimizes the distance between the template of reconstructed face image and the leaked template  $t$ .

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

$\mathbf{t}_{\text{real},i} = F_{\text{template}}(\mathbf{I}_{\text{real},i})$  from each face image  $\mathbf{I}_{\text{real},i}$  using the FR model  $F_{\text{template}}(\cdot)$ . We assume that the adversary does not have any information about the training dataset of  $F_{\text{template}}(\cdot)$  and  $F_{\text{target}}(\cdot)$ , and thus use another dataset for training the face reconstruction model. Since we do not have the true value of the *intermediate* latent space  $\mathcal{W}_{\text{GNeRF}}$  of the GNeRF model for the real face images in  $\{\mathbf{I}_{\text{real},i}\}_{i=0}^N$ , we consider training our mapping network using the real training data as *unsupervised* learning. For the real training data, we train our mapping  $M_{\text{rec}}(\cdot)$  within a GAN-based framework based on Wasserstein GAN (WGAN) [188] algorithm to learn the distribution of *intermediate* latent space  $\mathcal{W}_{\text{GNeRF}}$  of the GNeRF model. In this framework, our mapping network  $M_{\text{rec}}$  acts as the generator of our WGAN training and generates a latent code  $\hat{\mathbf{w}} = M_{\text{rec}}([\mathbf{n}, \mathbf{t}])$  from a random vector  $\mathbf{n} \in \mathcal{N}$  and the facial template  $\mathbf{t}$ . In our WGAN framework, we can also generate the real latent code  $\mathbf{w} = M_{\text{GNeRF}}(\mathbf{z}) \in \mathcal{W}_{\text{GNeRF}}$  using the GNeRF mapping function  $M_{\text{GNeRF}}$  and a random vector  $\mathbf{z} \in \mathcal{Z}$ . Then, we can use a critic network  $C(\cdot)$  to score the latent codes generated by GNeRF mapping (as real) and our mapping (as fake). Hence, we can train our mapping  $M_{\text{rec}}$  along with the critic network  $C(\cdot)$  in the WGAN framework using the following loss functions:

$$\mathcal{L}_C^{\text{WGAN}} = \mathbb{E}_{\mathbf{w} \sim M_{\text{GNeRF}}(\mathbf{z})} [C(\mathbf{w})] - \mathbb{E}_{\hat{\mathbf{w}} \sim M_{\text{rec}}([\mathbf{n}, \mathbf{t}])} [C(\hat{\mathbf{w}})] \quad (3.21)$$

$$\mathcal{L}_{M_{\text{rec}}}^{\text{WGAN}} = \mathbb{E}_{\hat{\mathbf{w}} \sim M_{\text{rec}}([\mathbf{n}, \mathbf{t}])} [C(\hat{\mathbf{w}})] \quad (3.22)$$

In addition to the WGAN training, we feed the generated latent code  $\hat{\mathbf{w}} = M_{\text{rec}}([\mathbf{n}, \mathbf{t}])$  to the GNeRF model to generate the face image  $\hat{\mathbf{I}} = G(\hat{\mathbf{w}}, \mathbf{c})$ , and then use the generated face image  $\hat{\mathbf{I}}$  to optimize our mapping network  $M_{\text{rec}}(\cdot)$  using the following multi-term loss function:

$$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{ID}}, \quad (3.23)$$

where  $\mathcal{L}^{\text{Pixel}}$  and  $\mathcal{L}^{\text{ID}}$  are pixel loss and ID loss, respectively, and are defined as:

$$\mathcal{L}^{\text{Pixel}} = \mathbb{E}_{\hat{\mathbf{w}} \sim M_{\text{rec}}([\mathbf{n}, \mathbf{t}])} [\|\mathbf{I} - G(\hat{\mathbf{w}}, \mathbf{c})\|_2^2] \quad (3.24)$$

$$\mathcal{L}^{\text{ID}} = \mathbb{E}_{\hat{\mathbf{w}} \sim M_{\text{rec}}([\mathbf{n}, \mathbf{t}])} [\|F_{\text{proxy}}(\mathbf{I}) - F_{\text{proxy}}(G(\hat{\mathbf{w}}, \mathbf{c}))\|_2^2] \quad (3.25)$$

The pixel loss  $\mathcal{L}^{\text{Pixel}}$  minimizes the pixel-level reconstruction error and the ID loss  $\mathcal{L}^{\text{ID}}$  optimizes the model to generate face images that have similar facial templates (extracted by  $F_{\text{proxy}}$ ) to the templates of the original image  $\mathbf{I}$ .

**Supervised learning using synthetic training data** To train our mapping network  $M_{\text{rec}}(\cdot)$  with the synthetic training face images, we use the pretrained GNeRF model to generate a set of random face images  $\{\mathbf{I}_{\text{syn},i}\}_{i=0}^K$ . Therefore, as opposed to real training data, we have the true value of *intermediate* latent space  $\mathbf{w} \in \mathcal{W}_{\text{GNeRF}}$  to generate the same synthetic face image, and therefore can directly learn the GNeRF *intermediate* latent code  $\mathbf{w} = M_{\text{GNeRF}}(\mathbf{z})$  from template  $\mathbf{t}_{\text{syn},i} = F_{\text{template}}(\mathbf{I}_{\text{syn},i})$ . Hence, we consider training our mapping network using the synthetic

data as *supervised* learning. In addition to directly learning the *intermediate* latent code  $\mathbf{w}$ , we use the generated face image to optimize our mapping network by minimizing the following multi-term loss function:

$$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^w + \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{ID}}, \quad (3.26)$$

where  $\mathcal{L}^{\text{Pixel}}$  and  $\mathcal{L}^{\text{ID}}$  are the pixel loss (Eq. 3.24) and ID loss (Eq. 3.25), respectively. Moreover,  $\mathcal{L}^w$  is  $w$ -loss to directly learn the latent space of GNeRF by minimizing the mean squared error between  $\mathbf{w}$  and  $\hat{\mathbf{w}} = M_{\text{rec}}([\mathbf{n}, \mathbf{t}])$  as follows:

$$\mathcal{L}^w = \mathbb{E}_{\mathbf{w} \sim M_{\text{GNeRF}}(\mathbf{z})} [\|\mathbf{w} - M_{\text{rec}}([\mathbf{n}, \mathbf{t}])\|_2^2] \quad (3.27)$$

To train our networks, we use Adam [178] optimizer and optimize the parameters of our new mapping network  $M_{\text{rec}}(\cdot)$  for  $\mathcal{L}_{\text{real}}^{\text{rec}}$  (i.e., Eq. 3.23) and  $\mathcal{L}_{\text{syn}}^{\text{rec}}$  (i.e., Eq. 3.26) losses in every iteration of our training process (also shown in Fig. 3.22). However, in the WGAN framework, we update weights of our new mapping network  $M_{\text{rec}}(\cdot)$  and critic network  $C(\cdot)$  every  $n_M^{\text{WGAN}}$  (for minimizing  $\mathcal{L}_{M_{\text{rec}}}^{\text{WGAN}}$  in Eq. 3.22) and every  $n_C^{\text{WGAN}}$  (for minimizing  $\mathcal{L}_C^{\text{WGAN}}$  in Eq. 3.21) iterations, respectively. Algorithm 3 represents our training process. We should note that our mapping network  $M_{\text{rec}}$  has 2 fully-connected layers with Leaky ReLU activation function.

#### 3.4.1.3 Camera Parameters Optimization

After generating a 3D reconstruction of face from the facial template using our proposed method described in Section 3.4.1.2, the adversary needs to select a pose to generate a 2D reconstructed face image to inject into the system or perform a presentation attack. To this end, during the inference stage we can optimize the camera parameters to find a pose that increases the success attack rate (SAR). In other words, having the 3D reconstruction of a face, we would like to find the camera parameters so that the 2D generated face image has a facial template that is more similar to the leaked templates than the templates of any other pose. Among different camera parameters  $\mathbf{c}$ , we consider the parameters that corresponds to the camera rotations and therefore can change the pose of the generated face image. It is noteworthy that by changing the camera rotations, we want to vary the pitch and yaw rotations of the reconstructed face and do not want to modify the roll rotation. As a matter of fact, the effect of any roll rotation will be eliminated in the FR system through the face alignment in the pre-processing step of the feature extraction. We consider two different approaches to optimize camera parameters as follows:

**Grid Search (GS)** In our grid search approach, we consider pre-defined steps to change the camera pitch  $\theta \in \Theta$  and yaw  $\psi \in \Psi$  and generate corresponding camera parameters  $\mathbf{c}$ . We generate the 2D face images for all values of camera rotation steps ( $\theta_{\text{step}}$  and  $\psi_{\text{step}}$ ) and find the facial templates for each generated image. Finally, we select the face image  $\hat{\mathbf{I}} = G(M_{\text{rec}}([\mathbf{n}, \mathbf{t}]), \mathbf{c})$  which has a template  $\hat{\mathbf{t}} = F_{\text{template}}(\hat{\mathbf{I}})$  that minimizes the mean squared

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

---

**Algorithm 3** Training process of our new mapping network.

---

**Require:**  $\theta_M$ , parameters of  $M_{\text{rec}}(\cdot)$  network.  $\theta_C$ , parameters of network  $C(\cdot)$ .

**Require:**  $n_{\text{epoch}}$ , no. epochs.  $n_{\text{iteration}}$ , no. iterations in each epoch.  $n_M^{\text{WGAN}}$ , no. training iterations after which to optimize  $\theta_M$  in WGAN.  $n_C^{\text{WGAN}}$ , no. training iterations after which to optimize  $\theta_C$  in WGAN.  $\delta$ , the WGAN clipping parameter.

**Require:**  $\alpha_M^{\text{real}}$ , learning rate for optimizing  $\theta_M$  based on  $\mathcal{L}_{\text{real}}^{\text{rec}}$ .  $\alpha_M^{\text{syn}}$ , learning rate for optimizing  $\theta_M$  based on  $\mathcal{L}_{\text{syn}}^{\text{rec}}$ .  $\alpha_M^{\text{WGAN}}$ , learning rate for optimizing  $\theta_M$  in WGAN.  $\alpha_C^{\text{WGAN}}$ , learning rate for optimizing  $\theta_C$  in WGAN.

**Require:**  $\mathcal{D}_{\text{real}}$ , a dataset of real face images and corresponding facial templates extracted using  $F_{\text{template}}$ .

```

1: procedure TRAINING
2:   Initialize  $\theta_C$  and  $\theta_M$ 
3:   for epoch = 1, ...,  $n_{\text{epoch}}$  do
4:     for itr = 1, ...,  $n_{\text{iteration}}$  do
5:       Sample a batch from  $\mathcal{Z}$  and calculate:
6:        $g_{\theta_M}^{\text{syn}} \leftarrow \nabla_{\theta_M} \mathcal{L}_{\text{syn}}^{\text{rec}}$ 
7:        $\theta_M \leftarrow \theta_M - \alpha_M^{\text{syn}} \cdot \text{Adam}(\theta_M, g_{\theta_M}^{\text{syn}})$ 
8:       Sample a batch from  $\mathcal{D}_{\text{real}}$  and calculate:
9:        $g_{\theta_M}^{\text{real}} \leftarrow \nabla_{\theta_M} \mathcal{L}_{\text{real}}^{\text{rec}}$ 
10:       $\theta_M \leftarrow \theta_M - \alpha_M^{\text{real}} \cdot \text{Adam}(\theta_M, g_{\theta_M}^{\text{real}})$ 
11:      if itr mod  $n_M^{\text{WGAN}} = 0$  then
12:         $g_{\theta_M}^{\text{WGAN}} \leftarrow \nabla_{\theta_M} \mathcal{L}_M^{\text{WGAN}}$ 
13:         $\theta_M \leftarrow \theta_M - \alpha_M^{\text{WGAN}} \cdot \text{Adam}(\theta_M, g_{\theta_M}^{\text{WGAN}})$ 
14:      end if
15:      if itr mod  $n_C^{\text{WGAN}} = 0$  then
16:        Sample a batch  $w \sim \mathcal{W}_{\text{GNeRF}}$  and calculate:
17:         $g_{\theta_C}^{\text{WGAN}} \leftarrow \nabla_{\theta_C} \mathcal{L}_C^{\text{WGAN}}$ 
18:         $\theta_C \leftarrow \theta_C - \alpha_C^{\text{WGAN}} \cdot \text{Adam}(\theta_C, g_{\theta_C}^{\text{WGAN}})$ 
19:         $\theta_C \leftarrow \text{clip}(\theta_C, -\delta, \delta)$ 
20:      end if
21:    end for
22:  end for
23: end procedure

```

---

error with the leaked template  $t$ :

$$\min_{\theta, \psi} \|\hat{t} - t\|_2^2, \quad (3.28)$$

Note that the grid search can be applied in both *whitebox* and *blackbox* scenarios (i.e., all attacks defined in Section 3.4.1.1) using the FR model  $F_{\text{template}}$ .

**Continuous Optimization (CO)** For continuous optimization, we start from the frontal camera parameters and use the Adam [178] optimizer to solve the following minimization

using the mapped latent code  $\hat{\mathbf{w}} = M_{\text{rec}}([\mathbf{n}, \mathbf{t}])$ :

$$\min_{\theta, \psi} \|F_{\text{template}}(G(\hat{\mathbf{w}}, \mathbf{c})) - \mathbf{t}\|_2^2, \quad (3.29)$$

By solving this optimization, we can find the  $\theta$  and  $\psi$  rotations and the corresponding camera parameters  $\mathbf{c}$  that lead to a face image with the template close to the leaked template  $\mathbf{t}$ . In contrast to the grid search, the continuous optimization approach can be applied only when the adversary has the *whitebox* knowledge of  $F_{\text{template}}$  (i.e., attack 1 and attack 2).

### 3.4.2 Experiments

In this section, we evaluate the vulnerability of SOTA FR systems to our TI attacks defined in Section 3.4.1. First, in Section 3.4.2.1 we describe our experimental setup. In Section 3.4.2.2, we consider the case where the adversary can inject the reconstructed face image as a query to the system to impersonate, and present our experimental results. In Section 3.4.2.3, we consider the situation where the adversary uses the reconstructed face images to perform presentation attacks and evaluate the vulnerability of SOTA FR systems. Finally, we discuss our findings in Section 3.4.2.4.

#### 3.4.2.1 Experimental Setup

**Face recognition models** In our experiments, we evaluate the vulnerability of different SOTA FR models to our TI attacks. We consider two SOTA models, including ArcFace [132], ElasticFace [133], as the models from which templates are leaked (i.e.,  $F_{\text{template}}$ ) and use our proposed method to reconstruct face images. Then, to evaluate the transferability of reconstructed face images, we also use four different FR models with SOTA backbones from FaceX-Zoo [156] for the target FR system (i.e.,  $F_{\text{target}}$ ), including AttentionNet [141], HRNet [138], RepVGG [145], and Swin [147]. The recognition performances of these models are reported in Table A.1 of Appendix A.

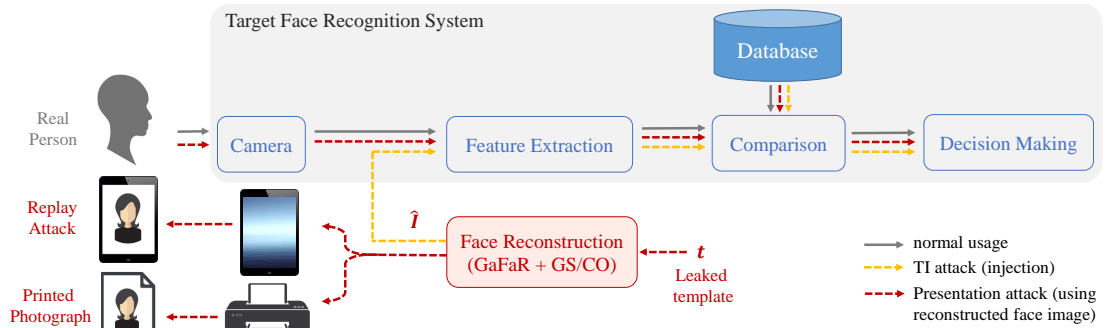


Figure 3.23: Block diagram of a FR system and data flows in normal usage (gray solid arrows), TI attack by injecting the reconstructed face image (orange dashed arrows), and performing presentation attack using the reconstructed face image (red dashed arrows).

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

---

**Datasets** All the FR models used in our experiments are trained on the MS-Celeb1M dataset [157]. However, we assume that the adversary does not have knowledge about the training data of the FR network (either  $F_{\text{template}}$  or  $F_{\text{target}}$ ), and uses another dataset for training the face reconstruction model. We use the Flickr-Faces-HQ (FFHQ) dataset [186], which consists of 70,000 high-resolution (i.e.,  $1024 \times 1024$ ) face images crawled from the internet (without identity labels), for training our 3D face reconstruction model. We randomly split the FFHQ dataset to train (90%) and validation (10%) subsets.

To evaluate the vulnerability of FR systems to TI attacks, we consider two other different face image datasets with identity labels, including the MOBIO [158] and Labeled Faces in the Wild (LFW) [159] datasets. The MOBIO dataset includes face images captured using mobile devices from 150 people in 12 sessions (6-11 samples in each session). The LFW dataset includes 13,233 face images of 5,749 people collected from the internet, where 1,680 people have two or more images.

**Evaluation Protocol** To implement each of the attacks described in Section 3.4.1.1, we build one or two separate FR systems using the same or two different SOTA feature extractor models (based on the attack type). If the target FR system is the *same* as the system from which the template is leaked (i.e.,  $F_{\text{template}} = F_{\text{target}}$ , as in attack 1 and attack 3), we have only one FR system. Otherwise, if the target system is *different* than the system from which the template is leaked (i.e.,  $F_{\text{template}} \neq F_{\text{target}}$ , as in attack 2, attack 4, and attack 5), we have two FR systems with *two* different feature extractors. We should note that in the transferability evaluations, we need that the subjects whose templates are leaked to be enrolled in the target system too. Therefore, to implement any of the attacks which require two FR systems (i.e., attack 2, attack 4, and attack 5), we use one of our evaluation datasets to build both FR systems (i.e.,  $F_{\text{template}}$  and  $F_{\text{target}}$ ).

To evaluate the vulnerability to all our TI attacks, we assume that the target FR system is configured at the threshold corresponding to a false match rate (FMR) of  $10^{-2}$  or  $10^{-3}$ , and we evaluate the adversary's success attack rate (SAR) in entering that system. In our experiments, we consider two situations, where the adversary can inject the reconstructed face image as a query to the FR system (Section 3.4.2.2), or use the reconstructed face image to perform a presentation attack (Section 3.4.2.3). Fig. 3.23 depicts and compares two scenarios of injecting the reconstructed face image or performing a presentation attack. In our evaluation of TI attacks by injecting the reconstructed face image (Section 3.4.2.2), we directly inject the reconstructed face images into the feature extractor of the FR system and evaluate the TI attack in terms of SAR. However, in our evaluation of the presentation attack using the reconstructed face image (Section 3.4.2.3), we present the reconstructed face image (using either a digital screen or a printed photograph) in front of the camera and evaluate the attack in terms of SAR.

**Implementation Details and Source Code** For the GNeRF model, we use the pretrained model of EG3D with StyleGAN [66] backbone to generate 3D faces with  $512 \times 512$  high-resolution images from any arbitrary view.

To train our 3D face reconstruction networks, we consider  $n_{\text{epoch}} = 15$ ,  $n_C^{\text{WGAN}} = 4$  and  $n_M^{\text{WGAN}} = 2$  in Algorithm 3. Furthermore, the input noise vectors to the mapping network of GNeRF’s pretrained network (i.e.,  $\mathbf{z} \in \mathcal{Z}$ ) and to our mapping network  $M_{\text{rec}}$  (i.e.,  $\mathbf{n} \in \mathcal{N}$ ) are both from the standard normal distribution and with 512 and 16 dimensions, respectively. The *intermediate* latent space of GNeRF model has  $14 \times 512$  dimensions, i.e.,  $\mathcal{W}_{\text{GNeRF}} \subset \mathbb{R}^{14 \times 512}$ . The templates extracted by the FR models in Table A.1 of Appendix A have 512 dimensions. For simplicity in training our mapping network, we assume that our training face images from the FFHQ dataset (i.e., real data) are frontal.

In our experiments, we use the continuous optimization (in *whitebox* attacks only) and grid search optimization (in both *whitebox* and *blackbox* attacks) in the inference stage, as described in Section 3.4.1.3, to optimize camera parameters. In the grid search approach, we consider  $\psi \in [-45^\circ, +45^\circ]$  and  $\theta \in [-30^\circ, +30^\circ]$  for a  $11 \times 11$  grid with step sizes of  $\psi_{\text{step}} = 9^\circ$  and  $\theta_{\text{step}} = 6^\circ$ . For the continuous optimization, we use Adam optimizer [178] with the learning rate of  $10^{-2}$  and 121 iterations. An ablation study on the effect of these hyperparameters and the corresponding execution times are reported in Section 3.4.2.4.

We should note that the source code and the captured images for our presentation attack evaluation are publicly available<sup>21</sup>.

#### 3.4.2.2 TI Attack by Injecting Reconstructed Face Images

In this section, we consider the situation where the adversary can inject the reconstructed face image to the feature extractor of the target FR system. We consider SOTA FR models and evaluate the vulnerability of these systems to different TI attacks described in Section 3.4.1.1 in the *whitebox* (attacks 1-2) and *blackbox* (attacks 3-5) scenarios.

**Whitebox Scenario** In attacks 1-2, we assume that the adversary has the *whitebox* knowledge of the FR system from which the template is leaked (i.e.,  $F_{\text{template}}$ ) and uses the same feature extraction model for training (i.e.,  $F_{\text{proxy}}$ ) the face reconstruction network. We considered ArcFace and ElasticFace models for the system from which the template is leaked (i.e.,  $F_{\text{template}}$ ) and evaluate the vulnerability of SOTA FR systems as the target FR systems against attacks 1-2. Table 3.15 compares the vulnerability of different target systems to attacks 1-2 using our method<sup>22</sup> in terms of adversary’s SAR at the system’s FMR of  $10^{-3}$ . As this table shows, our proposed face reconstruction method achieves considerable SAR values against ArcFace and ElasticFace target FR systems in attack 1. Comparing the SAR values between attack 1 and

<sup>21</sup>Project page: <https://www.idiap.ch/paper/gafar>

<sup>22</sup>Note that as reported in Table 2.1, none of the *whitebox* face reconstruction methods in the literature has an available source code, and we neither could reproduce their results.



Figure 3.24: Sample face images from the FFHQ dataset (first row) and their corresponding frontal face reconstruction (second row) as well as reconstructed face images within the camera parameters sub-grid (third row) using our method in the *whitebox* TI attacks (i.e., attacks 1-2) against ArcFace. The values below each image show the cosine similarity between templates of original and frontal reconstructed face images.



### 3.4 3D Face Reconstruction

Table 3.15: Evaluation of whitebox attacks (i.e., attacks 1-2) against SOTA FR models in terms of adversary’s success attack rate (SAR) when injecting reconstructed face image generated using our face reconstruction methods evaluated on the MOBIO and LFW datasets. All the values are in percentage and SAR values correspond to the threshold where the target system has  $FMR = 10^{-3}$ . **M1: GaFaR [ours]**, **M2: GaFaR+GS [ours]**, and **M3: GaFaR+CO [ours]**. Cells are color-coded according to the type of attack as defined in Section 3.4.1 for attack 1 (dark green) and attack 2 (light green).

$F_{\text{template}}$	$F_{\text{target}}$	MOBIO			LFW		
		M1	M2	M3	M1	M2	M3
<b>ArcFace</b>	ArcFace	84.29	86.67	<b>89.52</b>	79.74	82.38	<b>84.43</b>
	ElasticFace	78.10	78.10	<b>80.00</b>	65.19	67.81	<b>70.37</b>
	AttentionNet	65.24	67.14	<b>69.05</b>	30.20	33.43	<b>35.36</b>
	HRNet	62.86	61.43	<b>67.14</b>	30.41	33.26	<b>35.42</b>
	RepVGG	45.24	49.05	<b>55.24</b>	18.38	20.32	<b>21.39</b>
	Swin	70.95	71.90	<b>77.14</b>	51.18	53.91	<b>55.91</b>
<b>ElasticFace</b>	ArcFace	51.43	59.52	<b>61.43</b>	53.07	58.89	<b>61.08</b>
	ElasticFace	78.10	83.33	<b>84.29</b>	63.06	68.94	<b>71.78</b>
	AttentionNet	48.10	51.43	<b>56.67</b>	21.69	25.35	<b>26.92</b>
	HRNet	51.43	50.95	<b>54.29</b>	22.39	26.45	<b>28.23</b>
	RepVGG	37.14	40.95	<b>48.10</b>	12.80	14.72	<b>15.97</b>
	Swin	54.29	54.29	<b>60.00</b>	40.47	43.29	<b>45.59</b>

attack 2, the SAR values degrade for different target FR models in attack 2. However, the reconstructed face images are transferable and can still be used to enter a target system with a different feature extractor. It is also noteworthy that considering the recognition performances in Table A.1 of Appendix A, we can conclude that the target FR system with a higher recognition accuracy is generally more vulnerable to attack 2. For example, when ArcFace is used for  $F_{\text{template}}$  in Table 3.15, attacks against ElasticFace and Swin as target FR systems result in the highest SAR, and there is the same order for their recognition performance in Table A.1 of Appendix A. Comparing the frontal reconstructed face images by our proposed method (GaFaR) with our camera parameter optimizations methods (GaFaR+GS and GaFaR+CO), the results show that camera parameter optimization methods improve SAR in both attack 1 and attack 2. Therefore, camera parameter optimization methods not only enhance the attack against the same system (i.e., attack 1), but are also transferable to other FR systems (i.e., attack 2). Comparing the grid search and continuous optimization methods for camera parameter optimization, the results show that the continuous optimization method achieves higher SAR values, and therefore further enhances our TI attack. Fig. 3.24 illustrates sample face images and their corresponding frontal face reconstruction as well as a sub-grid of reconstructed face images with different poses from ArcFace templates in the *whitebox* TI attacks (i.e., attacks 1-2). We should note that the reconstructed face images in attack 1 and attack 2 are the same, however, they are used to enter different target FR systems.

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.16: Evaluation of blackbox attacks (i.e., attacks 3-5) against SOTA FR models in terms of adversary’s success attack rate (SAR) when injecting reconstructed face image generated using different face reconstruction methods evaluated on the MOBIO and LFW datasets. All the values are in percentage and SAR values correspond to the threshold where the target system has  $FMR = 10^{-3}$ . **M1**: NbNetB-M [53], **M2**: NbNetB-P [53], **M3**: NbNetA-M [53], **M4**: NbNetA-P [53], **M5**: Dong *et al.* [57], **M6**: Vendrow and Vendrow [60], **M7**: GaFaR [ours]. and **M8**: GaFaR+GS [ours]. Cells are color-coded according to the type of attack as defined in Section 3.4.1 for attack 3 ( yellow ), attack 4 ( orange ), and attack 5 ( red ).

$F_{\text{template}}$	$F_{\text{loss}}$	$F_{\text{target}}$	MOBIO								LFW							
			M1	M2	M3	M4	M5	M6	M7	M8	M1	M2	M3	M4	M5	M6	M7	M8
ArcFace	El.Face	ArcFace	5.24	15.24	0.0	1.90	3.33	26.19	50.95	<b>62.86</b>	16.83	40.25	4.32	10.97	13.24	57.44	52.72	<b>61.71</b>
		ElasticFace	4.29	10.95	0.0	1.43	3.81	17.14	52.38	<b>55.24</b>	13.09	34.41	3.32	8.56	6.25	29.06	43.59	<b>47.57</b>
		AttentionNet	3.81	6.67	0.0	2.86	2.86	5.71	29.05	<b>36.19</b>	1.89	7.21	0.51	1.22	2.12	9.79	14.58	<b>17.13</b>
		HRNet	4.29	6.19	0.0	1.43	3.33	10.48	31.90	<b>41.90</b>	2.03	7.77	0.41	1.44	1.70	9.51	14.77	<b>17.35</b>
		RepVGG	1.90	2.38	0.0	1.90	2.38	3.81	28.10	<b>32.86</b>	0.86	4.06	0.23	0.76	1.39	4.48	8.42	<b>10.11</b>
		Swin	4.29	13.33	0.0	0.95	4.29	13.81	43.33	<b>50.00</b>	8.44	23.82	1.60	4.79	6.22	20.75	29.69	<b>33.16</b>
El.Face	ArcFace	ArcFace	5.71	18.57	0.0	2.38	3.81	11.43	74.29	<b>77.62</b>	20.38	48.66	7.50	15.33	12.23	36.80	71.48	<b>74.30</b>
		ElasticFace	16.19	43.81	2.38	3.33	8.10	38.10	84.76	<b>88.10</b>	26.96	58.15	10.88	21.45	12.69	53.06	74.77	<b>78.18</b>
		AttentionNet	1.43	18.1	0.0	1.90	4.29	7.14	62.38	<b>65.24</b>	3.85	16.37	1.53	2.89	3.16	11.16	34.28	<b>37.34</b>
		HRNet	6.19	20.0	0.0	0.48	4.76	11.43	61.90	<b>65.71</b>	4.10	18.36	1.74	3.45	2.47	11.81	35.87	<b>39.19</b>
		RepVGG	7.62	13.81	0.0	0.0	4.29	5.71	47.62	<b>51.43</b>	1.75	9.14	0.66	1.54	2.01	6.04	21.12	<b>22.84</b>
		Swin	16.19	26.19	0.0	0.95	6.67	17.62	67.14	<b>70.95</b>	15.72	38.76	4.13	9.18	8.51	24.22	55.51	<b>58.12</b>

**Blackbox Scenario** In attacks 3-5, we assume that the adversary has the *blackbox* knowledge of the feature extractor of the FR system from which the template is leaked (i.e.,  $F_{\text{template}}$ ) and uses another feature extraction model for training (i.e.,  $F_{\text{proxy}}$ ). Similar to whitebox experiments, we consider ArcFace and ElasticFace models for  $F_{\text{template}}$  and evaluate the vulnerability of SOTA FR systems in the target FR systems against attacks 3-5. In each case, we also use the other model for  $F_{\text{proxy}}$  (i.e., ArcFace as  $F_{\text{template}}$  and ElasticFace as  $F_{\text{proxy}}$  or ElasticFace as  $F_{\text{template}}$  and ArcFace as  $F_{\text{proxy}}$ ). Table 3.16 compares the performance of our method with *blackbox* methods in the literature [53], [57], [60] for attacks 3-5 in terms of adversary’s SAR at system’s FMR of  $10^{-3}$ . As the results in this table show, the frontal face reconstruction by our method (i.e, GaFaR) achieves superior performance than previous methods in the literature. Moreover, when we apply camera parameter optimization (i.e., GaFaR+GS) the performance of our attack improves up to 11.91%, 3.98%, and 10.00% compared to our frontal face reconstruction (i.e, GaFaR) in attack 3, attack 4, and attack 5, respectively. Comparing the use of ArcFace and ElasticFace as  $F_{\text{proxy}}$ , the results show that the SAR values in attacks with the ArcFace model are higher. This can be due to the fact that according to Table A.1 of Appendix A, ArcFace has a better recognition performance than ElasticFace.

Table 3.16 also shows that SOTA FR systems are vulnerable to our TI attacks in the *blackbox* scenario. In particular, in attack 5 which is the hardest TI attack, where  $F_{\text{target}}$ ,  $F_{\text{template}}$ , and  $F_{\text{proxy}}$  are different, the results show that SOTA FR models (as the target FR system) are still vulnerable to our TI attack. The results of attack 5 for our proposed method also show the transferability of our attack to different FR systems. In addition, similar to the *whitebox* scenario, we can also observe that for TI attacks in the *blackbox* scenario, the FR model with a higher recognition performance is generally more vulnerable to our TI attacks. Comparing the results in Table 3.16 and Table 3.15 and as expected, attack 1 is the easiest attack with the



Figure 3.25: Sample face images from the FFHQ dataset (first row) and their corresponding frontal (second row) reconstructed face images using our method in the *blackbox* attack against ElasticFace using ArcFace as  $F_{\text{proxy}}$ . The values below each image show the cosine similarity between templates of original and frontal reconstructed face images.



Figure 3.26: Our evaluation setup for performing different types of presentation and capturing presentation using mobile devices (a) replay attack using Apple iPad Pro, and (b) presentation attack using printed photograph

highest SAR, where  $F_{\text{template}}$ ,  $F_{\text{proxy}}$ , and  $F_{\text{target}}$  are the same, and attack 5 is the most difficult attack, where  $F_{\text{template}}$ ,  $F_{\text{proxy}}$ , and  $F_{\text{target}}$  are different. Fig. 3.25 shows sample face images and their corresponding frontal face reconstruction as well as their sub-grids of reconstructed face images with different poses from ElasticFace templates in the *blackbox* TI attack (i.e., attacks 3-5) using ArcFace as  $F_{\text{proxy}}$ . Similar to attacks 1-2, the reconstructed face images in attacks 3-5 are the same, however, they are used to enter different target FR system.

### 3.4.2.3 Practical Presentation Attack using Reconstructed Face Images

In this section, we consider the situation where the adversary uses the reconstructed face image to perform a presentation attack to enter the target FR system. We consider reconstructed face images from ArcFace templates using our proposed face reconstruction method and camera parameter optimizations (i.e., GaFaR, GaFaR+GS, and GaFaR+CO) in both *white-box* and *blackbox* scenarios, and use the reconstructed face images in each case to perform presentation attacks. We perform our presentation attacks against different SOTA FR systems based on the various TI attacks described in Section 3.4.1.1. Therefore, we similarly have five different presentation attacks according to the adversary's knowledge of the FR system from which the template is leaked (i.e.,  $F_{\text{template}}$ ) and the target FR system (i.e.,  $F_{\text{target}}$ ). We also assume that the adversary can use the reconstructed face images to perform two types of attacks as follows:

- *Presentation attack via digital replay (replay attack)*: In this type of presentation attack, the adversary presents the reconstructed face image using a digital display in front of the camera. To perform this attack, we use a tablet (Apple iPad Pro) showing the reconstructed face image and put it in front of the camera of the target FR system.
- *Presentation attack via printed photograph*: In this type of presentation attack, the adversary prints the reconstructed face image and presents the printed photograph. To perform this attack, we print the reconstructed face images with a colorful laser printer (Develop Ineo+C364e) on typical papers and present the printed photograph in front of the camera of the target FR system.

To perform the presentation attacks (with either digital replay or printed photograph), the reconstructed image should be presented in front of the camera of the target FR system. For each of these cases, we considered three different mobile devices, including Apple iPhone 12, Xiaomi Redmi 9A, and Samsung Galaxy S9, as the camera of the target FR system and capture images from the presentations. Fig. 3.26 shows our evaluation setup for capturing presentation attacks from tablet and printed photographs using different mobile cameras. It is noteworthy that we used the default display scale on the digital screen (i.e., iPad), in which the reconstructed face images with  $512 \times 512$  resolution do not cover all the screen. However, the face area in the captured images is still larger than the required resolution to feed to be used in the target FR systems.

Fig. 3.27 illustrates a sample face image from the MOBIO dataset, its reconstructed face images from ArcFace templates using our different methods (GaFaR, GaFaR+GS, and GaFaR+CO) in the *whitebox* and *blackbox* (using ElasticFace as  $F_{\text{proxy}}$ ) scenarios, and captured images from the reconstructed face images using different mobile devices in replay attacks and presentation attacks using printed photographs. As this figure shows, the captured images from replay attacks are more similar to the reconstructed face images, while the ones from printed photographs suffer from quality degradation. In addition, different mobile devices introduce different sensor qualities, and therefore different image qualities for the captured images in our experiment. We use the captured images<sup>23</sup> by each mobile device from presentation attacks as inputs to different SOTA FR systems as target FR systems, and evaluate the vulnerability of these FR systems to the presentation attack using the reconstructed face images.

Table 3.17 reports the result of the vulnerability evaluation against SOTA FR systems to TI attacks (by injecting the reconstructed face images in our simulation), and different presentation attacks (digital replay attack and printed photograph) in the *whitebox* and *blackbox* scenarios in terms of SAR<sup>24</sup>. It is noteworthy that based on the presentation type, we have two types of

<sup>23</sup>The reconstructed face images and all captured images for our presentation attack evaluation are publicly available.

<sup>24</sup>According to the ISO/IEC 30107-3 standard [203], the adversary's success attack rate in the evaluation of presentation attack is reported in terms of the Impostor Attack Presentation Match Rate (IAPMR). However, for consistency with our experiments in Section 3.4.2.2, we use "SAR" to report the success attack rate in the evaluation of our presentation attacks using reconstructed face images too.



### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

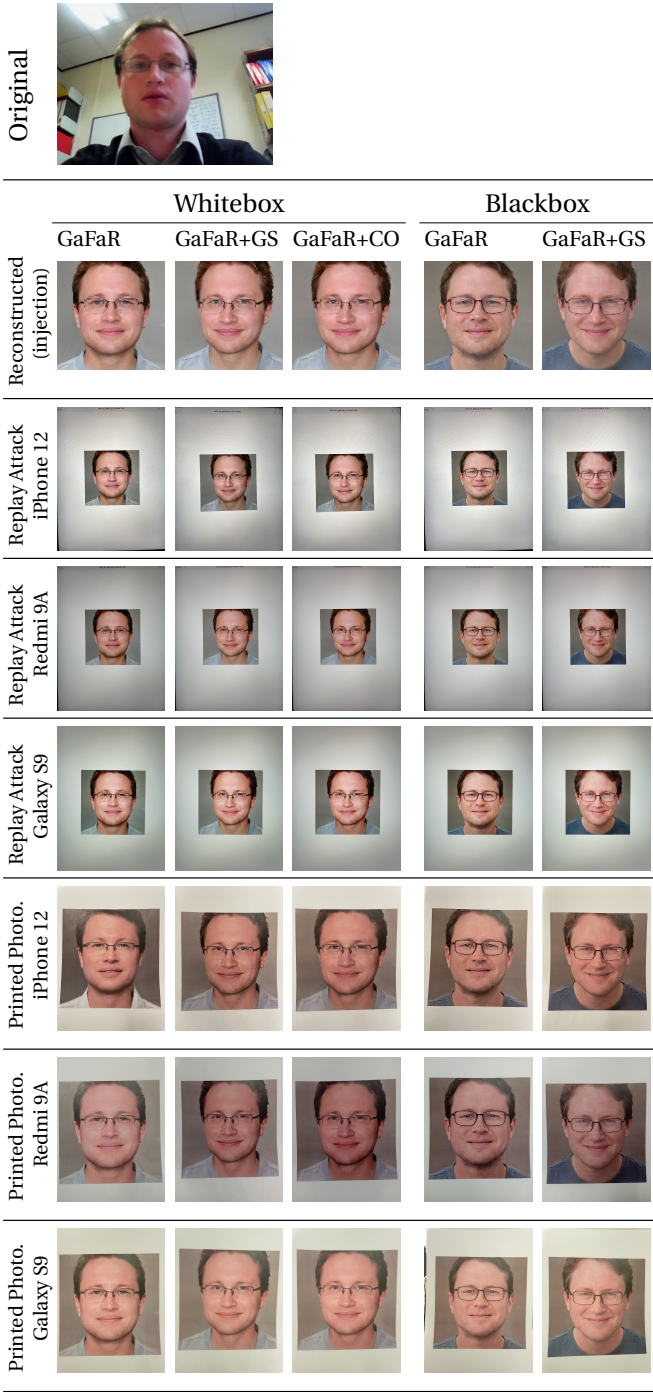


Figure 3.27: A sample image from the MOBIO dataset, its corresponding reconstructed face images using our face reconstruction methods (i.e. GaFaR, GaFaR+GS, and GaFaR+CO) in the *whitebox* and *blackbox* scenarios, the corresponding digital replay attacks and presentation attacks using printed photographs captured with different mobile devices.

### 3.4 3D Face Reconstruction

Table 3.17: Vulnerability evaluation of the simulation (i.e., injection) and practical *whitebox* and *blackbox* TI attacks using ArcFace templates against different FR systems as target in terms of SAR/IAPMR for FR systems with FMR of  $10^{-3}$  evaluated on the MOBIO dataset. The values are in percentage and the best values of SAR for different reconstruction methods are embolden in each attack. Cells are color-coded according to the type of attack as defined in Section 3.4.1 for attack 1 (dark green), attack 2 (light green), attack 3 (yellow), attack 4 (orange), and attack 5 (red).

Scenario	Attack type	Device	Reconstruction Method	F <sub>tagret</sub> (SAR/IAPMR)					
				ArcFace	ElasticFace	AttentionNet	HRNet	RepVGG	Swin
whitebox	injection	N/A	GaFaR	84.29	78.10	65.24	62.86	45.24	70.95
			GaFaR+GS	86.67	78.10	67.14	61.43	49.05	71.9
			GaFaR+CO	<b>89.52</b>	<b>80.00</b>	<b>69.05</b>	<b>67.14</b>	<b>55.24</b>	<b>77.14</b>
		iPhone 12	GaFaR	80.48	75.71	61.90	59.52	47.14	68.57
			GaFaR+GS	<b>85.71</b>	<b>79.05</b>	66.19	<b>61.43</b>	<b>50.95</b>	71.90
			GaFaR+CO	83.81	76.67	<b>68.10</b>	60.95	50.00	<b>72.86</b>
	Replay Attack	Redmi 9A	GaFaR	80.00	76.67	62.86	61.43	47.14	69.52
			GaFaR+GS	<b>86.19</b>	<b>79.05</b>	67.62	<b>65.71</b>	50.00	<b>74.29</b>
			GaFaR+CO	<b>86.19</b>	78.10	<b>70.48</b>	65.24	<b>51.90</b>	75.24
		Galaxy S9	GaFaR	75.71	72.38	58.10	49.52	40.95	60.95
			GaFaR+GS	80.95	73.81	62.86	<b>55.24</b>	42.38	63.81
			GaFaR+CO	<b>81.90</b>	<b>75.24</b>	<b>64.76</b>	<b>55.24</b>	<b>43.33</b>	<b>64.29</b>
		iPhone 12	GaFaR	65.24	56.19	49.52	49.05	37.62	53.33
			GaFaR+GS	82.86	71.43	<b>66.67</b>	61.43	46.67	68.10
			GaFaR+CO	<b>83.81</b>	<b>73.81</b>	64.76	<b>62.38</b>	<b>50.00</b>	<b>71.43</b>
	Print Photograph	Redmi 9A	GaFaR	74.76	66.19	57.14	54.76	44.29	64.29
			GaFaR+GS	<b>85.24</b>	73.33	65.71	<b>63.33</b>	47.14	68.10
			GaFaR+CO	83.81	<b>74.76</b>	<b>67.62</b>	62.86	<b>51.90</b>	<b>69.05</b>
		Galaxy S9	GaFaR	71.90	64.29	58.57	54.76	42.86	64.76
			GaFaR+GS	83.33	70.48	<b>65.71</b>	60.48	48.10	68.57
			GaFaR+CO	<b>83.33</b>	<b>72.86</b>	64.76	<b>61.90</b>	<b>51.43</b>	<b>69.05</b>
blackbox	injection	N/A	GaFaR	50.95	52.38	29.05	31.90	28.10	43.33
			GaFaR+GS	<b>62.86</b>	<b>55.24</b>	<b>36.19</b>	<b>41.90</b>	<b>32.86</b>	<b>50.00</b>
		iPhone 12	GaFaR	47.14	<b>51.43</b>	30.95	32.38	26.19	42.38
			GaFaR+GS	<b>54.76</b>	50.95	<b>38.10</b>	<b>39.05</b>	<b>32.86</b>	<b>47.14</b>
	Replay Attack	Redmi 9A	GaFaR	48.10	50.48	28.57	33.33	26.67	43.81
			GaFaR+GS	<b>58.57</b>	<b>52.86</b>	<b>36.19</b>	<b>39.05</b>	<b>31.43</b>	<b>47.62</b>
		Galaxy S9	GaFaR	42.86	<b>47.62</b>	27.62	28.57	23.81	41.9
			GaFaR+GS	<b>50.95</b>	46.67	<b>34.29</b>	<b>36.19</b>	<b>27.14</b>	<b>42.86</b>
		iPhone 12	GaFaR	42.86	<b>46.19</b>	30.95	32.86	25.24	43.81
			GaFaR+GS	<b>51.90</b>	<b>46.19</b>	<b>38.57</b>	<b>35.71</b>	<b>34.29</b>	<b>49.52</b>
	Print Photograph	Redmi 9A	GaFaR	41.90	47.62	29.05	28.10	28.10	40.95
			GaFaR+GS	<b>54.29</b>	<b>49.05</b>	<b>38.10</b>	<b>36.67</b>	<b>33.33</b>	<b>47.14</b>
		Galaxy S9	GaFaR	44.76	<b>48.10</b>	28.10	30.95	<b>32.86</b>	44.76
			GaFaR+GS	<b>54.29</b>	47.14	<b>36.19</b>	<b>36.19</b>	<b>32.38</b>	<b>50.00</b>

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

presentation attacks (replay attack and printed photograph), and based on the adversary’s knowledge of the FR system from which the template is leaked (i.e.,  $F_{\text{template}}$ ) and the target FR system (i.e.,  $F_{\text{target}}$ ), we have five different TI attacks (as described in Section 3.4.1.1) and thus five different corresponding presentation attacks. The results in Table 3.17 show that SOTA FR models as target systems are vulnerable to our attacks. In general, and as also seen in Section 3.4.2.2, attack 1 is the easiest attack, and as the adversary’s knowledge becomes more limited, the attack gets more difficult in attack 2, attack 3, attack 4, and attack 5, respectively. Comparing our different reconstruction methods (i.e., GaFaR, GaFaR+GS, and GaFaR+CO), we can observe that camera parameter optimizations improve SAR values. The results also show that replay attacks achieve higher SAR values compared to presentation attacks using printed photographs. Comparing the results in Table 3.17 for different mobile devices, the SAR values are comparable across different methods and in different attack types.

We also compare the performance of our method with two best *blackbox* methods in the literature from Table 3.16 (i.e., NBNetB-P [53] and Vebdrow and Vendrow [60]) in presentation attacks based on TI attacks 3-5 against SOTA FR models. Table 3.18 reports this evaluation for digital replay presentation attack (captured by Apple iPhone 12) based on TI attacks using ArcFace templates against SOTA FR models in terms of adversary’s SAR at the system’s FMR of  $10^{-3}$  on the MOBIO dataset. The results in this table show that our method still achieves superior performance than previous methods in the literature. Comparing this table with Table 3.16, we can see there are in average -4.7%, 0%, -0.87%, and -2.69% changes in the SAR values in presentation attacks than the injection of reconstructed face images (Table 3.16) for NBNetB-P [53], Vebdrow and Vendrow [60], GaFaR, GaFaR+GS, respectively.

Table 3.18: Comparison of our proposed method with previous *blackbox* TI methods in practical presentation attacks (replay attacks captured by iPhone 12) using ArcFace templates against different FR system (i.e., attacks 3-5) in terms of SAR/IAPMR at FMR of  $10^{-3}$  on the MOBIO dataset. The values are in percentage and the best values are embolden in attack against each FR system. Cells are color-coded according to the type of attack as defined in Section 3.4.1 for attack 3 ( yellow ), attack 4 ( orange ), and attack 5 ( red ).

Reconstruction Method	$F_{\text{target}}$ (SAR/IAPMR)					
	ArcFace	ElasticFace	AttentionNet	HRNet	RepVGG	Swin
NBNetB-P [53]	9.05	2.38	3.81	3.81	0.95	6.19
Vendrow & Vendrow [60]	25.24	10.48	7.14	10.95	7.62	15.71
GaFaR [ours]	47.14	<b>51.43</b>	30.95	32.38	26.19	42.38
GaFaR+GS [ours]	<b>54.76</b>	50.95	<b>38.10</b>	<b>39.05</b>	<b>32.86</b>	<b>47.14</b>

#### 3.4.2.4 Discussion

Our experiments in Section 3.4.2.2 show that our proposed method outperforms previous methods in the literature in TI attacks against FR systems. To evaluate the effect of each part in our proposed method, we perform an ablation study and train different models. To this end, we evaluate the effect of *semi-supervised* learning approach in our method compared to fully



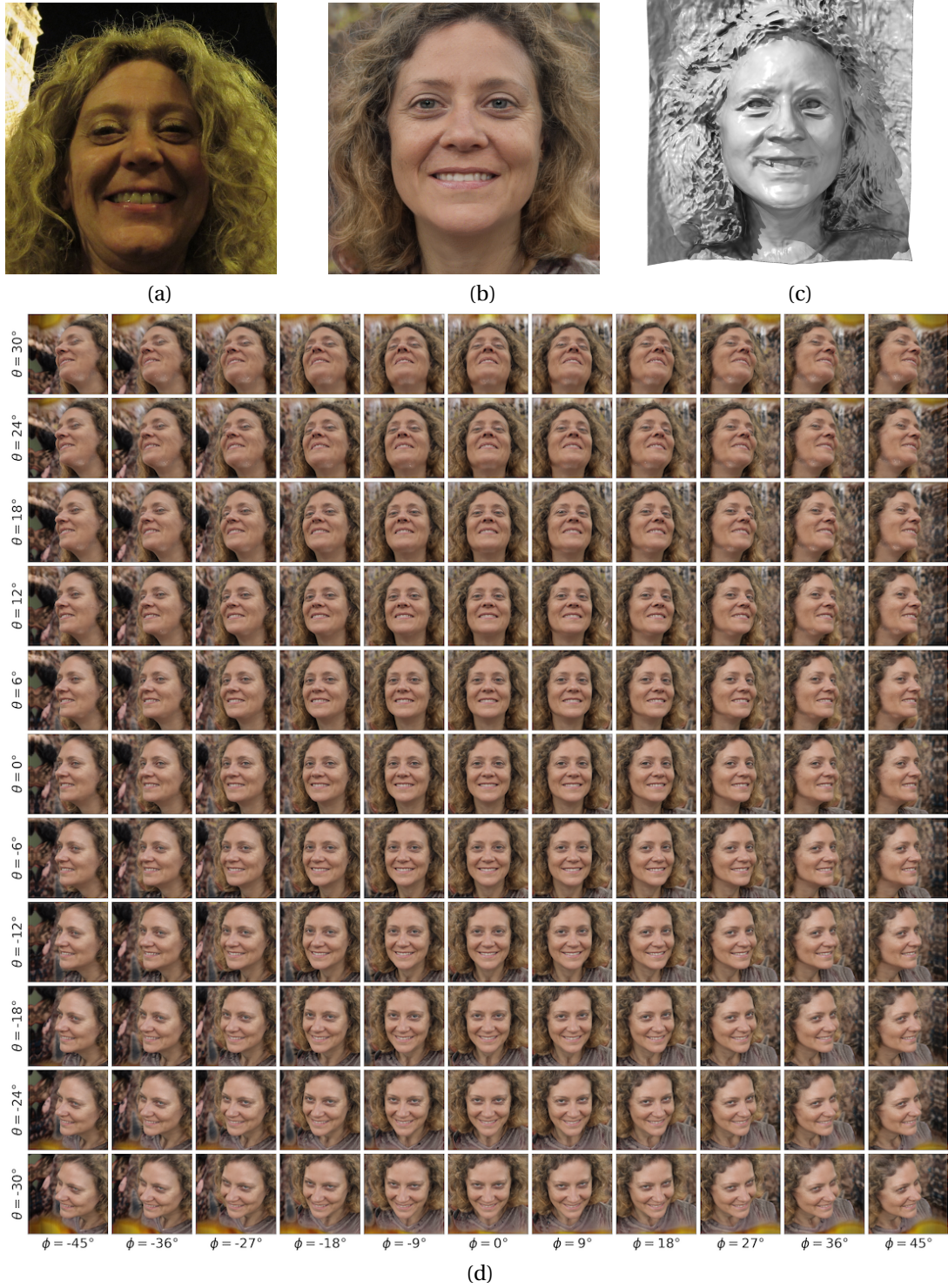


Figure 3.28: (a) a sample face image from the FFHQ dataset, (b) its frontal reconstructed face image, (c) its 3D face reconstruction, and (d) the corresponding reconstructed face images with camera parameters grid using our method in the *whitebox* attack against ArcFace. The cosine similarity between templates of original (a) and frontal (b) reconstructed face images is 0.679.

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.19: Ablation study on the proposed *semi-supervised* learning approach and evaluation of the effect of loss terms in attack 1 against ArcFace model in terms of success attack rate (SAR) on the MOBIO and LFW datasets. The SAR values are in percentage and for an attack without any camera parameter optimization (i.e., GS/CO).

approach	Loss Functions	MOBIO		LFW	
		FMR=10 <sup>-2</sup>	FMR=10 <sup>-3</sup>	FMR=10 <sup>-2</sup>	FMR=10 <sup>-3</sup>
supervised	$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^w + \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{ID}}$	90.96	82.38	83.80	69.467
	$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^w + \mathcal{L}^{\text{Pixel}}$	43.81	8.57	31.75	13.92
	$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^w + \mathcal{L}^{\text{ID}}$	0	0	0.86	0.30
	$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^w$	32.38	9.52	33.69	15.43
unsupervised	$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{ID}}$ [without WGAN]	0	0	0.44	0.15
	$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{ID}}$ [with WGAN]	70.48	31.90	67.72	45.76
	$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}^{\text{ID}}$ [with WGAN]	52.86	19.52	54.51	30.83
	$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}^{\text{Pixel}}$ [with WGAN]	0	0	2.21	0.40
<b>semi-supervised Eqs. 3.21,3.22,3.23,3.26</b>		<b>95.71</b>	<b>82.86</b>	<b>89.27</b>	<b>79.84</b>

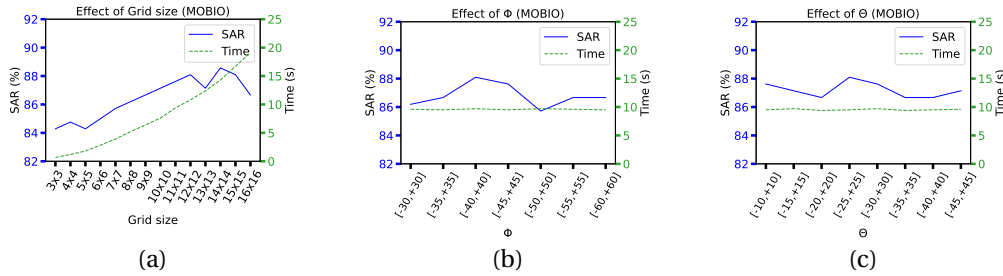


Figure 3.29: Ablation study on the effect of different hyperparameters in grid search for camera parameters optimization in terms of success attack rate (SAR) and average execution time for each image reconstruction for whitebox attack (i.e., attack 1) against a FR system based on ArcFace configured at FMR=10<sup>-3</sup> on the MOBIO dataset: a) grid size, b) interval of  $\Phi$ , and c) interval of  $\Theta$ .

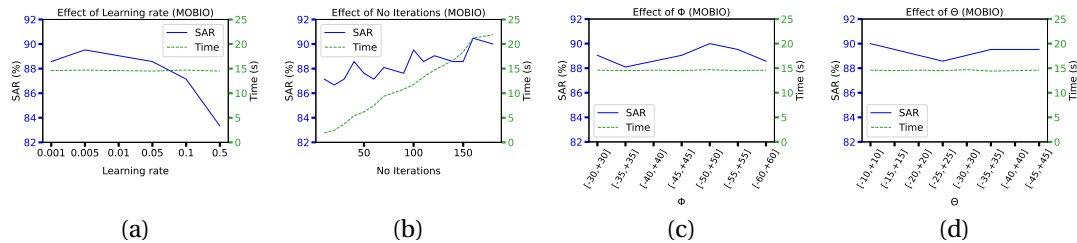


Figure 3.30: Ablation study on the effect of different hyperparameters in continuous optimization for camera parameters in terms of success attack rate (SAR) and average execution time for each image reconstruction for whitebox attack (i.e., attack 1) against a FR system based on ArcFace configured at FMR=10<sup>-3</sup> on the MOBIO dataset: a) learning rate, b) number of iterations, c) interval of  $\Phi$ , and d) interval of  $\Theta$ .

*supervised* learning (i.e., using only synthetic data where we have the corresponding latent code for each template) and fully *unsupervised* learning approach (i.e., using only real data where we do not have the corresponding latent code for each template). In each of fully *supervised* learning and fully *unsupervised* learning approaches, we also evaluate the effect of each loss function. In the case of the fully *unsupervised* learning approach, we also evaluate the effect of adversarial learning in our method. Table 3.19 reports our ablation study on the effect of each part in our proposed method in attack 1 (injection) against ArcFace model on the MOBIO and LFW datasets in terms of SAR at system's FMR of  $10^{-2}$  and  $10^{-3}$ . As the results of our ablation study show, the proposed *semi-supervised* approach has a better reconstruction performance (in terms of SAR) than fully *supervised* learning and fully *unsupervised* learning approaches. Moreover, our ablation study on the effect of loss terms shows that each of the loss terms has an important impact on the performance of our face reconstruction network. In particular, using WGAN for our *unsupervised* learning (i.e., using real training data where we don't have the true value of *intermediate* latent codes for each training data) helps our mapping network  $M_{\text{rec}}$  to learn the distribution of GNeRF *intermediate* latent space  $\mathcal{W}_{\text{GNeRF}}$ . However, if we do not use WGAN in training with real data, our mapping network  $M_{\text{rec}}$  cannot learn the distribution of GNeRF *intermediate* latent space  $\mathcal{W}_{\text{GNeRF}}$ , and therefore the generated latent codes by our mapping network  $M_{\text{rec}}$  will be out of distribution  $\mathcal{W}_{\text{GNeRF}}$ . This will cause the generator part of GNeRF to generate non-face-like images. In addition to WGAN training, the results in Table 3.19 show that each of the pixel loss and ID loss terms enhances the reconstruction performance of our method in training with either synthetic (*supervised* learning) or real (*unsupervised* learning) data.

As another ablation study, we evaluate the effect of hyperparameters in the camera parameter optimization for our proposed grid search (GS) and continuous optimization (CO) approaches. For the grid search optimization approach, in our experiments in Sections 3.4.2.2 and 3.4.2.3, we considered  $\psi \in [-45^\circ, +45^\circ]$  and  $\theta \in [-30^\circ, +30^\circ]$  for a  $11 \times 11$  grid with step sizes of  $\psi_{\text{step}} = 9^\circ$  and  $\theta_{\text{step}} = 6^\circ$ . Fig. 3.28 illustrates a sample face image from the FFHQ dataset and its frontal and 3D reconstruction as well as the grid of reconstruction with the size of  $11 \times 11$  and camera parameters  $\psi \in [-45^\circ, +45^\circ]$  and  $\theta \in [-30^\circ, +30^\circ]$ . For our ablation study, we use the same hyperparameters and only change one of these hyperparameters (i.e., grid size, interval of  $\Phi$ , and interval of  $\Theta$ ) to evaluate its effect on the performance of our method in terms of SAR and average execution time. Fig. 3.29 reports our ablation study in the attack 1 (injection) against the ArcFace FR system configured at FMR= $10^{-3}$  on the MOBIO dataset. The results in this figure show that the intervals of  $\Phi$  and  $\Theta$  are not required to be very large. Moreover, by increasing the size of our search grid (i.e., the number of steps) we can achieve a better SAR with the cost of a higher execution time. For the continuous optimization approach, in our experiments in Sections 3.4.2.2 and 3.4.2.3, we considered  $\psi \in [-45^\circ, +45^\circ]$  and  $\theta \in [-30^\circ, +30^\circ]$  and used Adam optimizer [178] with 121 iterations and the learning rate of  $10^{-2}$ . Similarly, for the ablation study, we use the same hyperparameters and only change one of these hyperparameters (i.e., learning rate, number of iterations, interval of  $\Phi$ , and interval of  $\Theta$ ) to evaluate its effect on the performance of our method in terms of SAR and average execution time. Fig. 3.30 reports

our ablation study in the attack 1 (injection) against the ArcFace FR system configured at  $\text{FMR}=10^{-3}$  on the MOBIO dataset. According to these results, similar to the ablation study for the grid search optimization, the intervals of  $\Phi$  and  $\Theta$  should not be necessarily very large. In addition, similar to the effect of the grid size in the grid search optimization, by increasing the number of iterations we can achieve a better SAR with the cost of a higher execution time.

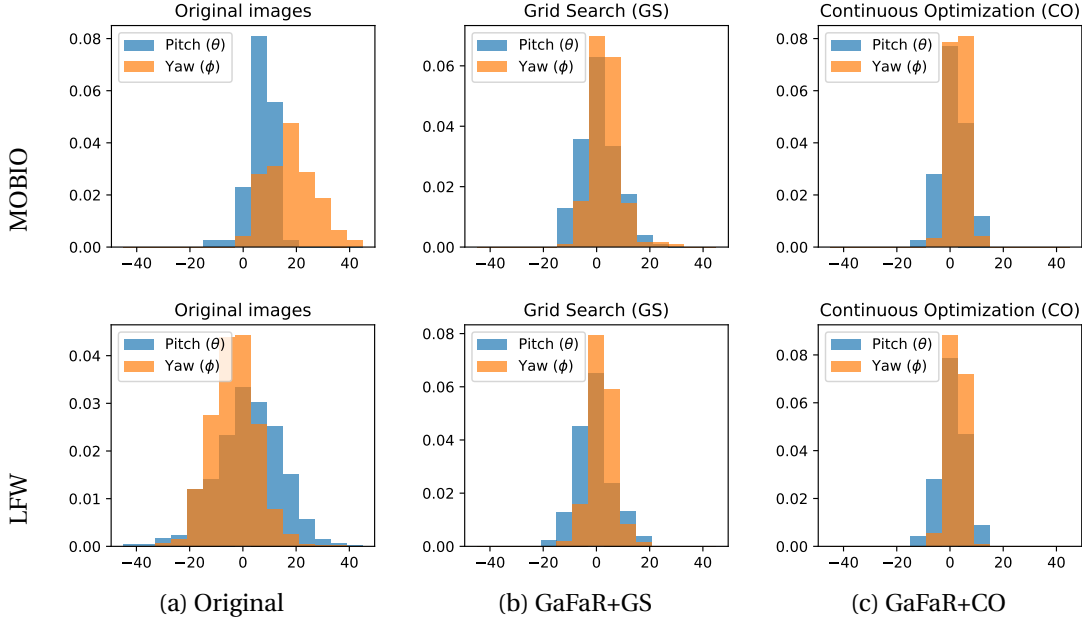


Figure 3.31: Histogram of pitch and yaw in (a) original, (b) GaFaR+GS, (c) GaFaR+CO for attack 1 against ArcFace on the MOBIO (first row) and LFW (second row) datasets. Note that for GaFaR without any camera parameter optimization, the reconstructed face images are frontal (i.e., pitch and yaw values are zero), and thus the histogram for GaFaR is not depicted in this figure.

According to the results in Tables 3.15, 3.16, and 3.17, our camera parameter optimization methods improve the performance of our face reconstruction network. In particular, we observe that GaFaR+GS and GaFaR+CO also improve the SAR in attacks against different target FR systems (i.e., transferability evaluation in attacks 2, 4, and 5) too. This shows that our camera parameter optimization methods improve the attacks in the way that the reconstructed face images have more similar templates to templates of the original face images, even if extracted by a different FR model. Achieving such improvements in attacks against different target FR systems shows the transferability of our pose-optimized reconstructed face images.

We further investigate the effect of our camera parameter optimization methods on our attacks. In attack 1 against ArcFace, our grid search method increases the similarity between templates of original and reconstructed face images for 89.52% and 88.70% of cases on the MOBIO and LFW datasets, respectively. Moreover, our continuous optimization method increases the similarity between templates for 99.04% and 98.66% of reconstructed face images on the MOBIO and LFW datasets, respectively<sup>25</sup>. We also use the pose estimation model in [204] to

<sup>25</sup>These results can also explain the superiority of GaFaR+CO compared to GaFaR+GS in Table 3.15 and Table 3.17.



find the histograms of the pose of original and reconstructed face images in attack 1 against<sup>26</sup> ArcFace on the MOBIO and LFW datasets. As the histograms in this figure show, most of the pose-optimized reconstructed face images have a small variation around the frontal pose. This observation is also consistent with our ablation study in Fig. 3.29 and Fig. 3.30, where we see that the intervals of  $\Phi$  and  $\Theta$  are not required to be very large. In addition, Fig. 3.31 also shows that the pose of reconstructed face images does not have the same distribution as that of the original face images. This demonstrates that our camera parameter optimization methods (CO or GS) do not try to find the same pose as the original images, but rather try to find a pose that has a template with higher similarity to the leaked template. Our transferability evaluations in Tables 3.15, 3.16, and 3.17 (i.e., attacks 2, 4, and 5) also confirm that the pose-optimized reconstructed face images also achieve better performance in attacks (either inject or even presentation attack) against different FR systems. Therefore, 3D reconstruction is essentially more useful than 2D reconstruction to generate better 2D reconstructed face images in our attacks. Fig. 3.32 shows sample reconstructed face images from the MOBIO dataset in *whitebox* and *blackbox* (using ElasticFace) TI attacks using our different reconstruction methods. We can observe that our camera parameter optimization leads to different poses to increase SAR.

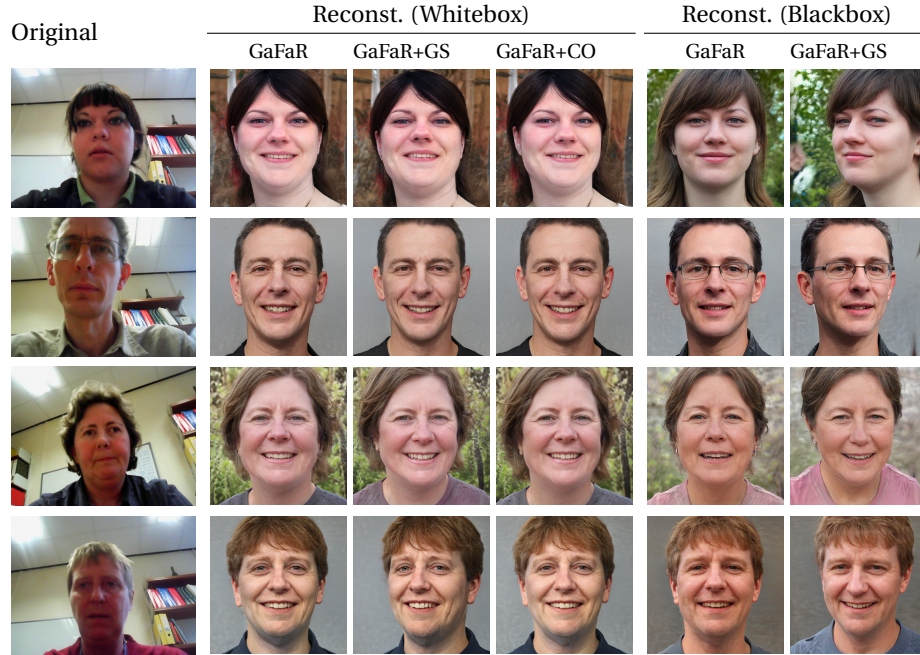


Figure 3.32: Reconstruction of sample images from the MOBIO dataset in *whitebox* and *blackbox* (using ElasticFace) TI attacks against ArcFace templates using our methods.

Comparing our result in *whitebox* (Table 3.15) and *blackbox* (Table 3.16) attacks in Section 3.4.2.2, we observe that our proposed face reconstruction network, GaFaR, achieves better performance in *whitebox* attacks (attacks 1-2) than *blackbox* attacks (attacks 1-2) when inverting ArcFace templates (i.e., ArcFace as  $F_{\text{template}}$ ). However, in inverting ElasticFace tem-

<sup>26</sup>We should note that since we use the same reconstructed face images for injection and presentation attacks, the histograms in Fig. 3.31 are valid for both injection and presentation attacks.

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

Table 3.20: Success attack rates in whitebox (attack 1) and blackbox (attack 3) TI attacks with our method, GaFaR, against different target FR systems in terms of SAR at FMR of  $10^{-3}$  on the MOBIO and LFW datasets. In whitebox attacks the same model, and in blackbox attacks the ArcFace model is used as  $F_{\text{proxy}}$ .

	MOBIO					LFW				
	Elas.Face	Att.Net	HRNet	RepVGG	Swin	Elas.Face	Att.Net	HRNet	RepVGG	Swin
whitebox	78.10	64.29	71.43	53.81	<b>94.76</b>	63.06	27.00	31.87	17.33	<b>74.08</b>
blackbox	<b>84.76</b>	<b>72.38</b>	<b>76.67</b>	<b>72.86</b>	89.05	<b>74.77</b>	<b>33.59</b>	<b>37.80</b>	<b>25.40</b>	67.11

plates, the results show that GaFaR achieves better performance in *blackbox* attacks (attacks 3-5) than *whitebox* attacks (attacks 1-2). As a matter of fact, the difference in *whitebox* and *blackbox* attacks in our method is the FR model used as  $F_{\text{proxy}}$  for training our network. In *blackbox* attacks against ElasticFace templates, the ArcFace model is used as  $F_{\text{proxy}}$  while in *whitebox* attacks, the ElasticFace model is used as  $F_{\text{proxy}}$ . Similarly, Table A.1 of Appendix A also shows that ArcFace has a superior recognition performance than ElasticFace, and thus it can more help the training of the face reconstruction network. To further investigate the effect of  $F_{\text{proxy}}$  for difference attacks, as another experiment we compare the performance of our method in *whitebox* attacks (attack 1) and *blackbox* attacks (attack 3 using ArcFace as  $F_{\text{proxy}}$ ) against different FR systems on the MOBIO and LFW datasets. As the results in Table 3.20 show, in all cases except attacks against Swin, *blackbox* attacks with ArcFace as  $F_{\text{proxy}}$  achieve superior performance than *whitebox* attacks for templates of different FR models. In contrast to other FR models in our experiments which are CNN-based, Swin is a transformer-based FR model, which can be the reason why in *blackbox* attacks with Swin templates using ArcFace (which is a CNN-based FR model) as  $F_{\text{proxy}}$  could not lead to superior performance.

In our experiments in Section 3.4.2.3, we used the reconstructed face images to perform presentation attacks against FR systems. For our presentation attacks, we assumed that the FR system does not have any presentation attack detection<sup>27</sup> (PAD) module, and therefore the captured images from the presentation attack were directly given to the feature extractor of the FR system. In Appendix C, we further consider a FR system equipped with a commercial PAD model and investigate the performance of PAD system in detecting the presentation attacks based on the reconstructed face images. The results show that presentation attacks can still pass the commercial PAD system, demonstrating the vulnerability of FR systems.

In drawing our discussion to a close, our experiments in Section 3.4.2.2 show the vulnerability of SOTA FR systems to TI attacks using our face reconstruction methods (GaFaR, GaFaR+GS, and GaFaR+CO). Similarly, our experiments in Section 3.4.2.3 show that the reconstructed face images by our proposed methods can be used for presentation attacks against the same FR system or different FR systems that the corresponding user is enrolled (i.e., transferability of the reconstructed face images). In fact, our experiments show potential threats that can seriously jeopardize the security and privacy of users if the facial templates are leaked. In addition to the experiments in Section 3.4.2.2 and Section 3.4.2.3, we should note that our

<sup>27</sup>also referred to as anti-spoofing.

proposed method can generate 3D face from facial templates (as shown in Fig 3.19 and Fig. 3.28). Such 3D reconstruction can be used for more sophisticated presentation attacks (e.g., 3D face mask, etc.) against FR systems, which require further studies in future works.

### 3.5 Conclusion

In this chapter, we considered different scenarios and extensively investigated the vulnerability of face recognition systems to template inversion attacks. Our experiments demonstrate critical vulnerability in these systems. First, the template inversion provides a good estimation of the face image of underlying user. The reconstructed face images not only reveal important information of users but can also be recognised as the same user by the system, raising a critical security issue. In addition, we showed that an adversary can reconstruct high-resolution or 3D face from face templates. The reconstructed face images achieved high success attack rates.

In Section 3.1, we proposed a face reconstruction network based on a new block, DSCasConv, and trained our network with a multi-term loss function. We measured the vulnerability of FR systems to our TI attack in terms of the SAR. We evaluated the vulnerability of SOTA FR models (with different backbones and different heads) to our TI method on the MOBIO, LFW, and AgeDB datasets. The experiments show that FR models with higher recognition performance tend to be more vulnerable to this type of attack. Furthermore, changing the backbone may have more effect than changing the head on the vulnerability of the FR models. Our experiments also confirm that the reconstructed face images may reveal important information about each user, including race, gender, age, etc. Therefore, a TI attack, in addition to being a security threat to the FR system itself, can be also considered as a privacy threat to FR systems' users. In Appendix B, we consider the case where the adversary gains access to a portion of face templates and investigate the vulnerability of FR systems to TI attacks based on partially leaked facial templates.

In Section 3.2, we considered StyleGAN, as a pretrained face generation network, and trained a neural network to map face templates to the *intermediate* latent space of StyleGAN. We trained our mapping network with a GAN-based framework to learn the distribution of the *intermediate* latent space of StyleGAN. Then, we used the synthesis network of the pretrained StyleGAN to generate realistic and high-resolution face images from our generated code in the *intermediate* latent space of StyleGAN corresponding to the underlying face template. In addition to the adversarial losses, we also used a multi-term loss reconstruction function, including a pixel loss (minimize pixel-level reconstruction error) and an ID loss (preserve the identity of synthesized images).

In Section 3.3, we used *synthetic* data and proposed a new method to reconstruct high-resolution (i.e.,  $1024 \times 1024$ ) face images from facial templates in TI attacks against FR systems. We used a face generator network to generate synthetic face images and extracted their facial templates to build our training dataset. Then, we used our generated training dataset to learn a mapping from facial templates to the intermediate latent space of the face generator

### Chapter 3. Vulnerability Analysis of Unprotected Systems: Face Reconstruction from Facial Templates

---

network using a multi-term loss function. The experimental results show the vulnerability of FR systems to TI attacks based on the reconstructed face images with our model (trained only with synthetic train data) on real face datasets.

In Section 3.4, we presented a comprehensive vulnerability evaluation of SOTA FR systems to TI attacks using 3D face reconstruction from facial templates. We proposed a new method (called GaFaR) to reconstruct 3D faces from facial templates using a geometry-aware face generation network based on GNeRF. We learned a mapping from facial templates to the *intermediate* latent space of the GNeRF model with a *semi-supervised* learning approach using real and synthetic training data. For the real data, where we do not have correct *intermediate* latent code, we used a GAN-based training to learn the distribution of *intermediate* latent space of the GNeRF model (*unsupervised* learning). For the synthetic data, we have the corresponding *intermediate* latent code and directly learn the mapping (*supervised* learning). In addition, we proposed two optimization methods on the camera parameters in GNeRF to find a pose that improves the TI attack: grid search and continuous optimization. In the grid search method, we considered a grid for pitch and yaw rotations of the reconstructed face, and in continuous optimization, we used a gradient-based optimizer to optimize camera parameters.

We proposed our method in the *whitebox* and *blackbox* attacks against face recognition systems and comprehensively evaluated the vulnerability of SOTA FR systems to our method. Considering *whitebox* and *blackbox* blackbox scenarios and adversary’s knowledge of target FR system, we defined five types of TI attacks and evaluated the *transferability* of our reconstructed face images across other FR systems on the MOBIO and LFW datasets. We evaluated the TI attacks by injecting reconstructed face images as queries to the target FR systems. In addition, we performed practical presentation attacks against SOTA FR systems using digital screen replay and printed photographs of reconstructed frontal and pose-optimized face images. Our experiments showed the vulnerability of SOTA FR models to our TI attacks and also presentation attacks using our reconstructed face images. Last but not least, our proposed method can generate 3D faces from facial images, and we used the 3D reconstruction to find a pose that improves the adversary’s success attack rate. However, 3D reconstruction of users’ faces paves the way for new types of attacks (e.g., 3D face masks, etc.), which need to be investigated in the future. We further evaluated the performance of a commercial PAD system in detecting our presentation attacks using the reconstructed face images from our TI attacks in Appendix C. The results show that the presentation attacks can still pass the commercial PAD system, demonstrating the critical vulnerability in FR systems and motivates the necessity to protect biometric templates.



## 4 Protection of Biometric Templates

In chapter 3, we showed the vulnerability of unprotected face recognition systems to template inversion attacks. Our experiments demonstrate critical vulnerability in these systems which hinder security and privacy of users. Such vulnerabilities motivates to develop systems based on biometric template protection mechanisms. In this chapter, we focus on template protection and propose new schemes to protect biometric templates. In section 4.1, we propose MLP-Hash, which is a new cancelable biometric scheme based on random multi-layer perceptrons (MLP). We also present a new hybrid template protection scheme using Homomorphic Encryption (HE) and cancelable biometrics in section 4.2, which leverages advantages of cancelable biometrics and HE with faster execution time compared to HE. The proposed methods in section 4.1 and section 4.2 are presented for face characteristics, but can be extended to other modalities as well<sup>1</sup>. In section 4.3, we focus on finger vein images and propose a new framework to enhance and protect finger vein recognition systems.

### 4.1 MLP-Hash: Protecting Biometric Templates via Randomized MLP

In this section, we propose a new cancelable biometric template protection scheme, dubbed MLP-Hash, which includes a non-linear projection step through a user-specific randomly-weighted multi-layer perceptron (MLP), followed by a binarization step. We employ the user's private key to initialize the MLP with random orthonormal values. Then, we project the templates to a new space through the initialized MLP, which contains nonlinear activation functions. Finally, at the output layer, we binarize the final layer of the MLP to generate the protected template.

We evaluate the unlinkability and irreversibility properties of our template protection method to fulfill the ISO/IEC 24745 standard [9] requirements. We also consider two scenarios when evaluating the method's recognition accuracy: the *normal* scenario (which is the expected

---

<sup>1</sup>In 5.1 of chapter 5, we evaluate the performance of MLP-Hash on voice, iris, and finger vein data

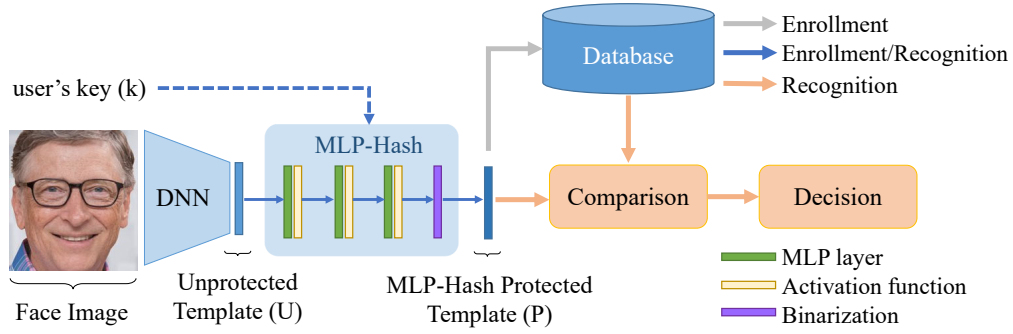


Figure 4.1: Block diagram of MLP-Hash protected face recognition system

scenario in practice) and the *stolen token* scenario (which is the case when the user's MLP-Hash key is stolen). Then, we evaluate the protected templates of three SOTA face recognition methods (i.e., ArcFace [132], FaceNet [73], and InceptionResnetV2-CenterLoss [134]) on the Labeled Faced in the Wild (LFW) [159] and MOBIO [158] datasets. Our experiments show that MLP-Hash achieves promising performance in protecting SOTA face recognition systems.

The rest of this section is organized as follows. First, we describe our biometric template protection method in section 4.1.1. Then, in section 4.1.2, we evaluate MLP-Hash in terms of unlinkability, irreversibility, and recognition accuracy.

#### 4.1.1 Proposed method

Figure 4.1 represents the block diagram of an MLP-Hash protected face recognition system. As depicted in this figure, MLP-Hash uses unprotected features, which are extracted from the user's face image, along with the user's key, to generate the protected template. In section 4.1.1.1, we describe the MLP-Hash algorithm in detail. During the enrollment stage, the protected templates are stored in the system's database and are later compared with the probe template during the recognition stage as described in section 4.1.1.2. We should note that compared to BTP schemes which use neural networks and require training, e.g. [205]–[207], our proposed method does not require training and the weights are specified using the user's key (as described in section 4.1.1.1).

##### 4.1.1.1 MLP-Hash Algorithm

Let  $U$  indicate an unprotected biometric template (i.e., embedding) extracted by a face recognition model. The MLP-Hash protected template,  $P$ , can be generated by algorithm 4 using the user's key,  $k$ , and the unprotected template,  $U$ , in two steps. First,  $U$  is fed into an MLP with  $H$  hidden layers, activation function<sup>2</sup>  $F(\cdot)$ , and the pseudo-random orthonormal weights initialized with seed  $k$ . To generate pseudo-random orthonormal matrix  $\mathbf{M}_{\perp\ell}$  in layer  $\ell$  of the MLP, we first generate a pseudo-random matrix  $\mathbf{M}_{\ell}$ , and then apply the Gram-Schmidt

<sup>2</sup>We use the Rectified Linear Unit (ReLU) activation function which is a non-linear and many-to-one function.

#### 4.1 MLP-Hash: Protecting Biometric Templates via Randomized MLP

---

##### Algorithm 4 MLP-Hash algorithm

---

```

1: Inputs:
2:    $U$  : unprotected biometric template (i.e., embedding)
3:    $H$  : number of MLP hidden layers
4:    $L_{\text{MLP}}$  : set of lengths of MLP layers ( $L_{\text{MLP}}^{(\ell)}$ ), including input layer ( $\ell = 0$ ), hidden layers ( $1 \leq \ell \leq H$ ),
   and output layer ( $\ell = H + 1$ )
5:    $F(\cdot)$  : activation function
6:    $k$  : user's key
7: Output:
8:    $P = \{p_i | i = 1, 2, \dots, L_{\text{MLP}}^{(H+1)}\}$  binary MLP-Hash protected template
9: Procedure:
10:  Step 1: Passing through pseudo-random MLP
11:    Set initial value of  $\Gamma$  with  $U$ 
12:    for  $\ell$  in  $\{1, \dots, H + 1\}$  do
13:      Generate a pseudo-random matrix  $\mathbf{M}_\ell$  based on the user's seed ( $k$ ):  $\mathbf{M}_\ell \in \mathbb{R}^{L_{\text{MLP}}^{(\ell-1)} \times L_{\text{MLP}}^{(\ell)}}$ .
14:      Apply the Gram-Schmidt process on the rows of the generated pseudo-random matrix  $\mathbf{M}_\ell$  to
      transform it into an orthonormal matrix  $\mathbf{M}_{\perp \ell}$ 
15:      Update value of  $\Gamma$  with matrix product of  $\Gamma$  and  $\mathbf{M}_{\perp \ell}$ 
16:      Update value of  $\Gamma$  by applying activation function  $F(\Gamma)$ 
17:    end for
18:  Step 2: Binarizing the output of MLP
19:    Compute  $L_{\text{MLP}}^{(H+1)}$  bits MLP-Hash  $\{p_i | i = 1, 2, \dots, L_{\text{MLP}}^{(H+1)}\}$ 
    from
        
$$p_i = \begin{cases} 0 & \text{if } \Gamma_i \leq \tau \\ 1 & \text{if } \Gamma_i > \tau \end{cases}, \quad i = 1, \dots, L_{\text{MLP}}^{(H+1)},$$

    where  $\tau$  is the average of  $\Gamma$  elements.
20: End Procedure

```

---

orthonormalization process on the rows of  $\mathbf{M}_\ell$ . After feeding the  $U$  into the MLP with the pseudo-random orthonormal weights, in the second step, we binarize the output of MLP to generate the protected template,  $P$ .

##### 4.1.1.2 Comparing MLP-Hash Templates

In the enrollment stage, the reference MLP-Hash templates,  $P$ , should be stored in the system database (ideally separately). In the recognition stage, we use *Hamming* distance to calculate the score between each pair of *probe* and *reference* MLP-Hashed templates. In the subsequent experiments, we consider the MLP-Hash protected face recognition systems operating in verification mode only.

##### 4.1.2 Experiments

In this section, we describe our experiments and evaluate the properties of MLP-Hash as a biometric template protection scheme in accordance with the ISO/IEC 30136 standard. First, in section 4.1.2.1, we describe our experimental setup and the baselines used. Next, we evaluate the unlinkability, irreversibility and recognition accuracy of MLP-Hash in sections

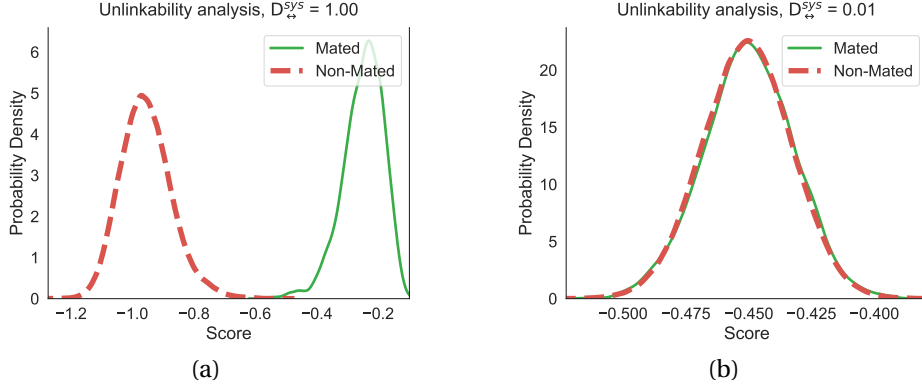


Figure 4.2: Unlinkability evaluation of unprotected and MLP-Hash protected ArcFace templates on the MOBIO dataset: a) Unprotected templates, b) MLP-Hash protected templates.

4.1.2.2, 4.1.2.3, and 4.1.2.4, respectively. We should note that cancelability is inherently satisfied in the MLP-Hash algorithm, since like other *cancelable* BTP methods, we can easily revoke the compromised template in the database, assign a new key for the user, and register the user with a new protected template. Finally, we discuss our experiments in section 4.1.2.5.

### 4.1.2.1 Experimental Setup and Baselines

In our experiments we use the MOBIO [158] and Labeled Faced in the Wild (LFW) [159] databases to evaluate the recognition accuracy of MLP-Hash on SOTA face recognition models. We use three SOTA face recognition models, including ArcFace [132], FaceNet [73], and InceptionResnetV2-CenterLoss [134]. We compare the performance of our template protection method on the same face recognition systems with the BioHashing [76] method and two methods based on Index-of-Maximum (IoM) Hashing [77] (i.e., Gaussian random projection-based hashing, shortly GRP, and uniformly random permutation-based hashing, shortly URP). In each case, we generate protected templates whose length is equal to the length of the embedding (i.e., number of elements in the embedding) for each face recognition model. We set all the hidden layers of MLP-Hash to twice the length of the embeddings for each face recognition model. The number of hidden layers ( $H$ ) was 3 in our experiments.

### 4.1.2.2 Unlinkability Evaluation

To evaluate the unlinkability criterion, we use the framework proposed in [117]. This framework, which is explained in Section 5.1.1.2, uses the score distributions of the *mated* templates (i.e., different templates from the same user) and *non-mated* templates (i.e., templates from different users) to measure unlinkability with respect to the overlap of these two distributions. More particularly, with this evaluation, we expect that in the case of linkable templates, the *mated* and *non-mated* templates score distributions will be separated. However, in the case of unlinkable templates, these distributions should completely overlap. Figure 4.2 compares

#### 4.1 MLP-Hash: Protecting Biometric Templates via Randomized MLP

Table 4.1: Unlinkability evaluation of MLP-Hash, BioHash, IoM-GRP, and IoM-URP protected templates of the ArcFace embeddings in terms of the system’s global unlinkability measure ( $D_{\rightarrow}^{sys}$ ).

MLP-Hash	BioHash	IoM-GRP	IoM-URP
0.010	0.009	0.011	0.007

Table 4.2: Irreversibility evaluation of MLP-Hash, BioHash, IoM-GRP, and IoM-URP protected templates of the ArcFace embeddings in terms of Success Attack Rate (%) on the MOBIO dataset at FMR of  $10^{-2}$  and  $10^{-3}$ .

Configuration	MLP-Hash	BioHash	IoM-GRP	IoM-URP
<b>FMR = <math>10^{-2}</math></b>	39.05	43.81	35.71	14.29
<b>FMR = <math>10^{-3}</math></b>	9.05	10.48	7.14	1.43

the unlinkability of original (unprotected) and MLP-Hash protected ArcFace templates on the MOBIO dataset using this evaluation framework. To calculate the distribution of *mated* scores in this figure, we generate different templates for the same user using different keys, then calculate the scores between these templates. However, for the distribution of *non-mated* scores, we generate protected templates for different users (with different keys) and compute the scores between them. As shown in this figure, while the distributions of *mated* scores and *non-mated* scores are fully separated for unprotected templates, they almost completely overlap for the MLP-Hash protected templates. Furthermore, the value of the system’s global unlinkability measure ( $D_{\rightarrow}^{sys}$ ) is reduced from 1.0 (for the unprotected system) to 0.01 (for the MLP-Hash protected system) by deploying our template protection method, showing that the resulting protected templates are almost fully unlinkable. Table 4.1 compares the unlinkability of MLP-Hash, BioHash, IoM-GRP, and IoM-URP protected templates of the ArcFace embeddings on the MOBIO database. As this table shows, all these template protection schemes have comparable unlinkability and they are almost fully unlinkable.

##### 4.1.2.3 Irreversibility Evaluation

To evaluate the irreversibility of the proposed template protection scheme, we consider the worst-case and most difficult threat model in ISO/IEC 30136 standard (referred to as *full disclosure threat model*), where the attacker knows everything about the system, including algorithms, secret keys, etc. We assume that the attacker would invert the protected template, then use the inverted template to enter a similar unprotected system. Accordingly, we evaluate the irreversibility in term of Success Attack Rate (SAR), which indicates the attacker’s success rate in entering the unprotected system using the inverted templates. Hence, a higher SAR shows that the templates are more invertible, while a lower (or zero) SAR indicates that the protected templates are harder to invert.

## Chapter 4. Protection of Biometric Templates

Table 4.3: Comparison of MLP-Hash-protected, BioHash-protected, IoM-GRP-protected, IoM-URP-protected, and unprotected (Baseline) SOTA Face Recognition models, in terms of TMR (%) in the *normal* and the *stolen* scenarios on the MOBIO and LFW datasets. The threshold in each system is selected individually at an FMR of  $10^{-3}$ . The results are reported as (mean $\pm$ std) for 10 different experimental trials.

Dataset	Model	Baseline	<i>normal scenario</i>				<i>stolen scenario</i>			
			MLP-Hash	BioHash	IoM-GRP	IoM-URP	MLP-Hash	BioHash	IoM-GRP	IoM-URP
MOBIO	ArcFace	100.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	99.59 $\pm$ 0.08	100.00 $\pm$ 0.00	99.95 $\pm$ 0.04	99.98 $\pm$ 0.03	98.88 $\pm$ 0.13
	FaceNet	97.87	99.05 $\pm$ 0.48	99.93 $\pm$ 0.04	99.99 $\pm$ 0.01	95.56 $\pm$ 0.50	76.40 $\pm$ 6.19	89.38 $\pm$ 2.12	94.41 $\pm$ 0.83	87.51 $\pm$ 1.03
	IncResNetV2	96.69	99.98 $\pm$ 0.04	99.99 $\pm$ 0.01	100.00 $\pm$ 0.00	99.96 $\pm$ 0.03	65.12 $\pm$ 5.07	76.46 $\pm$ 7.49	92.56 $\pm$ 2.73	91.38 $\pm$ 1.35
LFW	ArcFace	98.73	98.86 $\pm$ 0.13	98.84 $\pm$ 0.05	99.19 $\pm$ 0.06	88.78 $\pm$ 1.37	95.54 $\pm$ 0.54	98.56 $\pm$ 0.06	98.62 $\pm$ 0.06	84.79 $\pm$ 1.98
	FaceNet	93.17	90.90 $\pm$ 0.90	96.81 $\pm$ 1.12	99.38 $\pm$ 0.09	69.29 $\pm$ 2.99	59.42 $\pm$ 5.02	83.19 $\pm$ 5.32	85.78 $\pm$ 4.92	50.77 $\pm$ 6.96
	IncResNetV2	93.33	99.42 $\pm$ 0.44	99.95 $\pm$ 0.05	100.00 $\pm$ 0.00	98.06 $\pm$ 0.35	47.40 $\pm$ 14.22	64.13 $\pm$ 19.89	83.38 $\pm$ 6.27	81.38 $\pm$ 2.75

To evaluate such an attack, we use a numerical solver (implemented in the SciPy package<sup>3</sup>) to find an estimate of the original template, which is mapped to the same output through the template protection module. The solver starts from an initial guess, and through an iterative process, tries to find an answer which gives the same output (as the given protected template) when passed as the input to the MLP-Hash with the same key. We also assumed that the attacker knows the distribution of unprotected templates, and uses this distribution to extract 10 samples as initial guesses in separate attempts. In each attempt, in the case of convergence of the solver, the inverted template is used to enter an unprotected system with a match threshold at a False Match Rate (FMR) of  $10^{-3}$  (using the same feature extraction module).

Table 4.2 compares the irreversibility of MLP-Hash, BioHash, IoM-GRP, and IoM-URP protected templates of the ArcFace embeddings on the MOBIO database in terms of the SAR. As this table shows, the irreversibility of MLP-Hash is comparable to that of the BioHash and IoM-GRP methods. However, IoM-URP protected templates are more difficult to invert using our adopted inversion technique.

### 4.1.2.4 Recognition Accuracy Evaluation

To evaluate the recognition accuracy of MLP-Hash, we considered two scenarios: the *normal* scenario and the *stolen token* scenario. In the *normal* scenario, which is the expected scenario for most cases, each user's key is assumed to be secret. However, in the *stolen-token* scenario (or briefly *stolen* scenario), we assume that the impostor has access to the user's secret key and uses this key with the impostor's own unprotected template. To implement the *stolen* scenario, in the verification stage we use the same key as the genuine's key for other users in the database to generate their MLP-Hash templates.

Table 4.3 compares the MLP-Hash-protected, BioHash-protected, IoM-GRP-protected, IoM-URP-protected, and unprotected (baseline) templates of the SOTA face recognition models, in terms of True Match Rate (TMR) in the *normal* and the *stolen* scenarios on the MOBIO

<sup>3</sup><https://scipy.org/>

## 4.2 Hybrid Protection of Deep Templates using Cancelable Biometrics and Homomorphic Encryption

Table 4.4: Complexity comparison of template protection methods in terms of average execution time (milliseconds). The results are reported as (mean $\pm$ std) for 1000 different experimental trials.

MLP-Hash	BioHash	IoM-GRP	IoM-URP
61.9 $\pm$ 0.5	12.5 $\pm$ 0.5	77.6 $\pm$ 0.2	36.2 $\pm$ 0.9

and LFW datasets. The threshold in each system is selected individually at an FMR of  $10^{-3}$ . As this table shows, in the normal scenario, all the protection schemes achieve comparable performance on the MOBIO dataset. However, on the LFW dataset, IoM-URP clearly has the worst performance. In the stolen scenario, IoM-GRP appears to perform the best across all three face recognition models and both evaluation datasets.

### 4.1.2.5 Discussion

Table 4.1, Table 4.2, and Table 4.3 compare the unlinkability, irreversibility and recognition accuracy, respectively, of our proposed template protection method with the BioHash, IoM-GRP, and IoM-URP algorithms. Table 4.4 also compares the complexity of the aforementioned methods in generating protected templates from the ArcFace model in terms of average execution time (milliseconds) on a system equipped with an Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz. Based on these results, IoM-URP is the most irreversible algorithm, however it clearly has the worst performance in the normal scenario (which is the expected scenario in practice). IoM-GRP has slightly better irreversibility than MLP-Hash, and its recognition accuracy is the best in most cases. However, it has the longest execution time amongst the studied protection methods. BioHashing has comparable recognition accuracy with MLP-Hash, and has slightly worse irreversibility. However, BioHashing has the shortest execution time. All in all, our experiments show that while all these template protection schemes have comparable unlinkability, there is a trade-off between irreversibility, recognition accuracy, and complexity.

## 4.2 Hybrid Protection of Deep Templates using Cancelable Biometrics and Homomorphic Encryption

Despite several important advantages of HE (such as preservation of biometric recognition accuracy, as well as provable security guarantees), there are two main drawbacks in the application of HE as a BTP scheme. First, if the private (decryption) key is leaked, then the templates can be easily decrypted and inverted, which means that the security of the system solely depends on the secrecy of the keys. Second, the computational complexity of arithmetic operations on the ciphertexts is significant. To address these shortcomings, we propose a hybrid BTP scheme using CB methods and FHE to protect biometric templates (illustrated in Fig. 4.3). The proposed hybrid scheme tackles the security challenge in HE-

## Chapter 4. Protection of Biometric Templates

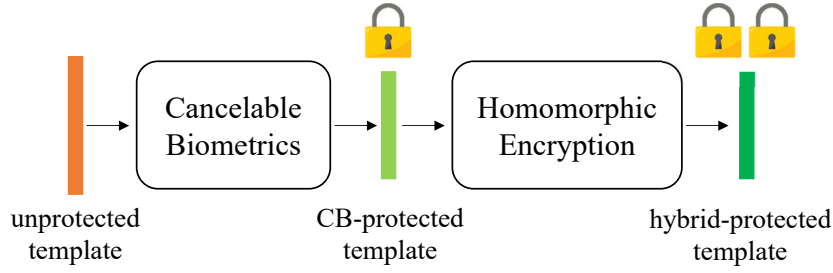


Figure 4.3: General scheme of the proposed hybrid protection method

Table 4.5: Protection of cancelable biometrics (CB), homomorphic encryption (HE), and hybrid (CB+HE) protection against three different threat models in the ISO/IEC 30136 standard.

Protection method	Naive/ Standard	Full disclosure
Cancelable Biometrics	✓	(✓)
Homomorphic Encryption	✓*	✗
Hybrid (CB+HE)	✓*	(✓)

\*provable secure

based systems when the private key is disclosed. In such cases, the CB provides more security for the system and helps to ensure that the protected templates remain irreversible even if an attacker manages to successfully decrypt the HE-protected templates. Table 4.5 compares the protection of cancelable biometrics (CB), homomorphic encryption (HE), and hybrid (CB+HE) protection against three different threat models introduced in the ISO/IEC 30136 standard [10]:

- *Naive threat model* is the case where the adversary has black box knowledge about the protection method, with no further information about the underlying algorithm and any associated secrets. We can also assume that the adversary has access to a small set of protected templates (not a large biometric database).
- *Standard threat model* is the case where the adversary has full knowledge of the protection algorithm, but does not know the secrets and, therefore, cannot execute submodules that require the secrets.
- *Full disclosure threat model* refers to the case where the adversary knows everything about the system, including all the submodules and secrets.

In addition to improving the security of the protected biometric system, our experiments show that CB methods can additionally reduce the dimensionality of templates before applying HE, thereby decreasing the complexity of operations performed on the ciphertexts. The results in [26] also showed that we can reduce dimensionality of the output of BioHashing and still achieve the recognition performance of the baseline system (which uses unprotected templates). In the experiment, we use the following CB methods to generate protected templates



## 4.2 Hybrid Protection of Deep Templates using Cancelable Biometrics and Homomorphic Encryption

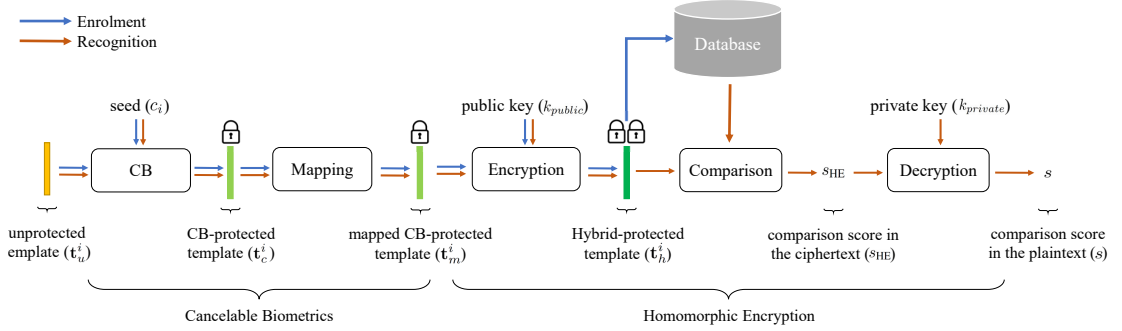


Figure 4.4: Block diagram of the proposed hybrid protection method

prior to the application of HE: BioHashing [76], Multi-Layer Perceptron (MLP) hashing [23], and Index-of-Maximum (IoM) hashing [77]. For each of these CB schemes, we propose a method for computing scores between probe and reference templates in the encrypted domain. We evaluate our proposed hybrid scheme using different state-of-the-art (SOTA) face recognition models (i.e., ArcFace [132], ElasticFace [133], and FaceNet [73]) on the Labeled Faced in the Wild (LFW) [159] and MOBIO [158] datasets.

The remainder of this section is structured as follows. First, we describe our hybrid biometric template protection method in section 4.2.1. Next, in section 4.2.2, we present the experiments and discuss our results.

### 4.2.1 Hybrid template protection method

In general, the input to the proposed template protection method can be the biometric templates extracted from different biometric modalities (e.g., face, speech, fingerprint, iris, finger vein, etc.) and with different data formats (e.g., binary, integer, float, etc.). In section 4.2.1.1, we describe the general formulation of our proposed hybrid protection method. Next, in section 4.2.1.2, we consider the combination of different CB methods (including BioHashing, MLP-Hashing, and IoM Hashing) with an HE algorithm, and we describe our hybrid template protection scheme.

#### 4.2.1.1 Notations and Formulation

Let  $t_u^i$  denote the unprotected template extracted from the biometric data of the subject  $i$ . We can generate the CB-protected template  $t_c^i$  using the CB method  $\mathcal{P}(\cdot, \cdot)$  applied on the unprotected template  $t_u^i$  along with the seed  $c_i$ :

$$t_c^i = \mathcal{P}(t_u^i, c_i) \quad (4.1)$$

To encrypt the CB-protected template  $t_c^i$  using HE, we may need to perform a pre-processing

## Chapter 4. Protection of Biometric Templates

---

step prior to encoding. Therefore, we can define a mapping function  $M_{\mathcal{P}}(.)$  to change the representation of the CB-protected template  $\mathbf{t}_c^i$  and generate the mapped CB-protected template  $\mathbf{t}_m^i$ :

$$\mathbf{t}_m^i = M_{\mathcal{P}}(\mathbf{t}_c^i) \quad (4.2)$$

Next, we can generate the hybrid-protected template  $\mathbf{t}_h^i$  (i.e., the ciphertext) by applying HE-based encryption function  $\text{Enc}_{\text{HE}}(.,.)$  on the mapped CB-protected template  $\mathbf{t}_m^i$ , using the public key  $k_{\text{public}}$ :

$$\mathbf{t}_h^i = \text{Enc}_{\text{HE}}(\mathbf{t}_m^i, k_{\text{public}}) \quad (4.3)$$

In the enrolment stage, the hybrid-protected template  $\mathbf{t}_h^i$  is then stored in the system database as the reference template. In the recognition stage, the hybrid-protected template of the probe should be compared to the reference templates in the homomorphically encrypted domain (i.e., the comparison should be between the corresponding ciphertexts). To calculate the comparison score between the hybrid-protected probe template  $\mathbf{t}_h^{\text{probe}}$  and each hybrid-protected reference template  $\mathbf{t}_h^{\text{ref}}$ , we should employ an appropriate function  $\text{Comp}_{\text{HE}}^{\mathcal{P}}(.,.)$ , which corresponds to the utilized CB method  $\mathcal{P}$ , in the encrypted domain. Hence, we need to compute the score between the reference and probe ciphertexts as follows:

$$s_{\text{HE}} = \text{Comp}_{\text{HE}}^{\mathcal{P}}(\mathbf{t}_h^{\text{probe}}, \mathbf{t}_h^{\text{ref}}) \quad (4.4)$$

Finally, we can decrypt the encrypted score  $s_{\text{HE}}$  to the plaintext using the private key  $k_{\text{private}}$  as below:

$$s = \text{Dec}_{\text{HE}}(s_{\text{HE}}, k_{\text{private}}), \quad (4.5)$$

where  $\text{Dec}_{\text{HE}}(.,.)$  denotes the decryption function of HE. Fig.4.4 illustrates the block diagram of the proposed hybrid BTP scheme. In the subsequent experiments, we will evaluate the proposed protection method on face recognition systems operating in verification mode only.

### 4.2.1.2 Combinations of different CB methods with HE

In the proposed hybrid protection scheme, we can generally use different CB methods and different HE algorithms. We employ three different CB methods, including BioHashing [76], Multi-Layer Perceptron (MLP) Hashing [23], and Index-of-Maximum (IoM) Hashing [77]. For HE, we use the Brakerski/Fan-Vercauteren (BFV) scheme [98], which supports homomorphic operations on integer templates. Since the aforementioned CB methods generate binary and integer values, we do not need to perform quantization on the CB-protected templates (unlike when applying HE on unprotected templates that may contain floating point values).

## 4.2 Hybrid Protection of Deep Templates using Cancelable Biometrics and Homomorphic Encryption

However, we might perform a mapping (i.e.,  $M_{\mathcal{D}}(.)$ ) to change the representation of the CB-protected templates prior to applying HE so that the corresponding comparison function  $\text{Comp}_{\text{HE}}^{\mathcal{P}}(.,.)$  can be properly applied on the hybrid-protected templates in the encrypted domain. Hereunder, we describe the application of BioHashing, MLP-Hashing, and IoM Hashing in our proposed method:

**BioHashing and MLP-Hashing** BioHashing and MLP-Hashing CB methods generate binary-valued templates and use Hamming distance for calculating the comparison scores during recognition [23]. Hence, we propose to encrypt the binary-valued templates generated by these CB methods directly during the HE protection stage, with no further mapping (i.e.,  $\mathbf{t}_m^i = M_{\mathcal{D}}(\mathbf{t}_c^i) = \mathbf{t}_c^i$ ). Then, we can apply equivalent homomorphic operations to calculate the sum squared error for  $\text{Comp}_{\text{HE}}^{\mathcal{P}}(.,.)$  on the hybrid-protected templates.

**IoM Hashing** The IoM Hashing CB scheme generates integer-valued templates and uses the average number of collisions for calculating the comparison scores during recognition [23]. Therefore, we propose to represent each integer element of IoM-hashed templates using one-hot encoding prior to encrypting them via HE (i.e., by one-hot encoding each integer element of IoM-Hash is mapped to a vector of zeros and a single one, where the index of the single one corresponds to the value of the IoM-Hash element). Therefore,  $M_{\mathcal{D}}(.)$  will be a one-hot encoding (i.e.,  $M_{\mathcal{D}}(t_c^i) = \text{OneHot}(t_c^i)$ ). Then, for comparison function  $\text{Comp}_{\text{HE}}^{\mathcal{P}}(.,.)$  we can apply a series of homomorphic operations, which is equivalent to calculating the sum squared error between the probe and reference hybrid-protected templates.

### 4.2.2 Experiments

In this section, we describe our experiments and evaluate the proposed hybrid BTP scheme. In section 4.2.2.1, we first detail our experimental setup. In section 4.2.2.2, we analyze the recognition performance and execution time of the proposed method in different scenarios and with different configurations. Finally, we discuss our experimental results in section 4.2.2.3. We should note that we do not evaluate the renewability, unlinkability, and irreversibility characteristics of our hybrid method, since these requirements have already been shown to be satisfied by the adopted CB methods and HE in the literature (e.g., [23], [77], [100]).

#### 4.2.2.1 Experimental Setup

**Baseline methods** In our experiments, we use three SOTA face recognition models<sup>4</sup>, including ArcFace [132], ElasticFace [133], and FaceNet [73]. As our baseline methods, we consider applying HE on the extracted embeddings (without first applying CB). Therefore, for the BFV HE algorithm, we need to quantize the embeddings prior to HE in order to obtain integer

<sup>4</sup>The implementation of each face recognition model is available at <https://gitlab.idiap.ch/bob/bob.bio.face>

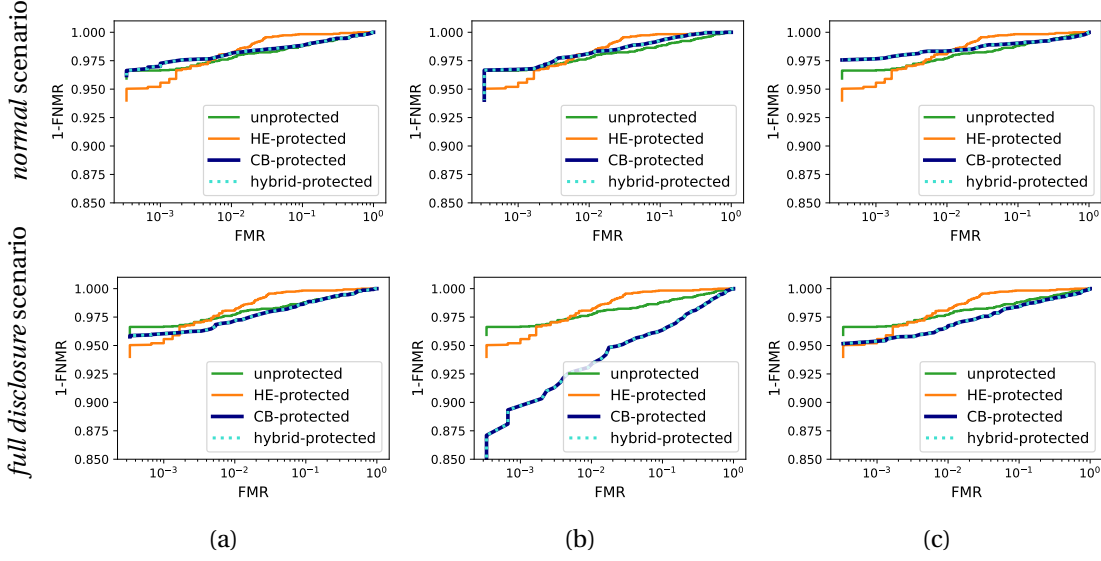


Figure 4.5: ROC curves of the unprotected, HE-protected, CB-protected, and hybrid-protected versions of features extracted by the ArcFace model on the LFW dataset in the (a) *normal* (first row) and *full disclosure* (second row) scenarios using (a) BioHashing, (b) MLP-Hashing, and (c) IoM Hashing.

values. In our experiments, we use the equal-probable quantile quantization scheme [208] with 4 quantization levels.

**Evaluation Datasets** We use the MOBIO [158] and Labeled Faced in the Wild (LFW) [159] databases to evaluate the recognition performance of the proposed hybrid BTP method on SOTA face recognition models.

**Evaluation Scenarios** To evaluate the recognition performance of our hybrid BTP method, we consider two scenarios in our experiments: the *normal* scenario and the *full disclosure* scenario. The *normal* scenario is the expected scenario in practice, where users' keys (for the CB scheme) and HE keys are secret. On the other hand, the *full disclosure* scenario (corresponds to the *full disclosure threat model* in the ISO/IEC 30136 standard [10]) is the case where we assume that everything about the system (including the protection algorithm, as well as all submodules and secrets) is disclosed. In particular, the HE keys are leaked, and we also assume that the adversary knows the users' keys for the CB schemes.

#### 4.2.2.2 Analysis

Fig. 4.5 compares the Receiver Operating Characteristic (ROC) curves of unprotected, HE-protected, CB-protected, and hybrid-protected (using our proposed BTP scheme) templates of ArcFace on the LFW dataset for different CB methods (i.e., BioHashing, MLP-Hashing, and IoM

## 4.2 Hybrid Protection of Deep Templates using Cancelable Biometrics and Homomorphic Encryption

Table 4.6: The average execution time (milliseconds) and recognition performance (in terms of TMR at FMR = 0.001) of HE and the proposed hybrid method, when applying **BioHashing** in the *normal* and *full disclosure* scenarios, on the MOBIO and LFW datasets using different face recognition models. In each model, the first row indicates HE protection (no CB) and the other rows show our hybrid template protection.

FR Model	$\alpha$	$\ell_{t_m}$	Average Execution Time (ms)					<i>normal scenario</i>		<i>full disclosure scenario</i>	
			CB	Encoding	Comparison	Decoding	Total	MOBIO	LFW	MOBIO	LFW
ArcFace ( $\ell_{t_u}=512$ )	-	-	-	$1.24 \pm 0.55$	$329.04 \pm 3.96$	$0.38 \pm 0.00$	$330.66 \pm 4.50$	100.00%	95.20%	100.00%	95.20%
	1.00	512	$15.08 \pm 2.42$	$1.24 \pm 0.55$	$329.04 \pm 3.96$	$0.38 \pm 0.00$	$345.74 \pm 5.11$	100.00%	97.00%	100.00%	95.93%
	0.75	384	$9.47 \pm 1.26$	$1.28 \pm 0.04$	$263.67 \pm 1.71$	$0.42 \pm 0.01$	$274.82 \pm 2.13$	100.00%	96.20%	100.00%	95.13%
	0.50	256	$6.24 \pm 1.92$	$1.19 \pm 0.04$	$166.21 \pm 4.40$	$0.38 \pm 0.00$	$174.01 \pm 4.83$	100.00%	94.83%	100.00%	93.53%
	0.25	128	$1.77 \pm 0.76$	$1.19 \pm 0.01$	$84.43 \pm 0.79$	$0.38 \pm 0.00$	$87.75 \pm 1.10$	100.00%	84.83%	99.68%	84.67%
ElasticFace ( $\ell_{t_u}=512$ )	-	-	-	$1.24 \pm 0.55$	$329.04 \pm 3.96$	$0.38 \pm 0.00$	$330.66 \pm 4.50$	99.96%	87.97%	99.96%	87.97%
	1.00	512	$15.08 \pm 2.42$	$1.24 \pm 0.55$	$329.04 \pm 3.96$	$0.38 \pm 0.00$	$345.74 \pm 5.11$	100.00%	96.47%	100.00%	93.43%
	0.75	384	$9.47 \pm 1.26$	$1.28 \pm 0.04$	$263.67 \pm 1.71$	$0.42 \pm 0.01$	$274.82 \pm 2.13$	100.00%	95.77%	100.00%	95.13%
	0.50	256	$6.24 \pm 1.92$	$1.19 \pm 0.04$	$166.21 \pm 4.40$	$0.38 \pm 0.00$	$174.01 \pm 4.83$	100.00%	94.63%	99.88%	88.43%
	0.25	128	$1.77 \pm 0.76$	$1.19 \pm 0.01$	$84.43 \pm 0.79$	$0.38 \pm 0.00$	$87.75 \pm 1.10$	99.68%	86.93%	86.27%	80.67%
FaceNet ( $\ell_{t_u}=128$ )	-	-	-	$1.19 \pm 0.01$	$84.27 \pm 0.08$	$0.38 \pm 0.00$	$85.83 \pm 0.08$	98.41%	88.97%	98.41%	88.97%
	1.00	128	$1.2 \pm 0.69$	$1.19 \pm 0.01$	$84.27 \pm 0.08$	$0.38 \pm 0.00$	$87.03 \pm 0.70$	99.92%	95.33%	92.86%	78.33%
	0.75	96	$0.49 \pm 0.65$	$1.19 \pm 0.01$	$64.22 \pm 2.91$	$0.38 \pm 0.00$	$66.28 \pm 2.98$	99.64%	88.73%	73.93%	75.60%
	0.50	64	$0.36 \pm 0.17$	$1.19 \pm 0.06$	$43.71 \pm 1.11$	$0.38 \pm 0.02$	$45.64 \pm 1.13$	98.69%	73.87%	83.69%	52.47%
	0.25	32	$0.12 \pm 0.01$	$1.19 \pm 0.01$	$23.14 \pm 0.05$	$0.38 \pm 0.00$	$24.82 \pm 0.05$	85.83%	48.3%	32.26%	23.17%

Hashing) in the *normal* and *full disclosure* scenarios. As this figure shows, the proposed hybrid method achieves exactly the same performance as the CB-protected templates in the *normal* and *full disclosure* scenarios for all CB methods. In addition, the hybrid-protected templates have a marginal improvement to the unprotected templates in the *normal* scenario. Comparing with HE-protected templates, in the *normal* scenario the hybrid-protected templates have slightly better performance for high values of the False Match Rate (FMR) and slightly worse performance for low values of the FMR. However, in each case the performance attainable using hybrid-protected templates is fairly close to that attainable using HE-protected templates. In the *full disclosure* scenario, while the recognition performance of HE-protected templates remains similar to the *normal* scenario, the performance of CB-protected templates degrades.

Table 4.6 reports the average execution time (over 100 executions) and recognition performance of HE and also the proposed hybrid method in the *normal* and *full disclosure* scenarios on the MOBIO and LFW datasets, when the adopted CB method is BioHashing and the length of the CB-protected templates (i.e., BioHashes) varies. In this table,  $\ell_{t_m}$  indicates the length of the mapped CB-protected template and  $\alpha$  denotes the ratio of the length of the CB-protected template  $\ell_{t_c}$  to the length of the unprotected template  $\ell_{t_u}$  (i.e.,  $\alpha = \ell_{t_c} / \ell_{t_u}$ ). Tables 4.7 and 4.8 also report similar evaluation when applying MLP-Hashing and IoM Hashing, respectively, in our proposed hybrid method. As these tables show, in general, the hybrid-protected templates can achieve superior recognition performance compared to the HE-protected templates in the *normal* scenario. In the *full disclosure* scenario, hybrid-protected templates (with  $\alpha = 1$ ) have competitive performance with HE-protected templates. Notwithstanding the good

## Chapter 4. Protection of Biometric Templates

Table 4.7: The average execution time (milliseconds) and recognition performance (in terms of TMR at FMR = 0.001) of HE and the proposed hybrid method, when applying **MLP-Hashing** in the *normal* and *full disclosure* scenarios, on the MOBIO and LFW datasets using different face recognition models. In each model, the first row indicates HE protection (no CB) and the other rows show our hybrid template protection.

FR Model	$\alpha$	$\ell_{t_m}$	Average Execution Time (ms)					<i>normal scenario</i>		<i>full disclosure scenario</i>	
			CB	Encoding	Comparison	Decoding	Total	MOBIO	LFW	MOBIO	LFW
ArcFace ( $\ell_{t_u}=512$ )	-	-	-	$1.24 \pm 0.55$	$329.04 \pm 3.96$	$0.38 \pm 0.00$	$330.66 \pm 4.50$	100.00%	95.20%	100.00%	95.20%
	1.00	512	$56.07 \pm 10.83$	$1.19 \pm 0.04$	$328.30 \pm 0.32$	$0.39 \pm 0.00$	$385.93 \pm 10.83$	100.00%	96.73%	99.84%	88.10%
	0.75	384	$48.39 \pm 10.25$	$1.19 \pm 0.04$	$247.20 \pm 1.87$	$0.39 \pm 0.01$	$297.16 \pm 10.42$	100.00%	96.43%	99.76%	88.33%
	0.50	256	$39.86 \pm 5.09$	$1.24 \pm 0.02$	$172.30 \pm 0.76$	$0.41 \pm 0.01$	$213.80 \pm 5.15$	100.00%	93.27%	98.61%	80.87%
	0.25	128	$37.59 \pm 6.23$	$1.23 \pm 0.03$	$87.03 \pm 1.56$	$0.40 \pm 0.01$	$126.24 \pm 6.43$	98.77%	86.57%	85.40%	53.37%
ElasticFace ( $\ell_{t_u}=512$ )	-	-	-	$1.24 \pm 0.55$	$329.04 \pm 3.96$	$0.38 \pm 0.00$	$330.66 \pm 4.50$	99.96%	87.97%	99.96%	87.97%
	1.00	512	$56.07 \pm 10.83$	$1.19 \pm 0.04$	$328.30 \pm 0.32$	$0.39 \pm 0.00$	$385.93 \pm 10.83$	100.00%	94.50%	99.68%	87.97%
	0.75	384	$48.39 \pm 10.25$	$1.19 \pm 0.04$	$247.20 \pm 1.87$	$0.39 \pm 0.01$	$297.16 \pm 10.42$	100.00%	95.07%	99.60%	82.43%
	0.50	256	$39.86 \pm 5.09$	$1.24 \pm 0.02$	$172.30 \pm 0.76$	$0.41 \pm 0.01$	$213.80 \pm 5.15$	100.00%	92.00%	97.90%	69.17%
	0.25	128	$37.59 \pm 6.23$	$1.23 \pm 0.03$	$87.03 \pm 1.56$	$0.40 \pm 0.01$	$126.24 \pm 6.43$	99.01%	78.17%	74.56%	44.37%
FaceNet ( $\ell_{t_u}=128$ )	-	-	-	$1.19 \pm 0.01$	$84.27 \pm 0.08$	$0.38 \pm 0.00$	$85.83 \pm 0.08$	98.41%	88.97%	98.41%	88.97%
	1.00	128	$1.62 \pm 0.24$	$1.17 \pm 0.01$	$84.11 \pm 0.08$	$0.37 \pm 0.01$	$87.27 \pm 0.25$	99.33%	86.53%	47.34%	50.53%
	0.75	96	$1.51 \pm 0.27$	$1.17 \pm 0.01$	$63.76 \pm 0.06$	$0.36 \pm 0.00$	$66.8 \pm 0.28$	98.57%	79.90%	47.54%	35.07%
	0.50	64	$1.43 \pm 0.29$	$1.17 \pm 0.01$	$43.44 \pm 0.04$	$0.36 \pm 0.00$	$46.40 \pm 0.29$	92.34%	55.73%	49.33%	25.50%
	0.25	32	$1.31 \pm 0.29$	$1.17 \pm 0.01$	$23.06 \pm 0.03$	$0.36 \pm 0.00$	$25.90 \pm 0.29$	58.69%	31.43%	20.99%	9.90%

Table 4.8: The average execution time (milliseconds) and recognition performance (in terms of TMR at FMR = 0.001) of HE and the proposed hybrid method, when applying **IoM Hashing** in the *normal* and *full disclosure* scenarios, on the MOBIO and LFW datasets using different face recognition models. In each model, the first row indicates HE protection (no CB) and the other rows show our hybrid template protection.

FR Model	$\alpha$	$\ell_{t_m}$	Average Execution Time (ms)					<i>normal scenario</i>		<i>full disclosure scenario</i>	
			CB	Encoding	Comparison	Decoding	Total	MOBIO	LFW	MOBIO	LFW
ArcFace ( $\ell_{t_u}=512$ )	-	-	-	$1.24 \pm 0.55$	$329.04 \pm 3.96$	$0.38 \pm 0.00$	$330.66 \pm 4.50$	100.00%	95.20%	100.00%	95.20%
	1.00	1536	$26.89 \pm 2.54$	$1.19 \pm 0.03$	$981.46 \pm 3.23$	$0.39 \pm 0.00$	$1009.92 \pm 4.12$	100.00%	97.67%	99.76%	95.30%
	0.75	1152	$23.06 \pm 8.31$	$1.19 \pm 0.01$	$736.24 \pm 0.82$	$0.38 \pm 0.00$	$760.86 \pm 8.36$	100.00%	97.17%	99.76%	94.17%
	0.50	768	$14.92 \pm 2.07$	$1.19 \pm 0.01$	$491.80 \pm 0.41$	$0.38 \pm 0.00$	$508.28 \pm 2.11$	100.00%	95.73%	99.76%	94.17%
	0.25	384	$6.67 \pm 0.50$	$1.19 \pm 0.01$	$248.39 \pm 9.48$	$0.38 \pm 0.00$	$256.62 \pm 9.49$	100.00%	91.33%	98.93%	90.37%
ElasticFace ( $\ell_{t_u}=512$ )	-	-	-	$1.24 \pm 0.55$	$329.04 \pm 3.96$	$0.38 \pm 0.00$	$330.66 \pm 4.50$	99.96%	87.97%	99.96%	87.97%
	1.00	1536	$26.89 \pm 2.54$	$1.19 \pm 0.03$	$981.46 \pm 3.23$	$0.39 \pm 0.00$	$1009.92 \pm 4.12$	100.00%	96.83%	98.10%	92.63%
	0.75	1152	$23.06 \pm 8.31$	$1.19 \pm 0.01$	$736.24 \pm 0.82$	$0.38 \pm 0.00$	$760.86 \pm 8.36$	100.00%	95.43%	98.10%	92.30%
	0.50	768	$14.92 \pm 2.07$	$1.19 \pm 0.01$	$491.80 \pm 0.41$	$0.38 \pm 0.00$	$508.28 \pm 2.11$	100.00%	94.07%	98.10%	91.23%
	0.25	384	$6.67 \pm 0.50$	$1.19 \pm 0.01$	$248.39 \pm 9.48$	$0.38 \pm 0.00$	$256.62 \pm 9.49$	100.00%	91.53%	98.21%	81.90%
FaceNet ( $\ell_{t_u}=128$ )	-	-	-	$1.19 \pm 0.01$	$84.27 \pm 0.08$	$0.38 \pm 0.00$	$85.83 \pm 0.08$	98.41%	88.97%	98.41%	88.97%
	1.00	384	$1.51 \pm 0.02$	$1.19 \pm 0.01$	$247.58 \pm 1.71$	$0.39 \pm 0.00$	$250.65 \pm 1.71$	99.96%	97.20%	95.44%	77.83%
	0.75	288	$1.13 \pm 0.01$	$1.19 \pm 0.01$	$186.34 \pm 0.63$	$0.38 \pm 0.00$	$189.03 \pm 0.63$	99.84%	95.37%	93.61%	74.10%
	0.50	192	$0.75 \pm 0.01$	$1.19 \pm 0.01$	$125.09 \pm 0.12$	$0.38 \pm 0.00$	$127.40 \pm 0.12$	99.33%	88.67%	87.38%	60.73%
	0.25	96	$0.38 \pm 0.00$	$1.19 \pm 0.01$	$63.91 \pm 0.10$	$0.38 \pm 0.00$	$65.85 \pm 0.10$	91.39%	67.97%	56.39%	45.00%

### 4.3 Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition

---

performance of HE-protected templates in the *full disclosure* scenario, we should note that in this scenario the adversary can easily reconstruct the unprotected templates using the HE private (decryption) key (i.e, very poor protection). However, for hybrid-protected templates, the adversary can only reconstruct the (mapped) CB-protected templates using the HE private key, but it is still difficult for the adversary to reconstruct the unprotected templates from the CB-protected templates. We can also see that with  $\alpha = 1$ , the hybrid protection requires a longer execution time than HE. However, we can adjust the value of  $\alpha$  so that the hybrid protection achieves a shorter execution time with comparable recognition performance.

#### 4.2.2.3 Discussion

Our experiments in section 4.2.2.2 show that the proposed hybrid scheme achieves exactly the same recognition performance as the corresponding CB method. Our experiments also show that in the *normal* scenario, the proposed hybrid method (with  $\alpha = 1$ ) achieves superior performance compared to HE. In the *full disclosure* scenario, hybrid-protected templates (with  $\alpha = 1$ ) achieve comparable performance with HE-protected templates for ArcFace and ElasticFace, but HE-protected templates perform better than hybrid-protected templates for FaceNet. Having said that, it is important to keep in mind that HE-protected templates can be easily inverted to recover the original (unprotected) templates, whereas hybrid-protected templates are not easily invertible due to the extra layer of protection provided by the CB method that is applied prior to HE.

Tables 4.6-4.8 also show that there is a trade-off between the execution time and recognition performance when using the proposed hybrid protection method. This trade-off can be controlled with  $\alpha$ . For  $\alpha = 1$ , hybrid-protected templates require longer execution times than HE-protected templates. However, with smaller  $\alpha$ , CB can in practice reduce the dimensionality of features prior to HE. Therefore, we can achieve a shorter execution time compared to HE. In particular, for a proper choice of  $\alpha$ , for the hybrid-protected templates we can simultaneously achieve a shorter execution time and comparable performance to the HE-protected templates. For example, for ElasticFace and BioHashing, we could set  $\alpha = 0.25$ , whereas for FaceNet at the same setting hybrid-protected templates have worse performance than HE-protected templates. Therefore, in this case, it would be better to set  $\alpha$  to a higher value such as  $\alpha = 0.75$ . The suitable lengths of BioHash-protected templates (and therefore  $\alpha$ ), which maintain the recognition performance of unprotected templates, are investigated in [26] for different SOTA face recognition models.

### 4.3 Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition

In this section, we propose a new framework to protect and enhance vascular biometric recognition systems. We use three existing finger vein recognition (FVR) approaches, including

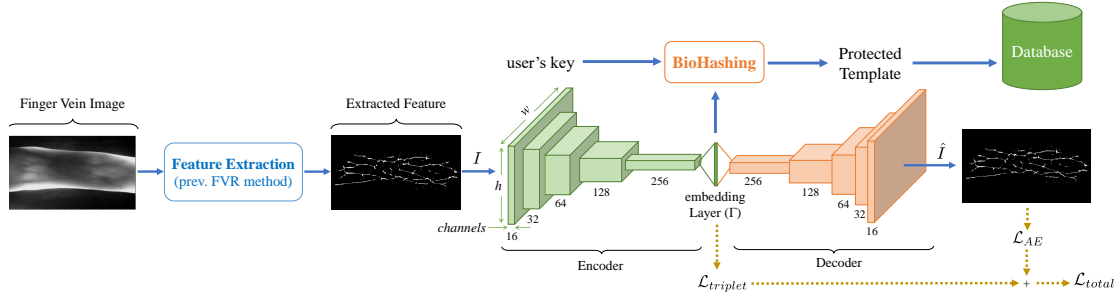


Figure 4.6: Block diagram of the proposed framework to enhance and protect vascular recognition.

Wide Line Detector (WLD) [163], Repeated Line Tracking (RLT) [164], and Maximum Curvature (MC) [165], and apply our framework to protect them and enhance their performance. The experimental results indicate that the protected templates generated by our framework achieve superior performance than BioHash protected templates of the raw features in the normal scenario. Furthermore, in the scenario where the BioHashing key is stolen, called *stolen* scenario, our framework achieves have far better performance than BioHashing protected templates from the raw features. In addition, we deploy raw finger vein images in our framework and use our framework as a secure FVR method. We also deploy our proposed framework using previous feature extractor on different vascular images to evaluate the generalization of our framework on other vascular biometric modalities (e.g., palm and wrist).

In the following, we first describe our proposed framework in section 4.3.1, and provide experimental analysis in section 4.3.2.

### 4.3.1 Proposed Framework

As illustrated in the figure 4.6, our proposed deep-learning-based framework consists of a deep convolutional auto-encoder (AE) which extracts deep features at its bottleneck layer (so called embedding). In 4.3.1.1, we describe in further details the network structure, our multi-term loss function and the training process. After deep features are calculated at the embedding layer, we use the BioHashing algorithm to generate protected templates. Finally, for the recognition stage, the BioHash templates should be compared and scored, which is described in 4.3.1.2. Indeed, a new protected template can be generated any time using BioHashing with a new key. We use the proposed framework to protect and enhance previous FVR methods. To this end, we train our proposed convolutional auto-encoder with the features extracted by the corresponding FVR methods. Indeed, by using the auto-encoder structure, we can considerably reduce the dimension of extracted features without losing significant information. The new deep features can be protected by BioHashing efficiently since the dimension of features is small enough. The protected templates are also cancelable, since BioHashing generates cancelable templates [76], [209].



### 4.3 Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition

In addition to protecting existing FVR systems, we use the proposed framework to introduce a new secure FVR method. To this end, we use raw finger vein images to train the proposed convolutional auto-encoder and directly extract features from finger vein images.

#### 4.3.1.1 Auto-encoder

**Network Structure** We use a convolutional auto-encoder that reduces the size of its input to the bottleneck layer (encoder), and then reconstruct the image (decoder). The encoder network consists of five convolutional layers with 16, 32, 64, 128, 256 filters, respectively. We use  $3 \times 3$  kernel with stride 2 in each layer, which divides the spatial size by factor 2. Additionally, we use Batch normalization [179] after each convolution operation. Finally, we use a fully connected layer to get the embedding layer. For the decoder network, we use the transpose convolution layers. Except for the final layer, which has sigmoid function, we use the rectified linear unit (ReLU) for the other layers.

We should note that the size of the input image given to the network is the size of the extracted features by the corresponding FVR model. Further information about the size of the features extracted by different FVR methods are reported in section 4.3.2.

**Multi-term Loss Function** To train the proposed network, we use a multi-term loss function. Let's consider  $I$ ,  $\hat{I}$ ,  $\Gamma$  as the input image, the reconstructed image, and the embedding layer, respectively. The total loss is

$$\mathcal{L}_{total} = (1 - \alpha)\mathcal{L}_{AE} + \alpha\mathcal{L}_{triplet}, \quad (4.6)$$

where  $\alpha$  is a hyper-parameter (in  $[0, 1]$  interval) to control the contribution of  $\mathcal{L}_{AE}$  and  $\mathcal{L}_{triplet}$ , where  $\mathcal{L}_{AE}$  and  $\mathcal{L}_{triplet}$  are the auto-encoder loss, and the embedding triplet loss, respectively. For the auto-encoder loss, we use different loss functions considering the input of our framework. If the proposed framework is used to protect and enhance the performance of a FVR method which has binary features, we use Binary Cross Entropy (BCE) as the auto-encoder loss:

$$\mathcal{L}_{AE} = \mathcal{L}_{BCE} = - \sum_{i=1}^w \sum_{j=1}^h I(i, j) \log \hat{I}(i, j), \quad (4.7)$$

where  $w$  and  $h$  are the width and height of the input image,  $I$ , respectively. However, if the proposed framework is used as our proposed FVR, we use a different term which works with continuous values of finger vein images. In this case, we use the  $l_2$  norm of the auto-encoder error:

$$\mathcal{L}_{AE} = \|I - \hat{I}\|_2. \quad (4.8)$$

Furthermore, the embedding triplet loss is defined as:

$$\mathcal{L}_{triplet} = [\|\Gamma_a - \Gamma_p\|^2 - \|\Gamma_a - \Gamma_n\|^2 + \beta]_+ \quad (4.9)$$

## Chapter 4. Protection of Biometric Templates

---

where  $\Gamma_a$ ,  $\Gamma_p$ , and  $\Gamma_n$  are the values of the embedding layer for anchor, positive, and negative images, respectively [73], and  $\beta$  is also a margin which is enforced between positive and negative pairs which is set to 1 in our experiments.

**Training Process** To train the proposed auto-encoder with our multi-term loss function, we use Adam [178] optimizer. We use the initial learning rate of  $10^{-3}$ , and decrease the learning rate every 10 epochs. We use the Pytorch framework for the experiments.

For our experiments, we use the UTFVP finger vein dataset [169] which contains 1440 finger vein images with  $672 \times 380$  resolution that have been collected from 60 individuals. We apply random data augmentation technique to the training set by randomly adjusting each finger vein image, including random rotation [range:  $< 7$  degree], width shift [range:  $< 0.025 \times$  image width], height shift [range:  $< 0.025 \times$  image height], channel shift (i.e., offset) [range:  $< 0.075$ ], and zoom [range: (0.95,1.05)] transformations.

### 4.3.1.2 Protecting Deep Templates

After training the auto-encoder, we can apply a template protection scheme to generate protected templates. We use the BioHashing algorithm [76] to generate the protected templates<sup>5</sup>. In the subsequent experiments, we will consider that FVR operated in verification mode only. In the enrolling stage, the protected templates for every individual are stored at the system database. For the verification stage, either verification or identification, the probe templates should be compared with the templates in the database. To find the score between the probe template and the model template, we use the Hamming distance between the BioHash templates.

### 4.3.2 Experiments

#### 4.3.2.1 Experimental Setup

We use the publicly available finger vein UTFVP dataset [169] in our experiments. This dataset contains in total 1440 finger vein images which have been collected from 60 subjects. We used the training (subjects 1-10, 240 images), development (subjects 11-28, 432 images) and evaluation (subjects 29-60, 768 images) subsets of this dataset<sup>6</sup>. The training subset is used for training the convolutional auto-encoder in our framework, the development subset is used for threshold estimation in the recognition stage, and the evaluation subset is used for reporting the final results and further comparisons. We implemented Wide Line Detector (WLD) [163], Repeated Line Tracking (RLT) [164], and Maximum Curvature (MC) [165] algorithms to extract biometric features from finger vein images, and then apply our

---

<sup>5</sup>Note that other cancelable biometric schemes can similarly be used instead of BioHashing.

<sup>6</sup>The implementation of this division for the UTFVP dataset is available under NOM protocol at <https://gitlab.idiap.ch/bob/bob.db.utfvp>

### 4.3 Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition

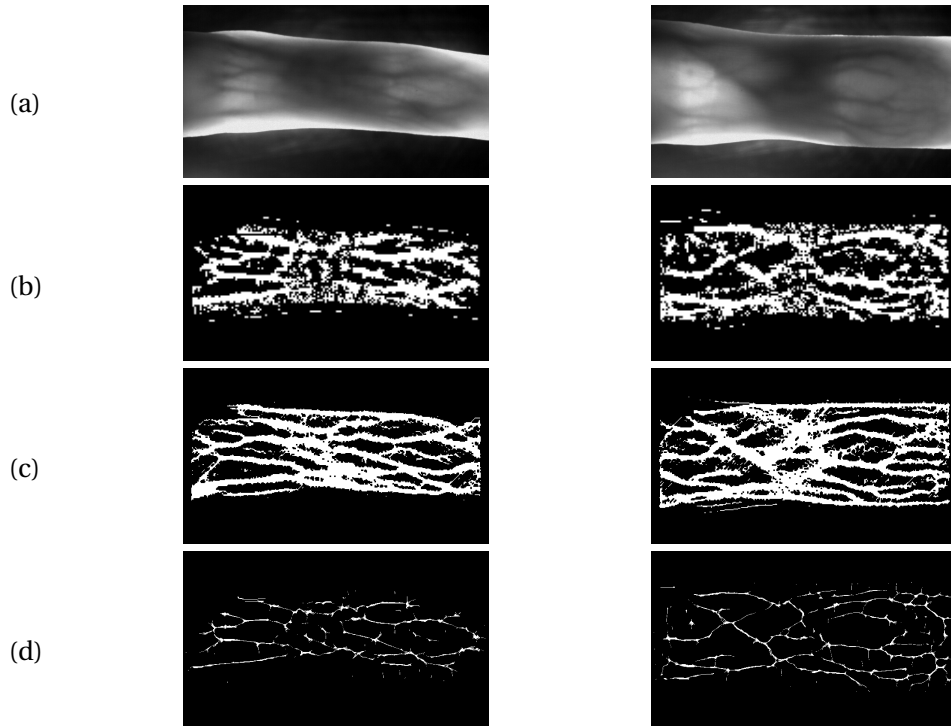


Figure 4.7: Two sample finger vein images from two individuals in UTFVP dataset [169] and their corresponding Wide Line Detector (WLD)[163], Repeated Line Tracking (RLT)[164], and Maximum Curvature (MC)[165] features: (a) Finger vein image, (b) WLD, (c) RLT, (d) MC.

Table 4.9: Size of the features extracted by WLD, RLT, and MC methods and their execution times on UTFVP dataset

	WLD	RLT	MC
Feature size	$164 \times 94$	$409 \times 234$	$682 \times 390$
Execution Time	0.17	22.6	3.25

framework on the features extracted by these methods. Figure 4.7 shows two sample finger vein images and the corresponding WLD, RLT, and MC features for two individuals in the UTFVP dataset. Furthermore, table 4.9 reports the size of these features and also the execution time on a system with an Intel i7-7700K 4.2 GHz to extract these features from each image in UTFVP dataset.

We consider two scenarios in our experiments: the *normal* scenario and the *stolen-token* (shortly, *stolen*) scenario. In the normal scenario, which is the expected scenario for most cases, each user's key is considered to be secret and not been disclosed. However, in the stolen scenario, the impostor has access to the genuine user's secret key and use it with the impostor's own finger vein template. While such a scenario is expected to happen rarely in practice, the system's vulnerability relies on the leakage of the user's secret key. To implement the stolen

## Chapter 4. Protection of Biometric Templates

---

scenario, in the verification stage, we calculate the BioHash code of other users in the database using the same key as the genuine's key.

We evaluate the performance of our proposed framework in the normal scenario and the stolen scenario, and we do not evaluate cancelability, unlinkability, and non-invertibility characteristics. Because, our framework relies on BioHashing algorithm, and evaluation of these characteristics of BioHashing have been addressed already by [76], [112], [209]–[212].

In our experiments, we have three different hyper-parameters including the length of the embedding layer in the AE ( $L_{embedding}$ ), the length of BioHash templates ( $L_{BioHash}$ ), and the value of  $\alpha$  in equation 4.6 for controlling the contribution of different loss terms. For simplicity in our experiments, we consider  $L_{embedding}$ ,  $L_{BioHash}$ , and  $\alpha$  equal with 100, 100, and 0.1, respectively. Afterwards, we provide an ablation study to evaluate the effect of each of these hyper-parameters.

After evaluating the performance of our framework on previous FVR methods, in another experiment, we deploy our framework on the raw finger vein images (without any preprocessing) of UTFVP dataset and compare the results with enhanced versions of previous FVR methods.

In addition, in another experimental setup, we evaluate the performance of our framework on other vascular biometric modalities. To this end, we use PUT Vein dataset [170] which includes palm vein and wrist vein images. This dataset consists of 2400 images, where half of images contains palm vein images (1200 images) and another half contains wrist vein images (another 1200 images) which were acquired from both hands of 50 individuals. We consider the images from "right" hands of this dataset and divide it into two part, the first part (subject 1-25, 600 images) for training and development, and the second part (subject 26-50, 600 images) for evaluation<sup>7</sup>.

### 4.3.2.2 Performance Evaluation for Previous FVR Methods

Figure 4.8 compares the receiver operating characteristic (ROC) curves of the protected and enhanced version WLD, RLT, and MC methods via our framework, namely WLD+AE+BioHash, RLT+AE+BioHash, and MC+AE+BioHash, respectively, against the corresponding BioHash protected templates of these methods, namely WLD+BioHash, RLT+BioHash, and MC+BioHash, respectively, in the normal and the stolen scenarios on the evaluation subset of the UTFVP dataset. In all ROC curves, the marked points which are connected with the dashed lines correspond to the threshold that leads to False Match Rate (FMR)= $10^{-3}$  on the development subset. In figure 4.8, we also compare the mentioned ROC curves to WLD, RLT, and MC methods alone without template protection in the normal scenario. Furthermore, since the AE does the dimensionality reduction in our framework, we also compare the performance of our framework with the Principal Component Analysis (PCA) as a traditional dimensionality reduction

---

<sup>7</sup>The implementation of this division for the PUT Vein dataset is available under R\_1 protocol at <https://gitlab.idiap.ch/bob/bob.db.putvein>

### 4.3 Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition

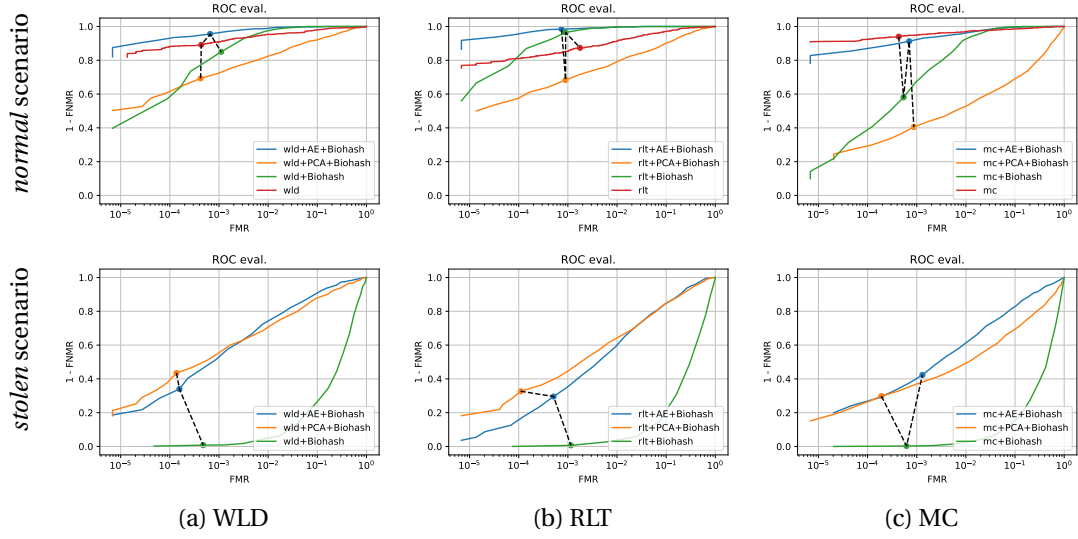


Figure 4.8: Comparison of ROC curves of previous FVR methods with their BioHash protected templates, BioHash protected of their PCA transformation, and their protected version via our proposed framework in normal scenario (first row) and stolen scenario (second row): a)WLD, b)RLT, c)MC. The marked points which are connected with the dashed lines in each plot correspond to the threshold that leads to  $FMR=10^{-3}$  on the development subset.

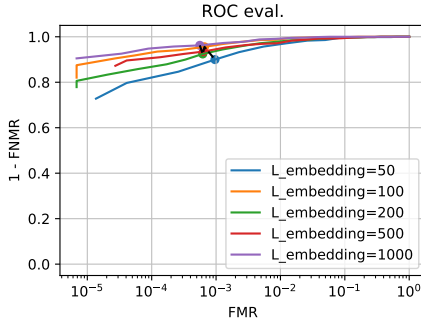
technique. To this end, we use the features extracted by WLD, RLT, and MC methods and use the PCA algorithm prior to BioHashing, namely WLD+PCA+BioHash, RLT+PCA+BioHash, and MC+PCA+BioHash, respectively. We should note that for a fair comparison, we considered the dimension reduced by PCA to be consistent with that of AE. As this figure shows, the proposed framework achieves superior performance than BioHash protected versions of WLD, RLT, and MC methods, both in normal and in stolen scenarios. In the case of WLD and RLT methods, our method even outperforms unprotected versions in the normal scenario. Comparing our method with the PCA algorithm, while our method has competitive performance with PCA algorithm in the stolen scenario, our method achieves far better performance in the normal scenario.

In addition to the ROC curves, we compare the performance of the aforementioned methods in terms of False Match Rate (FMR), False Non-Match Rate (FNMR), and Equal Error Rate (EER) for the evaluation subset of the UTFVP dataset in normal scenario and stolen scenario, which is reported in the table 4.10. Please note that for the values in this table, the threshold for each method is selected individually in the way that we achieve minimum EER on the development subset. This table also shows that our method achieves the best performance in the stolen scenario in terms of FMR, FNMR, and EER. However, in the normal scenario, our method has competitive performance with BioHash protected versions of the mentioned FVR methods.

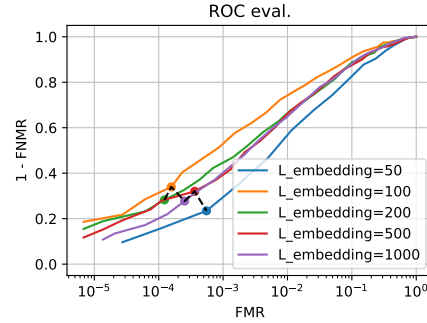
## Chapter 4. Protection of Biometric Templates

Table 4.10: Comparing the performance of previous FVR methods with their BioHash protected templates and their enhanced version via our proposed framework in normal scenario and stolen scenario in terms of FMR, FNMR, and EER on the evaluation subset of UTFVP dataset. The best performance is emboldened.

method	Normal Scenario			Stolen Scenario		
	FMR	FNMR	EER	FMR	FNMR	EER
WLD	3.8%	3.8%	3.8%	-	-	-
WLD + BioHash	1.4%	1.7%	1.5%	34.3%	44.0%	39.1%
WLD + PCA + BioHash	6.8%	8.7%	7.8%	12.4%	11.1%	11.7%
<b>WLD + AE + BioHash</b>	<b>0.8%</b>	<b>1.3%</b>	<b>1.1%</b>	<b>8.5%</b>	<b>10.3%</b>	<b>9.4%</b>
RLT	4.6%	4.6%	4.6%	-	-	-
RLT + BioHash	<b>0.5%</b>	<b>0.7%</b>	<b>0.6%</b>	46.8%	34.4%	40.6%
RLT + PCA + BioHash	9.7%	10.0%	9.9%	13.5%	13.2%	13.3%
<b>RLT + AE + BioHash</b>	<b>0.5%</b>	0.9%	0.7%	<b>12.7%</b>	<b>13.5%</b>	<b>13.1%</b>
MC	<b>2.3%</b>	2.3%	<b>2.3%</b>	-	-	-
MC + BioHash	2.9%	2.3%	2.6%	42.1%	53.0%	47.5%
MC + PCA + BioHash	24.1%	22.4%	23.3%	21.8%	22.7%	22.2%
<b>MC + AE + BioHash</b>	<b>2.3%</b>	<b>2.2%</b>	<b>2.3%</b>	<b>14.2%</b>	<b>13.4%</b>	<b>13.8%</b>



(a) normal scenario



(b) stolen scenario

Figure 4.9: Evaluating the effect of  $L_{embedding}$ : a) normal scenario, b) stolen scenario. The marked points which are connected with the dashed lines in each plot correspond to the threshold that leads to  $FMR=10^{-3}$  on the development subset.

### 4.3.2.3 Ablation Study

In section 4.3.2.2, we evaluated the performance of the proposed framework in protecting and enhancing the performance of previous FVR methods. We considered values of the hyper-parameters (including the length of the embedding layer in the AE ( $L_{embedding}$ ), the length of BioHash templates ( $L_{BioHash}$ ), and the value of  $\alpha$  in equation 4.6) to be fixed. In this section, we evaluate the effect of each of these hyper-parameters on the performance of our framework by investigating the ROC curves in the normal and the stolen scenarios on the evaluation subset of the UTFVP dataset. To this end, we consider WLD method, which has a smaller feature size and is faster to be calculated (see table 4.9), and build our experiments upon WLD.

### 4.3 Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition

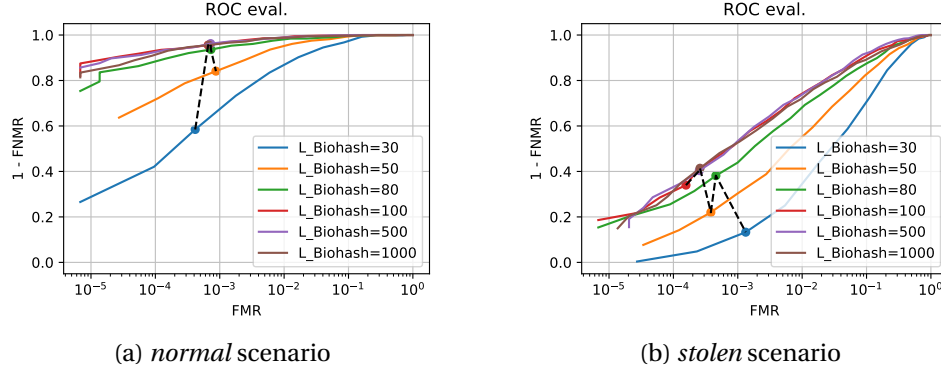


Figure 4.10: Evaluating the effect of  $L_{BioHash}$ : a) normal scenario, b) stolen scenario. The marked points which are connected with the dashed lines in each plot correspond to the threshold that leads to  $FMR=10^{-3}$  on the development subset.

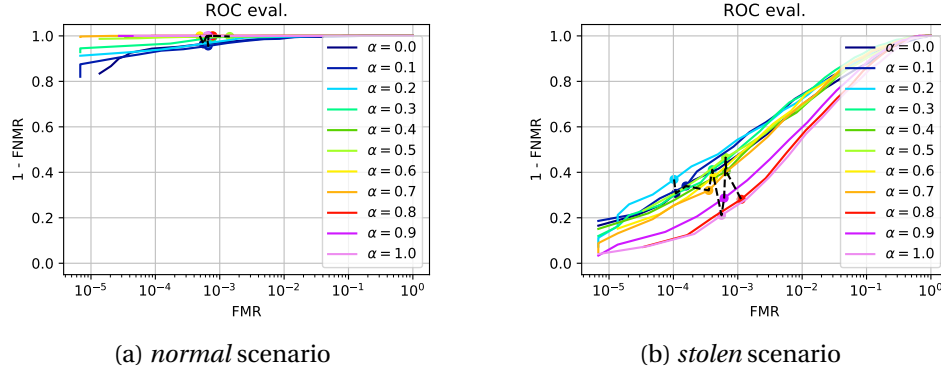


Figure 4.11: Evaluating the effect of  $\alpha$ : a) normal scenario, b) stolen scenario. The marked points which are connected with the dashed lines in each plot correspond to the threshold that leads to  $FMR=10^{-3}$  on the development subset.

**Evaluating the effect of  $L_{embedding}$**  To evaluate the effect of the length of the embedding layer in the AE ( $L_{embedding}$ ), we evaluate the performance of our framework with values 50, 100, 200, 500, and 1000 for  $L_{embedding}$ . Figure 4.9, represents the ROC curves of our framework for different values of  $L_{embedding}$ . As this figure shows, generally, a higher  $L_{embedding}$  leads to superior performance in the normal scenario but inferior performance in the stolen scenario except for  $L_{embedding} = 50$  and 100. This figure also shows that for  $L_{embedding} = 100$ , we achieve superior performance for both normal and stolen scenarios. Meanwhile, the performance for  $L_{embedding} = 50$  is subordinate for both normal and stolen scenarios. This might be due to the fact that embedding layer of length 50 is not enough to represent finger vein features.

**Evaluating the effect of  $L_{BioHash}$**  Similar to  $L_{embedding}$ , we evaluate the effect of the length of BioHash templates ( $L_{BioHash}$ ) by varying its value between 30, 50, 80, 100, 500, and 1000. Figure 4.10, illustrates the ROC curves of our framework for different values of  $L_{BioHash}$ . As

this figure shows, increasing the value of  $L_{BioHash}$  above 100 does not significantly change the performance of our method neither in the normal scenario nor in the stolen scenario. However, decreasing the value of  $L_{BioHash}$  lower than 100 degrades the performance of our framework in both normal and stolen scenarios.

**Evaluating the effect of  $\alpha$  in equation 4.6** The hyper-parameter  $\alpha$  in equation 4.6 controls the contribution of different terms in the loss function. To evaluate the effect of each loss term on the performance of our framework, we vary the value of  $\alpha$  in  $[0, 1]$  interval. Figure 4.11, illustrates the ROC curves of our framework for different values of  $\alpha$ . As this figure indicates, in the normal scenario, increasing the value of  $\alpha$  enhances the performance of our framework. On the other hand, in the stolen scenario, increasing the value of  $\alpha$  decreases the performance of our method. Therefore, the value of  $\alpha$  which leads to the best performance in the normal scenario has the worst performance in the stolen scenario. We should also note that  $\alpha = 1.0$  practically eliminates the effect of the decoder part of our auto-encoder network in the training process. Therefore, as depicted in figure 4.11, it leads to high performance in the normal scenario, but very poor performance in the stolen scenario.

### 4.3.2.4 Using our Framework as a FVR Method

In another experiment, we use raw finger vein images as the input to our convolutional auto-encoder and used the extracted features in the embedding layer for generating protected templates using BioHashing. Figure 4.12 compares the ROC curves of this setup, namely img+AE+BioHash, with protected and enhanced version WLD, RLT, and MC methods via our framework. As this figure shows, using raw finger vein images leads to superior performance in the normal scenario, but competitively inferior performance in the stolen scenario.

### 4.3.2.5 Palm and Wrist Vein Recognition

As mentioned earlier, to evaluate the generalization of our framework for other modalities, in another experimental setup, we use our framework for palm and wrist images from the PUT Vein database. Table 4.11 compares the performance of the proposed for WLD, RLT, MC methods as well as raw images on the evaluation subset of PUT Vein dataset in terms of FMR, FNMR, and EER. As this table shows, in general, our framework enhances the performance of BioHash protected versions of WLD, RLT, MC methods for palm and wrist data in both normal and stolen scenarios.

### 4.3.2.6 Discussion

As shown in figure 4.8, BioHashing decreases the performance of WLD, RLT, and MC methods. In the normal scenario, we observe a considerable drop in the performance of BioHash protected versions of these FVR methods than their unprotected versions for low FMR thresholds.



### 4.3 Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition

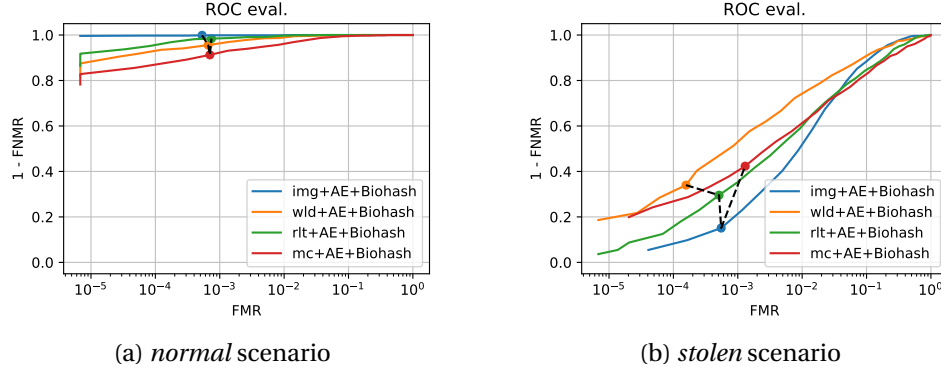


Figure 4.12: Comparison of ROC curves of our framework in two modes: 1) given the raw finger vein images as the input of auto-encoder, and 2) given the features extracted from WLD, RLT, and MC as the input: a) normal scenario, b) stolen scenario. The marked points which are connected with the dashed lines in each plot correspond to the threshold that leads to  $FMR=10^{-3}$  on the development subset.

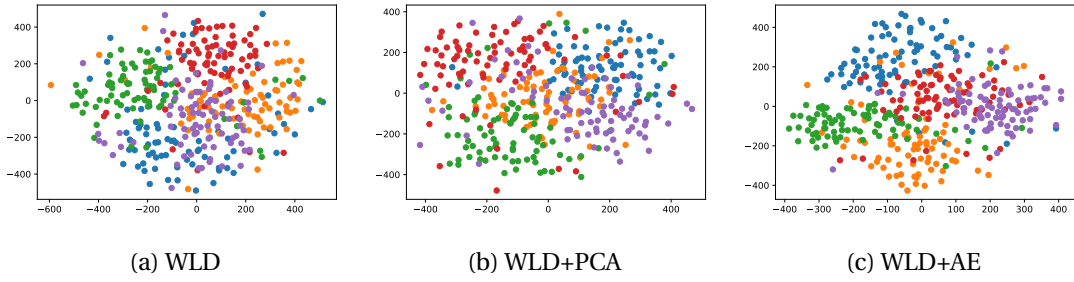


Figure 4.13: 2D representation of extracted features for 5 different identities with a) WLD, b) WLD+PCA, and c) WLD+AE methods. Different colors illustrate different identities. The axes denote the reduced dimensions to represent each feature using the T-SNE technique.

Besides, in the stolen scenario, the poor performance of BioHash protected templates indicates serious vulnerability of such systems to the reveal of users' keys. Comparing our proposed framework with BioHashing in the normal scenario, our framework achieves competitive performance for high FMR thresholds and also much superior performance for low FMR thresholds. In addition, in the stolen scenario, our method has far better performance than BioHash protected versions of the mentioned FVR methods. In fact, we reduced the dimension of extracted features through our framework that helps prevent the enormous dimensionality reduction gap caused by directly applying BioHashing to pre-processed images. While traditional dimensionality reduction techniques such as PCA can help to generate features in the lower dimension, experiments show the superiority of auto-encoder in the recognition performance. For instance, in the case of WLD, as shown in figure 4.8 and table 4.10, our framework achieves better performance than WLD, WLD+BioHash, and WLD+PCA+BioHash. To interpret the performance of our method, we use T-SNE technique to visualize the features prior to BioHashing in WLD+BioHash, WLD+PCA+BioHash, and WLD+AE+BioHash. Figure

## Chapter 4. Protection of Biometric Templates

Table 4.11: Comparing the performance of the proposed framework on PUT Vein dataset. The best performance is emboldened.

Data	method	Normal Scenario			Stolen Scenario		
		FMR	FNMR	EER	FMR	FNMR	EER
Palm	WLD	11.6%	11.6%	11.6%	-	-	-
	WLD + BioHash	3.5%	3.5%	3.5%	43.9%	50.0%	47.0%
	<b>WLD + AE + BioHash</b>	<b>1.7%</b>	<b>2.1%</b>	<b>1.9%</b>	<b>29.3%</b>	<b>30.0%</b>	<b>29.7%</b>
	RLT	37.5%	37.5%	37.5%	-	-	-
	RLT + BioHash	1.4%	1.1%	1.2%	41.3%	57.0%	49.1%
	<b>RLT + AE + BioHash</b>	<b>0.1%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>33.9%</b>	<b>38.5%</b>	<b>36.2%</b>
	MC	34.5%	34.5%	34.5%	-	-	-
	MC + BioHash	2.1%	1.8%	1.9%	46.7%	51.4%	49.1%
	<b>MC + AE + BioHash</b>	<b>0.0%</b>	<b>0.1%</b>	<b>0.1%</b>	<b>43.0%</b>	<b>38.6%</b>	<b>40.8%</b>
	<b>img + AE + BioHash</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	40.5%	43.1%	41.8%
Wrist	WLD	27.0%	27.0%	27.0%	-	-	-
	WLD + BioHash	3.2%	4.5%	3.9%	44.8%	52.5%	48.7%
	<b>WLD + AE + BioHash</b>	<b>0.7%</b>	<b>0.5%</b>	<b>0.6%</b>	<b>30.7%</b>	<b>36.9%</b>	<b>33.8%</b>
	RLT	36.4%	36.4%	36.4%	-	-	-
	RLT + BioHash	1.4%	1.8%	1.6%	41.5%	53.9%	47.7%
	<b>RLT + AE + BioHash</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>35.4%</b>	<b>37.6%</b>	<b>36.5%</b>
	MC	34.5%	34.5%	34.5%	-	-	-
	MC + BioHash	2.9%	2.0%	2.5%	<b>45.2%</b>	53.0%	49.1%
	<b>MC + AE + BioHash</b>	<b>0.1%</b>	<b>0.0%</b>	<b>0.0%</b>	46.1%	<b>39.2%</b>	<b>42.7%</b>
	<b>img + AE + BioHash</b>	0.7%	0.6%	0.7%	48.9%	46.1%	47.5%

4.13 illustrates the 2D representation of WLD, WLD+PCA, and WLD+AE of 5 different identities in the UTFVP dataset. As this figure shows, WLD could not completely separate the identity of finger vein images. Nonetheless, the features in WLD+PCA and WLD+AE could better determine the identity. In particular, the identities are better distinguished in WLD+AE. Therefore, it is expectable to achieve better performance with the features extracted in the embedding layer of our auto-encoder.

As seen in section 4.3.2.3, adapting hyper-parameters changes the performance of our proposed framework. Experiments show that adapting and choosing suitable values for  $L_{embedding}$  and  $\alpha$  is indeed a trade-off between the performance in the normal and stolen scenarios (see figure 4.9 and figure 4.11, respectively). Meanwhile, we notice that  $L_{embedding}$  should be greater than 50 to achieve sufficient performance in both normal and stolen scenarios. However, our experiments in section 4.3.2.3 suggest that we can find a lower band for  $L_{BioHash}$  where decreasing the value of  $L_{BioHash}$  less than that value degrades the performance of our framework in both normal and stolen scenarios while increasing the value of  $L_{BioHash}$  above that lower band does not significantly change the performance of our method neither in the normal scenario nor in the stolen scenario. To empirically find that lower band, in another experiment, we change  $L_{embedding}$  between 50, 100, 200, and 500, and in each case, we evaluate the performance of our framework for different values of  $L_{BioHash}$  between 25, 50, 75, 100, 200, 500, and 1000. Figure 4.14 represents the ROC curves of our framework for

### 4.3 Deep Auto-encoding and BioHashing for Secure and Protected Vascular Recognition

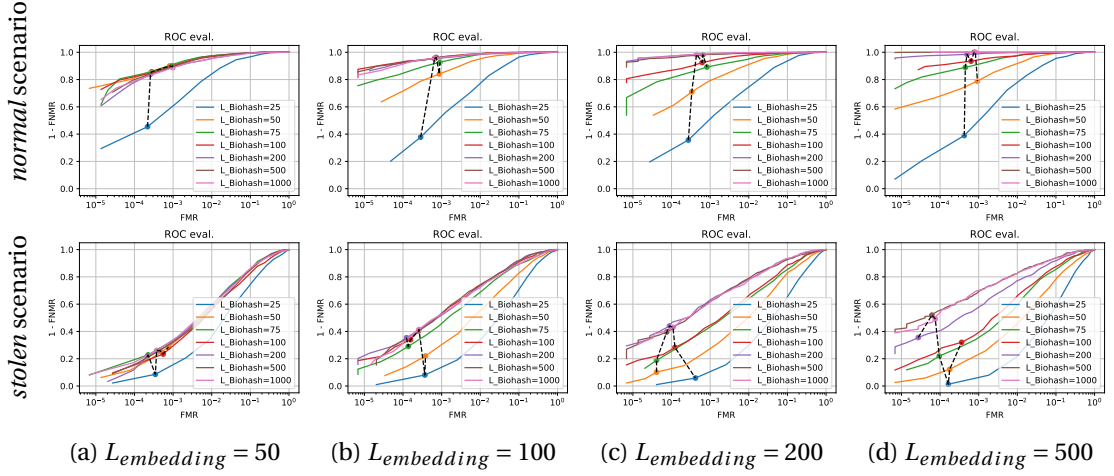


Figure 4.14: Evaluating the performance of our framework with different values of  $L_{embedding}$  and  $L_{BioHash}$  in protecting WLD features on UTFVP dataset in normal scenario (first row) and stolen scenario (second row): a)  $L_{embedding} = 50$ , b)  $L_{embedding} = 100$ , c)  $L_{embedding} = 200$ , d)  $L_{embedding} = 500$ . The marked points which are connected with the dashed lines in each plot correspond to the threshold that leads to  $FMR=10^{-3}$  on the development subset.

the aforementioned configurations<sup>8</sup>. As this figure demonstrates, for each  $L_{embedding}$ , opting the value of  $L_{BioHash}$  less than  $L_{embedding}$  degrades the performance of our framework in both normal and stolen scenarios. Therefore, this experiment shows that the lower band for  $L_{BioHash}$  in our framework to have the higher performance with smaller length of BioHashing is  $L_{embedding}$ .

Our experiment in section 4.3.2.4 also shows that using raw image in the input of the auto-encoder leads to superior performance in the normal scenario. Nevertheless, in the stolen scenario, the pre-processed images achieve better performance when they are used as the input of our framework. Meanwhile, we should note that each of the pre-processing methods requires more execution time, as reported in table 4.9.

It is noteworthy that the execution of the encoder in our framework is very fast. Table 4.12 reports the execution time to get the embedding features from the encoder in our proposed framework and the required memory for the encoder network in our framework for WLD, RLT, and MC methods on the images from UTFVP dataset on a system equipped with an Intel i7-7700K 4.2 GHz CPU and an NVIDIA 1080 Ti GPU. Comparing the execution times in table 4.12 with the required time to extract each feature reported in table 4.9, we can conclude that our proposed framework is quite fast and its execution time is almost negligible respecting the run time for feature extraction in mentioned FVR methods. In a nutshell, considering the enhancement in both normal and stolen scenarios and the high speed of our framework, it is much worth applying our framework to previous FVR methods.

Last but not least, experiments in section 4.3.2.5 show that our framework enhanced the

<sup>8</sup>In this experiment, we use WLD features extracted from images of UTFVP dataset ( $\alpha = 0.1$ ).

## Chapter 4. Protection of Biometric Templates

Table 4.12: The average execution time (second) to get embedding features from the encoder and the required memory for encoder network in our framework

	WLD	RLT	MC
Encoder's Exe. Time*	0.003 (0.0002)	0.02 (0.0008)	0.06 (0.004)
Encoder's req. Memory	3.1 MB	10.9 MB	27.5 MB

\*The values are for the CPU (GPU) implementation.

performance of protected versions of WLD, RLT, and MC methods in both normal and stolen scenarios. Moreover, this experiment confirms the generalization capability of our framework for other vascular biometric modalities such as palm and vein images.

### 4.4 Conclusion

In this chapter, we proposed different methods to protected biometric templates. In section 4.1, we proposed a new cancelable biometric template protection scheme, called MLP-Hash, which uses a user-specific randomly-weighted multi-layer perceptron (MLP) with non-linear activation functions, followed by binarization of the output. We evaluated the unlinkability, irreversibility and recognition accuracy of MLP-Hash as per the ISO/IEC 30136 standard requirements, using SOTA face recognition models. Our protection method was found to satisfy these criteria to a high degree. In addition, we compared MLP-Hash with the BioHashing and IoM Hashing (IoM-GRP and IoM-URP) protection algorithms on the same SOTA face recognition systems, in terms of the recognition accuracy, unlinkability, and irreversibility criteria. Our experiments indicate that while all these template protection schemes are almost unlinkable, there is a trade-off between irreversibility, recognition accuracy, and complexity.

In section 4.2, we proposed a generic hybrid BTP scheme for biometric templates by combining Cancelable Biometrics (CB) and Homomorphic Encryption (HE). We showed that the comparison methods of CB schemes (such as BioHashing, MLP-Hashing, and IoM Hashing) can be adapted to perform equivalent computations in the HE domain, and therefore our hybrid scheme was found to achieve equal recognition performance with the corresponding CB. Our experiments further showed that the proposed hybrid method is able to achieve better performance compared to HE alone in the *normal* scenario. In the *full disclosure* threat model (where algorithms and secret keys are disclosed to an adversary), the hybrid-protected templates were found to have comparable performance with HE-protected templates in most cases, when the length of the CB-protected templates was equal to the length of the unprotected templates. As the length of the CB-protected templates was decreased, the performance of the hybrid-protected templates was found to also decrease, so for much smaller lengths the performance of templates protected using HE alone was sometimes found to be better than that of hybrid-protected templates. However, the main drawback of HE-protected templates is that they can be easily inverted by an adversary with access to the secret decryption key, while

hybrid-protected templates remain irreversible in this case. Besides the additional template protection offered by our hybrid BTP method, it is also useful for reducing the dimensionality of the biometric templates with CB, prior to applying HE, which can decrease the amount of computation on the encrypted templates (ciphertexts). In particular, by appropriately tuning the length of CB-protected templates, we could achieve comparable recognition performance with HE, but with a faster execution time.

In section 4.3, we considered vascular images as biometric data and proposed a deep-learning-based framework to protect and enhance the previous vascular recognition methods by reducing the dimension of biometric features using a DNN and then protecting the reduced-dimension features with BioHashing. To this end, we used the raw finger vein images and the extracted features from previous FVR methods to train a deep convolutional auto-encoder with a multi-term loss function. We used the auto-encoder to extract reduced-dimension features in the bottleneck layer (embedding layer). Finally, we generated protected templates from deep features using BioHash. The experimental results indicate that the protected templates generated by our framework achieve superior performance than both BioHash protected templates of the raw features in the normal scenario. Furthermore, in the stolen scenario, our framework has far better performance than BioHash protected templates of the raw features. In addition to improving previous FVR methods, we trained our auto-encoder directly on finger vein images as a new FVR system. We also evaluated the generalization of our proposed framework on other vascular biometric modalities (i.e., palm and wrist).



## 5 Evaluation of Biometric Template Protection Schemes

In Chapter 3, we showed the vulnerability of unprotected systems, and with the motivation of mitigating such attacks, in Chapter 4 we presented different template protection methods. As described in Chapter 2, there are also different template protection schemes proposed in the literature. Each of the proposed template protection scheme are expected to provide protection for the biometric systems, and it is essential to evaluate the security and performance of the protected systems. With this introduction, this chapter focuses on the evaluation of protected systems, and investigate requirements of template protection mechanisms. First, in Section 5.1, we present a benchmark of several popular template protection schemes based on cancelable biometrics (CB) using some metrics from the literature and discuss the limitations of evaluation in each case. In Section 5.2, we propose a learning-based approach for reconstructing face image from protected templates, as a general method to evaluate irreversibility of protected facial templates. In Section 5.3, we focus on the linkability of protected templates and proposed a new measure to evaluate linkability in protected biometric systems. The proposed linkability metric is based on maximal leakage, which is a well-studied measure in information-theoretic literature. We show that the resulting linkability measure has a number of important theoretical properties and an operational interpretation in terms of statistical hypothesis testing. In Section 5.4, we extend our linkability metric for the case where the adversary gains access to multiple protected templates and evaluate the linkability of multiple protected templates, which has not been investigated in the literature.

### 5.1 Benchmarking Cancelable Biometric Protection Schemes for Deep Templates

In this section, we focus on CB schemes and benchmark several existing methods based on the requirements defined in the ISO/IEC 24745 standard on biometric information protection. We consider BioHashing [76], Multi-Layer Perceptron (MLP) Hashing [23], Bloom Filters [78], and two schemes based on Index-of-Maximum (IoM) Hashing [77] (i.e., uniformly random permutation-based hashing, shortly IoM-URP, and Gaussian random projection-based hash-

## Chapter 5. Evaluation of Biometric Template Protection Schemes

---

ing, shortly IoM-GRP). In addition to the mentioned CB schemes, we introduce a CB scheme (as a baseline), called Rand-Hash, based on user-specific random transformation, including random permutation, random scale, and random sign flip, followed by binarization.

Each of the mentioned previous CB schemes from the literature is proposed and evaluated on a particular biometric characteristic. Furthermore, the evaluation of privacy protection schemes on different biometric characteristics is often limited to the recognition performance. Most of these schemes are also not evaluated on the state-of-the-art feature extractor methods in the literature. Therefore, a comparison of these CB schemes based on the requirements defined in the ISO/IEC 24745 standard for state-of-the-art feature extractors of different biometric characteristics is still challenging. In our experiments, we consider different physiological and behavioral biometric characteristics, including face, voice, finger vein, and iris. For each biometric characteristic, we use state-of-the-art (SOTA) feature extraction models in the field which are based on deep neural networks (DNNs).

In section 5.1.1, we describe the metrics used to evaluate the recognition performance, unlinkability, and irreversibility in our benchmark. We apply the same metrics to all CB schemes, which allows for a direct comparison across different characteristics. To obtain a fair comparison of CB schemes, as far as possible, we generate protected templates of the same length across different CB schemes in each characteristic. In section 5.1.2, we report the experimental results of our benchmark for different CB schemes on various biometric modalities.

### 5.1.1 Evaluation metrics

In this section, we describe the metrics used for evaluating recognition performance, unlinkability, and irreversibility in our benchmark.

#### 5.1.1.1 Recognition Performance

To evaluate the recognition performance of protected templates, we only consider verification and calculate the Equal Error Rate (EER) as well as the False Non-Match Rate (FNMR) at the decision thresholds corresponding to False Match Rates (FMRs) of 1% and 0.1%. We also plot the Detection Error Trade-off (DET) curves. We evaluate the recognition performance in two scenarios:

- *unknown-key scenario*: it is the expected case in practice, where we generate protected templates with user-specific keys.
- *known-key scenario*: we assume that keys are disclosed, hence we evaluate the recognition performance considering the same key for each user<sup>1</sup>.

To evaluate the recognition performance, we consider all possible combinations of samples for

---

<sup>1</sup>Also referred to as *stolen-token* scenario in the literature.



## 5.1 Benchmarking Cancelable Biometric Protection Schemes for Deep Templates

mated comparisons. For non-mated comparisons, we consider all possible pairs of subjects and use the first sample for each subject in the dataset.

### 5.1.1.2 Unlinkability

To evaluate unlinkability of CB schemes, we first generate mated and non-mated template pairs with sample-specific keys, and then we calculate the general unlinkability measure introduced in [117]. The linkability of two templates is measured in terms of the difference of conditional probabilities of two hypotheses of being mated,  $H_m$ , and non-mated,  $H_{nm}$ , for a given comparison score  $s$  between two given templates:

$$D_{\leftrightarrow}(s) = p(H_m|s) - p(H_{nm}|s). \quad (5.1)$$

Then, by finding conditional expectation of this local measure  $D_{\leftrightarrow}(s)$  over all comparison scores, we can find a global measure,  $D_{\leftrightarrow}^{sys}$ , which is considered as the system unlinkability measure:

$$D_{\leftrightarrow}^{sys} = \int p(s|H_m) D_{\leftrightarrow}(s) ds. \quad (5.2)$$

The value of  $D_{\leftrightarrow}^{sys}$  is in interval  $[0,1]$ , with lower values indicating smaller possibilities to link templates of the same subject.

### 5.1.1.3 Irreversibility

Because each template protection method uses different mechanism, the inversion of protected templates may differ in different BTP schemes. Therefore, the irreversibility of each BTP method has been evaluated based an ad-hoc approach, and the literature lacks a general approach to measure the irreversibility of protected templates.

While information-theoretic metrics, such as *conditional entropy* and *mutual information*, can be used to quantify the uncertainty in estimating original data from the protected templates these metrics are difficult to compute in practice, especially if biometric templates contain an high number of features [3]. To simplify the computation of entropy and mutual information (MI), we consider the set of original (unprotected) biometric templates  $X$  as well as their corresponding set of protected biometric templates  $Y$ , and apply Principal Component Analysis (PCA) to our sets  $X$  and  $Y$ . Let us assume that  $X$  and  $Y$  are matrices with initial dimensions  $s \times u$  and  $s \times p$ , respectively, where  $s$ ,  $u$ ,  $p$  are the number of samples, the number of features in unprotected templates, and the number of features in protected templates, respectively. After applying PCA to the matrices  $X$  and  $Y$ , we obtain the reduced matrices  $X_r = PCA(X)$  and  $Y_r = PCA(Y)$ , with dimensions  $s \times r$ , where  $r$  is the number of reduced features (possibly different across matrices). While decreasing the number of features, PCA retains the most significant information of biometric templates. That is, reduced matrices are

## Chapter 5. Evaluation of Biometric Template Protection Schemes

Table 5.1: Summary of feature extraction models, datasets, numbers of mated and non-mated comparisons, and biometric performance in terms of Equal Error Rate (EER) achieved for different biometric characteristics in biometric verification.

Characteristic	Model	Dataset	# Subjects	# Mated	# Non-mated	EER
Face	ArcFace	MOBIO (face)	150	1,516,300	22,952	0.02%
Voice	ECAPA-TDNN	MOBIO (voice)	150	1,516,300	22,952	6.64%
Finger Vein	Modified Densenet-161	SDUMLA	318	9,540	100,806	0.32%
Iris	Modified Densenet-201	CASIA Thousand	457	13,710	207,956	2.05%

suitable to account for the partial reversibility of protected biometric data, which in many cases is sufficient to obtain access in biometric recognition systems.

To obtain a fair comparison between the different CB schemes, we apply PCA to the matrices of unprotected templates  $X$  and protected templates  $Y_i$  resulting from different CB schemes  $i$ , always considering a fixed number of features  $r = 100$  for the reduced matrices. Then, we approximate to multivariate Gaussian the distribution of features of the reduced matrices. For each matrix  $Y_{r,i}$ , we can compute the MI between  $X_r$  and  $Y_{r,i}$  as follows:

$$MI(X_r, Y_{r,i}) = H(X_r) + H(Y_{r,i}) - H(X_r, Y_{r,i}), \quad (5.3)$$

where  $H(\cdot)$  is the measure of entropy, quantified with the Shannon's entropy formula [213].

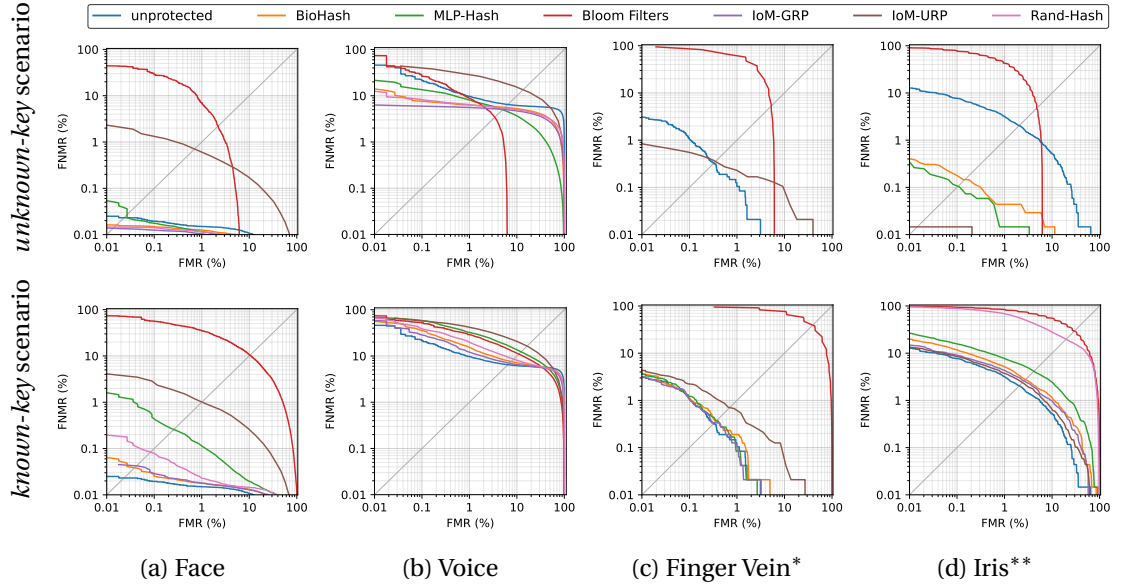
### 5.1.2 Experimental Results

In this section, we report the experimental results of our benchmark framework for the evaluation of different CB schemes: BioHashing [76], Multi-Layer Perceptron (MLP) Hashing [23], Bloom Filters [78], and two schemes based on Index-of-Maximum (IoM) Hashing [77] (i.e., IoM-URP and IoM-GRP). In addition, as a baseline, we consider a CB scheme named Rand-Hash, based on user-specific random transformations (including random permutation, random scale, and random sign flip) followed by binarization. We evaluate these CB schemes on different biometric characteristics. Table 5.1 summarises the feature extraction model, the dataset used for each biometric characteristic, the number of mated and not-mated comparisons, and the system verification performance achieved in terms of Equal Error Rate (EER).

#### 5.1.2.1 Recognition Performance Evaluation

Figure 5.1 depicts the DET curves for different CB schemes on different biometric characteristics. Table 5.2 also compares the recognition performance in terms of Equal Error Rate (EER) as well as False Non-Match Rate (FNMR) at a False Match Rate (FMR) of 1% and 0.1%. In general, Bloom Filters (which was not initially proposed to protect DNN-based features)

## 5.1 Benchmarking Cancelable Biometric Protection Schemes for Deep Templates



\* In the *unknown-key* scenario, DET curves of protected templates with BioHash, MLP-Hash, IoM-GRP, and Rand-Hash are not visible because EER=0. Similarly, in the *known-key* scenario, DET curves of Rand-Hash are not visible.

\*\* In the *unknown-key* scenario, DET curve of protected templates with IoM-GRP is not visible because EER=0.

Figure 5.1: System performance evaluation on the *unknown-key* and *known-key* scenarios for different physiological and behavioral biometric traits.

has the lowest recognition performance for face, finger vein, and iris. For voice recognition however, IoM-URP has the worst recognition performance. Also, we observe that in face recognition, BioHash, MLP-Hash, IoM-GRP, IoM-URP, and Rand-Hash have comparable performance in the *unknown-key* scenario. In voice recognition, IoM-GRP achieves the best performance in the *unknown-key* and *known-key* scenarios. In finger vein recognition, BioHash, MLP-Hash, IoM-GRP, IoM-URP, and Rand-Hash have comparable performance in the *unknown-key* scenario. Nevertheless, Rand-Hash achieves the best recognition performance in the *known-key* scenario for finger vein recognition. Similarly in iris recognition, we observe that BioHash, MLP-Hash, IoM-GRP, IoM-URP, and Rand-Hash have comparable performance, and IoM-GRP achieves the best performance. However, in the *known-key* scenario, IoM-URP achieves the best recognition performance for iris recognition. It is noteworthy that generally for each biometric characteristic, protected templates with some of the CB schemes achieve better recognition accuracy than unprotected templates in the *unknown-key* scenario. The improvement in the accuracy compared to unprotected templates in such cases is obtained with the cost of using user-specific keys.

In Table 5.2, we observe that it is not possible to rank CB schemes according to their recognition performance with the different biometric modalities. Some CB schemes perform poorly with specific biometric modalities, for instance Bloom Filters in finger vein and iris recognition,

## Chapter 5. Evaluation of Biometric Template Protection Schemes

Table 5.2: Recognition performance evaluation in benchmarking CB schemes.

CB scheme	Characteristic	unknown-key scenario [%]			known-key scenario [%]		
		EER	FNMR@FMR=1%	FNMR@FMR=0.1%	EER	FNMR@FMR=1%	FNMR@FMR=0.1%
<b>Unprotected</b>	Face	0.02	0.01	0.02	–	–	–
	Voice	6.4	9.71	22.53	–	–	–
	Finger Vein	0.32	0.10	1.05	–	–	–
	Iris	2.05	3.18	7.69	–	–	–
<b>BioHash</b> [76]	Face	0.02	0.01	0.01	0.04	0.02	0.03
	Voice	5.28	6.31	8.31	7.64	15.84	36.60
	Finger Vein	0.00	0.00	0.00	0.35	0.19	1.13
	Iris	0.14	0.04	0.18	2.77	5.15	11.80
<b>MLP-Hash</b> [23]	Face	0.02	0.01	0.02	0.02	0.13	0.54
	Voice	5.25	8.30	14.19	12.16	33.12	61.70
	Finger Vein	0.00	0.00	0.00	0.37	0.08	1.22
	Iris	0.11	0.01	0.10	4.14	7.91	15.80
<b>Bloom Filters</b> [78]	Face	2.19	7.10	30.53	35.40	56.67	43.33
	Voice	3.42	9.01	28.70	11.03	28.97	53.03
	Finger Vein	4.97	65.41	91.28	36.49	94.76	94.76
	Iris	4.69	42.57	78.15	29.26	84.11	93.17
<b>IoM-GRP</b> [77]	Face	0.02	0.01	0.01	0.04	0.02	0.03
	Voice	5.36	5.58	5.96	7.14	12.57	29.34
	Finger Vein	0.00	0.00	0.00	0.33	0.10	1.05
	Iris	0.00	0.00	0.00	2.58	4.25	9.25
<b>IoM-URP</b> [77]	Face	0.72	0.68	1.51	1.10	1.18	2.87
	Voice	12.35	31.93	43.91	16.52	44.40	61.14
	Finger Vein	0.35	0.23	0.55	0.69	0.67	2.24
	Iris	0.02	0.00	0.01	2.38	3.66	8.56
<b>Rand-Hash</b>	Face	0.02	0.01	0.01	0.08	0.02	0.08
	Voice	5.70	6.48	8.40	8.35	20.81	40.66
	Finger Vein	0.00	0.00	0.00	0.00	0.00	0.00
	Iris	0.00	0.00	0.00	19.76	68.11	88.55

or IoM-URP in voice recognition. Numerous reasons may be behind this behaviour: feature vectors of different biometric modalities share the same length but contain different information, variability, and entropy. In addition, some CB schemes can handle variability better than others, and as a result some biometric modalities may be more challenging than others for specific CB schemes.

Last but not least, we should note that the evaluation of recognition accuracy in the *known-key* scenario reflects the security of the system when the adversary uses the leaked key with their own biometric data, without any more knowledge of the enrolled subject or the biometric system. However, in a different threat model, an adversary may have also more knowledge of the system or enrolled subject that can be used to enter the system. Such scenarios can be further explored based on the adversary's knowledge and capabilities.

## 5.1 Benchmarking Cancelable Biometric Protection Schemes for Deep Templates

Table 5.3: Unlinkability evaluation of in benchmarking CB schemes

CB scheme	Face	Voice	Finger Vein	Iris
<b>BioHash</b> [76]	0.0110	0.0078	0.0140	0.0106
<b>MLP-Hash</b> [23]	0.0088	0.0160	0.0099	0.0122
<b>Bloom Filter</b> [78]	0.0545	0.0735	0.0091	0.0131
<b>IoM-GRP</b> [77]	0.0086	0.0065	0.0130	0.0072
<b>IoM-URP</b> [77]	0.0090	0.0053	0.0136	0.0090
<b>Rand-Hash</b>	0.0084	0.0061	0.0107	0.0099

### 5.1.2.2 Unlinkability Evaluation

Table 5.3 compares the system unlinkability measure proposed in [117] (i.e.,  $D_{\rightarrow}^{\text{sys}}$  as in E.q. 2) for different CB schemes when protecting templates extracted from different biometric modalities. This measure evaluates unlinkability of protected templates based on the overlap between the distribution of scores of mated templates and the distribution of scores of non-mated templates protected with different keys. Therefore, if the distribution of scores of mated templates and the distribution of scores of non-mated templates largely overlap, based on the hypothesis test in this measure, it is hard to link templates. Therefore, protected templates are considered to be unlinkable and the global measure  $D_{\rightarrow}^{\text{sys}}$  will be close to zero. Accordingly, as Table 5.3 shows, all CB schemes achieve low values for  $D_{\rightarrow}^{\text{sys}}$  for different biometric characteristics, and thus are *almost unlinkable* across different biometric characteristics. Comparing different CB schemes in this table, Bloom Filter has the highest linkability (i.e., the least unlinkability) for all biometric characteristics except for finger vein. Meanwhile, the most unlinkable CB scheme is different across different biometric characteristics. We should note that while the unlinkability measure  $D_{\rightarrow}^{\text{sys}}$  can be used to compare two different protected systems, it remains difficult to interpret differences in the values of the unlinkability measure  $D_{\rightarrow}^{\text{sys}}$  for different cases, e.g., one biometric characteristic when protected with different CB schemes or different biometric characteristics when protected with one CB scheme. Hence, this aspect of unlinkability evaluation of protected biometric templates requires further studies. We propose a new measure in Section 5.3 which addresses this limitation and have also other advantages compared to measure in [117].

### 5.1.2.3 Irreversibility Evaluation

In Table 5.4, for each biometric characteristic, CB scheme, and scenario, we report the MI between the reduced matrices of unprotected and protected templates. By comparing the MI values obtained between the *unknown-key* and *known-key* scenarios, we observe a clear increase of MI in the *known-key* scenario. In the latter scenario, the key required by CB schemes is no more user-specific and it can be simply considered as a parameter of the CB scheme. As a consequence, from protected templates in the *known-key* scenario it is easier to extract information about the original (unprotected) templates. We also observe that, in general, for the *unknown-key* scenario face is the biometric characteristic that provides

Table 5.4: Irreversibility evaluation in benchmarking CB schemes.

CB scheme	Characteristic	unknown-key scenario	known-key scenario
<b>BioHash</b> [76]	Face	39.63	98.81
	Voice	12.97	53.74
	Finger Vein	18.80	115.99
	Iris	8.63	63.99
<b>MLP-Hash</b> [23]	Face	35.42	58.00
	Voice	10.74	26.37
	Finger Vein	19.35	110.04
	Iris	7.92	38.65
<b>Bloom Filters</b> [78]	Face	40.18	21.37
	Voice	20.26	29.60
	Finger Vein	12.32	8.89
	Iris	8.14	8.56
<b>IoM-GRP</b> [77]	Face	31.33	48.91
	Voice	8.68	22.83
	Finger Vein	18.18	57.85
	Iris	8.31	38.29
<b>IoM-URP</b> [77]	Face	8.79	9.06
	Voice	1.69	3.10
	Finger Vein	14.01	16.06
	Iris	6.55	24.59
<b>Rand-Hash</b>	Face	39.26	97.07
	Voice	12.47	52.31
	Finger Vein	19.65	113.96
	Iris	9.77	26.43

the highest MI, while for the *known-key* scenario finger vein is the biometric characteristic that provides the highest MI. As observed for recognition performance, it is not possible to rank CB schemes according to the irreversibility evaluation across different biometric modalities. In particular, there is a high difference between the values of MI obtained for the different biometric characteristics, and also it is difficult to explain the values and their differences. Meanwhile, the results are aligned with the recognition performance, and the aforementioned reasons that cause different recognition performance of CB schemes across biometric modalities also apply here.

## 5.2 Inversion of Protected Biometric Templates

As mentioned in Section 5.1, it is always challenging to investigate the invertability of protected templates since BTP schemes have different mechanisms, and therefore for each BTP scheme, a specific inversion method has been used in the literature. In addition, despite general measures in the literature to evaluate linkability of protected templates (such as [24]), there is no general method to investigate invertability of protected templates.

In this section, we focus on the inversion of face images from protected facial templates. We consider a scenario where the adversary gains knowledge of the template protection scheme as well as its secrets<sup>2</sup> and tries to reconstruct the face image using a leaked protected template. We consider different template protection schemes, including BioHashing, MLP-Hashing, and Homomorphic Encryption (HE), and reconstruct face images from protected templates. We also use different state-of-the-art face recognition models in both *whitebox* (where the adversary has a complete knowledge of feature extractor) and *blackbox* (where the adversary has a blackbox knowledge of feature extractor) scenarios. To our knowledge, this is the first work on the reconstruction of face images from protected facial templates, which is independent of the template protection scheme and can be applied against different protection schemes.

As discussed in Section 2.1 of Chapter 2, several works explored the reconstruction of face images from facial templates, particularly from raw (unprotected) templates [11], [12], [14], [15], [17], [19], [53], [61]. In contrast to most work on the reconstruction of face images from raw templates, [214] used the network in [53] to reconstruct binarised facial features. However, no template protection mechanism was considered, and the authors only considered a simple binarisation transformation being applied to raw templates.

### 5.2.1 Face Reconstruction from Protected Templates

We first describe our threat model, where the adversary gains access to a protected template and aims to reconstruct the underlying face image. Then, we describe our network to reconstruct face images from leaked protected templates.

#### 5.2.1.1 Threat Model

We consider the protected face recognition system with the situation where an adversary gains access to a protected facial template and aims to reconstruct a face image from the leaked protected template and use the reconstructed face image to impersonate. We consider the following properties for the adversary:

---

<sup>2</sup>which is the case in the *full-disclosure* scenario defined in the ISO/IEC 30136 standard [10] for evaluating the invertability of protected templates.

## Chapter 5. Evaluation of Biometric Template Protection Schemes

- *Adversary's goal:* The adversary aims to impersonate a user enrolled in the FR system.
- *Adversary's knowledge:* The adversary has the following information:
  1. A leaked protected face template  $\mathbf{t}_{\text{btp}}$  of a user enrolled in the database of the face recognition system.
  2. Complete knowledge of template protection scheme  $P$  and its secrets  $\mathbf{k}_{\text{btp}}$ .
  3. The whitebox knowledge (including parameters and internal functioning) or black-box knowledge of the feature extraction model  $F(\cdot)$  of the face recognition system. In the case of the blackbox scenario, the adversary is assumed to have whitebox knowledge of an alternative feature extraction model  $F_{\text{adv}}(\cdot)$ .
- *Adversary's capability:* The adversary can inject the reconstructed face image into the feature extractor of the target face recognition system and bypass the camera. For simplicity and to verify how similar is the reconstructed face image to the original image, we assume that injection is made to a similar system without protection.
- *Adversary's strategy:* Under the above assumptions, the adversary can use the leaked protected template and underlying reconstruct face image  $\hat{\mathbf{I}}$  using a face reconstruction method. Then, the adversary can use the reconstructed face image  $\hat{\mathbf{I}}$  to inject a query to impersonate.

### 5.2.1.2 Face Reconstruction

Our network for reconstructing face images from the leaked *protected* templates stems from the network proposed in [12] for reconstructing unprotected facial templates. We consider a dataset of face images  $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$ , with  $N$  images. We extract facial templates  $\mathbf{t} = F(\mathbf{I})$  from each face image  $\mathbf{I}$ , and then generate the protected version  $\mathbf{t}_{\text{btp}} = P(\mathbf{t}, \mathbf{k}_{\text{btp}})$  with the template protection scheme  $P$  using the leaked secret  $\mathbf{k}_{\text{btp}}$ . The secret  $\mathbf{k}_{\text{btp}}$  can be user-specific or identical for all subjects (i.e., application-based key). In either case, it is assumed to be known by the adversary. In case of encryption-based template protection (such as HE), the protected templates  $\mathbf{t}_{\text{btp}}$  are in the ciphertext, however since the adversary is assumed to have access to secret key, the adversary can decrypt the protected template into the plaintext<sup>3</sup> as  $\mathbf{t}_{\text{btp,adv}}$  and use it for the inversion attack. In other cases where the protection is not based on encryption (such as cancelable biometric schemes), the adversary can directly use the protected template for the inversion attack, i.e.,  $\mathbf{t}_{\text{btp,adv}} = \mathbf{t}_{\text{btp}}$ . Then, the adversary can build the training dataset  $\mathcal{D} = \{(\mathbf{t}_{\text{btp,adv},i}, \mathbf{I}_i)\}_{i=1}^N$  with  $N$  pairs of protected templates  $\mathbf{t}_{\text{btp},i}$  and their corresponding face images  $\mathbf{I}_i$ . We use the network structure in [12], composed of enhanced deconvolution using cascaded convolution and skip connections (shortly, DSCasConv) blocks, and use the

<sup>3</sup>In such cases, the protected templates in the ciphertext  $\mathbf{t}_{\text{btp,ciphertext}} = \text{Enc}(M(\mathbf{t}), \mathbf{k}_{\text{btp}})$  are generated by encrypting the mapped template  $M(\mathbf{t})$ , where  $M(\cdot)$  is a transformation function which is specific to the encryption algorithm (e.g., quantization). Therefore, decrypting the protected version into the plaintext will lead to  $\mathbf{t}_{\text{btp,plaintext}} = \text{Dec}(\mathbf{t}_{\text{btp,ciphertext}}, \mathbf{k}_{\text{btp}}) = M(\mathbf{t})$ , which is the mapped version of unprotected template, and thus  $\mathbf{t}_{\text{btp,adv}} = M(\mathbf{t})$ . We should note that  $\text{Enc}(\cdot, \cdot)$  and  $\text{Dec}(\cdot, \cdot)$  denote encryption and decryption, respectively.



protected template (instead of the unprotected template) as the input. We optimize our model with a multi-term loss function, including:

- *Mean Absolute Error (MAE)*: To reduce the pixel level reconstruction error, we minimize the  $\ell_1$  of reconstruction error:

$$\mathcal{L}_{\text{MAE}}(\hat{\mathbf{I}}, \mathbf{I}) = \|\hat{\mathbf{I}} - \mathbf{I}\|_1, \quad (5.4)$$

where  $\hat{\mathbf{I}}$  and  $\mathbf{I}$  are the reconstructed and original face images, respectively.

- *Dissimilarity Structural Index Metric (DSSIM)*: To enhance the quality of the reconstructed image in terms of the Similarity Structural Index Metric (SSIM) [175], we further minimize the DSSIM loss term [215] as follows:

$$\mathcal{L}_{\text{DSSIM}}(\hat{\mathbf{I}}, \mathbf{I}) = \frac{1 - \text{SSIM}(\hat{\mathbf{I}}, \mathbf{I})}{2} \quad (5.5)$$

- *ID loss*: To preserve the identity information in the reconstructed face image, we use a feature extractor  $F_{\text{adv}}$  and minimize the distance between the features extracted from the original face  $\mathbf{I}$  and reconstructed face  $\hat{\mathbf{I}}$  images. We minimize the  $\ell_1$ -norm distance and cosine distance of the extracted templates as follows:

$$\begin{aligned} \mathcal{L}_{\text{ID}}(\hat{\mathbf{I}}, \mathbf{I}) &= \mathcal{L}_{\text{ID}, \ell_1}(\hat{\mathbf{I}}, \mathbf{I}) + \mathcal{L}_{\text{ID}, \cos}(\hat{\mathbf{I}}, \mathbf{I}) \\ &= \|F_{\text{adv}}(\hat{\mathbf{I}}) - F_{\text{adv}}(\mathbf{I})\|_1 + \frac{-F_{\text{adv}}(\hat{\mathbf{I}}) \cdot F_{\text{adv}}(\mathbf{I})}{\|F_{\text{adv}}(\hat{\mathbf{I}})\|_2 \cdot \|F_{\text{adv}}(\mathbf{I})\|_2} \end{aligned} \quad (5.6)$$

Similar to [11], [15], [17], in the whitebox scenario we consider  $F_{\text{adv}} = F$ , but in the blackbox scenario we assume that the adversary has access to an alternative model  $F_{\text{adv}}$  and uses it in the loss function.

We use a weighted summation of these loss terms as our total loss:

$$\mathcal{L} = \mathcal{L}_{\text{MAE}} + \gamma_1 \mathcal{L}_{\text{DSSIM}} + \gamma_2 \mathcal{L}_{\text{ID}} \quad (5.7)$$

We experimentally found that the choice of  $\gamma_1 = 0.75$ , and  $\gamma_2 = 0.025$  achieves the best performance.

### 5.2.2 Experimental Results

In this section, we present our experimental results and discuss our findings. First, we describe our experimental setup. Then, we present our experimental result in the reconstruction of protected templates. Finally, we discuss our findings.

## Chapter 5. Evaluation of Biometric Template Protection Schemes

Table 5.5: Performance of reconstructed face images from protected templates in attacking a face recognition system with same feature extractor evaluated on the MOBIO, LFW, and AgeDB datasets for the false match rate of  $10^{-2}$ . The ArcFace model is used as  $F_{adv}$ , and thus the attacks against ArcFace are **whitebox** but against other face recognition models are in **blackbox** (denoted as cell color in gray). The values are in percentage.

Dataset	BTP	Face Recognition					
		ArcFace	ElasticFace	AttentionNet	HRNet	RepVGG	Swin
MOBIO	BioHashing	100.0	100.0	98.57	99.05	95.71	100.0
	MLP-Hash	96.67	91.9	84.29	82.86	75.24	94.76
	HE	100.0	100.0	100.0	99.05	98.1	99.52
LFW	BioHashing	95.74	96.34	79.73	85.73	69.31	90.18
	MLP-Hash	88.2	91.0	58.09	66.04	47.62	77.61
	HE	97.18	96.57	82.87	87.52	73.07	91.65
AgeDB	BioHashing	83.23	88.13	73.51	71.63	67.44	89.79
	MLP-Hash	62.89	64.73	32.94	32.74	28.42	58.8
	HE	92.75	90.88	77.99	78.61	72.93	91.8

### 5.2.2.1 Experimental Setup

**Biometric template protection schemes:** We consider different biometric template protection schemes, including two cancelable biometric schemes as well as a template protection method based on Homomorphic Encryption (HE). For cancelable biometric, we consider BioHashing [76] (which is a simple and popular scheme) and MLP-Hash [23] (which is a recently proposed scheme). We consider the protected systems with these schemes to be operating with a user-specific key setting, and thus, the adversary knows the key for the leaked facial template. For the HE-based method, different algorithms have been used for biometric template protection. For instance, HE based on the CKKS scheme supports floating-point encryption, and thus decryption of the protected template using the leaked template will lead to the original unprotected template. In contrast, some other schemes, such as BFV, support integers and, therefore, require quantization before encryption. That means the decryption of the protected templates leads to quantized templates in plaintext. In our experiments, we consider the protection based on HE schemes that require quantized templates. We should note that in the HE-based protection, the secret key (i.e., private key) is often the same for all subjects.

**Face recognition models:** We use state-of-the-art face reconstruction models including ArcFace [132] and ElasticFace [133] as well as four different face recognition models with state-of-the-art backbones from FaceX-Zoo [156], including AttentionNet [141], HRNet [138], RepVGG [145], and Swin [147]. We use the pretrained models of each of these network trained on the MS-Celeb-1M dataset [157]. The recognition performance of these models on the MOBIO, LFW, and AgeDB datasets is reported in Table A.1 of Appendix A.

**Dataset:** We use the FFHQ [161] dataset for training our face reconstruction model. We

## 5.2 Inversion of Protected Biometric Templates

Table 5.6: Performance of reconstructed face images from protected templates in attacking a face recognition system with same feature extractor evaluated on the MOBIO, LFW, and AgeDB datasets for the false match rate of  $10^{-3}$ . The ArcFace model is used as  $F_{adv}$ , and thus the attack against ArcFace is **whitebox** and against other face recognition models are in **blackbox** (denoted as cell color in gray). The values are in percentage.

Dataset	BTP	Face Recognition					
		ArcFace	ElasticFace	AttentionNet	HRNet	RepVGG	Swin
MOBIO	BioHashing	99.52	100.0	91.43	95.24	89.05	100.0
	MLP-Hash	78.57	80.95	56.67	59.05	47.62	85.71
	HE	100.0	99.52	97.62	97.14	96.19	99.52
LFW	BioHashing	92.69	92.79	66.4	68.29	56.3	85.21
	MLP-Hash	77.51	77.1	30.44	30.83	24.49	63.62
	HE	95.92	93.91	72.54	73.51	60.58	87.52
AgeDB	BioHashing	62.47	76.81	49.34	44.57	50.92	73.54
	MLP-Hash	36.37	45.34	13.05	11.8	15.03	31.68
	HE	82.35	82.1	56.34	53.43	57.4	79.63

evaluate our models on the MOBIO [158], LFW [159], and AgeDB [160] datasets. We build protected face recognition systems using the mentioned face recognition model and BTP schemes on each of our evaluation datasets. Then, we use our corresponding reconstruction model trained on FFHQ to invert enrolled protected templates and reconstruct face images. We inject the reconstructed face image as a query to the system to evaluate the performance of face reconstruction in terms of an adversary’s Success Attack Rate (SAR) in entering an unprotected face recognition system with the same feature extractor when the system is configured at False Match Rate (FMR) of  $10^{-3}$ .

### 5.2.2.2 Face Reconstruction from Protected Templates

We consider face recognition systems protected with BioHashing, MLP-Hash, and HE and assume that the adversary knows the template protection scheme and its secrets. We train our face reconstruction network and use the protected templates stored in the database of the face recognition system to reconstruct the face images. We use ArcFace as  $F_{adv}$  and evaluate the performance of our method in attack against protected templates of different face recognition models. Table 5.5 and Table 5.6 report the adversary’s success attack rate in entering a face recognition with the same feature extractor on false match rates (FMRs) of  $10^{-2}$  and  $10^{-3}$  on the MOBIO, LFW, and AgeDB datasets. We should note that since we use ArcFace as  $F_{adv}$ , the attacks against ArcFace are whitebox attacks but against other face recognition models are blackbox attacks. As the results in these tables shows, the reconstructed face images by inverting protected templates using our method achieve significant performance in attacks against protected templates. Fig. 5.2 also shows sample reconstructed face images using our method in the reconstruction of ElasticFace templates protected with different template

## Chapter 5. Evaluation of Biometric Template Protection Schemes

Table 5.7: Performance of reconstructed face images from HE-protected templates in attacking a face recognition system with same feature extractor using different models as  $F_{adv}$ , evaluated on the MOBIO, LFW, and AgeDB datasets for the false match rate of  $10^{-3}$ . In case  $F_{adv}$  is the same as target face recognition model, the adversary is assumed to have knowledge of the target model and thus the attack is *whitebox*; otherwise, the attack is *blackbox* (denoted as cell color in gray). The values are in percentage.

Dataset	$F_{adv}$	Face Recognition					
		ArcFace	ElasticFace	AttentionNet	HRNet	RepVGG	Swin
MOBIO	ArcFace	100.0	99.52	97.62	97.14	96.19	99.52
	ElasticFace	100.0	99.52	97.62	97.14	96.19	99.52
	same (i.e., whitebox)	100.0	99.52	95.24	97.14	96.19	99.52
LFW	ArcFace	95.92	93.91	72.54	73.51	60.58	87.52
	ElasticFace	95.92	93.91	71.25	73.51	60.58	87.02
	same (i.e., whitebox)	95.92	93.91	71.38	73.51	60.58	87.61
AgeDB	ArcFace	82.35	82.1	56.34	53.43	57.4	79.63
	ElasticFace	82.35	82.1	54.05	53.42	57.4	79.0
	same (i.e., whitebox)	82.35	82.1	55.21	53.42	57.4	79.41

protection schemes in blackbox attacks using ArcFace as  $F_{adv}$ . As the sample reconstructed face images show, inversion of protected templates can reveal important information about underlying subjects.

To further explore the effect of  $F_{adv}$ , we consider HE-protected templates and use ElasticFace as  $F_{adv}$ . In addition, we consider the whitebox scenario, where the adversary has access to the feature extractor of the face recognition model and uses it as  $F_{adv}$ . As the results in Table 5.7 show, using the same feature extractor (i.e., whitebox attack) or different feature extractor (i.e., ArcFace or ElasticFace in blackbox attacks), the reconstructed face images achieve very similar performances. Even in some cases, such as attacks against AttentionNet, we can see that the blackbox attack using ArcFace as  $F_{adv}$  achieves better performance than the whitebox attack. This observation can be interpreted considering the superior performance of ArcFace compared to other face recognition models used in our experiments, as reported in Table A.1 of Appendix A. Therefore, we can expect that ArcFace enhances the reconstruction when it used as  $F_{adv}$ .

### 5.2.2.3 Discussion

Our experiments show that if the template protection scheme and its secrets are known, then an adversary can reconstruct face images from protected facial templates. We considered different feature extractors protected with different template protection schemes and evaluated them on the MOBIO, LFW, and AgeDB datasets. However, the reconstructed face images from different feature extractors have different performances when comparing for the same protection scheme and the same dataset. For most cases, the model with higher recognition

## 5.2 Inversion of Protected Biometric Templates

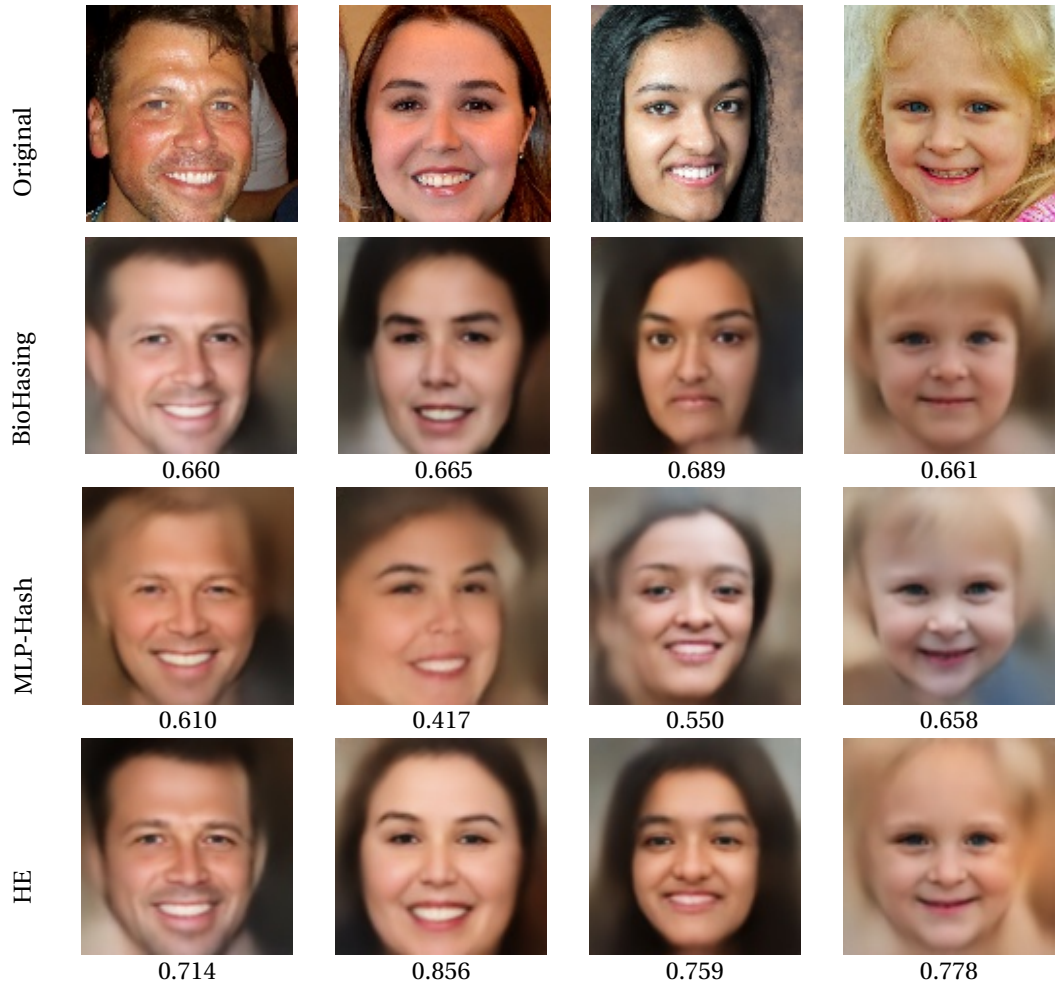


Figure 5.2: Sample face images from the FFHQ dataset (first row) and their corresponding reconstructed face images from ElasticFace templates protected with BioHashing (second row), MLP-Hash (third row), and Homomorphic Encryption (fourth row) in blackbox attacks. The values indicate cosine similarity between templates of the original and reconstructed face images. The decision threshold corresponding to  $\text{FMR} = 10^{-3}$  is 0.29 for ELasticFace on the LFW dataset.

performance in Table A.1 of Appendix A is more vulnerable to reconstruction attack. Similarly, if the performance of a model is better on a dataset in Table A.1 of Appendix A, the attack rates are higher on that dataset in Table 5.5 and Table 5.6. Comparing different template protection schemes, we can see that in most cases, protected templates with HE are the most invertible, but protected templates with MLP-Hash are more robust to inversion and lead to the lowest invertibility.

One of the limitations of our face reconstruction method is that the adversary needs to train a face reconstruction network for each set of secrets. In protected systems that have the same secrets for all subjects (such as in HE), the adversary needs only to train a single face reconstruction network. However, to attack a protected system with a user-specific key setting, the adversary needs to train different face reconstruction models for each leaked protected template. This limitation, however, can be resolved if the adversary applies optimization-based approaches, described in Section 2.1 of Chapter 2, that do not require gradients in their method.

### 5.3 Measuring Linkability of Protected Templates using Maximal Leakage

In this section, we propose a new measure for evaluating the linkability of protected biometric templates. Our proposed measure combines the work on maximal leakage from information-theoretic literature [126], [127] with the perspective on global linkability introduced in [117]. Since our proposed measure is based on a well-studied information measure, it inherits many of the theoretic properties of this measure. In addition, we show that the proposed linkability measure has an appealing operational interpretation in terms of hypothesis testing that the adversary could perform on a pair of protected templates. This hypothesis testing interpretation of our proposed measure makes it consistent with the definition of linkability in the ISO/IEC 30136 standard [10]. We further compare our proposed measure to a similar measure based on differential privacy [122] and show that the differential privacy-based measure is too strict for the linkability application. Finally, the experimental implementation of our proposed measure shows that it gives intuitively correct linkability scores across different BTP schemes, biometric characteristics, and scoring functions.

In Section 5.3.1, we define our proposed measure, as well as discuss its operational interpretations and its properties. In Section 5.3.2 we compare the proposed measure to two other linkability measures: the global measure introduced in [117] and a similar measure based on differential privacy. Finally, in Section 5.3.3, we evaluate the unlinkability of different biometric recognition systems based on different biometric characteristics and protected with different BTP schemes.

### 5.3.1 Maximal Linkability

We first introduce our notation and overview the maximal leakage information measure. Then, we define our measure of linkability as a maximal leakage of information about the mated and non-mated hypothesis, as well as review its properties. We end by interpreting the new measure in terms of statistical hypothesis testing.

#### 5.3.1.1 Notations and the Definition of Maximal Leakage

We use capital letters to denote random variables, calligraphic letters to denote support sets of these random variables (and sets in general), and lower case letters to denote realizations of these random variables. For example,  $X$  is a random variable taking values on  $\mathcal{X}$  while  $x \in \mathcal{X}$  is a possible realization of this random variable. We use the notation  $X \leftrightarrow Y \leftrightarrow Z$  to denote that  $X$ ,  $Y$ , and  $Z$  form a Markov chain. We use  $p_X$  to denote the probability mass function (if  $\mathcal{X}$  is discrete) or the probability density function (if  $\mathcal{X}$  is continuous) of  $X$ . If the associated random variable is clear from context, we omit the subscript: for example,  $p(y|s)$ . We use sanserif font to indicate functions, for example  $f: \mathcal{X} \rightarrow \mathcal{Y}$  denotes a function from  $\mathcal{X}$  to  $\mathcal{Y}$ . Finally, all the logarithms in this thesis will be assumed to have base two.

Maximal leakage is an information leakage measure introduced in [126], [127]. Specifically, [126] defined this measure as follows. Let  $X$  and  $Y$  be two jointly-distributed random variables, where  $X$  represents some secret information which may be of interest to an adversary, while  $Y$  represents the actual observations of an adversary. The maximal leakage of information from  $X$  to  $Y$  is defined as

$$\mathcal{L}(X \rightarrow Y) = \sup_{U \leftrightarrow X \leftrightarrow Y \leftrightarrow \hat{U}} \log \frac{\mathbb{P}(U = \hat{U})}{\max_{u \in \mathcal{U}} p_U(u)} \quad (5.8)$$

where  $U, \hat{U}$  are random variables over some common finite alphabet. The auxiliary random variable  $U$  in Eq. 5.8 denotes some, possibly random, mapping of secret information  $X$ , while  $\hat{U}$  denotes the best guess an adversary could make about  $U$ . Thus, the ratio  $\frac{\mathbb{P}(U = \hat{U})}{\max_{u \in \mathcal{U}} p_U(u)}$  captures how much an adversary's ability to guess any hidden mapping of data  $U$  improves by observing  $Y$ . The whole quantity in Eq. 5.8 measures multiplicative improvement of the adversary's ability to guess any possible function of the secret  $X$ .

Maximal leakage was independently introduced in [127] where it was defined as

$$\mathcal{L}(X \rightarrow Y) = \sup_{p_X} \log \frac{\mathbb{P}(X = \hat{X})}{\max_{x \in \mathcal{X}} p_X(x)} \quad (5.9)$$

where  $X \leftrightarrow Y \leftrightarrow \hat{X}$ . When  $X$  has full support, both definitions in Eq. 5.8) and Eq. 5.9 are equivalent [126].

Although it is not immediately clear that Eq. 5.8) and Eq. 5.9 are computable, it is shown in [126,

## Chapter 5. Evaluation of Biometric Template Protection Schemes

---

Theorem 1] that, for discrete  $(X, Y)$ , maximal leakage could be evaluated via the following simple formula

$$\mathcal{L}(X \rightarrow Y) = \log \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X} : p_X(x) > 0} p_{Y|X}(y|x). \quad (5.10)$$

This result could be extended to more general settings [126, Theorem 7]. For example, a setting that will be of interest to us is when  $\mathcal{Y}$  is continuous,  $\mathcal{X}$  is discrete, and the probability density functions  $p_{Y|X}(y|x)$  exist. In this case, the maximal leakage reduces to

$$\mathcal{L}(X \rightarrow Y) = \log \int_{\mathcal{Y}} \max_{x \in \mathcal{X}} p_{Y|X}(y|x) dy. \quad (5.11)$$

Finally, it is shown in [126] that

$$\mathcal{L}(X \rightarrow Y) = I_{\infty}(X; Y) \quad (5.12)$$

where  $I_{\infty}(X; Y)$  denotes the *Sibson's mutual information* of order infinity [216], [217]. In other words,  $\mathcal{L}(X \rightarrow Y)$  could be viewed as a generalization of Shannon's mutual information in the same way that Rényi entropy is a generalization of Shannon's entropy [218].

Because maximal leakage is a well-defined information measure, it has a number of mathematical properties. We highlight some of the most important properties here:

- First, maximal leakage is non-negative, that is

$$\mathcal{L}(X \rightarrow Y) \geq 0. \quad (5.13)$$

It is zero if and only if  $X$  and  $Y$  are statistically independent.

- Secondly, it satisfies the *data processing inequality* which states that

$$\mathcal{L}(X \rightarrow Z) \leq \mathcal{L}(X \rightarrow Y) \quad (5.14)$$

where  $X \leftrightarrow Y \leftrightarrow Z$  form a Markov chain.

- Finally, for a discrete random variable  $X$ ,

$$\mathcal{L}(X \rightarrow Y) \leq \log |\mathcal{X}|. \quad (5.15)$$

Proofs of these properties and additional properties of maximal leakage could be found in [126].



### 5.3.1.2 Maximal Linkability of Biometric Templates

The proposed linkability metric uses maximal leakage to measure the amount of information revealed by two templates about the two possible hypotheses: the templates are mated, and the templates are not mated. Specifically, given two biometric systems, let  $\mathcal{T}_1$  be the space of all possible protected templates that could be produced by the first system and  $\mathcal{T}_2$  be the space of all possible protected templates that could be produced by the second system. Given two templates  $(t_1, t_2) \in \mathcal{T}_1 \times \mathcal{T}_2$  we can define the following hypothesis:

$$\begin{aligned} h_m &= \{\text{templates } t_1 \text{ and } t_2 \text{ belong to mated instances}\} \\ h_{nm} &= \{\text{templates } t_1 \text{ and } t_2 \text{ belong to non-mated instances}\}. \end{aligned}$$

Moreover, let  $(T_1, T_2)$  be random variables each taking values on  $\mathcal{T}_1 \times \mathcal{T}_2$  and let  $H$  be a random variable taking values on  $\mathcal{H} = \{h_m, h_{nm}\}$ . In other words,  $H$  denotes the true hypotheses about templates  $T_1$  and  $T_2$ .

**Definition 1** (Maximal Linkability). Maximal linkability of two systems producing templates  $(T_1, T_2)$  is defined as

$$M_{\leftarrow}^{\text{sys}} = \mathcal{L}(H \rightarrow (T_1, T_2)) \quad (5.16)$$

$$= \log \sum_{(t_1, t_2) \in \mathcal{T}_1 \times \mathcal{T}_2} \max\{p(t_1, t_2|h_m), p(t_1, t_2|h_{nm})\}. \quad (5.17)$$

We can make two observations about maximal linkability in light of Eq. 5.17. First, since maximal linkability depends only on the conditional distributions  $p(t_1, t_2|h_m)$  and  $p(t_1, t_2|h_{nm})$ , it is independent of the distribution of the hypothesis  $H$ . This is a desirable property for a linkability measure since it means that  $M_{\leftarrow}^{\text{sys}}$  depends on the BTP scheme itself, and not on any assumptions on the distributions of mated and non-mated pairs of templates.

Secondly, from an information-theoretic perspective, it is important to define  $M_{\leftarrow}^{\text{sys}}$  as we do in Definition 1. This measure is the ‘true’ linkability score of the system. That is, as we will see in Lemma 3, this score gives us the most general guarantees with fewest assumptions on the behaviour of the adversary. However, to compute  $M_{\leftarrow}^{\text{sys}}$ , it is necessary to know  $p(t_1, t_2|h_m)$  and  $p(t_1, t_2|h_{nm})$  for all possible values of  $(t_1, t_2) \in \mathcal{T}_1 \times \mathcal{T}_2$ . This means that if  $M_{\leftarrow}^{\text{sys}}$  is to be estimated from data, we need to generate a number of samples on the order of  $|\mathcal{T}_1||\mathcal{T}_2|$  and this is prohibitive in most practical settings. In other words, estimating  $M_{\leftarrow}^{\text{sys}}$  accurately is an intractable problem as is shown in the following lemma:

**Lemma 1.** Let  $(T_1^n, T_2^n)$  be  $n$  i.i.d. samples of protected template pairs and let  $\hat{M}(T_1^n, T_2^n)$  be any estimate of maximal linkability that could be made from these samples.

Define,

$$S_{\delta, \epsilon} = \min\{n: \mathbb{P}(|M_{\leftarrow}^{\text{sys}} - \hat{M}(T_1^n, T_2^n)|) > \delta\} < \epsilon \quad (5.18)$$

for all distributions of  $H$  and  $(T_1, T_2)$  on  $\mathcal{H} \times \mathcal{T}_1 \times \mathcal{T}_2$ .

## Chapter 5. Evaluation of Biometric Template Protection Schemes

---

Fix  $\epsilon \leq 0.1$  and  $\delta < \frac{1}{2}$ . Then, for some constant  $c$ ,

$$S_{\delta, \epsilon} \geq c \frac{N}{\log N} \log^2 \frac{1}{\delta} \quad (5.19)$$

where  $N = |\mathcal{T}_1 \times \mathcal{T}_2|$ .

*Proof.* This is application of [126, Theorem 10] for maximal leakage to maximal linkability. The constant  $c$  is given in [219, Theorem 2].  $\square$

In other words, if we do not make any assumptions about the distribution of templates under the mated and non-mated hypothesis, estimating linkability scales with the size of the template pair space. The main issue is that the space  $\mathcal{T}_1 \times \mathcal{T}_2$  is very large. For example, for ArcFace with BioHash-protected templates,  $N = 2^{1024}$ . According to Theorem 1, to estimate maximal linkability with accuracy  $\delta = 0.1$  and probability of error  $\epsilon = 0.1$ , approximately  $20 \cdot 2^{1014}$  samples would be needed. Likewise, for Finger vein with BioHash-protected templates  $N = 2^{384}$ . According to Theorem 1, to estimate maximal linkability with accuracy  $\delta = 0.1$  and probability of error  $\epsilon = 0.1$ , approximately  $20 \cdot 2^{380}$  samples would be needed.

We remark that the fundamental issue with estimating  $M_{\leftrightarrow}^{\text{sys}}$  will arise with any measure of privacy that is sufficiently strong. For example, estimating differential privacy<sup>4</sup> will behave even worse than the present approach. This is because any strong measure would look at the worst case elements in some way. This would inevitably require exploring the whole template space to approximate the measure well.

### 5.3.1.3 On Estimating System Linkability

To circumvent the issue with estimating the probability distributions, we follow [117] and propose a linkability measure based on a similarity function. That is, we assume that there is a similarity function

$$s: \mathcal{T}_1 \times \mathcal{T}_2 \rightarrow \mathcal{S} \quad (5.20)$$

which captures the relevant information about the similarity of the two templates. This similarity function could then be used to approximate the linkability score proposed in Definition 1. To this end, we define another linkability measure with respect to a fixed similarity function.

**Definition 2** (Maximal  $s$ -Linkability). *Let  $S = s(T_1, T_2)$  be a similarity score for templates  $T_1$  and  $T_2$ , and a similarity function  $s$ . Maximal  $s$ -linkability of two systems producing templates  $(T_1, T_2)$  is defined as*

$$M_{\leftrightarrow}^s = \mathcal{L}(H \rightarrow S). \quad (5.21)$$

---

<sup>4</sup>We discuss more about differential privacy for measuring linkability of protected templates in Section 5.3.2.

### 5.3 Measuring Linkability of Protected Templates using Maximal Leakage

Then, for discrete  $S$ ,

$$M_{\leftrightarrow}^s = \log \sum_{s \in \mathcal{S}} \max\{p(s|h_m), p(s|h_{nm})\}, \quad (5.22)$$

and for continuous  $S$ ,

$$M_{\leftrightarrow}^s = \log \int_{\mathcal{S}} \max\{p(s|h_m), p(s|h_{nm})\} ds. \quad (5.23)$$

Maximal  $s$ -linkability generalizes maximal linkability in the following sense. It measures the amount of information revealed by the similarity score  $S$  about the two possible hypotheses: the templates are mated, and the templates are not mated. If  $s$  is taken to be the identity function, maximal  $s$ -linkability reduces to maximal linkability. Thus, just like in [117], the linkability of the system should be evaluated for several similarity functions and the worst-case score should be considered.

**Lemma 2.** *Let  $s$  be any similarity function on  $\mathcal{T}_1 \times \mathcal{T}_2$ . Then*

$$0 \leq M_{\leftrightarrow}^s \leq M_{\leftrightarrow}^{sys} \leq 1. \quad (5.24)$$

*Proof.* Eq. 5.24 follows from Eq. 5.13, 5.14, and 5.15. Specifically, the first inequality follows from Definition 2 and from Eq. 5.13. In other words, since  $M_{\leftrightarrow}^s$  is an information measure, it cannot be negative. The second inequality follows from the data processing inequality (i.e., Eq. 5.14) since we have a Markov chain  $H \leftrightarrow (T_1, T_2) \leftrightarrow S$ . Finally, the last inequality follows from Definition 1 and Eq. 5.15 since  $H$  is a binary-valued random variable.  $\square$

Just like the linkability measure proposed in [117], our measure is supported on  $[0, 1]$ . If  $M_{\leftrightarrow}^{sys} = 0$  then the system is completely unlinkable. That is, templates  $T_1$  and  $T_2$  reveal nothing about the hypothesis  $h_m$  and  $h_{nm}$ . On other hand,  $M_{\leftrightarrow}^s = 1$  means that the system is completely linkable and the adversary could always determine the correct hypothesis after observing  $T_1$  and  $T_2$ .

#### 5.3.1.4 Maximal Linkability and Hypothesis Testing

In this section, we interpret  $M_{\leftrightarrow}^{sys}$  and  $M_{\leftrightarrow}^s$  in terms of Neyman-Pearson hypothesis testing. Recall that in this framework, the goal is to design a hypothesis test based on the available data while trading-off two types of errors: *false alarm* error and *missed detection* error. In the present case, the adversary's goal is to distinguish between two hypotheses  $\{h_m, h_{nm}\}$ , while keeping the two errors small. In the biometrics literature, the false alarm error is also known as *false match rate* (FMR), while the missed detection error is also known as the *false non-match rate* (FNMR). The maximal linkability metrics provide impossibility bounds on the adversary's ability to design well-performing hypothesis tests. If an adversary has access to the protected templates  $(T_1, T_2)$ , the relevant bound is derived in terms of  $M_{\leftrightarrow}^{sys}$ . On the other hand, if an

## Chapter 5. Evaluation of Biometric Template Protection Schemes

---

adversary has access to similarity score  $S = s(T_1, T_2)$  only, the relevant bound is derived in terms of  $M_{\leftarrow}^s$ . These impossibility bounds are formalized in the following lemmas.

**Lemma 3.** *Suppose  $\hat{H}$  is a decision rule for the hypothesis  $H$  based on observing  $(T_1, T_2)$  and taking values on  $\{h_m, h_{nm}\}$ . In other words,  $H \leftrightarrow (T_1, T_2) \leftrightarrow \hat{H}$ . Let*

$$\begin{aligned} FMR &= \mathbb{P} [\hat{H} = h_m | H = h_{nm}] \\ \text{and } FNMR &= \mathbb{P} [\hat{H} = h_{nm} | H = h_m] \end{aligned}$$

*be the False Match and False Non-match Rates for this decision rule. Let  $M_{\leftarrow}^{\text{sys}}$  be the maximal linkability score of the system. Then*

$$(1 - FMR) + (1 - FNMR) \leq 2^{M_{\leftarrow}^{\text{sys}}}. \quad (5.25)$$

The proof of Lemma 3 is given in the Appendix D. The Proof of the following Lemma 4 is identical to the proof of Lemma 3 with the key difference being that the adversary's hypothesis testing is assumed to be done on the similarity score  $S$  and not on the protected templates  $(T_1, T_2)$ .

**Lemma 4.** *Suppose  $\hat{H}$  is a decision rule for the hypothesis  $H$  based on observing  $S = s(T_1, T_2)$  and taking values on  $\{h_m, h_{nm}\}$ . In other words,  $H \leftrightarrow S \leftrightarrow \hat{H}$ . Let*

$$\begin{aligned} FMR &= \mathbb{P} [\hat{H} = h_m | H = h_{nm}] \\ \text{and } FNMR &= \mathbb{P} [\hat{H} = h_{nm} | H = h_m] \end{aligned}$$

*be the False Match and False Non-match Rates for this decision rule. Let  $M_{\leftarrow}^s$  be the maximal s-linkability score of the system. Then*

$$(1 - FMR) + (1 - FNMR) \leq 2^{M_{\leftarrow}^s}. \quad (5.26)$$

We see from Lemma 3 that a low value of  $M_{\leftarrow}^{\text{sys}}$  guarantees that an adversary cannot perform any meaningful hypothesis testing on observed templates  $T_1$  and  $T_2$  to decide if they are mated or non-mated. Likewise, we see from Lemma 4 that a low value of  $M_{\leftarrow}^s$  guarantees that an adversary cannot perform any meaningful hypothesis testing on an observed similarity score  $S$  to decide if it comes from mated or non-mated templates. These results give an operational interpretation to  $M_{\leftarrow}^{\text{sys}}$  and  $M_{\leftarrow}^s$  an addition to those already provided in [126], see Figure 5.3.

Figure 5.4 further illustrates different examples of synthetic scores with Gaussian distributions, and the corresponding ROC curves. For almost overlapping distributions (e.g., Figure 5.4a) our measure returns a low value (i.e, near zero), while for distributions with less overlap (e.g., Figure 5.4d) our measure returns a higher value. In addition, we see in all four cases that our measure provides a good upper bound on the true ROC curve of an optimal hypothesis test performed by the adversary.

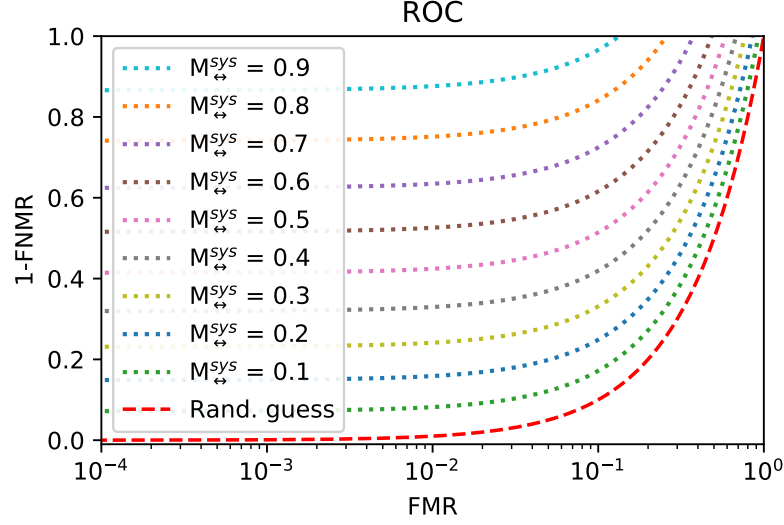


Figure 5.3: Bounds on adversary's ability to perform hypothesis testing for different maximal linkability scores. For example, for  $M_{\leftrightarrow}^{sys} = 0.1$ , a ROC curve for any hypothesis test that could be performed by an adversary on  $(T_1, T_2)$  will be between the dashed red (random guess) and the dotted green ( $M_{\leftrightarrow}^{sys} = 0.1$ ) curves.

### 5.3.2 Comparison With Other Measures

In this section, we compare the proposed measure to other approaches to measuring linkability. In Section 5.3.2.1, we discuss the implications of using differential privacy as an information measure in the definition of linkability. In Section 5.3.2.2, we compare our proposed measure to the one from [117], as the most relevant linkability measure in the literature for protected biometric templates.

#### 5.3.2.1 On Linkability via Differential Privacy

The main insight behind the proposed linkability measure is to measure the amount of information leaked by a pair protected biometric templates about whether these templates are mated or not mated. Definitions 1 and 2 use maximal leakage as a measure of such information leakage. This raises the question of whether other measures of privacy loss could be used instead of maximal leakage. In this section, we consider the most prominent such measure: differential privacy [122].

We will show that for  $\epsilon$ -differential privacy the resulting linkability measure does not differentiate between the four distinct examples in Figure 5.4. That is, it assigns the value of infinity to all four examples and classifies all four systems as completely linkable. Another possible approach is to apply a common relaxation of  $\epsilon$ -differential privacy known as  $(\epsilon, \delta)$ -differential privacy. We will show as well, from the example of Figure 5.4, that this approach does not provide us with a single linkability measurement. Instead, it provides us with a curve trading

## Chapter 5. Evaluation of Biometric Template Protection Schemes

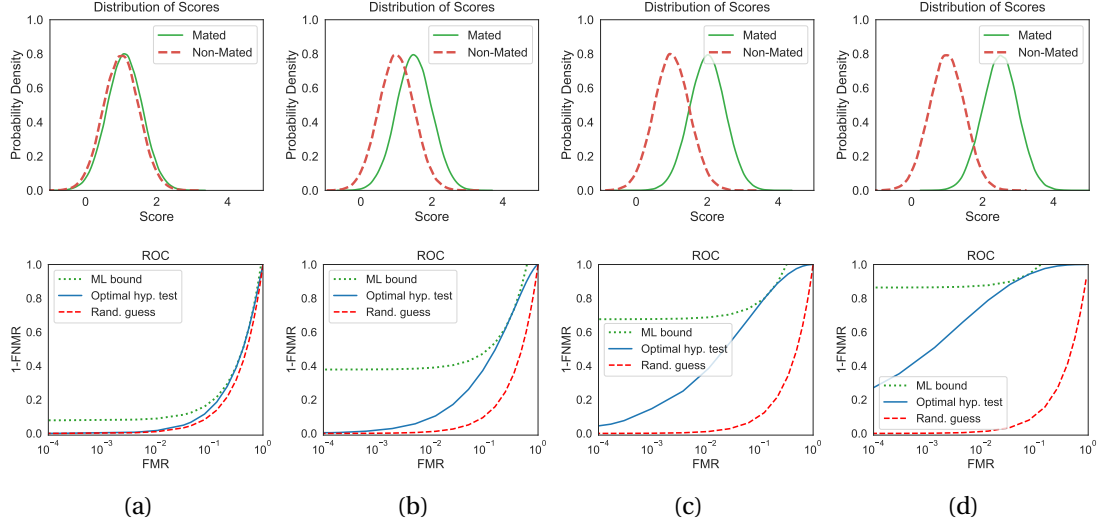


Figure 5.4: Synthetic distributions of mated and non-mated scores (first row) and their corresponding ROC plots (second row): (a) mated:  $\mathcal{N}(1.1, 0.5)$ , non-mated:  $\mathcal{N}(1, 0.5)$ , and  $M_{\leftarrow}^S = 0.1077$ , (b) mated:  $\mathcal{N}(1.5, 0.5)$ , non-mated:  $\mathcal{N}(1, 0.5)$ , and  $M_{\leftarrow}^S = 0.4626$ , (c) mated:  $\mathcal{N}(2.0, 0.5)$ , non-mated:  $\mathcal{N}(1, 0.5)$ , and  $M_{\leftarrow}^S = 0.7450$ , (d) mated:  $\mathcal{N}(2.5, 0.5)$ , non-mated:  $\mathcal{N}(1, 0.5)$ , and  $M_{\leftarrow}^S = 0.8980$ . In the ROC plots, the green dotted curves indicate the maximal likability bound for the adversary hypothesis test, the solid blue curves show the optimal possible hypothesis test by the adversary, and the dashed red curves depict the random guess accuracy.

off between the  $\epsilon$  and the  $\delta$  privacy parameters.

**$\epsilon$ -Differential Privacy** Differential privacy is the most prominent approach to privacy that was designed for a private data release problem [122]. In this discussion, we view  $\epsilon$ -differential privacy as an information measure between our true hypothesis  $H$  and an observed template pair  $T_1, T_2$ , and apply it in the manner similar to Definition 1. In other words, we seek to measure how differentially private the mapping  $H$  to  $(T_1, T_2)$  is. In this way, we can define a new measure of linkability:

$$\mathcal{DP}(H \rightarrow (T_1, T_2)) = \max_{\substack{(t_1, t_2) \in \mathcal{T}_1 \times \mathcal{T}_2, \\ h, \hat{h} \in \{h_m, h_{nm}\}}} \log \frac{p(t_1, t_2 | h)}{p(t_1, t_2 | \hat{h})}. \quad (5.27)$$

Likewise, for a given similarity function  $s$  with continuous scores  $S$ , we can define a measure of linkability:

$$\mathcal{DP}(H \rightarrow S) = \sup_{\substack{s \in \mathcal{S}, \\ h, \hat{h} \in \{h_m, h_{nm}\}}} \log \frac{f(s | h)}{f(s | \hat{h})}. \quad (5.28)$$

where  $f(s | h)$  denotes the probability density function of  $S$  given  $h \in \{h_m, h_{nm}\}$ .

### 5.3 Measuring Linkability of Protected Templates using Maximal Leakage

As it turns out, these definitions do not distinguish between any of the cases in Table 5.8 and instead classify all of them as fully linkable. In other words,  $\epsilon$ -differential privacy is too pessimistic for the linkability application. For example, let the score distribution of mated and non-mated templates be any of the four normally distributed pairs in Table 5.8. Then

$$\mathcal{DP}(H \rightarrow S) = \infty. \quad (5.29)$$

This is because the four synthetic distributions in Table 5.8 are all examples of a Gaussian additive mechanism applied to a database  $\{h_m, h_{nm}\}$ . These do not satisfy  $\epsilon$ -DP according to [122, Theorem A.1]. To be more precise, we can take  $f: \{h_m, h_{nm}\} \rightarrow \{\mu_m, \mu_{nm}\}$  where, for example,  $\mu_m = 1.1$  and  $\mu_{nm} = 1$  as in Figure 5.4a. Setting  $\delta = 0$  in [122, Theorem A.1] we see that  $\epsilon = \infty$ .

**$(\epsilon, \delta)$ -Differential Privacy**  $(\epsilon, \delta)$ -Differential privacy is a well-studied relaxation of differential privacy which introduces a second parameter  $\delta$ . We could also consider treating this as an information measure between our true hypothesis  $H$  and an observed template pair  $(T_1, T_2)$ , and apply it in the manner similar to Definition 1. Or, we could consider treating this as an information measure between our true hypothesis  $H$  and an observed similarity score  $S$ , and apply it in the manner similar to Definition 2. However, in both of these cases we would need to estimate two parameters:  $\epsilon$  and  $\delta$ . In general, a BTP scheme will not satisfy  $(\epsilon, \delta)$ -differential privacy for a single pair  $(\epsilon, \delta)$ , but would instead satisfy it for an  $(\epsilon, \delta)$  curve.

As an example, take the score distribution of mated and non-mated templates be normally distributed  $\mathcal{N}(1.1, 0.5)$  and  $\mathcal{N}(1, 0.5)$  as in Figure 5.4a. Let  $c \in [0, \infty]$  be any non-negative constant. Then, mapping from  $H$  to  $S$  induced by the BTP scheme satisfies  $(\epsilon, \delta)$ -differential privacy with

$$\epsilon > \frac{0.1c}{\sqrt{0.5}} \quad \text{and} \quad \delta > 1.25e^{-0.5c^2}. \quad (5.30)$$

This is again an examples of a Gaussian additive mechanism applied to a database  $\{h_m, h_{nm}\}$  where we take  $f: \{h_m, h_{nm}\} \rightarrow \{\mu_m, \mu_{nm}\}$  with  $\mu_m = 1.1$  and  $\mu_{nm} = 1$  as in Figure 5.4a. Applying [122, Theorem A.1] with  $\Delta_1 f = 0.1$  and  $\sigma = \sqrt{0.5}$  we obtain lower bounds on  $\epsilon$  and  $\delta$  in terms of  $c \in [0, \infty]$ .

As we see from the above discussion, differential privacy does not appear to be an appropriate information measure for the linkability application. In the case of  $\epsilon$ -differential privacy, it does not differentiate between the simple synthetic examples in Table 5.8 and labels all of them completely linkable. On the other hand, in the case of  $(\epsilon, \delta)$ -differential privacy, it is not clear how to obtain a single linkability score.

### 5.3.2.2 Comparison with Gomez-Barrero *et al.* Measure

Recall that the first quantitative measure of linkage was introduced in [117]. The main idea of [117] is to base the measure on the distributions of mated and non-mated hypotheses conditioned on a similarity score.

**Overview of Gomez-Barrero *et al.* Measure** As mentioned in Section 2.3.2 of Chapter 2, Gomez-Barrero *et al.* [117] proposed two quantitative measures (local and global) based on score distributions. They considered a similarity function  $s$  to find the score  $s = s(t_1, t_2) \in \mathcal{S}$  between two templates  $t_1$  and  $t_2$ , and found distributions of mated and non-mated pairs. Next, they defined their local measure for each score  $s$  in [117, Eq. 4] as:

$$D_{\leftrightarrow}(s) = p(h_m|s) - p(h_{nm}|s). \quad (5.31)$$

With some assumptions and simplification, they define their local unlinkability measure in [117, Eq. 14] as:

$$D_{\leftrightarrow}(s) = \begin{cases} 0 & \text{if } LR(s) \cdot \omega \leq 1 \\ 2 \frac{LR(s) \cdot \omega}{1 + LR(s) \cdot \omega} - 1 & \text{if } LR(s) \cdot \omega > 1 \end{cases}, \quad (5.32)$$

where  $LR(s) = p(s|h_m)/p(s|h_{nm})$  is the likelihood ratio and  $\omega = p(h_m)/p(h_{nm})$  denotes the ratio between the prior probabilities of the mated and non-mated samples. The value of  $\omega = 1$ , i.e.  $p(h_m) = p(h_{nm})$ , is proposed as the worst-case scenario. Finally, the global measure  $D_{\leftrightarrow}^{sys}$  is found by calculating the conditional expectation of the local measure  $D_{\leftrightarrow}(s)$  over all comparison scores in [117, Eq. 19] as:

$$D_{\leftrightarrow}^{sys} = \int p(s|H_m) D_{\leftrightarrow}(s) ds. \quad (5.33)$$

The global measure  $D_{\leftrightarrow}^{sys}$  was the first quantitative evaluation that measures the degree of unlinkability of the biometric systems. In addition to the mathematical definition of  $D_{\leftrightarrow}^{sys}$ , [117, Section V] proposes a general protocol for evaluating linkability from data.

**Comparison with Maximal Linkability** Both  $D_{\leftrightarrow}^{sys}$  (as in Eq. 5.33) and  $M_{\leftrightarrow}^s$  are based on the similarity score of biometric templates. As discussed in Section 5.3.1, the true linkability of the system is given by  $M_{\leftrightarrow}^{sys}$ . However, as this is computationally infeasible in most real-world biometric systems, we follow [117] and focus on computing  $M_{\leftrightarrow}^s$  as proxies for the true linkability. Just like in [117], it is thus important to compute  $M_{\leftrightarrow}^s$  for a number of different similarity scores. In addition, maximal linkabilities  $M_{\leftrightarrow}^{sys}$  and  $M_{\leftrightarrow}^s$  as well as  $D_{\leftrightarrow}^{sys}$ , are bounded in  $[0, 1]$ , where 0 indicates full unlinkability and 1 indicates fully linkability. However, maximal linkability and  $D_{\leftrightarrow}^{sys}$  do have a number of significant differences which are highlighted next.

First, while the values of both measures are bounded in  $[0, 1]$ , the value of maximal linkability is always higher. This result is formalized in the following lemma.



### 5.3 Measuring Linkability of Protected Templates using Maximal Leakage

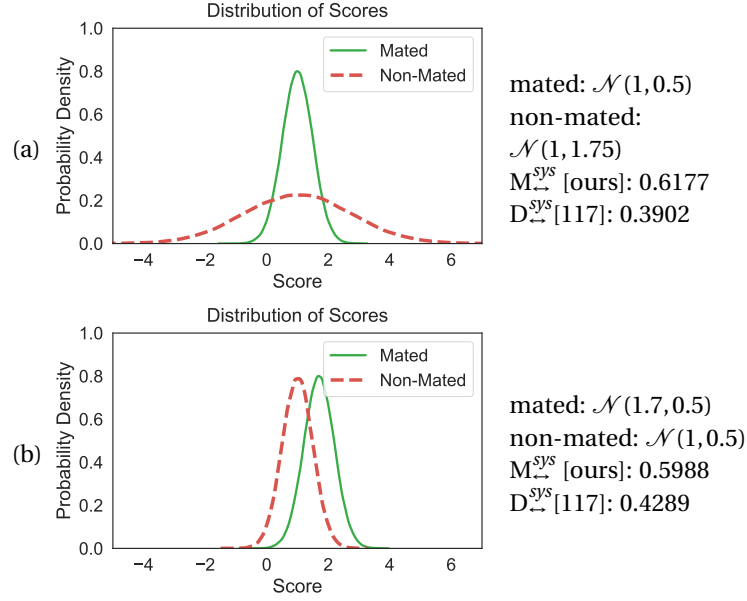


Figure 5.5: Distributions of scores for mated and non-mated templates: (a) mated:  $\mathcal{N}(1, 0.5)$ , non-mated:  $\mathcal{N}(1, 1.75)$ , with linkability value of 0.6177 (i.e., somewhat linkable) by our measure (i.e,  $M_{\leftarrow}^{\text{sys}}$  as in Eq. 5.17) and 0.3902 (i.e., somewhat unlinkable) by the measure in [117] (i.e,  $D_{\leftarrow}^{\text{sys}}$  as in Eq. 5.33). (b) mated:  $\mathcal{N}(1.7, 0.5)$ , non-mated:  $\mathcal{N}(1, 0.5)$ , with linkability value of 0.5988 (i.e., somewhat linkable) by our measure (i.e,  $M_{\leftarrow}^{\text{sys}}$  as in Eq. 5.17) and 0.4289 (i.e., somewhat unlinkable) by the measure in [117] (i.e,  $D_{\leftarrow}^{\text{sys}}$  as in Eq. 5.33). Note that these two systems are ranked differently by our measure and the one in [117].

**Lemma 5.** Assume that  $D_{\leftarrow}^{\text{sys}}$  is computed using similarity function  $s$  and  $\omega \leq 1$ . Then

$$0 \leq D_{\leftarrow}^{\text{sys}} \leq M_{\leftarrow}^{\text{s}} \leq M_{\leftarrow}^{\text{sys}} \leq 1. \quad (5.34)$$

The proof for  $D_{\leftarrow}^{\text{sys}} \leq M_{\leftarrow}^{\text{s}}$  is given in Appendix D, while the other inequalities follow from Lemma 2 and [117]. We highlight that even though  $M_{\leftarrow}^{\text{s}}$  is always higher than  $D_{\leftarrow}^{\text{sys}}$ , it is possible for the two measures to give different rankings to two biometric systems. As an example, consider distributions of scores for mated and non-mated pairs as depicted in Figure 5.5. In this example, the linkability of mated and non-mated templates is 0.6177 by our measure (i.e,  $M_{\leftarrow}^{\text{sys}}$  as in Eq. 5.17) and 0.3902 by the measure in [117] (i.e,  $D_{\leftarrow}^{\text{sys}}$  as in Eq. 5.33) for system (a). For system (b), the linkability of mated and non-mated templates is 0.5988 by our measure and 0.4289 by the measure in [117].

Secondly, according to Lemmas 3 and 4, maximal linkability has a clear operational interpretation in terms of hypothesis testing capabilities of an adversary. This makes it consistent with the definition of unlinkability in the ISO/IEC 30136 standard [10] presented in Section 2.3.2 of Chapter 2. The measure  $D_{\leftarrow}^{\text{sys}}$  does not appear to have such a hypothesis testing interpretation. Considering again the example in Figure 5.5, we see that from the hypothesis testing

## Chapter 5. Evaluation of Biometric Template Protection Schemes

Table 5.8: Linkability of synthetic distributions of scores for mated and non-mated templates in Figure 5.4 using the measure in [117] (i.e,  $D_{\leftarrow}^{\text{sys}}$  as in Eq. 5.33) with different values of  $\omega$  and our measure (i.e,  $M_{\leftarrow}^{\text{sys}}$  as in Eq. 5.17).

Figure	Mated	Non-Mated	[117] measure			Proposed measure
			$\omega = 0.1$	$\omega = 1$	$\omega = 10$	
Fig. 5.4a	$\mathcal{N}(1.1, 0.5)$	$\mathcal{N}(1, 0.5)$	0	0.0434	0.8188	0.1077
Fig. 5.4b	$\mathcal{N}(1.5, 0.5)$	$\mathcal{N}(1, 0.5)$	0.0063	0.2890	0.8357	0.4626
Fig. 5.4c	$\mathcal{N}(2.0, 0.5)$	$\mathcal{N}(1, 0.5)$	0.2265	0.6111	0.8941	0.7450
Fig. 5.4d	$\mathcal{N}(2.5, 0.5)$	$\mathcal{N}(1, 0.5)$	0.6027	0.8310	0.9507	0.8980

perspective of Lemmas 3 and 4 it is correct to label system (a) as more linkable than system (b). The rationale for labeling system (b) as more linkable than system (a) (as is done by  $D_{\leftarrow}^{\text{sys}}$ ) is less apparent. In addition, unlike maximal linkability,  $D_{\leftarrow}^{\text{sys}}$  has a built-in asymmetry where it prioritizes the linkability of mated templates in its definition. While according to the definition of unlinkability in the ISO/IEC 30136 standard [10] given in Section 2.3.2, a linkability measure should take into account the difficulty of arriving at both, mated and non-mated, hypotheses. From an information-theoretic perspective, understanding that two templates are non-mated could also leak information to the adversary and should not be overlooked by a linkability measure. A closely related issue is that, to prevent (5.32) from being negative, it is rounded up to zero in certain cases. This rounding again leads to a similar loss of information.

A third difference is that maximal linkability appears to be numerically more stable. For example, to estimate  $M_{\leftarrow}^{\text{sys}}$  we simply need to estimate the area under the curve of the maximum of mated and non-mated probability density function as in Eq. 5.23. On the other hand, to calculate  $D_{\leftarrow}(s)$  in Eq. 5.32, it is necessary to estimate the likelihood ratio  $LR(s) = p(s|h_m)/p(s|h_{nm})$ , which is numerically unstable for low values of  $p(s|h_{nm})$ . In addition, for estimating  $LR(s)$  in practical evaluation in the case of  $p(s|h_{nm}) = 0$ , the authors considered  $LR(s) = 1$  in their open-source implementation<sup>5</sup> which is theoretically incorrect.

Finally, maximal linkability is independent of the prior probabilities of mated and non-mated hypotheses. By contrast,  $D_{\leftarrow}^{\text{sys}}$  requires the ratio of prior probabilities of the mated and non-mated samples ( $\omega$ ). We further discuss the effect of this assumption in Section 5.3.2.2.

**Different Values of  $\omega$**  As mentioned in Section 5.3.2.2, the measure in [117] requires the ratio of prior probabilities of the mated and non-mated samples (i.e.,  $\omega = p(H_m)/p(H_{nm})$ ). If we vary the value of  $\omega$  in this measure, we get counter intuitive results. For small values of  $\omega$ , clearly linkable systems are characterized as unlinkable. On the other hand, for large values of  $\omega$ , clearly unlinkable systems are characterized as linkable. Table 5.8 reports the linkability measurement of synthesized distributions in Figure 5.4 using the measure in [117] with different values of  $\omega$  and our measure. As this table shows, while our linkability measure is independent of prior probabilities, the linkability measure  $D_{\leftarrow}^{\text{sys}}$  is sensitive to the value  $\omega$

<sup>5</sup>Available at <https://github.com/dasec/unlinkability-metric>

and thus depends on the prior distributions of mated and non-mated template pairs. This may be an issue for two reasons. First, estimating this prior probability could, in general, be hard. While the authors in [117] considered  $\omega = 1$  as the worst-case scenario, such an assumption is not necessarily realistic in many practical cases. In particular, the adversary might have some knowledge about the prior probabilities. For instance, in many practical cases, it is reasonable to assume that non-mated pairs have a higher probability than mated pairs. Secondly, a linkability measure should depend on the BTP scheme and not on the prior belief about the distribution of the hypothesis. Arguably, it makes sense to consider measures that do not depend on the prior probability of  $H$ .

#### 5.3.3 Experiments

In this section, we describe the experimental results of evaluating the linkability of protected biometric templates using the proposed measure. First, we describe our experimental setup. Next, we analyze the numerical results of linkability measurement for different BTP schemes, different scoring functions, different characteristics, different feature extractors, and also examples of linkable templates. Finally, we discuss our experimental results.

##### 5.3.3.1 Experimental Setup

In our experiments, we evaluate the linkability of different BTP schemes on different characteristics (face, voice, and finger vein). We also considered DNN-based (face and voice) and hand-crafted (finger vein) feature extractors in our experiments.

**BTP schemes** We measure the linkability of biometric templates, which are protected using different BTP schemes, including BioHashing [76], Multi-Layer Perceptron (MLP) Hashing[23], Bloom Filters[78], two methods based on Index-of-Maximum (IoM) Hashing [77] (i.e., Gaussian random projection-based hashing, shortly GRP, and uniformly random permutation-based hashing, shortly URP), and Homomorphic Encryption (HE) based on Brakerski/Fan-Vercauteren (BFV) [98] algorithm. Table 5.9 summarizes the list of BTP schemes and compares their outputs and corresponding scoring functions.

**Biometric Characteristics** In our experiments, we use different biometric characteristics, including face, voice, and finger vein. We build different biometric recognition systems based on the aforementioned characteristics as follows. Table 5.10 summarises different biometric recognition systems used in our experiments.

## Chapter 5. Evaluation of Biometric Template Protection Schemes

Table 5.9: Summary of BTP schemes used in our linkability experiments.

BTP scheme	output	score function
BioHashing [76]	binary	Hamming distance*
MLP-Hash [23]	binary	Hamming distance*
Bloom Filters [78]	binary	Normalized <sup>†</sup> Hamming distance*
IoM-GRP [77]	integer	number of collisions
IoM-URP [77]	integer	number of collisions
HE [98]	ciphertext	Euclidean distance* (in ciphertext)

\*To have similarity values, distance functions are multiplied by -1.

<sup>†</sup>Normalized by the total number of ones in two templates.

Table 5.10: Summary of biometric recognition systems used in our linkability experiments.

Characteristic	Feature Extractor	Feature Type	Dataset	# Subjects	# Sessions	# Mated	# Non-mated
Face	ArcFace, ElasticFace, FaceNet	DNN-based	MOBIO (face)	150	12	1,516,300	2,235,000
Voice	ECAPA-TDNN	DNN-based	MOBIO (voice)	150	12	1,516,300	2,235,000
Finger Vein	WLD	Hand-crafted	UTFVP	360	2	216,000	2,067,840

### 5.3.3.2 Analyze

In this section, we describe our experiments on different biometric recognition systems. We evaluate the linkability of protected templates with different BTP schemes (in Section 5.3.3.2), based on different scoring functions (in Section 5.3.3.2), across different characteristics (in Section 5.3.3.2), and from different feature extractors (in Section 5.3.3.2). In each experiment, we try to fix all biometrics modules, except only one module<sup>6</sup>. We also evaluate the linkability of exemplary linkable templates in Section 5.3.3.2, including linkable protected templates (Section 5.3.3.2) and linkable unprotected templates (Section 5.3.3.2).

**Linkability measurement of different BTP schemes** In this experiment, we consider the features extracted from face images using the ArcFace model, and apply different BTP schemes, including BioHashing, MLP-Hashing, Bloom Filters, IoM-GRP, IoM-URP, and HE. Table 5.11 reports the linkability measurement of protected templates using the measure in [117] and our proposed measure. As this table shows, protected templates by these BTP schemes are almost unlinkable. This table also compares the rank of each BTP scheme compared to other schemes in terms of unlinkability by both measures (ranks are reported in parentheses). As this table shows, both methods rank these schemes the same in terms of the unlinkability of protected templates. However, the values of the measure [117] do not have any interpretation, and it is not clear how significant is the difference in unlinkability of these BTP schemes based on their

<sup>6</sup>We consider BioHash-protected templates in our experiments in Sections 5.3.3.2-5.3.3.2, since BioHashing is the simplest BTP scheme in Table 5.9. Similarly, we use face templates in our experiments in Sections 5.3.3.2, 5.3.3.2, 5.3.3.2, and 5.3.3.2 since face is one of the most popular biometric characteristics. However, we should note that similar experiments with other BTP schemes and other biometric characteristics can be implemented using our open-source paper package mentioned in Appendix E.

### 5.3 Measuring Linkability of Protected Templates using Maximal Leakage

Table 5.11: Linkability of different BTP schemes for ArcFace templates (values in the parentheses indicate the rank of the corresponding BTP scheme compared to other schemes).

BTP scheme	[117] measure	Proposed measure
BioHashing	0.0058 (6)	0.0162 (6)
MLP-Hash	0.0034 (5)	0.0096 (5)
Bloom Filters	0.0002 (1)	0.0007 (1)
IoM-GRP	0.0009 (3)	0.0027 (3)
IoM-URP	0.0008 (2)	0.0022 (2)
HE	0.0018 (4)	0.0053 (4)

unlinkability values by measure [117]. Whereas the values of our measure can be interpreted by Lemma 4 by providing an upper bound given the unlinkability value which guarantees that the adversary cannot perform any better hypothesis test than that upper bound. Therefore, each of these BTP schemes leads to a different upper bound for the accuracy of the adversary’s hypothesis testing (similar to the upper bounds illustrated in the ROC plots of Figure 5.3 and Figure 5.4).

**Linkability measurement with different scoring functions** Recall that our proposed measure and the one proposed in [117] are both based on score distributions of mated and non-mated templates. Therefore, as also discussed in Section 5.3.1, different scoring functions can provide different levels of linkability for protected templates. To evaluate the effect of the scoring function, in this experiment, we generate BioHash-protected templates from the features extracted by the ArcFace model from face images. Then, we apply different scoring functions<sup>7</sup>, including Hamming distance, Euclidean distance, Cosine distance, Kulsinski distance, Russell-Rao distance, Sokal-Michener distance, and Correlation distance. Table 5.12 represents the linkability measurement of BioHash-protected templates using the measure in [117] and our proposed measure based on different scoring functions. This table shows that the different scoring functions can lead to different levels of linkability of templates. Therefore, it is necessary to consider different scoring functions when measuring the linkability of protected templates.

**Linkability measurement across different biometric characteristics** To explore the application of our measure on different biometric characteristics, in this experiment, we evaluate the linkability of BioHash-protected templates across different biometric characteristics, including face (ArcFace), voice (ECAPA-TDNN), and finger vein (WLD). Table 5.13 compares the linkability measurement of BioHash-protected templates using the measure in [117] and our proposed measure across different biometric characteristics. This experiment confirms that our measure can be applied to templates with different biometric characteristics, and Table 5.13 show that BioHash-protected templates are almost unlinkable across different biometric

<sup>7</sup>Implementations of all these scoring functions are available in the SciPy package: <https://scipy.org>

## Chapter 5. Evaluation of Biometric Template Protection Schemes

Table 5.12: Linkability of BioHash-protected templates of AcrFace with different scoring functions.

Function	[117] measure	Proposed measure
Hamming distance	0.0058	0.0162
Euclidean distance	0.0058	0.0163
Cosine distance	0.0088	0.0245
Kulsinski distance	0.0098	0.0270
Russell-Rao distance	0.0102	0.0280
Sokal-Michener distance	0.0059	0.0166
Correlation distance	0.0059	0.0165

Table 5.13: Linkability of BioHash-protected templates across different biometric characteristics.

Characteristic	Feat. Extractor	[117] measure	Proposed measure
Face	ArcFace	0.0058	0.0162
Voice	ECAPA-TDNN	0.0053	0.0149
Finger Vein	WLD	0.0054	0.0153

characteristics.

**Linkability measurement for different feature extractors** To evaluate the effect of the feature extractor, in this experiment, we evaluate the linkability of BioHash-protected templates of face data extracted using different feature extractors, including ArcFace, ElasticFace, and FaceNet. Table 5.14 compares the linkability measurement of BioHash-protected templates using the measure in [117] and our proposed measure for different feature extractors. As this table shows, BioHash-protected templates are almost unlinkable for these feature extractors.

**Linkability measurement of linkable templates** In our experiments in Sections 5.3.3.2-5.3.3.2, we measured the linkability of protected biometric templates using different BTP schemes across different biometric recognition systems. Our experiments indicate that the protected templates with the aforementioned BTP schemes are almost fully unlinkable. In this section, we consider two examples of linkable protected templates and linkable unprotected templates:

**Linkable protected templates** As an example of linkable protected templates, we consider FaceNet features protected by the BioHashing scheme using *user-specific* keys. Note that in our experiments in Sections 5.3.3.2-5.3.3.2, we considered *sample-specific* keys for generating protected templates. While considering *user-specific* keys in this experiment may be assumed as a hypothetical scenario, it can reflect the situation where templates with the same key<sup>8</sup> for

<sup>8</sup>as in the typical operating of protected biometric systems

### 5.3 Measuring Linkability of Protected Templates using Maximal Leakage

Table 5.14: Linkability of BioHash-protected templates for different feature extractors.

Feat. Extractor	[117] measure	Proposed measure
ArcFace	0.0058	0.0162
ElasticFace	0.0049	0.0139
FaceNet	0.0109	0.0302

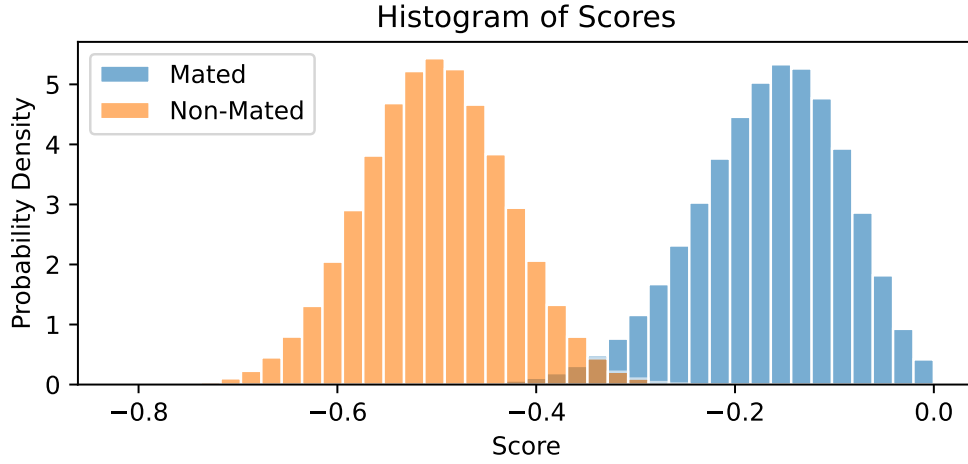


Figure 5.6: Histogram of mated and non-mated scores for linkable protected templates (FaceNet templates protected by BioHashing scheme using user-specific keys). The linkability of the mated and non-mated scores in this example is 0.9765 and 0.9574 by our and [117] measure, respectively.

each user are leaked. For instance, consider a biometric recognition system where multiple protected templates are stored for each user in the system's database (i.e., multiple reference templates). Then, an adversary gains access to all (or a portion of) the templates stored in the system's database, and aims to distinguish mated and non-mated pairs. In such a situation, since mated templates are generated using the *same* key corresponding to the user (i.e., *user-specific* key), there should be a high link between protected templates. Figure 5.6 depicts the histogram of scores for mated and non-mated templates for FaceNet features protected by the BioHashing scheme. The linkability of mated and non-mated templates in this example is 0.9765 and 0.9574 by our proposed measure and the measure in [117], respectively. Therefore, as also expected from the histogram of scores, these templates are almost fully linkable.

**Linkability of unprotected templates** In this experiment, we consider an unprotected system, and because no key is applied to generate templates in such systems, we expect to observe a high distinguishability between mated and non-mated templates (as expected from the normal operation of a biometric recognition system). As an example of such a case, we consider FaceNet features in this experiment. Figure 5.7 illustrates the histogram of scores for (unprotected) mated and non-mated templates. The linkability of templates for this case is

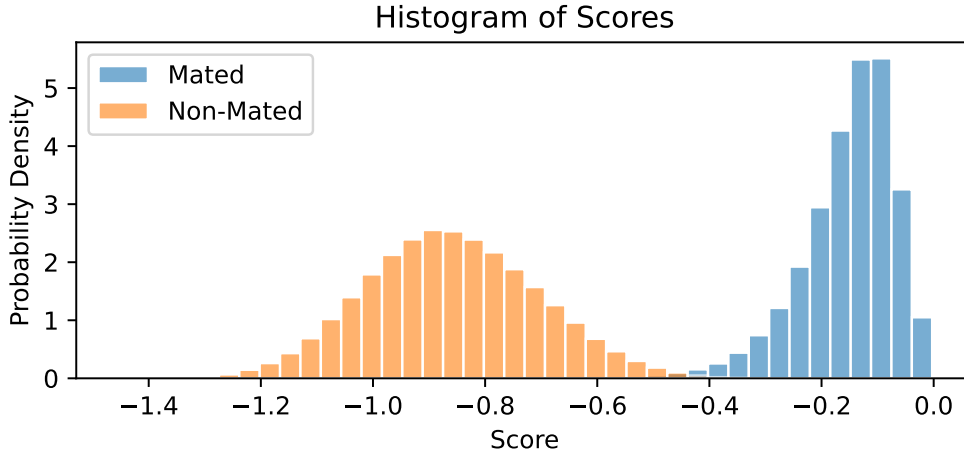


Figure 5.7: Histogram of mated and non-mated scores for unprotected templates (FaceNet). The linkability of the mated and non-mated scores in this example is 0.9912 and 0.9669 by our and [117] measure, respectively.

0.9912 by our proposed measure and 0.9669 by the one in [117]. Therefore, this experiment confirms that unprotected templates are almost fully linkable.

### 5.3.3.3 Discussion

In our experiments in Sections 5.3.3.2-5.3.3.2, we evaluated the linkability of protected biometric templates. In Section 5.3.3.2, we observed that our proposed measure and the one proposed in [117] return low values for linkability, and therefore the protected templates with different BTP schemes are almost unlikely based on both measures. Comparing the values for different BTP schemes in Table 5.11, both methods rank the evaluated BTP schemes similarly. While the values for different BTP schemes in each of these measures are close, there is theoretically no interpretation possible for the values of measure [117] and the significance of the difference between the two values in this measure. In contrast, the values of our measure can be interpreted according to Lemma 4, which provides an upper bound for the adversary's hypothesis testing (similar to the upper bounds depicted in the ROC plots of Figure 5.3 and Figure 5.4). For example, to compare the linkability of BioHashing and Bloom Filters, we have different values for the linkability measurement of these schemes in Table 5.11, and therefore we have different upper bounds according to Lemma 4. Comparing the corresponding bounds, we can say that if an adversary can gain access to BioHash-protected templates instead of templates protected with Bloom Filters, then the adversary can achieve up to  $2^{0.0162} - 2^{0.0007} = 0.0108 (\approx 1.1\%)$  more accuracy when performing hypothesis test (i.e., up to 1.1% more accuracy in distinguishing mated and non-mated templates). However, such an exercise cannot be done with [117] because there is no practical interpretation for the linkability values in [117].

The experiment in Section 5.3.3.2 showed that different scoring functions can provide different



---

## 5.4 Measuring Linkability of Multiple Protected Templates

levels of linkability for protected templates. This is reasonable since each scoring function compares two given templates differently, and thus provides different information from the similarity of the two templates. Hence, since our proposed measure and the one in [117] are based on score distributions of mated and non-mated templates, different scoring functions lead to different linkability values. Therefore, it is important to consider different scoring functions when evaluating the linkability of protected templates.

In our experiments in Section 5.3.3.2 and Section 5.3.3.2, we measured the linkability of BioHash-protected biometric templates across different biometric characteristics and for different feature extractors, respectively. These experiments show that the BioHash-protected biometric templates from different biometric characteristics and from different feature extractors are almost fully unlinkable. This experiment also confirms the application of our measure across different biometric characteristics and for different feature extractors.

In our experiments in Section 5.3.3.2, we measured the linkability of two systems that we expect to be linkable. In Section 5.3.3.2 we considered an example of linkable protected templates where we assumed that *user-specific* keys are used to generate protected templates. Since keys to generate protected templates for each user are the same in this scenario, we should have high linkability between templates, which is also confirmed by our results. As another example of linkable templates, we considered unprotected templates in Section 5.3.3.2. Similarly, in this case, we expect that the templates from the same user be similar and differ from templates of other users, which means a high level of linkability. The result of our linkability measurement also confirms that unprotected templates are almost fully linkable.

All in all, our experiments confirm that our proposed method can be deployed to measure the linkability of protected templates, and the results are intuitively correct. We evaluated the linkability of protected templates using our measure for different BTP schemes, scoring functions, biometric characteristics, and feature extractors. Furthermore, we evaluated two examples of linkable templates, where our measure also showed a high level of linkability. As discussed in Section 5.3.1 our measure has a solid theoretical background, and also the values of our measure have a practical interpretation according to Lemma 4, where our proposed measure can provide an upper bound for the accuracy of the adversary's hypothesis testing given score distributions for mated and non-mated templates.

## 5.4 Measuring Linkability of Multiple Protected Templates

In Section 5.3, score distribution of mated and non-mated templates was used with maximal leakage from information-theoretic literature [126] to propose a measure (called maximal linkability) for evaluating the linkability of protected biometric templates. The proposed measure is also properly defined and bounded in the  $[0, 1]$  interval. In particular, the proposed measure has a number of important theoretical properties and an appealing operational interpretation in terms of statistical hypothesis testing. More precisely, we showed that the proposed measure gives a theoretical upper bound on the adversary's hypothesis test and

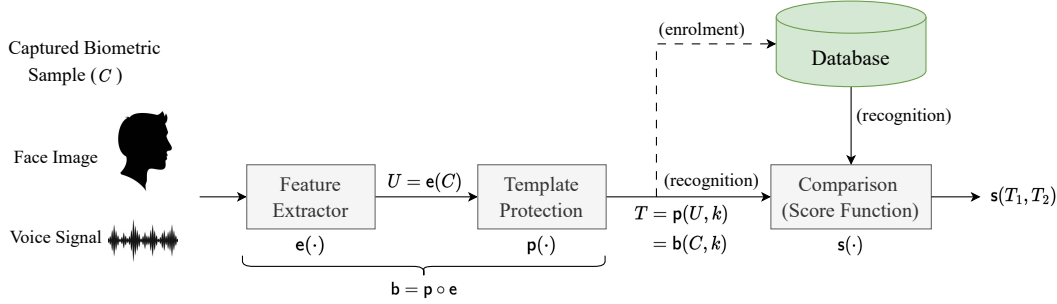


Figure 5.8: General block diagram of a biometric recognition system.

guarantees that an adversary cannot achieve higher accuracy than the resulting bound.

In this section, we build upon our measure proposed in Section 5.3 for measuring the linkability of protected biometric templates and extend it to situations where the adversary has access to multiple (two or more) similarity scores. We consider different scenarios where the adversary gains access to two or more protected templates within a single biometric system or across multiple biometric systems where the same user is enrolled. Then, the adversary will have more than a single score value for the hypothesis testing task of determining if the templates are mated or not. In general, this means that the adversary has more information and the associated linkability score should be higher. We investigate how this more general setting degrades the linkability guarantees of the system in terms of theoretical properties of the maximal linkability measure. Then, we compare the theoretical prediction with the actual linkability of the compositive systems. In our experiments, we use different BTP schemes, different biometric modalities (face and voice), and state-of-the-art deep neural network feature extractors, to evaluate the linkability of protected templates with multiple similarity scores. To our knowledge, this is the first work on measuring linkability of multiple protected biometric templates.

The remainder of this section is organized as follows. In Section 5.4.1, we define the problem of linkability evaluation based on multiple similarity scores and present the notations used in our problem formulation. In Section 5.4.2, we present our proposed method to measure the linkability having multiple similarity scores based on our measure proposed earlier in Section 5.3. In Section 5.4.3, we report our experiments for different scenarios defined in Section 5.4.1 for biometric systems and evaluate the linkability of the protected templates using our method.

### 5.4.1 Problem Definition and Formulation

#### 5.4.1.1 Notations

To facilitate understanding of our problem definition in this section, we first present our notation. We denote a biometric sample (e.g., face image or voice signal, etc.) captured by a

sensor with  $C$  and a feature extractor with  $e(\cdot)$ . We also denote the features extracted from  $C$  by the feature extractor with  $e(\cdot)$  as unprotected templates with  $U$ . In a protected biometric system, a template protection scheme  $p(\cdot, \cdot)$  is applied on each unprotected template  $U$  to generate protected template  $T = p(U, k)$ , where  $k$  is a key. Let us also denote a protected biometric system with  $b = p \circ e$ , which generates a protected template  $T = b(C, k)$  from the biometric sample  $C$  and key  $k$ . We also denote a scoring function with  $s(\cdot, \cdot)$  to compare two protected templates  $T_1$  and  $T_2$  and find the similarity score  $S = s(T_1, T_2)$ . We distinguish different templates of the same subject with different indices. For example,  $T_{1,1}$  and  $T_{1,2}$  indicate two templates of subject 1.

### 5.4.1.2 Different Scenarios with Multiple Similarity Scores

Fig. 5.8 shows a general block diagram of a protected biometric recognition system. Based on this block diagram, we consider different scenarios (denoted with Sc.), where the adversary may have multiple scores from templates leaked from different points in biometric systems and aims to find the linkability of the protected templates.

**Sc. 1: Different biometric modalities** We have two biometric modalities, e.g., face and voice. We have two samples captured for the face ( $C_{1,f}$  and  $C_{2,f}$ ) and two samples captured for voice ( $C_{1,v}$  and  $C_{2,v}$ ) and know that  $C_{1,f}$  and  $C_{1,v}$  are from the same person (subject 1), and also  $C_{2,f}$  and  $C_{2,v}$  are from the same person (subject 2). These samples are used in two biometric recognition systems<sup>9</sup>  $b_f(\cdot)$  (i.e., face recognition) and  $b_v(\cdot)$  (i.e., speaker recognition), yielding  $T_{1,1} = b_f(C_{1,f}, k_{1,1})$  and  $T_{1,2} = b_v(C_{1,v}, k_{1,2})$  as well as  $T_{2,1} = b_f(C_{2,f}, k_{2,1})$  and  $T_{2,2} = b_v(C_{2,v}, k_{2,2})$ . We use two scoring functions for these two biometric systems  $s_f(\cdot, \cdot)$  and  $s_v(\cdot, \cdot)$  and have  $S_1 = s_f(T_{1,1}, T_{2,1})$  and  $S_2 = s_v(T_{1,2}, T_{2,2})$ . We want to determine whether an adversary can say if  $C_{1,f}$  and  $C_{2,f}$  (and similarly  $C_{1,v}$  and  $C_{2,v}$ ) are for the same person or not, given  $S_1$  and  $S_2$ ? (i.e., subjects 1 and 2 are the same person or not?)

**Sc. 2: Different feature extractions** We have biometric samples (e.g., two face images)  $C_1$  and  $C_2$ , and extract features with two different feature extractors ( $e_1(\cdot)$  and  $e_2(\cdot)$ ), yielding  $U_{1,1} = e_1(C_1)$  and  $U_{1,2} = e_2(C_1)$  as well as  $U_{2,1} = e_1(C_2)$  and  $U_{2,2} = e_2(C_2)$ . We protect each extracted feature and have four protected templates  $T_{1,1} = p(U_{1,1}, k_{1,1})$ ,  $T_{1,2} = p(U_{1,2}, k_{1,2})$ ,  $T_{2,1} = p(U_{2,1}, k_{2,1})$ , and  $T_{2,2} = p(U_{2,2}, k_{2,2})$ . We use a scoring function  $s(\cdot, \cdot)$  and have  $S_1 = s(T_{1,1}, T_{1,2})$  and  $S_2 = s(T_{2,1}, T_{2,2})$ . We want to determine whether an adversary can find if  $C_1$  and  $C_2$  are for the same person or not, given  $S_1$  and  $S_2$ ?

**Sc. 3: Different template protection schemes** We have two unprotected templates,  $U_1$  and  $U_2$ , which are protected with two different BTP schemes  $p_1(\cdot, \cdot)$  and  $p_2(\cdot, \cdot)$ , yielding

<sup>9</sup>We can also consider a bi-modal biometric recognition system which uses face and voice data for recognition, and thus extracts separate templates from face and voice samples.

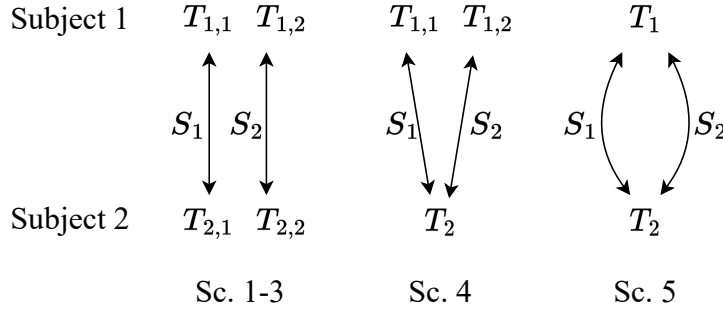


Figure 5.9: Different scenarios where we have two scores from different leaked templates.

$T_{1,1} = p_1(U_1, k_{1,1})$  and  $T_{1,2} = p_2(U_1, k_{1,2})$  as well as  $T_{2,1} = p_1(U_2, k_{2,1})$  and  $T_{2,2} = p_2(U_2, k_{2,2})$ . We use two scoring functions,  $s_1(\cdot, \cdot)$  and  $s_2(\cdot, \cdot)$  correspond to BTP scheme  $p_1(\cdot, \cdot)$  and  $p_2(\cdot, \cdot)$ , respectively, and have  $S_1 = s_1(T_{1,1}, T_{1,2})$  and  $S_2 = s_2(T_{2,1}, T_{2,2})$ . We aim to determine whether an adversary can find if  $U_1$  and  $U_2$  are for the same person or not, given  $S_1$  and  $S_2$ ?

**Sc. 4: Different protected templates** We have three protected templates  $T_{1,1}$ ,  $T_{1,2}$ , and  $T_2$  and also a scoring function  $s(\cdot, \cdot)$ . We know that  $T_{1,1} = p(U_1, k_{1,1})$ ,  $T_{1,2} = p(U_1, k_{1,2})$  are for the same person ( $U_1$ ) and aim to determine whether having  $S_1 = s(T_{1,1}, T_2)$  and  $S_2 = s(T_{1,2}, T_2)$ , an adversary can find if  $T_2 = p(U_2, k_2)$  is also for the same person or not? (i.e., subjects 1 and 2 are the same person or not?)

**Sc. 5: Different scoring functions** We have two protected templates  $T_1 = p(U_1, k_1)$  and  $T_2 = p(U_2, k_2)$  and also two scoring functions  $s_1(\cdot, \cdot)$  and  $s_2(\cdot, \cdot)$ . We want to determine whether having  $S_1 = s_1(T_1, T_2)$  and  $S_2 = s_2(T_1, T_2)$ , an adversary can find if  $T_1$  and  $T_2$  are for the same person (mated) or for different persons (non-mated)? Note that in contrast to scenarios 1-4, in this scenario the adversary does not have any additional knowledge about any potential links between leaked templates, but uses two scoring functions to facilitate the hypothesis testing task.

Fig. 5.9 illustrates different scenarios where we have two scores from different leaked templates. It is worth mentioning that each of these scenarios can be extended to any number of templates/scores or combined with other scenarios. For example, by combining Sc. 4 and Sc. 5, with two scoring functions we can have four similarity scores:  $s_1 = S_1(T_{1,1}, T_2)$ ,  $s_2 = S_1(T_{1,2}, T_2)$ ,  $s_3 = S_2(T_{1,1}, T_2)$ , and  $s_4 = S_2(T_{1,2}, T_2)$ . For simplicity, we do not discuss such combinations in this thesis, however, the proposed method can be extended for such scenarios too.

#### 5.4.2 Measuring linkability using multiple similarity scores

In this section, we discuss the properties of maximal linkability which follow from well known properties of maximal leakage and composition across multiple views, and address the be-

havior of maximal linkability with respect to the five composition scenarios outlined in Section 5.4.1.

Maximal linkability as introduced in Section 5.3 is based on an information-theoretic measure called maximal leakage [126, Theorem 1]. Maximal linkability measures the amount of information revealed by two templates about the two possible hypotheses: the templates are mated, and the templates are not mated. The maximal linkability  $M_{\rightarrow}^s$  is well defined even when the similarity scores is a vector. For example, if the similarity score  $s = (s_1, s_2)$  is a tuple, then Eq. (5.22) becomes

$$\log \sum_{(s_1, s_2) \in \mathcal{S}_1 \times \mathcal{S}_2} \max \{p(s_1, s_2 | h_m), p(s_1, s_2 | h_{nm})\}, \quad (5.35)$$

and Eq. (5.23) becomes

$$\log \int_{\mathcal{S}_1 \times \mathcal{S}_2} \max \{p(s_1, s_2 | h_m), p(s_1, s_2 | h_{nm})\} ds. \quad (5.36)$$

An example for a two-dimensional synthetic distribution is provided in Fig. 5.10. Likewise, the same hypothesis testing interpretation from Section 5.3 holds for the example in Fig. 5.10.

#### 5.4.2.1 Data Processing Inequality

Intuitively, the data processing inequality says that new information cannot be gained by processing an observation. It is well known that maximal leakage satisfies the data processing inequality, see for example [126]. That is, if  $X \leftrightarrow Y \leftrightarrow Z$  are random variables which form a Markov chain, then

$$\mathcal{L}(X \rightarrow Z) \leq \mathcal{L}(X \rightarrow Y), \quad (5.37)$$

and

$$\mathcal{L}(X \rightarrow Z) \leq \mathcal{L}(Y \rightarrow Z). \quad (5.38)$$

In terms of biometric systems, (5.37) could be alternatively stated as

$$\mathcal{L}(H \rightarrow S) \leq \mathcal{L}(H \rightarrow (T_1, T_2)). \quad (5.39)$$

This is because  $(T_1, T_2)$  is processed to obtain a scoring function output  $S$ ; thus, the amount of information in  $S$  about  $H$  cannot be greater than in  $(T_1, T_2)$ .

Given any similarity function  $s$  on  $\mathcal{T}_1 \times \mathcal{T}_2$ , the second inequality in Eq. 5.24 follows from the data processing inequality and in general  $M_{\rightarrow}^s$  underestimates the true linkability  $M_{\rightarrow}^{sys}$ . Now,

## Chapter 5. Evaluation of Biometric Template Protection Schemes

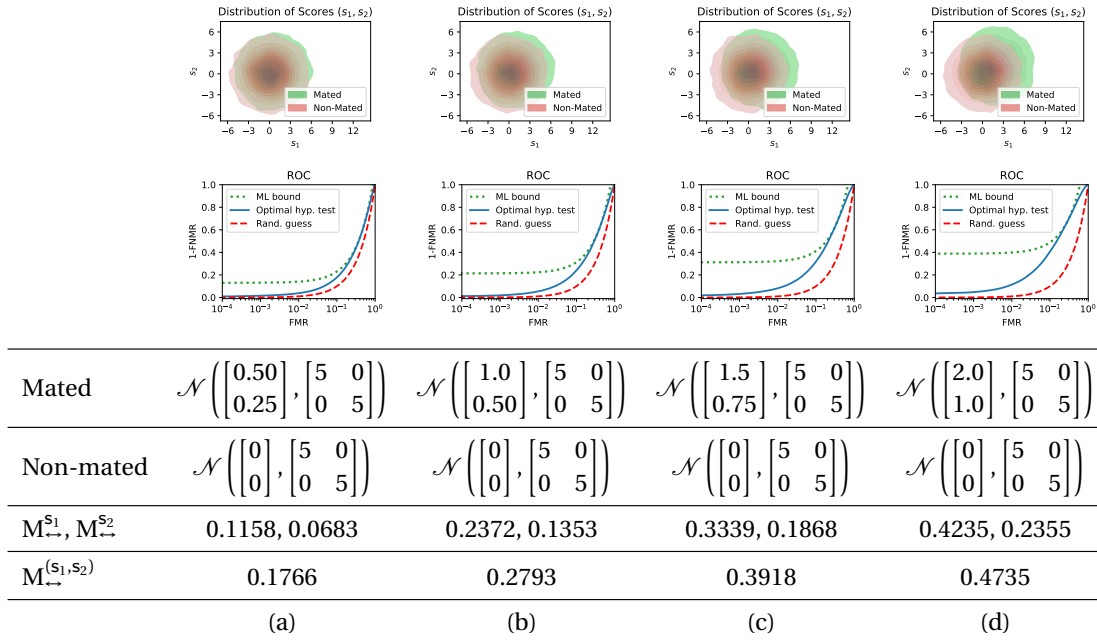


Figure 5.10: 2D histograms of synthetic distributions of mated and non-mated scores (first row) and their corresponding ROC plots (second row). The distribution of mated and non-mated scores are indicated in the third and fourth rows, respectively. In each case, the value of maximal linkability for each individual score (i.e.,  $M^{s_1}$  and  $M^{s_2}$ ) and for the tuple of  $s = (s_1, s_2)$ , (i.e.,  $M^{(s_1, s_2)}$ ) is also indicated in the fifth and sixth rows, respectively. The linkability  $M^{(s_1, s_2)}$  for the joint similarity score is always higher than the linkability for individual scores since in the joint case more information is available to the adversary. In the ROC plots, the green dotted curves indicate the maximal likability bound for the adversary hypothesis test (from Lemma 6), the solid blue curves show the optimal possible hypothesis test by the adversary, and the dashed red curves depict the random guess accuracy.

suppose that we have two scoring functions. That is  $S_1 = s_1(T_1, T_2)$  and  $S_2 = s_2(T_1, T_2)$ . Then

$$H \leftrightarrow (T_1, T_2) \leftrightarrow (S_1, S_2) \leftrightarrow S_1 \quad (5.40)$$

and

$$H \leftrightarrow (T_1, T_2) \leftrightarrow (S_1, S_2) \leftrightarrow S_2. \quad (5.41)$$

From this we have that

$$\max(M^{s_1}, M^{s_2}) \leq M^{(s_1, s_2)} \leq M^{s_1/s_2}. \quad (5.42)$$

In other words, considering two scoring functions gives us a better estimate of the true linkability than individual scores.

Similar to Lemma 4, we can bound the adversary hypothesis test using two similarity scores as follows:

## 5.4 Measuring Linkability of Multiple Protected Templates

**Lemma 6.** Suppose  $\hat{H}$  is a decision rule for the hypothesis  $H$  based on observing  $S_1 = s_1(T_1, T_2)$  and  $S_2 = s_2(T_1, T_2)$  and taking values on  $\{h_m, h_{nm}\}$ . In other words,  $H \leftrightarrow (S_1, S_2) \leftrightarrow \hat{H}$ . Let

$$\begin{aligned} FMR &= \mathbb{P}[\hat{H} = h_m | H = h_{nm}] \\ \text{and } FNMR &= \mathbb{P}[\hat{H} = h_{nm} | H = h_m] \end{aligned}$$

be the False Match and False Non-match Rates for this decision rule. Let  $M_{\leftrightarrow}^{(s_1, s_2)}$  be the maximal  $s$ -linkability score of the system. Then

$$(1 - FMR) + (1 - FNMR) \leq 2^{M_{\leftrightarrow}^{(s_1, s_2)}}. \quad (5.43)$$

Fig. 5.10 illustrates examples for two-dimensional synthetic distributions of scores and their corresponding maximal linkability.

### 5.4.2.2 Composition Theorems

Composition theorems track how much an information or privacy measure changes if multiple noisy views of the event of interest are available. For example, a composition theorem for maximal leakage [126, Lemma 6] says that if  $Z_1 \leftrightarrow X \leftrightarrow Z_2$  form a Markov chain, then

$$\mathcal{L}(X \rightarrow (Z_1, Z_2)) \leq \mathcal{L}(X \rightarrow Z_1) + \mathcal{L}(X \rightarrow Z_2). \quad (5.44)$$

On the other hand, if  $X, Z_1, Z_2$  do not satisfy the Markov chain condition, we can still say that [126, Corollary 2]

$$\mathcal{L}(X \rightarrow (Z_1, Z_2)) \leq \mathcal{L}(X \rightarrow Z_1) + \mathcal{L}(X \rightarrow Z_2 | Z_1), \quad (5.45)$$

where  $\mathcal{L}(X \rightarrow Z_2 | Z_1)$  denotes the so-called *conditional maximal leakage*.

In terms of biometric systems, if we have that

$$S_1 \leftrightarrow (\tilde{T}_1, \tilde{T}_2) \leftrightarrow H \leftrightarrow (T_1, T_2) \leftrightarrow S_2, \quad (5.46)$$

then from (5.44)

$$M_{\leftrightarrow}^{(s_1, s_2)} \leq M_{\leftrightarrow}^{s_1} + M_{\leftrightarrow}^{s_2}. \quad (5.47)$$

On the other hand, if we just have that

$$H \leftrightarrow (T_1, T_2) \leftrightarrow S_1 \text{ and } H \leftrightarrow (\tilde{T}_1, \tilde{T}_2) \leftrightarrow S_2, \quad (5.48)$$

then the Markov chain condition  $(\tilde{T}_1, \tilde{T}_2) \leftrightarrow H \leftrightarrow (T_1, T_2)$  may not hold. We need to use the

## Chapter 5. Evaluation of Biometric Template Protection Schemes

---

bound in (5.45) to obtain

$$M_{\leftrightarrow}^{(S_1, S_2)} \leq M_{\leftrightarrow}^{S_1} + \mathcal{L}(H \rightarrow S_2 | S_1), \quad (5.49)$$

where  $\mathcal{L}(H \rightarrow S_2 | S_1)$  is

$$\log \max_{s_1 \in \mathcal{S}_1} \sum_{s_2 \in \mathcal{S}_2} \max \{p(s_2 | s_1, h_m), p(s_2 | s_1, h_{nm})\}. \quad (5.50)$$

Therefore, by having access to two similarity scores, the adversary cannot learn more than the true linkability of the system. However, they may learn more than simply the sum of these two linkability.

### 5.4.2.3 Maximal Linkability for Multiple Similarity Scores

In this section we overview how the properties of maximal linkability could be applied to the five different scenarios in Section 5.4.1.

**Different scoring functions** In Sc. 5 described in Section 5.4.1, the adversary observes two scores,  $S_1$  and  $S_2$  from template pair  $(T_1, T_2)$ . By applying the data processing inequality we can thus obtain that this scenario satisfies Eq. 5.42. While it is not possible to get an upper bound on true linkability  $M_{\leftrightarrow}^{\text{sys}}$  by observing  $S_1$  and  $S_2$ , the joint linkability  $M_{\leftrightarrow}^{(S_1, S_2)}$  gives a better estimate of  $M_{\leftrightarrow}^{\text{sys}}$  than  $\max(M_{\leftrightarrow}^{S_1}, M_{\leftrightarrow}^{S_2})$ . Moreover, if  $S_1$  or  $S_2$  really capture the relevant information about the linkability of  $(T_1, T_2)$ , one would expect that considering them jointly would not lead to a significant increase in adversary's ability to link that two templates.

**Different protected templates** Sc. 4 described in Section 5.4.1 behaves similarly to Sc. 5. The difference is in Sc. 4 two protected templates are available for the first user instead of one template in Sc. 5. That is, the adversary has  $(T_{1,1}, T_{1,2})$  for the first user, and  $T_2$  for the second user. Thus,

$$H \leftrightarrow ((T_{1,1}, T_{1,2}), T_2) \leftrightarrow (T_{1,i}, T_2), \quad i \in \{1, 2\}. \quad (5.51)$$

From the data processing inequality we see that the linkability of such a system will be generally higher than  $M_{\leftrightarrow}^{\text{sys}}$  since, by definition,

$$M_{\leftrightarrow}^{\text{sys}} = \mathcal{L}(H \rightarrow (T_{1,1}, T_2)) = \mathcal{L}(H \rightarrow (T_{1,2}, T_2)). \quad (5.52)$$

It is also no longer possible to determine how the joint linkability  $M_{\leftrightarrow}^{(S_1, S_2)}$  will relate to  $M_{\leftrightarrow}^{\text{sys}}$ . On the one hand, it could be lower since some information is lost by considering the scoring functions instead of the templates. On the other hand, it could be higher since the adversary learns something about two different linked templates  $(T_{1,1}, T_{1,2})$ .



## 5.4 Measuring Linkability of Multiple Protected Templates

With respect to the data processing inequality, Sc. 2 (i.e., different feature extractors) and Sc. 3 (i.e., different template protection schemes) described in Section 5.4.1 also resemble Sc. 4. That is, the adversary has access to two templates from each user and thus access to more information. In general, the linkability of this overall composite systems will be higher than  $M_{\leftrightarrow}^{\text{sys}}$ , where

$$M_{\leftrightarrow}^{\text{sys}} = \mathcal{L}(H \rightarrow (T_{1,1}, T_{2,1})) = \mathcal{L}(H \rightarrow (T_{1,2}, T_{2,2})). \quad (5.53)$$

It is also not clear how the joint linkability  $M_{\leftrightarrow}^{(s_1, s_2)}$  will relate to  $M_{\leftrightarrow}^{\text{sys}}$ . Therefore, performing robustness analysis to see how the linkability changes with multiple views could be very interesting.

**Different biometric modalities** In Sc. 1 described in Section 5.4.1, we have that the same two individuals are compared with scoring functions derived from voice and face templates. Let  $M_{\leftrightarrow}^v$  denote the linkability of the voice system by itself, and  $M_{\leftrightarrow}^f$  denote the linkability of the image system by itself. Let  $M_{\leftrightarrow}^{\text{sys}}$  denote the linkability of the overall bi-modal system and observe that  $\max(M_{\leftrightarrow}^v, M_{\leftrightarrow}^f) \leq M_{\leftrightarrow}^{\text{sys}}$ . However, since the voice and image templates are already linked at an individual level, we cannot find an upper bound for  $M_{\leftrightarrow}^{\text{sys}}$  with the sum of individual linkabilities  $M_{\leftrightarrow}^v$  and  $M_{\leftrightarrow}^f$ . Thus, it is important to analyze the whole bi-modal system as a single system and not individual parts.

From the perspective of the scoring functions, we have

$$H \leftrightarrow ((T_{1,1}, T_{1,2}), (T_{2,1}, T_{2,2})) \leftrightarrow (S_1, S_2). \quad (5.54)$$

By applying the data processing inequality we can thus obtain that this scenario satisfies Eq. 5.42. That is, the adversary cannot learn more than the true linkability  $M_{\leftrightarrow}^{\text{sys}}$ . Once again, the joint linkability  $M_{\leftrightarrow}^{(s_1, s_2)}$  gives a better estimate of  $M_{\leftrightarrow}^{\text{sys}}$  than  $\max(M_{\leftrightarrow}^{s_1}, M_{\leftrightarrow}^{s_2})$ .

**Composition for all scenarios** We emphasize that, from the composition perspective, Eq. 5.47 is not guaranteed to hold for any of the five scenarios. That is, even a sum of two individual scores is not an upper bound on the true score. Instead, we need Eq. 5.50.

Note, however, we can modify the setting in the following way. Suppose two sets of face images  $(C_1, C_2)$  and  $(\tilde{C}_1, \tilde{C}_2)$  are known to be both mated or both non-mated (but, the adversary does not know which). However, the images themselves come from two (if mated) or four (in non-mated) different people. Then

$$S_1 \leftrightarrow (T_{1,1}, T_{2,1}) \leftrightarrow H \leftrightarrow (T_{1,2}, T_{2,2}) \leftrightarrow S_2. \quad (5.55)$$

In this case, Eq. 5.47 is a valid bound on the overall linkability.

### 5.4.3 Experiments

In this section, we present the experimental results for evaluating the linkability of *multiple* protected biometric templates based on our measure explained in Section 5.4.2. First, in Section 5.4.3.1, we describe our experimental setup for the used biometric systems and implementation details. In Section 5.4.3.2, we analyze the numerical results for different scenarios (defined in Section 5.4.1) in biometric systems where multiple information is available (more specifically, two similarity scores) for linkability measurement of protected biometric systems. In Section 5.4.3.3, we discuss the extension of the scenarios studied in Section 5.4.3.2 to the situation where the adversary can find three similarity scores to perform hypothesis tests. Finally, in Section 5.4.3.4, we further discuss our experimental findings of measuring the linkability of the biometric systems.

#### 5.4.3.1 Experimental Setup

We consider different BTP schemes, different modalities (face and voice), and SOTA DNN-based feature extractors, to evaluate the linkability of protected templates with multiple similarity scores.

**BTP schemes** In our experiments, we use different BTP schemes, including BioHashing [76], Multi-Layer Perceptron (MLP) Hashing [23], Index-of-Maximum (IoM) Hashing [77] (i.e., Gaussian random projection-based hashing, shortly GRP), and Homomorphic Encryption (HE) based on Brakerski/Fan-Vercauteren (BFV) [98] algorithm.

**Biometric Characteristics** In our experiments, we use two different biometric characteristics (modalities), including face and voice<sup>10</sup>.

#### 5.4.3.2 Analysis of different scenarios in biometric systems

In this section, we consider different scenarios in Section 5.4.1 and for each scenario we describe a case based on biometric systems based on biometric modalities and BTP schemes explained in Section 5.4.3.1. In all cases, we assume that the adversary could find two similarity scores to perform hypothesis tests, and we evaluate the linkability of multiple protected templates with different biometric modalities (in Section 5.4.3.2), different feature extractors (in Section 5.4.3.2), different BTP schemes (in Section 5.4.3.2), different keys (in Section 5.4.3.2), and different scoring functions (in Section 5.4.3.2). In our experiments in Sections 5.4.3.2, 5.4.3.2, and 5.4.3.2, we consider BioHash-protected templates since BioHashing is the simplest BTP scheme in our experiments. Similarly, we use face templates in our experiments

---

<sup>10</sup>Maximal linkability has been also used to evaluate the linkability of other biometric modalities, such as finger vein [24], for single biometric template, and it can be similarly applied to multiple leaked templates for different types of biometric modalities.

## 5.4 Measuring Linkability of Multiple Protected Templates

Table 5.15: Linkability of BioHash-protected templates for features from different biometric modalities.

mod. #1 ( $b_f$ )	mod. #2 ( $b_v$ )	$M_{\leftrightarrow}^{S_f}$	$M_{\leftrightarrow}^{S_v}$	$M_{\leftrightarrow}^{(S_f, S_v)}$
Face (ArcFace)	Voice (ECAPA)	0.0169	0.0162	0.0232
Face (ElasticFace)	Voice (ECAPA)	0.0143	0.0162	0.0212
Face (FaceNet)	Voice (ECAPA)	0.0302	0.0162	0.0344

Table 5.16: Linkability of BioHash-protected templates for different face feature extractors.

Feat. Ext. #1 ( $e_1$ )	Feat. Ext. #2 ( $e_2$ )	$M_{\leftrightarrow}^{S_1}$	$M_{\leftrightarrow}^{S_2}$	$M_{\leftrightarrow}^{(S_1, S_2)}$
ArcFace	ElasticFace	0.0156	0.0149	0.0227
ArcFace	FaceNet	0.0135	0.0149	0.0209
ElasticFace	FaceNet	0.0295	0.0149	0.0337

in Sections 5.4.3.2-5.4.3.2 since the face is one of the most popular biometric characteristics. However, we should note that similar experiments with other BTP schemes and other biometric characteristics can be implemented using our open-source paper package mentioned in Appendix E.

**Linkability of protected templates with different biometric modalities (Sc. 1)** In a multi-modal biometric recognition system the protected templates of each biometric modalities can be stored in the database of the system. In this experiment, we use voice signals and corresponding face images from the MOBIO dataset. We consider voice features extracted by ECAPA-TDNN and face features extracted by different models (ArcFace, ElasticFace, and FaceNet) and protect features of each modality separately using BioHashing. Table 5.15 reports the linkability of multiple protected biometric templates from voice and face modalities. As the results in this table show, pairs of voice and face protected templates have more linkability than their individual protected templates.

**Linkability of protected templates with different feature extractors (Sc. 2)** In some biometric systems, different feature extractors may be used and the final decision is made by fusing the scores from different templates available for each subject. In this experiment, we consider face images from the MOBIO dataset and extract features using different feature extractors, including ArcFace, ElasticFace, and FaceNet. We protect features extracted by each model separately using BioHashing. Table 5.16 reports the linkability of multiple protected biometric templates extracted from different face feature extractor models. As the results in this table show, protected pairs of features extracted from different models reveal more linkability than individual protected templates.

## Chapter 5. Evaluation of Biometric Template Protection Schemes

Table 5.17: Linkability of different BTP schemes for ArcFace templates.

BTP #1 ( $p_1$ )	BTP #2 ( $p_2$ )	$M_{\leftrightarrow}^{S_1}$	$M_{\leftrightarrow}^{S_2}$	$M_{\leftrightarrow}^{(S_1, S_2)}$
BioHashing	MLP-Hashing	0.0156	0.0096	0.0188
BioHashing	IoM-GRP	0.0156	0.0024	0.0171
BioHashing	HE	0.0156	0.0042	0.0178
MLP-Hashing	IoM-GRP	0.0096	0.0024	0.0107
MLP-Hashing	HE	0.0096	0.0042	0.0118
IoM-GRP	HE	0.0024	0.0042	0.0088*

\*  $M_{\leftrightarrow}^{(S_1, S_2)}$  is greater than  $M_{\leftrightarrow}^{S_1} + M_{\leftrightarrow}^{S_2}$ .

**Linkability of protected templates with different template protection schemes (Sc. 3)** Considering the required level of security, different BTP schemes may be used in different biometric systems. In a particular case, the same user can be enrolled in two systems and the adversary may get access to templates of the same users in both systems. In another case, different BTP schemes may be used in the same system. In this experiment, we extract ArcFace from face images in the MOBIO dataset and use different BTP schemes including BioHashing, MLP-Hashing, IoM-Hashing, and HE. The results of the linkability evaluation of multiple protected templates with different BTP schemes are reported in Table 5.17, and show that multiple templates leak more information. In particular, in the case of IoM-GRP and HE, we observe that the linkability of multiple protected templates is greater than the summation of the linkability of protected templates with individual BTP schemes.

**Linkability of protected templates with different keys (Sc. 4)** In a single biometric system, the same user may be registered with different keys at different times. This can happen because of the particular application or even in a typical system if the user is removed and registered again into the system. If the adversary gains access to both protected templates with different keys, the linkability of the pairs of protected templates is different than single protected templates. In this experiment, we use ArcFace features extracted from face images of the MOBIO dataset, and generate protected templates using different BTP schemes including BioHashing, MLP-Hashing, IoM-Hashing, and HE. For each BTP scheme, we generated two sets of protected templates using different keys. Table 5.18 compares the linkability of protected biometric templates if the adversary has access to single templates or multiple templates with different keys. The results in this table show that multiple protected templates with different keys have more linkability than single protected templates. Particularly, for HE the results in this table show that the linkability of multiple protected templates with different keys is greater than the summation of the linkability of single protected templates.

**Linkability of protected templates using different scoring functions (Sc. 5)** In every protected biometric system, if the protected templates are leaked, the adversary can use different scoring functions to perform a hypothesis test to identify if the two templates are mated or non-mated. In this experiment, we consider ArcFace features extracted from face images of the

## 5.4 Measuring Linkability of Multiple Protected Templates

Table 5.18: Linkability of protected templates with different keys for ArcFace templates.

<b>BTP</b>	$M_{\leftrightarrow}^{s_1}$	$M_{\leftrightarrow}^{s_2}$	$M_{\leftrightarrow}^{(s_1, s_2)}$
BioHashing	0.0156	0.0156	0.0156
MLP-Hashing	0.0096	0.0096	0.0096
IoM-GRP	0.0024	0.0024	0.0024
HE	0.0042	0.0031	0.0100*

\*  $M_{\leftrightarrow}^{(s_1, s_2)}$  is greater than  $M_{\leftrightarrow}^{s_1} + M_{\leftrightarrow}^{s_2}$ .

Table 5.19: Linkability of BioHash-protected templates of AcrFace with different scoring functions

<b>Func. #1 (<math>s_1</math>)</b>	<b>Func. #2 (<math>s_2</math>)</b>	$M_{\leftrightarrow}^{s_1}$	$M_{\leftrightarrow}^{s_2}$	$M_{\leftrightarrow}^{(s_1, s_2)}$
Hamming	Euclidean	0.0156	0.0162	0.0162
Hamming	Cosine	0.0156	0.0245	0.0272
Hamming	Correlation	0.0156	0.0158	0.0159
Hamming	Kulsinski	0.0156	0.0270	0.0287
Hamming	Russell-Rao	0.0156	0.0266	0.0277
Hamming	Sokal-Michener	0.0156	0.0166	0.0166
Euclidean	Cosine	0.0162	0.0245	0.0274
Euclidean	Correlation	0.0162	0.0158	0.0163
Euclidean	Kulsinski	0.0162	0.0270	0.0282
Euclidean	Russell-Rao	0.0162	0.0266	0.0276
Euclidean	Sokal-Michener	0.0162	0.0166	0.0166
Cosine	Correlation	0.0245	0.0158	0.0268
Cosine	Kulsinski	0.0245	0.0270	0.0270
Cosine	Russell-Rao	0.0245	0.0266	0.0268
Cosine	Sokal-Michener	0.0245	0.0166	0.0275
Correlation	Kulsinski	0.0158	0.0270	0.0288
Correlation	Russell-Rao	0.0158	0.0266	0.0276
Correlation	Sokal-Michener	0.0158	0.0166	0.0166
Kulsinski	Russell-Rao	0.0270	0.0266	0.0271
Kulsinski	Sokal-Michener	0.0270	0.0166	0.0283
Russell-Rao	Sokal-Michener	0.0266	0.0166	0.0276

MOBIO dataset and protected with BioHashing. We apply different scoring functions<sup>11</sup> (including Hamming distance, Euclidean distance, Cosine distance, Kulsinski distance, Russell-Rao distance, Sokal-Michener distance, and Correlation distance) for BioHash-protected templates and consider the scores available from all pairs of score functions. Table 5.19 reports the linkability of biometric templates when using two scoring functions. The results in this table show that in such a hypothesis test, the linkability of protected templates is higher than using each scoring function separately and the adversary gains a better hypothesis test. However, theoretical properties of maximal linkability discussed in Section 5.4.2 tell us that it is still less than the true linkability of the system.

<sup>11</sup>Implementations of all these scoring functions are available in the SciPy package: <https://scipy.org>

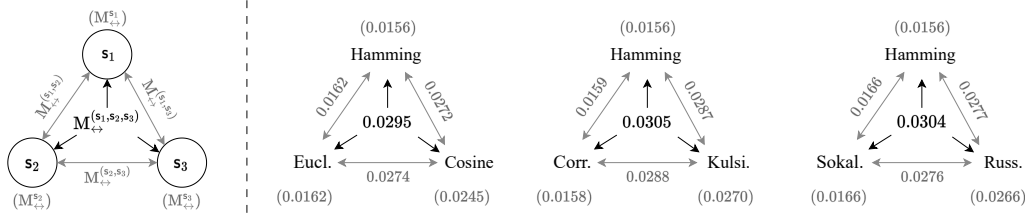


Figure 5.11: Linkability of protected templates using multiple scoring functions for BioHash-protected templates of ArcFace.

#### 5.4.3.3 Extending studied scenarios to three similarity scores

In our experiments in Sections 5.4.3.2-5.4.3.2, we considered the scenarios where the adversary could eventually find two similarity scores to perform hypothesis tests. For each of the scenarios analyzed in Section 5.4.3.2, we can extend our linkability measurement to the situation where there are more than two (e.g., three) scores available for the adversary's hypothesis tests. For example, let us consider the scenario discussed in Section 5.4.3.2 (i.e., Sc. 5) and extend to the situation where the adversary applies three scoring functions to find mated and non-mated protected templates. In such a case, we can use our method to find linkability based on three similarity scores. Fig. 5.11 illustrates the linkability of BioHash-protected of ArcFace templates if the adversary tries three different scoring functions in their hypothesis test. This figure also depicts the linkability of protected templates based on one and two similarity scores available for the adversary's hypothesis test. As we can observe by increasing the number of scoring functions, the adversary achieves a higher linkability. However, the value of maximal linkability is still slightly higher than the maximum of linkability based on a single similarity score in each case. Other scenarios (Sc. 1-4) can be similarly extended to the situations where the adversary can have three or more similarity scores between multiple templates.

#### 5.4.3.4 Discussion

Our experiments in Sections 5.4.3.2 and 5.4.3.3 evaluate the linkability of biometric systems when the adversary can find multiple similarity scores from different stages (based on defined scenarios) of a protected biometric system. With more available information, it is natural to assume that there is more linkability between protected templates, and thus an adversary can achieve better accuracy when performing hypothesis tests to distinguish mated and non-mated templates. Our experiments confirm that within biometric systems, multiple available similarity scores facilitate the linkability of templates. However, the unlinkability of the protected biometric systems degrades gradually with more available similarity scores. In most cases studied in Sections 5.4.3.2 and 5.4.3.3, we observe that linkability often does not

exceed the summation of linkability of each similarity score<sup>12</sup>, and in many cases, it is slightly greater than the maximum of linkability based on a single similarity score.

In general, we are interested in estimating  $M_{\leftrightarrow}^{sys}$  for each biometric system to find the true linkability of the system. However, as discussed in Section 5.4.2, it is not computationally possible to calculate  $M_{\leftrightarrow}^{sys}$  for real biometric systems in practice. An alternative approach is to use scoring functions to compute a proxy for  $M_{\leftrightarrow}^{sys}$  and using several scoring functions at once gives a better estimate of  $M_{\leftrightarrow}^{sys}$ . In particular, in our experiments in Sections 5.4.3.2 and 5.4.3.3, we showed that we can use two and three scoring functions, respectively, to better estimate the linkability of system with different scoring functions. Different scoring functions (as in Sc. 5) can be also applied to other scenarios (i.e., Sc. 1-4) defined in Section 5.4.1 to improve our estimation of the linkability of multiple biometric templates.

For a number of cases in Sc. 3 and Sc. 4 in our experiments in Section 5.4.3.2, we observe that maximal linkability using multiple similarity scores can be higher than the summation of maximal linkability of each individual score, which was expected from our theoretical analysis in Section 5.4.2. Therefore, as mentioned in Section 5.4.2, it is important to perform robustness analysis and evaluate the linkability of protected templates based on multiple similarity scores available for an adversary.

## 5.5 Conclusion

In Section 5.1, we presented a comprehensive benchmark by evaluating the recognition performance, unlinkability, and irreversibility of deep templates from different biometric characteristics, which are protected with different CB schemes. We used SOTA DNN models to extract features from face, voice, finger vein, and iris, and evaluated their protected templates using different CB schemes, including BioHashing, MLP-Hashing, Bloom Filters, IoM-URP, and IoM-GRP. In addition to the mentioned CB schemes, we introduced a CB scheme named Rand-Hash based on user-specific random transformations followed by binarization. Our experiments show that all the mentioned CB schemes achieve close to perfect unlinkability across different characteristics. We also evaluate the irreversibility in terms of mutual information between protected and unprotected templates. Our experimental results indicate limitations in evaluating linkability and invertibility of protected in the literature. Therefore, we dedicated the rest of the chapter to evaluate invertibility (Section 5.2) and linkability (Section 5.3 and Section 5.4) of protected templates.

In Section 5.2, we proposed a learning-based method for the inversion of protected templates, which can be used for different protection mechanisms. We considered a protected face recognition system in a situation where an adversary gains knowledge of the template protection scheme and its secrets and tries to reconstruct the face image using a leaked protected template. To this end, we trained a neural network to generate face images from protected

<sup>12</sup>Even if as discussed in Section 5.4.2, there is no theoretical guarantee not to exceed the summation for scenarios 1-4 defined in Section 5.4.1.

## Chapter 5. Evaluation of Biometric Template Protection Schemes

---

facial templates. In our experiments, we considered different template protection schemes, including BioHashing, MLP-Hashing, and Homomorphic Encryption (HE), and reconstructed face images from protected templates. We also used different state-of-the-art face recognition models and inverted their protected templates in both whitebox and blackbox scenarios. The experimental results show that our method can be used to reconstruct face images from templates protected by different template protection schemes and shed light on the vulnerability of protected face recognition systems to template inversion attacks. Considering the importance of template protection in the light of data protection regulations, our proposed method and experimental results pave the way for evaluating the invertibility of protected face recognition systems and shed light on the necessity of more research toward robust and protected biometric systems. To our knowledge, this is the first work on learning-based reconstruction of face images from protected facial templates.

In Section 5.3, we proposed a new method for measuring the linkability of protected biometric templates. We used maximal leakage, which is a well-studied measure in information-theoretic literature. Our proposed measure is based on hypothesis testing using the distributions of similarity scores of mated and non-mated protected templates. The proposed measure is consistent with the definition of linkability in the ISO/IEC 30136 standard and quantifies the linkability degree of protected templates. In particular, we showed that our measure can provide an upper bound on the accuracy of the adversary's hypothesis test given distributions of scores, and guarantees that an adversary cannot achieve better performance than the provided upper bound. The value of our measure is bounded in the  $[0, 1]$  interval, where a higher value indicates more linkability (i.e., 0 shows fully unlinkable and 1 shows fully linkable). The proposed method is also computationally stable and does not require any assumptions on prior probabilities of mated or non-mated hypotheses.

We also investigated the application of differential privacy to measure the linkability of protected biometric templates and showed that the differential privacy-based measure is too strict for the linkability application. Last but not least, in our experiments, we used the proposed measure to evaluate the linkability of biometric templates from different biometric characteristics (face, voice, and finger vein), different feature extractors, and protected with different BTP schemes. The experimental implementation of our proposed measure showed that it gives intuitively correct linkability scores across different BTP schemes, biometric characteristics, and scoring functions.

In Section 5.4, we focused on measuring the linkability of protected biometric templates when an adversary can access multiple protected templates from different biometric systems, a single multi-modal biometric system, or even a single unimodal biometric system. We defined maximal linkability for the case where an adversary can find multiple similarity scores from the available protected templates. We considered five different scenarios where the adversary gains access to multiple biometric templates from different biometric modalities, different feature extractors, different template protection schemes, or with different keys, and also two protected templates with different scoring functions. In each of these scenarios, the adversary



can find multiple similarity scores and can perform a hypothesis test to determine mated and non-mated biometric templates. In our experiments, we focused on the situation where the adversary can find two similarity scores for their hypothesis test and measured the linkability of protected templates. However, our approach can be extended to more than two similarity scores. In particular, we showcased measuring linkability of protected templates where the adversary can find three similarity scores for their hypothesis test. To our knowledge, the linkability of multiple protected biometric templates has not been studied in the literature.



## 6 Conclusion and Future Directions

In this thesis, we focused on biometric recognition systems and investigated the information in biometric templates. We first considered face recognition systems in Chapter 3, and explored the vulnerability of these systems to template inversion attacks, where an adversary gains access to the biometric templates stored in the database of the system and reconstructs the corresponding face image. We proposed different methods to reconstruct face images from facial templates, including low-resolution face reconstruction (Section 3.1), high-resolution face reconstruction using real data (Section 3.2), or high-resolution face reconstruction using synthetic data without real training data (Section 3.3), and 3D face reconstruction (Section 3.4). For each method, we explored whitebox and blackbox attacks based on the adversary's knowledge of the face recognition model. Our experiments show that an adversary can reconstruct face images from facial images in different conditions and the reconstructed face images can be used to enter the face recognition system. In addition, the reconstructed face images can reveal privacy-sensitive information of users, such as age, gender, race, etc. Therefore template inversion attacks jeopardize both the security and privacy of users in face recognition systems.

In Appendix B, we extended our template inversion attack in Section 3.1 to the situation where the adversary gains access to a portion of facial templates and investigated the reconstruction of face images based on partially leaked templates. We also explored the transferability of the reconstructed face images in Section 3.4 for attacking a different face recognition model (with a different feature extractor) where the same user is enrolled. The results show that the reconstructed face images can be recognized by other face recognition systems and can be used to enter other systems where the user is enrolled.

Similar to previous work in the literature on template inversion attacks against face recognition, in our experiments in Sections 3.1, 3.2, 3.3, and 3.4, we assumed the adversary can bypass the camera and use the reconstructed face images inject to the feature extractor of the target face recognition system. Based on this assumption, we evaluated the vulnerability of the face recognition system to injection attacks based on reconstructed face images. However, in many real-world cases, the adversary may not have enough access to inject the reconstructed face images into the feature extractor, but can only perform presentation attacks

## Chapter 6. Conclusion and Future Directions

---

using reconstructed face images. Therefore, as another experiment in Section 3.4, we used the reconstructed face images to perform a practical presentation attack against the face recognition system. To this end, we considered two types of presentation attacks, using printed photographs and digital replay attacks, and evaluated the success attack rate based on the captured images from presentation attacks. The results of our vulnerability evaluation based on presentation attacks demonstrate that the reconstructed face images can be used to enter the system in real-world attack scenarios. We comprehensively evaluated the vulnerability of the face recognition systems to our presentation attacks based on reconstructed face images from template inversion attacks in both whitebox and blackbox scenarios and also under the transferability evaluation. In Appendix C, we further explored our presentation attacks in attacking a face recognition system equipped with a commercial presentation attack detection (PAD) module. The results show that the presentation attacks can still pass the PAD system, and therefore seriously endanger the security of face recognition systems. We should note that our proposed method in Section 3.4 can generate 3D faces from facial templates, and therefore 3D reconstruction can be used for more sophisticated presentation attacks (e.g., 3D face mask, etc.) against FR systems, which require further studies in future works.

Our study on the vulnerability of face recognition systems to template inversion attacks motivates the necessity of protecting biometric templates. To this end, in Chapter 4 we focus on the protection of biometric templates and propose new template protection methods. In Section 4.1, we propose a new cancelable biometric scheme, called MLP-Hash. In our experiments in Section 4.1, we consider face templates and show that MLP-Hash satisfies the requirements of the ISO/IEC 24745 standard for BTP schemes. In addition to face templates, MLP-Hash can be also used to protect templates of other biometric characteristics, which is shown in our benchmark in Section 5.1 of Chapter 5.

In Section 4.2, we proposed a hybrid template protection method by combining cancelable biometrics and homomorphic encryption. The proposed hybrid protection scheme provides stronger protection for biometric templates than cancelable biometrics and homomorphic encryption, and also accelerates the computation in the homomorphic encryption part. We evaluated the performance of our proposed hybrid method for different face recognition models and showed that we can find suitable lengths for CB-protected templates that maintain recognition performance and achieve significantly faster execution time.

In Section 4.3, we focus on vascular recognition and propose a new method to protect and enhance the previous vascular recognition methods. We use a deep auto-encoder to reduce the dimension of vascular templates and then protect them with BioHashing. We also deploy our method on raw vascular images and extract protected templates for recognition. We show that our protected templates, in addition to further security, outperform previous vascular recognition methods. We explore the application of our method on finger vein, palm, and wrist recognition datasets. While we proposed three methods to generate protected methods in Chapter 4, this topic still requires further research in the future and new template protection schemes can be proposed. We should note that in our experiments in Chapter 4, we focused

---

on the verification scenario only and explored the application of our proposed template protection methods for the verification scenario. Therefore, the application of template protection schemes for the identification scenario remains a potential future direction.

After proposing different methods to protect biometric templates in Chapter 4, we focused on the evaluation in Chapter 5. First, we benchmarked several cancelable biometric schemes in Section 5.1 based on the ISO/IEC 24745 standard requirements. We consider different biometric characteristics and evaluate the recognition performance, irreversibility, and unlinkability of protected templates using different cancelable biometric schemes. Our benchmark showed some limitations in previous measures in the literature for evaluating the invertibility and linkability of protected templates. Therefore, in the remaining of Chapter 5, we focused on the evaluation of invertibility and linkability of protected templates.

In Section 5.2, we proposed the first learning-based method for the inversion of protected templates, which can be used for different protection mechanisms. We considered the full-disclosure scenario defined in the ISO/IEC 30136 standard, where the adversary is assumed to have access to the template protection scheme and its secrets. We considered different face recognition models in both whitebox and blackbox scenarios and evaluated the invertibility of their protected templates using different template protection schemes. Our experimental results indicate the vulnerability of protected templates and shed light on the necessity of more research toward robust and protected biometric systems. In addition, our proposed method paves the way for general methods for evaluating the invertibility of protected face recognition systems.

In Section 5.3, we proposed a new measure for evaluating the linkability of protected biometric templates. Our proposal is based on hypothesis testing using the distributions of similarity scores of mated and non-mated protected templates and uses maximal leakage to evaluate the leakage of information from protected templates. The proposed measure is consistent with the definition of linkability in the ISO/IEC 30136 standard and quantifies the linkability degree of protected templates. We showed that our measure can provide an upper bound on the accuracy of the adversary's hypothesis test given distributions of scores, and guarantees that an adversary cannot achieve better performance than the provided upper bound. The proposed method is also computationally stable and does not require any assumptions on prior probabilities of mated or non-mated hypotheses.

Our experiments for evaluating the linkability of biometric templates from different biometric characteristics, different feature extractors, and protection with different BTP schemes showed that our measure gives intuitively correct linkability scores. We conclude the discussion with some comments on an important question: how to estimate,  $M_{\leftarrow}^{sys}$ , the true linkability of the system. We adopted the approach of using  $M_{\leftarrow}^s$  as proxies for  $M_{\leftarrow}^{sys}$ . As we see in Lemma 2 in Chapter 5, the value of  $M_{\leftarrow}^s$  is always lower than the value of  $M_{\leftarrow}^{sys}$  and it is therefore important to take the highest available value of  $M_{\leftarrow}^s$  across different similarity scores. Other approaches to this problem include a stronger theoretical analysis of Eq. 5.14, as well as a more extensive

## Chapter 6. Conclusion and Future Directions

---

analysis of how well different similarity functions estimate  $M_{\leftarrow}^{\text{sys}}$ . Understanding how to better estimate the true linkability of a system is thus an important direction for future work.

In Section 5.4, we extended the linkability problem in Section 5.3 to the linkability of protected biometric templates when an adversary can access multiple protected templates from different biometric systems, a single multi-modal biometric system, or even a single unimodal biometric system. To this end, we defined maximal linkability for the case where an adversary can find multiple similarity scores from the available protected templates. We considered five different scenarios where the adversary gains access to multiple scores from different biometric templates and performs a hypothesis test to determine mated and non-mated biometric templates. In our experiments, we focused on the situation where the adversary could find two similarity scores and measured the linkability of protected templates. However, our approach can be used for more than two similarity scores, and we showcased measuring the linkability of protected templates based on three similarity scores. Our proposed measure can particularly be used to evaluate the linkability of protected templates at different stages within the same biometric system, across different biometric systems, and within multi-modal biometric systems. To our knowledge, the linkability of multiple protected biometric templates has not been studied in the literature, and thus this thesis paves the way for more comprehensive linkability studies of protected biometric templates.

# A Recognition Performance of Face Recognition Models

In our experiments in Sections 3.2.2, 3.3.2, and 3.4.2 of Chapter 3 as well as Section 5.2.2 of Chapter 5, we used ArcFace [132], ElasticFace [133], and also different FR models with SOTA backbones from FaceX-Zoo [156], including AttentionNet [141], HRNet [138], RepVGG [145], and Swin [147]. The recognition performances of these models on the MOBIO, LFW, and AgeDB datasets are reported in Table A.1. Furthermore, the Pearson Linear Correlation Coefficient (PLCC) of the comparison scores of these models for both genuines and impostors pairs on the LFW dataset is reported in Table A.2.

Table A.1: Recognition performance of face recognition models used in our experiments in terms of true match rate (TMR) at the thresholds correspond to false match rates (FMRs) of  $10^{-2}$  and  $10^{-3}$  evaluated on the MOBIO, LFW, and AgeDB datasets. The values are in percentage.

FR model	MOBIO		LFW		AgeDB	
	FMR= $10^{-2}$	FMR= $10^{-3}$	FMR= $10^{-2}$	FMR= $10^{-3}$	FMR= $10^{-2}$	FMR= $10^{-3}$
<b>ArcFace</b>	100.00	99.98	97.60	96.40	98.33	98.07
<b>ElasticFace</b>	100.00	100.00	96.87	94.70	98.20	97.57
<b>AttentionNet</b>	99.71	97.73	84.27	72.77	97.93	96.90
<b>HRNet</b>	98.98	98.23	89.30	78.43	97.67	96.23
<b>RepVGG</b>	98.75	95.80	77.20	58.07	95.93	93.93
<b>Swin</b>	99.75	98.98	91.70	87.83	98.03	97.10

## Appendix A. Recognition Performance of Face Recognition Models

---

Table A.2: Pearson Linear Correlation Coefficient (PLCC) of comparison scores for face recognition models used in our experiments for different pairs on the LFW dataset.

	<b>ArcFace</b>	<b>ElasticFace</b>	<b>AttentionNet</b>	<b>HRNet</b>	<b>RepVGG</b>	<b>Swin</b>
<b>ArcFace</b>	1.0	0.98	0.88	0.92	0.84	0.94
<b>ElasticFace</b>	0.98	1.0	0.88	0.92	0.84	0.94
<b>AttentionNet</b>	0.88	0.88	1.0	0.94	0.88	0.94
<b>HRNet</b>	0.92	0.92	0.94	1.0	0.87	0.96
<b>RepVGG</b>	0.84	0.84	0.88	0.87	1.0	0.88
<b>Swin</b>	0.94	0.94	0.94	0.96	0.88	1.0



## B Face Reconstruction from Partially Leaked Templates

In all TI attacks against FR systems investigated in the literature, it is assumed that the adversary gains access to the *complete* version of the facial templates and aims to reconstruct the underlying face image. However, it is also possible in some real-world scenarios that the adversary cannot find a complete template, but rather can reach a part of the template. This can happen due to the limited access of the adversary to the target template or the design of the FR system in which partial leakage is possible. As an example of the latter case, we can consider a FR system with a distributed database, where different parts of each face template are stored on different servers. In such a case, it is possible that an adversary can breach into one server and find *a part* of templates instead of *complete* templates. In this appendix, we focus on the inversion of *partially* leaked facial templates and investigate the amount of information required by the adversary for a successful TI attack. To our knowledge, the inversion of *partially* leaked facial templates has not been investigated in previous works.

Our proposed approach for the inversion of partially leaked face templates stems from our face reconstruction network proposed in Section 3.1 of Chapter 3. We train our previous face reconstruction network with the available part of facial templates and use the trained model to invert facial templates in the database of the system. we use the MOBIO, LFW, and AgeDB datasets and evaluate the vulnerability of SOTA FR systems to our TI attack. We consider different leakage percentages for the elements of each target template and investigate the required amount of information for a successful TI attack.

The remainder of this appendix is organized as follows. In Section B.1, we describe the threat model and explain our face reconstruction method. In Section B.2, we present our experimental results.

### B.1 Methodology

In this section, we present our proposed method to invert *partially* leaked facial templates. First, we describe our threat model in Section B.1.1, where the adversary gains access to a part

## Appendix B. Face Reconstruction from Partially Leaked Templates

---

of a facial template and aims to reconstruct the underlying face image to impersonate. Next, we describe our face reconstruction network in Section B.1.2.

### B.1.1 Threat Model

We consider the situation where the adversary gains access to a portion of a target facial template and aims to invert the *partially* leaked template to impersonate into the FR system. Let  $\mathbf{t}_c = F(\mathbf{I}) \in \mathbb{R}^d$  denote the *complete* facial template with  $d$  dimensions extracted from the face image  $\mathbf{I}$  using the feature extractor  $F(\cdot)$ . Also, let us assume that the adversary has access to a portion  $\tilde{\mathbf{t}}_{\mathcal{M}}$  of the *complete* facial template  $\mathbf{t}_c$  with indices  $M \subseteq \{1, 2, \dots, d\}$ . We define the following properties for the adversary:

- *Adversary's goal*: The adversary aims to impersonate a user enrolled in the FR system.
- *Adversary's knowledge*: The adversary is assumed to have the following information:
  1. A portion  $\tilde{\mathbf{t}}_{\mathcal{M}}$  of the target face template  $\mathbf{t}$  of a user enrolled in the system's database.
  2. The set  $\mathcal{M}$  of indices of the known elements in the *partially* leaked template  $\tilde{\mathbf{t}}_{\mathcal{M}}$ .
  3. The whitebox knowledge (including parameters and internal functioning) of the feature extraction model  $F(\cdot)$  of the FR system.
- *Adversary's capability*: The adversary can inject the reconstructed face image  $\hat{\mathbf{I}}$  from the inversion of the *partially* leaked template  $\tilde{\mathbf{t}}_{\mathcal{M}}$  directly into the feature extractor of the target system and bypass the camera.
- *Adversary's strategy*: Under the above assumptions, the adversary can invert the *partially* leaked template  $\tilde{\mathbf{t}}_{\mathcal{M}}$  and reconstruct face image  $\hat{\mathbf{I}}$  using a face reconstruction method. Then, the adversary can inject the reconstructed face image  $\hat{\mathbf{I}}$  as a query to enter the target FR system.

### B.1.2 Face Reconstruction

Our method to reconstruct face images from the *partially* leaked templates stems from our previous face reconstruction network using a *complete* leaked template proposed in Section 3.1 of Chapter 3. To train our network, we consider a dataset of face images  $\mathcal{S} = \{\mathbf{I}_i\}_{i=1}^N$ , where  $\mathbf{I}_i$  and  $N$  indicate the  $i$ th image and the total number of images, respectively. We use the data augmentation (randomly adjusting contrast and brightness, Gaussian blurring, and JPEG compression) and generate our training dataset  $\mathcal{D} = \{(\tilde{\mathbf{t}}_{\mathcal{M},i}, \mathbf{I}_i)\}_{i=1}^K$  with  $K$  pairs of *partial* templates  $\tilde{\mathbf{t}}_{\mathcal{M},i}$  and face images  $\mathbf{I}_i$ . To generate partial template  $\tilde{\mathbf{t}}_{\mathcal{M},i}$ , we first extract complete template  $\mathbf{t}_{c,i} = [F \circ A](\mathbf{I}_{a,i})$  from augmented face image  $\mathbf{I}_{a,i}$ , where  $A(\cdot)$  indicates the face alignment function. Then, we keep only elements of known indices  $\mathcal{M}$  from the complete template  $\mathbf{t}_{c,i}$  as the *partial* template  $\tilde{\mathbf{t}}_{\mathcal{M},i}$ .

We use a similar network structure based on DSCasConv blocks as proposed in Section 3.1 of Chapter 3, but the input is a partially leaked template (instead of *complete* template). We optimize our model with a multi-term loss function, including:

- *Mean Absolute Error (MAE)*: We use Mean Absolute Error (MAE) loss term on the reconstructed face images, to minimize the pixel level reconstruction error:

$$\mathcal{L}_{\text{MAE}}(\hat{\mathbf{I}}, \mathbf{I}) = \|\hat{\mathbf{I}} - \mathbf{I}\|_1, \quad (\text{B.1})$$

where  $\mathbf{I}$  and  $\hat{\mathbf{I}}$  are the original and reconstructed face images, respectively.

- *Dissimilarity Structural Index Metric (DSSIM)*: In addition to MAE loss, we enhance the objective quality of the reconstructed image in terms of the Similarity Structural Index Metric (SSIM) [175] by optimizing the DSSIM loss term [215] as follows:

$$\mathcal{L}_{\text{DSSIM}}(\hat{\mathbf{I}}, \mathbf{I}) = \frac{1 - \text{SSIM}(\hat{\mathbf{I}}, \mathbf{I})}{2} \quad (\text{B.2})$$

- *Perceptual Loss*: To further enhance the reconstruction quality, we minimize the  $\ell_1$ -norm distance between the features extracted from  $\mathbf{I}$  and  $\hat{\mathbf{I}}$  by a CNN trained on ImageNet. To this end, we extract the middle feature maps of a pre-trained VGG-16 [63] model. Let us denote the feature mapping of VGG-16 as  $P(\cdot)$ , then the perceptual loss is as follows:

$$\mathcal{L}_{\text{Perc}}(\hat{\mathbf{I}}, \mathbf{I}) = \|P(\hat{\mathbf{I}}) - P(\mathbf{I})\|_1 \quad (\text{B.3})$$

- *ID loss*: To preserve the identity in the reconstructed face image, we minimize the distance between the complete templates extracted from the reconstructed face  $\hat{\mathbf{I}}$  and original face  $\mathbf{I}$  images. To this end, we minimize the  $\ell_1$ -norm distance and maximize the cosine similarity of the extracted templates as follows:

$$\begin{aligned} \mathcal{L}_{\text{ID}}(\hat{\mathbf{I}}, \mathbf{I}) &= \mathcal{L}_{\text{ID}, \ell_1}(\hat{\mathbf{I}}, \mathbf{I}) + \mathcal{L}_{\text{ID}, \cos}(\hat{\mathbf{I}}, \mathbf{I}) \\ &= \underbrace{\|F(\hat{\mathbf{I}}) - F(\mathbf{I})\|_1}_{\text{minimizing } \ell_1\text{-norm}} + \underbrace{\frac{-F(\hat{\mathbf{I}}) \cdot F(\mathbf{I})}{\|F(\hat{\mathbf{I}})\|_2 \cdot \|F(\mathbf{I})\|_2}}_{\text{maximizing cosine similarity}} \end{aligned} \quad (\text{B.4})$$

It is noteworthy that in this loss, we extract *complete* template from each face image using face recognition model  $F$ .

We use a linear combination of the aforementioned loss terms as our total loss:

$$\mathcal{L} = \mathcal{L}_{\text{MAE}} + \gamma_1 \mathcal{L}_{\text{DSSIM}} + \gamma_2 \mathcal{L}_{\text{Perceptual}} + \gamma_3 \mathcal{L}_{\text{ID}} \quad (\text{B.5})$$

We experimentally found that the choice of  $\gamma_1 = 0.75$ ,  $\gamma_2 = 0.02$ , and  $\gamma_3 = 0.025$  performs the best. We train our network using the Adam [178] optimizer with the initial learning rate of  $10^{-3}$ , and we decrease the learning rate by a factor of 0.5 every 10 epochs.

## Appendix B. Face Reconstruction from Partially Leaked Templates

Table B.1: Recognition performance of face recognition models in terms of true match rate (TMR) at the thresholds correspond to false match rates (FMRs) of 1% and 0.1% evaluated on the MOBIO, LFW, and AgeDB datasets.

FR model	MOBIO		LFW		AgeDB	
	FMR=1%	FMR=0.1%	FMR=1%	FMR=0.1%	FMR=1%	FMR=0.1%
<b>ArcFace</b>	100.00	99.98	97.60	96.40	98.33	98.07
<b>ElasticFace</b>	100.00	100.00	96.87	94.70	98.20	97.57

## B.2 Experiments

In Section B.2.1, we describe our experimental setup, and then present our experimental results in Section B.2.2.

### B.2.1 Experimental Setup

We use two different SOTA FR models in our experiments, including ArcFace [132] and ElasticFace [133]. We consider different percentages of the elements of each facial template being leaked and evaluate the vulnerability of FR systems models based on these FR models on different datasets against our TI attacks using *partially* leaked templates. For each percentage of *partially* leaked templates, we randomly select some elements in the facial templates and remove them from the *complete* template to achieve the *partial* template. We use five different random seeds for implementing each percentage of *partially* leaked templates in our experiments.

We use the Flickr-Faces-HQ (FFHQ) dataset [186], to train our face reconstruction network. The FFHQ dataset consists of 70,000 face images (with no identity labels), and we randomly split the FFHQ dataset to train (90%) and validation (10%). After training our face reconstruction networks, we evaluate the trained models in TI attacks against FR systems on the MOBIO [158], Labeled Faces in the Wild (LFW) [159], and AgeDB [160] datasets. Table B.1 reports the recognition performance of ArcFace and ElasticFace on our evaluation datasets.

### B.2.2 Analysis

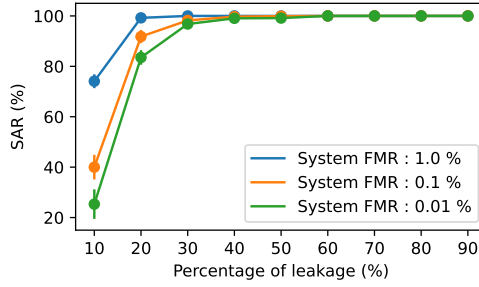
To evaluate the vulnerability of FR systems to TI attacks using partially leaked templates, we consider different leakage percentages and evaluate the adversary’s success attack rate (SAR). Table B.2 reports the SAR of TI attacks from partial templates of ArcFace and ElasticFace on the MOBIO, LFW, and AgeDB datasets for FR systems configured at false match rate (FMR) of 0.1%. As the results in this table show, if 30% of facial templates are leaked, the adversary can achieve a considerably high SAR value. For the leakage of less than 30%, the SAR is dropping while still achieving considerable SAR at 20% of leakage. This experiment demonstrates that a portion of facial templates can still be useful to represent and reconstruct face images.

Fig.B.1 illustrates the SAR for different leakage percentages of ArcFace and ElasticFace templates on the MOBIO dataset for FMR values of 1%, 0.1%, and 0.01%. This plot also confirms

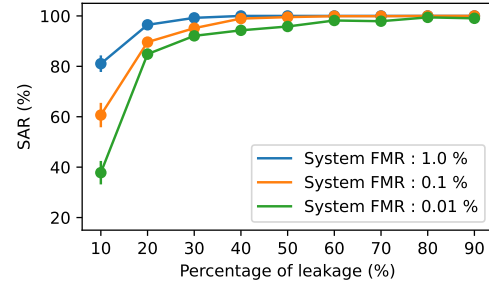
## B.2 Experiments

Table B.2: Vulnerability of face recognition systems to TI attack from different percentage of template leakage on the MOBIO, LFW, and AgeDB datasets (configured at FMR=0.1%). Cells are color coded according the SAR value between low SAR (indicated with light pink ) and high SAR (indicated with dark pink ).

Dataset	Face Recognition	SAR at different percentage of template leakage								
		10%	20%	30%	40%	50%	60%	70%	80%	90%
MOBIO	ArcFace	40.0±4.87	91.81±2.62	98.21±0.97	99.71±0.38	99.9±0.19	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
	ElasticFace	60.67±4.84	89.62±1.02	95.14±0.63	98.95±0.82	99.52±0.52	99.9±0.19	99.9±0.19	100.0±0.0	100.0±0.0
LFW	ArcFace	51.58±2.76	87.32±0.67	84.79±37.35	95.45±0.17	96.13±0.11	96.36±0.05	96.42±0.09	96.58±0.08	96.57±0.03
	ElasticFace	55.57±0.53	79.59±11.59	91.26±0.19	91.85±0.5	93.84±0.0	93.99±0.13	94.22±0.01	94.35±0.13	94.42±0.14
AgeDB	ArcFace	12.26±2.08	47.76±1.17	65.29±1.36	76.99±0.73	80.89±0.38	82.81±0.53	83.55±0.36	84.45±0.37	84.77±0.38
	ElasticFace	21.74±0.28	54.23±0.64	69.37±0.48	77.38±0.65	80.9±0.51	82.73±0.17	83.37±0.37	83.74±0.22	84.54±0.14



(a) ArcFace



(b) ElasticFace

Figure B.1: Vulnerability of face recognition systems to inversion of partially leaked templates on the MOBIO dataset for systems configured at different false match rate (FMR) values: (a) ArcFace and (b) ElasticFace.

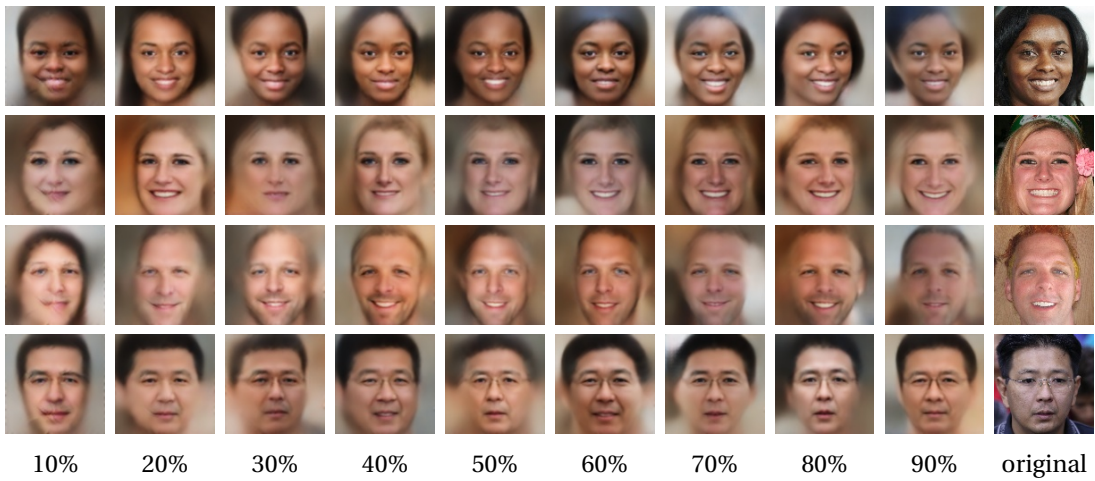


Figure B.2: Sample face images from the FFHQ dataset and their corresponding reconstructed face images from different percentages of leaked facial templates extracted by ArcFace.

## Appendix B. Face Reconstruction from Partially Leaked Templates

---

that the SAR is significant if the adversary can access at least 30% of templates. The partially leaked template can then be used by the adversary to reconstruct the underlying face image, which can be recognized as the same user by the FR model. Fig.B.2 depicts sample face images from the validation set of the FFHQ dataset and the corresponding reconstructed face images from partial templates with different leakage percentages. As the results in this figure show, the sample reconstructed face images can reveal privacy-sensitive information about the underlying user, such as age, gender, etc. As expected, the reconstruction is weakened if the adversary gains access to less portion of the template.

# C Vulnerability of an Anti-spoofing Face Recognition System to Reconstructed Face Images from TI Attacks

In our experiments in Section 3.4.2.3 of Chapter 3, we used the reconstructed face images to perform presentation attacks against FR systems. For our presentation attacks in Section 3.4.2.3, we assumed that the FR system does not have any presentation attack detection (PAD) module (also known as anti-spoofing), and therefore the captured images from the presentation attack were directly given to the feature extractor of the FR system. However, there has been extensive research on PAD methods to detect presentation attacks against FR systems [50], [220]–[222], and many real-world FR systems are equipped with PAD. Hence, as another experiment, we used the images captured in the presentation attacks using the reconstructed face images from our TI attacks, and evaluated the vulnerability of a commercial PAD system.

To this end, we considered the VeriLook SDK (Version 13.0) from Neurotechnology and used our captured images with iPhone 12 from presentation attacks with digital replay and printed photograph from our whitebox TI attacks against the ArcFace system. We considered the original images from the MOBIO dataset as bona fide, and evaluated the performance of the PAD model in detecting our presentation attacks in terms of Attack Presentation Classification Error Rate (APCER), Bona fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER). The decision threshold is calculated at the Equal Error Rate (EER). Table C.1 reports the results achieved by our presentation attacks and the performance of the commercial PAD system. As the results in this table show, the presentation attacks can still pass the commercial PAD system, demonstrating the effectiveness of our attack on systems with PAD. These results further elucidate the importance of TI attacks against FR

Table C.1: Evaluation of a PAD face recognition system for presentation attacks using the reconstructed face images from our TI attacks.

Metric	Value (%)
APCER (Printed Photography)	0.0
APCER (Digital Replay)	31.0
BPCER	15.7
ACER	23.3

## **Appendix C. Vulnerability of an Anti-spoofing Face Recognition System to Reconstructed Face Images from TI Attacks**

---

systems and prove the critical vulnerability in these systems.



# D Proofs for Chapter 5

## D.1 Proof for Section 5.3.1

*Lemma 3.* From Eq. 5.9 and Eq. 5.16 we obtain

$$M_{\leftrightarrow}^{\text{sys}} = \sup_{p_H} \log \frac{\mathbb{P}[H = \hat{H}]}{\max\{p_H(h_m), p_H(h_{nm})\}}. \quad (\text{D.1})$$

Fixing a distribution  $p_H(h_m) = p_H(h_{nm}) = 0.5$  on  $\{h_m, h_{nm}\}$ ,

$$M_{\leftrightarrow}^{\text{sys}} \geq \log \frac{\mathbb{P}[H = \hat{H}]}{\max\{p_H(h_m), p_H(h_{nm})\}} \quad (\text{D.2})$$

$$= \log(\mathbb{P}[H = \hat{H}|H = h_m] + \mathbb{P}[H = \hat{H}|H = h_{nm}]) \quad (\text{D.3})$$

$$= \log((1 - \text{FMR}) + (1 - \text{FNMR})) \quad (\text{D.4})$$

where Eq. D.3 is obtained by applying the law of total probability

$$\begin{aligned} \mathbb{P}[H = \hat{H}] &= \mathbb{P}[H = h_m] \mathbb{P}[H = \hat{H}|H = h_m] \\ &\quad + \mathbb{P}[H = h_{nm}] \mathbb{P}[H = \hat{H}|H = h_{nm}]. \end{aligned} \quad (\text{D.5})$$

□

## D.2 Proof for Section 5.3.2

*Lemma 5.* The first inequality is shown in [117], while the third and fourth are shown in Lemma 2. It remains to show that  $D_{\leftrightarrow}^{\text{sys}} \leq M_{\leftrightarrow}^{\text{s}}$ . Let  $\mathcal{F} = \{s: p(h_m|s) \geq p(h_{nm}|s)\}$  be the set of all

## Appendix D. Proofs for Chapter 5

---

the scores for which the mated hypothesis is at least as likely as the non-mated one. Then

$$D_{\leftrightarrow}(s) = p(h_m|s) - p(h_{nm}|s) \quad (D.6)$$

$$= p(s|h_m) \frac{p(h_m)}{p(s)} - p(s|h_{nm}) \frac{p(h_{nm})}{p(s)} \quad (D.7)$$

$$= \frac{p(s|h_m)\omega - p(s|h_{nm})}{p(s|h_m)\omega + p(s|h_{nm})} \quad (D.8)$$

$$\leq \frac{p(s|h_m) - p(s|h_{nm})}{p(s|h_m)} \quad (D.9)$$

where the last line holds only for  $\omega \leq 1$ . Then

$$D_{\leftrightarrow}^{\text{sys}} = \int p(s|h_m) D_{\leftrightarrow}(s) ds \quad (D.10)$$

$$= \int_{\mathcal{F}} p(s|h_m) D_{\leftrightarrow}(s) ds \quad (D.11)$$

$$\leq \int_{\mathcal{F}} [p(s|h_m) - p(s|h_{nm})] ds \quad (D.12)$$

$$\leq \int_{\mathcal{F}} p(s|h_m) ds + \int_{\bar{\mathcal{F}}} p(s|h_{nm}) ds - 1 \quad (D.13)$$

$$= \tilde{D} - 1 \quad (D.14)$$

where we defined  $\tilde{D} = \int_{\mathcal{F}} p(s|h_m) ds + \int_{\bar{\mathcal{F}}} p(s|h_{nm}) ds$ . Note that

$$M_{\leftrightarrow}^s = \log(\tilde{D}) \quad (D.15)$$

and thus,

$$M_{\leftrightarrow}^s \geq \log(1 + D_{\leftrightarrow}^{\text{sys}}) \geq D_{\leftrightarrow}^{\text{sys}}, \quad (D.16)$$

where recall that the logarithm has base two. □

# E Reproducibility

To build the biometrics systems in our experiments, we use the Bob<sup>1</sup> toolbox [223], [224] in the experiments in this thesis. For the implementation of iResNet100-ArcFace [132], ElasticsFace[133], FaceNet [73], and InceptionResnetV2-CenterLoss [134] we use their official pretrained model available in Bob<sup>2</sup>. For the other SOTA FR models with different backbones and different heads, we use the FaceX-Zoo<sup>3</sup> [156] toolbox. For other FR models, such as AdaFace [181], EdgeFace [40], and PocketNet [182], we use the pretrained models from their corresponding repositories.

For our experiments in Sections 3.2 and 3.3 of Chapter 3, we use the pretrained model of StyleGAN<sup>4</sup> to generate  $1024 \times 1024$  high-resolution images. For our experiments in Section 3.4 of Chapter 3, we use the pretrained model of EG3D<sup>5</sup> with StyleGAN [66] backbone to generate 3D faces with  $512 \times 512$  high-resolution images from any arbitrary view.

To implement the BFV algorithm (Homomorphic Encryption) for template protection experiments based on HE, we use the SEAL-Python<sup>6</sup> wrapper on Python 3.8, which uses the C++ SEAL open-source library [225].

The source code of our experiments are publicly available as different paper packages<sup>7</sup>, as follows:

- [J1] [https://gitlab.idiap.ch/bob/bob.paper.tpami2023\\_face\\_ti](https://gitlab.idiap.ch/bob/bob.paper.tpami2023_face_ti)
- [J2] [https://gitlab.idiap.ch/bob/bob.paper.tifs2024\\_face\\_ti](https://gitlab.idiap.ch/bob/bob.paper.tifs2024_face_ti)
- [J3] [https://gitlab.idiap.ch/bob/bob.paper.tbiom2024\\_face\\_ti](https://gitlab.idiap.ch/bob/bob.paper.tbiom2024_face_ti)

---

<sup>1</sup>Available at <https://www.idiap.ch/software/bob>

<sup>2</sup>Available at <https://gitlab.idiap.ch/bob/bob.bio.face>

<sup>3</sup>Available at <https://github.com/JDAI-CV/FaceX-Zoo>

<sup>4</sup>Available at <https://github.com/NVLabs/stylegan3>

<sup>5</sup>Available at <https://github.com/NVLabs/eg3d>

<sup>6</sup>Available at <https://github.com/Huelse/SEAL-Python>

<sup>7</sup>correspond to the published papers mentioned in Section 1.2 of Chapter 1.

## Appendix E. Reproducibility

---

- [C1] [https://gitlab.idiap.ch/bob/bob.paper.icip2022\\_face\\_reconstruction](https://gitlab.idiap.ch/bob/bob.paper.icip2022_face_reconstruction)
- [C2] [https://gitlab.idiap.ch/bob/bob.paper.neurips2023\\_face\\_ti](https://gitlab.idiap.ch/bob/bob.paper.neurips2023_face_ti)
- [C3] [https://gitlab.idiap.ch/bob/bob.paper.iccv2023\\_face\\_ti](https://gitlab.idiap.ch/bob/bob.paper.iccv2023_face_ti)
- [C4] [https://gitlab.idiap.ch/bob/bob.paper.icip2023\\_blackbox\\_face\\_reconstruction](https://gitlab.idiap.ch/bob/bob.paper.icip2023_blackbox_face_reconstruction)
- [C5] [https://gitlab.idiap.ch/bob/bob.paper.ijcb2023\\_face\\_ti](https://gitlab.idiap.ch/bob/bob.paper.ijcb2023_face_ti)
- [C6] [https://gitlab.idiap.ch/bob/bob.paper.icassp2024\\_face\\_ti\\_partial](https://gitlab.idiap.ch/bob/bob.paper.icassp2024_face_ti_partial)
- [J4] [https://gitlab.idiap.ch/bob/bob.paper.tbiom2021\\_protect\\_vascular\\_dnn\\_biohash](https://gitlab.idiap.ch/bob/bob.paper.tbiom2021_protect_vascular_dnn_biohash)
- [C7] [https://gitlab.idiap.ch/bob/bob.paper.eusipco2023\\_mlphash](https://gitlab.idiap.ch/bob/bob.paper.eusipco2023_mlphash)
- [C8] [https://gitlab.idiap.ch/bob/bob.paper.ijcb2022\\_hybrid\\_btp](https://gitlab.idiap.ch/bob/bob.paper.ijcb2022_hybrid_btp)
- [C9] [https://gitlab.idiap.ch/bob/bob.paper.icassp2021\\_deepae\\_biohashing\\_securefvr](https://gitlab.idiap.ch/bob/bob.paper.icassp2021_deepae_biohashing_securefvr)
- [J5] [https://gitlab.idiap.ch/bob/bob.paper.tifs2023\\_linkability\\_ml](https://gitlab.idiap.ch/bob/bob.paper.tifs2023_linkability_ml)
- [J6] [https://gitlab.idiap.ch/bob/bob.paper.access2024\\_linkability\\_multiple](https://gitlab.idiap.ch/bob/bob.paper.access2024_linkability_multiple)
- [C10] [https://gitlab.idiap.ch/bob/bob.paper.wifs2021\\_biohashing\\_sota\\_face](https://gitlab.idiap.ch/bob/bob.paper.wifs2021_biohashing_sota_face)
- [C11] [https://gitlab.idiap.ch/bob/bob.paper.fg2024\\_breaking\\_btp](https://gitlab.idiap.ch/bob/bob.paper.fg2024_breaking_btp)

The captured images for our presentation attack experiments in Section 3.4 of Chapter 3 are also publicly available at the corresponding project page: <https://www.idiap.ch/paper/gafar>

# Bibliography

- [1] A. K. Jain, A. Ross, and S. Pankanti, “Biometrics: A tool for information security”, *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, 2006.
- [2] A. K. Jain, A. Ross, and S. Prabhakar, “An introduction to biometric recognition”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.
- [3] K. Nandakumar and A. K. Jain, “Biometric template protection: Bridging the performance gap between theory and practice”, *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 88–100, 2015.
- [4] The Federal Assembly of the Swiss Confederation, *Federal act on data protection*, 2020.
- [5] European Council, *Regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation)*, Apr. 2016.
- [6] I. G. Assembly, *740 ilcs 14/biometric information privacy act*, 2008.
- [7] C. Rathgeb and A. Uhl, “A survey on biometric cryptosystems and cancelable biometrics”, *EURASIP Journal on Information Security*, vol. 2011, no. 1, pp. 1–25, 2011.
- [8] V. M. Patel, N. K. Ratha, and R. Chellappa, “Cancelable biometrics: A review”, *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 54–65, 2015.
- [9] *ISO/IEC 24745:2022(E) Information technology, cybersecurity and privacy protection – Biometric information protection*, International Standard, Switzerland, Feb. 2022.
- [10] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 30136:2018(e) information technology – security techniques – performance testing of biometric template protection schemes*, International Organization for Standardization, 2018.
- [11] H. O. Shahreza and S. Marcel, “Comprehensive vulnerability evaluation of face recognition systems to template inversion attacks via 3d face reconstruction”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [12] H. O. Shahreza, V. K. Hahn, and S. Marcel, “Vulnerability of state-of-the-art face recognition models to template inversion attack”, *IEEE Transactions on Information Forensics and Security*, 2024.

## Bibliography

---

- [13] H. O. Shahreza and S. Marcel, “Template inversion attack using synthetic face images against real face recognition systems”, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.
- [14] H. O. Shahreza, V. K. Hahn, and S. Marcel, “Face reconstruction from deep facial embeddings using a convolutional neural network”, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 1211–1215.
- [15] H. O. Shahreza and S. Marcel, “Face reconstruction from facial templates by learning latent space of a generator network”, 2023.
- [16] H. O. Shahreza and S. Marcel, “Template inversion attack against face recognition systems using 3d face reconstruction”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 19 662–19 672.
- [17] H. O. Shahreza and S. Marcel, “Blackbox face reconstruction from deep facial embeddings using a different face recognition model”, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 2435–2439.
- [18] H. O. Shahreza and S. Marcel, “Inversion of deep facial templates using synthetic data”, in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2023, pp. 1–8.
- [19] H. O. Shahreza and S. Marcel, “Face reconstruction from partially leaked facial embeddings”, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024.
- [20] H. O. Shahreza and S. Marcel, “Towards protecting and enhancing vascular biometric recognition methods via biohashing and deep neural networks”, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 394–404, 2021.
- [21] H. O. Shahreza and S. Marcel, “Deep auto-encoding and biohashing for secure finger vein recognition”, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 2585–2589.
- [22] H. O. Shahreza, C. Rathgeb, D. Osorio-Roig, V. K. Hahn, S. Marcel, and C. Busch, “Hybrid protection of biometric templates by combining homomorphic encryption and cancelable biometrics”, in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2022, pp. 1–10.
- [23] H. O. Shahreza, V. K. Hahn, and S. Marcel, “Mlp-hash: Protecting face templates via hashing of randomized multi-layer perceptron”, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, IEEE, 2023.
- [24] H. O. Shahreza, Y. Y. Shkel, and S. Marcel, “Measuring linkability of protected biometric templates using maximal leakage”, *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2262–2275, 2023. DOI: 10.1109/TIFS.2023.3266170.
- [25] H. O. Shahreza, Y. Y. Shkel, and S. Marcel, “On measuring linkability of multiple protected biometric templates using maximal leakage”, *IEEE Access*, 2024.

- [26] H. O. Shahreza, V. K. Hahn, and S. Marcel, “On the recognition performance of biohashing on state-of-the-art face recognition models”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2021, pp. 1–6.
- [27] H. O. Shahreza and S. Marcel, “Breaking template protection: Reconstruction of face images from protected facial templates”, in *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, 2024.
- [28] H. O. Shahreza, P. Melzi, D. Osorio-Roig, *et al.*, “Benchmarking of cancelable biometrics for deep templates”, *arXiv preprint arXiv:2302.13286*, 2023.
- [29] H. O. Shahreza, A. Veuthey, and S. Marcel, “Towards high-resolution face image generation from coded aperture camera”, *IEEE Sensors Letters*, 2023.
- [30] H. O. Shahreza, A. Veuthey, and S. Marcel, “Face recognition using lensless camera”, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 4385–4389.
- [31] D. Osorio-Roig, C. Rathgeb, H. O. Shahreza, C. Busch, and S. Marcel, “Indexing protected deep face templates by frequent binary patterns”, in *Proceedings of the International Joint Conference on Biometrics (IJCB)*, IEEE, 2022, pp. 1–8.
- [32] P. Melzi, H. O. Shahreza, C. Rathgeb, *et al.*, “Multi-ive: Privacy enhancement of multiple soft-biometrics in face embeddings”, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2023, pp. 323–331.
- [33] H. O. Shahreza, A. Bassit, S. Marcel, and R. Veldhuis, “Remote cancelable biometric system for verification and identification applications”, in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2023, pp. 1–5.
- [34] P. Melzi, R. Tolosana, R. Vera-Rodriguez, *et al.*, “Frcsyn-ongoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems”, *Information Fusion*, vol. 107, p. 102 322, 2024, ISSN: 1566-2535. DOI: <https://doi.org/10.1016/j.inffus.2024.102322>.
- [35] H. O. Shahreza, A. George, and S. Marcel, “Synthdistill: Face recognition with knowledge distillation from synthetic data”, in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2023, pp. 1–10.
- [36] H. O. Shahreza *et al.*, “SDFR: Synthetic data for face recognition competition”, in *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, 2024.
- [37] P. Melzi, R. Tolosana, R. Vera-Rodriguez, *et al.*, “FRCSyn challenge at WACV 2024: Face recognition challenge in the era of synthetic data”, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2024, pp. 892–901.

## Bibliography

---

- [38] I. DeAndres-Tame, R. Tolosana, P. Melzi, *et al.*, “Second edition frcsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.
- [39] D. Geissbühler, H. O. Shahreza, and S. Marcel, “Synthetic face datasets generation via latent space exploration from brownian identity diffusion”, *arXiv preprint arXiv:2405.00228*, 2024.
- [40] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel, “Edgeface: Efficient face recognition model for edge devices”, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.
- [41] J. N. Kolf, F. Boutros, J. Elliesen, *et al.*, “EFaR 2023: Efficient face recognition competition”, in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2023, pp. 1–12.
- [42] L. Colbois, H. O. Shahreza, and S. Marcel, “Approximating optimal morphing attacks using template inversion”, in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2023, pp. 1–9.
- [43] A. Unnervik, H. O. Shahreza, A. George, and S. Marcel, “Model pairing using embedding translation for backdoor attack detection on open-set classification tasks”, *arXiv preprint arXiv:2402.18718*, 2024.
- [44] A. E. Daryani, M. Mirmahdi, A. Hassanpour, H. O. Shahreza, B. Yang, and J. Fierrez, “Irl-net: Inpainted region localization network via spatial attention”, *IEEE Access*, 2023.
- [45] A. Hassanpour, Y. Kowsari, H. O. Shahreza, B. Yang, and S. Marcel, “Chatgpt and biometrics: An assessment of face recognition, gender detection, and age estimation capabilities”, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, 2024.
- [46] R. Cappelli, D. Maio, A. Lumini, and D. Maltoni, “Fingerprint image reconstruction from standard templates”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1489–1503, 2007.
- [47] C. Kauba, S. Kirchgasser, V. Mirjalili, A. Uhl, and A. Ross, “Inverse biometrics: Generating vascular images from binary templates”, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 464–478, 2021.
- [48] B. Biggio, P. Russu, L. Didaci, F. Roli, *et al.*, “Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective”, *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 31–41, 2015.
- [49] J. Galbally, C. McCool, J. Fierrez, S. Marcel, and J. Ortega-Garcia, “On the vulnerability of face verification systems to hill-climbing attacks”, *Pattern Recognition*, vol. 43, no. 3, pp. 1027–1038, 2010.
- [50] S. Marcel, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*. Springer, 2023.



- [51] A. Zhmoginov and M. Sandler, “Inverting face embeddings with convolutional neural networks”, *arXiv preprint arXiv:1606.04189*, 2016.
- [52] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, “Synthesizing normalized faces from facial identity features”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3703–3712.
- [53] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, “On the reconstruction of face images from deep face templates”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 5, pp. 1188–1202, 2018.
- [54] C. N. Duong, T.-D. Truong, K. Luu, K. G. Quach, H. Bui, and K. Roy, “Vec2face: Unveil human faces from their blackbox features in face recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6132–6141.
- [55] T.-D. Truong, C. N. Duong, N. Le, M. Savvides, and K. Luu, *Vec2face-v2: Unveil human faces from their blackbox features via attention-based network in face recognition*, 2022. DOI: 10.48550/ARXIV.2209.04920. [Online]. Available: <https://arxiv.org/abs/2209.04920>.
- [56] M. Akasaka, S. Maeda, Y. Sato, M. Nishigaki, and T. Ohki, “Model-free template reconstruction attack with feature converter”, in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2022, pp. 1–5.
- [57] X. Dong, Z. Jin, Z. Guo, and A. B. J. Teoh, “Towards generating high definition face images from deep templates”, in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2021, pp. 1–11.
- [58] M.-H. Le and N. Carlsson, “Iddecoder: A face embedding inversion tool and its privacy and security implications on facial recognition systems”, in *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, 2023, pp. 15–26.
- [59] M. Kansy, A. Raël, G. Mignone, *et al.*, “Controllable inversion of black-box face-recognition models via diffusion”, *arXiv preprint arXiv:2303.13006*, 2023.
- [60] E. Vendrow and J. Vendrow, “Realistic face reconstruction from deep embeddings”, in *Proc. of NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.
- [61] X. Dong, Z. Miao, L. Ma, *et al.*, “Reconstruct face from features based on genetic algorithm using gan generator as a distribution constraint”, *Computers & Security*, vol. 125, p. 103 026, 2023.
- [62] P. J. Burt and E. H. Adelson, “The laplacian pyramid as a compact image code”, in *Readings in computer vision*, Elsevier, 1987, pp. 671–679.
- [63] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, *arXiv preprint arXiv:1409.1556*, 2014.
- [64] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation”, *arXiv preprint arXiv:1710.10196*, 2017.

## Bibliography

---

- [65] Y. Jiang, S. Chang, and Z. Wang, “Transgan: Two pure transformers can make one strong gan, and that can scale up”, *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 745–14 758, 2021.
- [66] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.
- [67] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis”, *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
- [68] P. J. Van Laarhoven and E. H. Aarts, “Simulated annealing”, in *Simulated annealing: Theory and applications*, Springer, 1987, pp. 7–15.
- [69] M. Srinivas and L. M. Patnaik, “Genetic algorithms: A survey”, *computer*, vol. 27, no. 6, pp. 17–26, 1994.
- [70] A. Mignon and F. Jurie, “Reconstructing faces from their signatures using rbf regression”, in *Proceedings of the British Machine Vision Conference (BMVC)*, 2013, pp. 103–1.
- [71] P. Mohanty, S. Sarkar, and R. Kasturi, “From scores to face templates: A model-based approach”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2065–2078, 2007.
- [72] S. U. Hussain, T. Napoléon, and F. Jurie, “Face recognition using local quantized patterns”, in *Proceedings of the British Machine Vision Conference (BMVC)*, 2012, 11–pages.
- [73] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.
- [74] A. Sarkar and B. K. Singh, “A review on performance, security and various biometric template protection schemes for biometric authentication systems”, *Multimedia Tools and Applications*, pp. 1–56, 2020.
- [75] M. Sandhya and M. V. Prasad, “Biometric template protection: A systematic literature review of approaches and modalities”, in *Biometric Security and Privacy*, Springer, 2017, pp. 323–370.
- [76] A. T. B. Jin, D. N. C. Ling, and A. Goh, “Biohashing: Two factor authentication featuring fingerprint data and tokenised random number”, *Pattern Recognition*, vol. 37, no. 11, pp. 2245–2255, 2004.
- [77] Z. Jin, J. Y. Hwang, Y.-L. Lai, S. Kim, and A. B. J. Teoh, “Ranking-based locality sensitive hashing-enabled cancelable biometrics: Index-of-max hashing”, *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 2, pp. 393–407, 2017.
- [78] C. Rathgeb, F. Breiting, and C. Busch, “Alignment-free cancelable iris biometric templates based on adaptive bloom filters”, in *Proceedings of the International Conference on Biometrics (ICB)*, IEEE, 2013, pp. 1–8.

- 
- [79] A. Juels and M. Sudan, “A fuzzy vault scheme”, *Designs, Codes and Cryptography*, vol. 38, pp. 237–257, 2006.
  - [80] A. Juels and M. Wattenberg, “A fuzzy commitment scheme”, in *Proceedings of the 6th ACM Conference on Computer and Communications Security (ACM CCS)*, 1999, pp. 28–36.
  - [81] U. Uludag, S. Pankanti, S. Prabhakar, and A. K. Jain, “Biometric cryptosystems: Issues and challenges”, *Proceedings of the IEEE*, vol. 92, no. 6, pp. 948–960, 2004.
  - [82] C. Rathgeb, J. Merkle, J. Scholz, B. Tams, and V. Nesterowicz, “Deep face fuzzy vault: Implementation and performance”, *Computers & Security*, vol. 113, p. 102 539, 2022.
  - [83] M. Ao and S. Z. Li, “Near infrared face based biometric key binding”, in *Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3*, Springer, 2009, pp. 376–385.
  - [84] B. P. Gilkalaye, A. Rattani, and R. Derakhshani, “Euclidean-distance based fuzzy commitment scheme for biometric template security”, in *Proceedings of the International Workshop on Biometrics and Forensics (IWBF)*, IEEE, 2019, pp. 1–6.
  - [85] B. Tams, “Decodability attack against the fuzzy commitment scheme with public feature transforms”, *arXiv preprint arXiv:1406.1154*, 2014.
  - [86] A. Kholmatov and B. Yanikoglu, “Realization of correlation attack against the fuzzy vault scheme”, in *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, SPIE, vol. 6819, 2008, pp. 263–269.
  - [87] W. J. Scheirer and T. E. Boult, “Cracking fuzzy vaults and biometric encryption”, in *2007 Biometrics Symposium*, IEEE, 2007, pp. 1–6.
  - [88] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith, “Fuzzy extractors: How to generate strong keys from biometrics and other noisy data”, *SIAM Journal on Computing*, vol. 38, no. 1, pp. 97–139, 2008.
  - [89] B. Tams, P. Mihăilescu, and A. Munk, “Security considerations in minutiae-based fuzzy vaults”, *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 985–998, 2015.
  - [90] Y. J. Lee, K. Bae, S. J. Lee, K. R. Park, and J. Kim, “Biometric key binding: Fuzzy vault based on iris images”, in *Advances in Biometrics: International Conference, ICB 2007, Seoul, Korea, August 27-29, 2007. Proceedings*, Springer, 2007, pp. 800–808.
  - [91] W. Ponce-Hernandez, R. Blanco-Gonzalo, J. Liu-Jimenez, and R. Sanchez-Reillo, “Fuzzy vault scheme based on fixed-length templates applied to dynamic signature verification”, *IEEE Access*, vol. 8, pp. 11 152–11 164, 2020.
  - [92] P. Paillier, “Public-key cryptosystems based on composite degree residuosity classes”, in *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques*, Springer, 1999, pp. 223–238.
  - [93] T. ElGamal, “A public key cryptosystem and a signature scheme based on discrete logarithms”, *IEEE Transactions on Information Theory*, vol. 31, no. 4, pp. 469–472, 1985.

## Bibliography

---

- [94] D. Boneh, E.-J. Goh, and K. Nissim, “Evaluating 2-dnf formulas on ciphertexts”, in *Proceedings of the Theory of Cryptography Conference*, Springer, 2005, pp. 325–341.
- [95] A. C. Yao, “Protocols for secure computations”, in *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS)*, IEEE, 1982, pp. 160–164.
- [96] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, “(leveled) fully homomorphic encryption without bootstrapping”, *ACM Transactions on Computation Theory (TOCT)*, vol. 6, no. 3, pp. 1–36, 2014.
- [97] C. Gentry, *A fully homomorphic encryption scheme*. Stanford university, 2009.
- [98] J. Fan and F. Vercauteren, “Somewhat practical fully homomorphic encryption”, *Cryptography ePrint Archive*, 2012.
- [99] V. N. Boddeti, “Secure face matching using fully homomorphic encryption”, in *Proceedings of the International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2018, pp. 1–10.
- [100] P. Drozdowski, N. Buchmann, C. Rathgeb, M. Margraf, and C. Busch, “On the application of homomorphic encryption to face identification”, in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2019, pp. 1–5.
- [101] J. Kolberg, P. Drozdowski, M. Gomez-Barrero, C. Rathgeb, and C. Busch, “Efficiency analysis of post-quantum-secure face template protection schemes based on homomorphic encryption”, in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2020, pp. 1–4.
- [102] J. Kolberg, P. Bauspieß, M. Gomez-Barrero, C. Rathgeb, M. Dürmuth, and C. Busch, “Template protection based on homomorphic encryption: Computationally efficient application to iris-biometric verification and identification”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2019, pp. 1–6.
- [103] M. Gomez-Barrero, E. Maiorana, J. Galbally, P. Campisi, and J. Fierrez, “Multi-biometric template protection based on homomorphic encryption”, *Pattern Recognition*, vol. 67, pp. 149–163, 2017.
- [104] J. J. Engelsma, A. K. Jain, and V. N. Boddeti, “Hers: Homomorphically encrypted representation search”, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022.
- [105] D. Osorio-Roig, C. Rathgeb, P. Drozdowski, and C. Busch, “Stable hash generation for efficient privacy-preserving face identification”, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.
- [106] P. Drozdowski, F. Stockhardt, C. Rathgeb, D. Osorio-Roig, and C. Busch, “Feature fusion methods for indexing and retrieval of biometric data: Application to face recognition with privacy protection”, *IEEE Access*, vol. 9, pp. 139 361–139 378, 2021.

- [107] A. Nagar, K. Nandakumar, and A. K. Jain, "Biometric template transformation: A security analysis", in *Proc. Media Forensics and Security II*, vol. 7541, SPIE, Jan. 2010, pp. 237–251.
- [108] A. Nagar and A. K. Jain, "On the security of non-invertible fingerprint template transforms", in *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, 2009, pp. 81–85. DOI: 10.1109/WIFS.2009.5386477.
- [109] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 2382-37:2022(e) information technology – vocabulary – part 37: Biometrics*, International Organization for Standardization, 2022.
- [110] I. Buhan, J. Breebaart, J. Guajardo, K. d. Groot, E. Kelkboom, and T. Akkermans, "A quantitative analysis of indistinguishability for a continuous domain biometric cryptosystem", in *Data Privacy Management and Autonomous Spontaneous Security*, Springer, 2009, pp. 78–92.
- [111] E. J. Kelkboom, J. Breebaart, T. A. Kevenaar, I. Buhan, and R. N. Veldhuis, "Preventing the decodability attack based cross-matching in a fuzzy commitment scheme", *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 107–121, 2010.
- [112] A. Nagar, K. Nandakumar, and A. K. Jain, "Biometric template transformation: A security analysis", in *Media Forensics and Security II*, SPIE, vol. 7541, 2010, pp. 237–251.
- [113] E. Piciucco, E. Maiorana, C. Kauba, A. Uhl, and P. Campisi, "Cancelable biometrics for finger vein recognition", in *Proceedings of the First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, IEEE, 2016, pp. 1–5.
- [114] E. A. Rua, E. Maiorana, J. L. A. Castro, and P. Campisi, "Biometric template protection using universal background models: An application to online signature", *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 269–282, 2011.
- [115] M. Ferrara, D. Maltoni, and R. Cappelli, "A two-factor protection scheme for mcc fingerprint templates", in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2014, pp. 1–8.
- [116] S. Wang and J. Hu, "Design of alignment-free cancelable fingerprint templates via curtailed circular convolution", *Pattern Recognition*, vol. 47, no. 3, pp. 1321–1329, 2014.
- [117] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General framework to evaluate unlinkability in biometric template protection systems", *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1406–1420, 2017.
- [118] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey", *ACM Comput. Surv.*, vol. 51, no. 3, Jun. 2018, ISSN: 0360-0300. DOI: 10.1145/3168389. [Online]. Available: <https://doi.org/10.1145/3168389>.
- [119] M. Bloch, O. Günlü, A. Yener, *et al.*, "An overview of information-theoretic security and privacy: Metrics, limits and applications", *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 5–22, 2021. DOI: 10.1109/JSait.2021.3062755.

## Bibliography

---

- [120] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis”, in *Proceedings of the Third Conference on Theory of Cryptography*, ser. TCC’06, New York, NY: Springer-Verlag, 2006, pp. 265–284. DOI: 10.1007/11681878\_14. [Online]. Available: [http://dx.doi.org/10.1007/11681878\\_14](http://dx.doi.org/10.1007/11681878_14).
- [121] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, “Our data, ourselves: Privacy via distributed noise generation”, in *Advances in Cryptology - EUROCRYPT 2006*, S. Vaudenay, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 486–503.
- [122] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy”, *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014, ISSN: 1551-305X. DOI: 10.1561/04000000042. [Online]. Available: <http://dx.doi.org/10.1561/04000000042>.
- [123] M. Chamikara, P. Bertok, I. Khalil, D. Liu, and S. Camtepe, “Privacy preserving face recognition utilizing differential privacy”, *Computers & Security*, vol. 97, p. 101 951, 2020, ISSN: 0167-4048. DOI: <https://doi.org/10.1016/j.cose.2020.101951>.
- [124] L. Wasserman and S. Zhou, “A statistical framework for differential privacy”, *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 375–389, 2010.
- [125] P. Kairouz, S. Oh, and P. Viswanath, “The composition theorem for differential privacy”, *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 4037–4049, 2017. DOI: 10.1109/TIT.2017.2685505.
- [126] I. Issa, A. B. Wagner, and S. Kamath, “An operational approach to information leakage”, *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, 2020. DOI: 10.1109/TIT.2019.2962804.
- [127] C. Braun, K. Chatzikokolakis, and C. Palamidessi, “Quantitative notions of leakage for one-try attacks”, *Proceedings of the Conference on Mathematical Foundations of Programming Semantics (MFPS)*, vol. 249, 2009, pp. 75–91. DOI: <https://doi.org/10.1016/j.entcs.2009.07.085>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1571066109003077>.
- [128] J. Liao, L. Sankar, F. P. Calmon, and V. Y. F. Tan, “Hypothesis testing under maximal leakage privacy constraints”, in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 779–783. DOI: 10.1109/ISIT.2017.8006634.
- [129] F. Farokhi, “Noiseless privacy: Definition, guarantees, and applications”, *IEEE Transactions on Big Data*, pp. 1–1, 2021. DOI: 10.1109/TBDATA.2021.3104021.
- [130] A. R. Esposito, M. Gastpar, and I. Issa, “Generalization error bounds via rényi-, f-divergences and maximal leakage”, *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, 2021. DOI: 10.1109/TIT.2021.3085190.

- 
- [131] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 19795-1:2021. information technology – biometric performance testing and reporting – part 1: Principles and framework*, International Organization for Standardization and International Electrotechnical Committee, 2021, p. 56.
  - [132] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
  - [133] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, “Elasticface: Elastic margin loss for deep face recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1578–1587.
  - [134] T. de Freitas Pereira, A. Anjos, and S. Marcel, “Heterogeneous face recognition using domain specific units”, *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1803–1816, 2018.
  - [135] S. Chen, Y. Liu, X. Gao, and Z. Han, “Mobilefacenet: Efficient cnns for accurate real-time face verification on mobile devices”, in *Proceedings of the Chinese Conference on Biometric Recognition (CCBR)*, Springer, 2018, pp. 428–438.
  - [136] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
  - [137] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.
  - [138] J. Wang, K. Sun, T. Cheng, *et al.*, “Deep high-resolution representation learning for visual recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
  - [139] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks”, in *International Conference on Machine Learning (ICML)*, PMLR, 2019, pp. 6105–6114.
  - [140] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “Ghostnet: More features from cheap operations”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1580–1589.
  - [141] F. Wang, M. Jiang, C. Qian, *et al.*, “Residual attention network for image classification”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3156–3164.
  - [142] Y. Hu, X. Wu, and R. He, “Tf-nas: Rethinking three search freedoms of latency-constrained differentiable neural architecture search”, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, 2020, pp. 123–139.
  - [143] H. Zhang, C. Wu, Z. Zhang, *et al.*, “Resnest: Split-attention networks”, *arXiv preprint arXiv:2004.08955*, 2020.

## Bibliography

---

- [144] D. Han, S. Yun, B. Heo, and Y. Yoo, “Rethinking channel dimensions for efficient model design”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 732–741.
- [145] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, “Repvgg: Making vgg-style convnets great again”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 733–13 742.
- [146] X. Wu, R. He, Z. Sun, and T. Tan, “A light cnn for deep face representation with noisy labels”, *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.
- [147] Z. Liu, Y. Lin, Y. Cao, *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows”, *arXiv preprint arXiv:2103.14030*, 2021.
- [148] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification”, *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.
- [149] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, “Adacos: Adaptively scaling cosine logits for effectively learning deep face representations”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 823–10 832.
- [150] H. Liu, X. Zhu, Z. Lei, and S. Z. Li, “Adaptiveface: Adaptive margin and sampling for face recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 947–11 956.
- [151] Y. Sun, C. Cheng, Y. Zhang, *et al.*, “Circle loss: A unified perspective of pair similarity optimization”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6398–6407.
- [152] Y. Huang, Y. Wang, Y. Tai, *et al.*, “Curricularface: Adaptive curriculum learning loss for deep face recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5901–5910.
- [153] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, “Mis-classified vector guided softmax loss for face recognition”, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 12 241–12 248.
- [154] D. Zeng, H. Shi, H. Du, J. Wang, Z. Lei, and T. Mei, “Npcface: A negative-positive cooperation supervision for training large-scale face recognition”, *arXiv preprint arXiv:2007.10172*, 2020.
- [155] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “Magface: A universal representation for face recognition and quality assessment”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 225–14 234.
- [156] J. Wang, Y. Liu, Y. Hu, H. Shi, and T. Mei, “Facex-zoo: A pytorch toolbox for face recognition”, in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.



- 
- [157] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition”, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, 2016, pp. 87–102.
  - [158] C. McCool, R. Wallace, M. McLaren, L. El Shafey, and S. Marcel, “Session variability modelling for face authentication”, *IET Biometrics*, vol. 2, no. 3, pp. 117–129, Sep. 2013, ISSN: 2047-4938. DOI: 10.1049/iet-bmt.2012.0059.
  - [159] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments”, University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.
  - [160] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, “Agedb: The first manually collected, in-the-wild age database”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 51–59.
  - [161] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.
  - [162] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification”, in *Proc. of Interspeech 2020*, 2020, pp. 3830–3834.
  - [163] B. Huang, Y. Dai, R. Li, D. Tang, and W. Li, “Finger-vein authentication based on wide line detector and pattern normalization”, in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, IEEE, 2010, pp. 1269–1272.
  - [164] N. Miura, A. Nagasaka, and T. Miyatake, “Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification”, *Machine Vision and Applications*, vol. 15, no. 4, pp. 194–203, 2004.
  - [165] N. Miura, A. Nagasaka, and T. Miyatake, “Extraction of finger-vein patterns using maximum curvature points in image profiles”, *IEICE Transactions on Information and Systems*, vol. 90, no. 8, pp. 1185–1194, 2007.
  - [166] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.
  - [167] R. S. Kuzu, E. Maiorana, and P. Campisi, “Loss functions for cnn-based biometric vein recognition”, in *Proceedings of the 28th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 750–754.
  - [168] Y. Yin, L. Liu, and X. Sun, “Sdumla-hmt: A multimodal biometric database”, in *Proceedings of the Chinese Conference on Biometric Recognition (CCBR)*, Springer, 2011, pp. 260–268.

## Bibliography

---

- [169] B. T. Ton and R. N. J. Veldhuis, “A high quality finger vascular pattern dataset collected using a custom designed capturing device”, in *Proceedings of the International Conference on Biometrics (ICB)*, Madrid, Spain, Jun. 2013, pp. 1–5.
- [170] R. Kabaciński and M. Kowalski, “Vein pattern database and benchmark results”, *Electronics Letters*, vol. 47, no. 20, pp. 1127–1128, 2011.
- [171] *Chinese academy of sciences institute of automation. casia iris image database*, 2004. [Online]. Available: <http://biometrics.idealtest.org> (visited on 06/16/2022).
- [172] A. Hafner, P. Peer, Ž. Emeršič, and M. Vitek, “Deep iris feature extraction”, in *Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, IEEE, 2021, pp. 258–262.
- [173] P. Drozdowski, F. Struck, C. Rathgeb, and C. Busch, “Detection of glasses in near-infrared ocular images”, in *Proceedings of the International Conference on Biometrics (ICB)*, 2018, pp. 202–208. DOI: 10.1109/ICB2018.2018.00039.
- [174] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [175] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity”, *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [176] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Ieee, 2009, pp. 248–255.
- [177] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, in *Proceedings of the International Conference on Learning Representations (ICLR)*, Computational and Biological Learning Society, 2015.
- [178] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization”, in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California., USA, May 2015.
- [179] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, in *Proceedings of the International Conference on Machine Learning (ICML)*, Lille, France, Jul. 2015, pp. 448–456.
- [180] H. Zhang, C. Wu, Z. Zhang, *et al.*, “Resnest: Split-attention networks”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2736–2746.
- [181] M. Kim, A. K. Jain, and X. Liu, “Adaface: Quality adaptive margin for face recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 750–18 759.

- [182] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper, “Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation”, *IEEE Access*, vol. 10, pp. 46 823–46 833, 2022.
- [183] Z. Zhu, G. Huang, J. Deng, *et al.*, “Webface260m: A benchmark unveiling the power of million-scale deep face recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 492–10 502.
- [184] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch”, *arXiv preprint arXiv:1411.7923*, 2014.
- [185] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium”, *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [186] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.
- [187] T. Karras, M. Aittala, S. Laine, *et al.*, “Alias-free generative adversarial networks”, *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [188] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks”, in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, 2017, pp. 214–223.
- [189] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis”, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 405–421.
- [190] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, “Graf: Generative radiance fields for 3d-aware image synthesis”, *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 154–20 166, 2020.
- [191] Q. Meng, A. Chen, H. Luo, *et al.*, “Gnerf: Gan-based neural radiance field without posed camera”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6351–6361.
- [192] J. Zhang, E. Sangineto, H. Tang, *et al.*, “3d-aware semantic-guided generative model for human synthesis”, *arXiv preprint arXiv:2112.01422*, 2021.
- [193] S. Cai, A. Obukhov, D. Dai, and L. Van Gool, “Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3981–3990.
- [194] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, “Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5799–5809.

## Bibliography

---

- [195] M. Niemeyer and A. Geiger, “Giraffe: Representing scenes as compositional generative neural feature fields”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 453–11 464.
- [196] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman, “StyleSDF: High-resolution 3d-consistent image and geometry generation”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 503–13 513.
- [197] J. Gu, L. Liu, P. Wang, and C. Theobalt, “StyleNeRF: A style-based 3d aware generator for high-resolution image synthesis”, in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022, pp. 1–25.
- [198] Y. Xu, S. Peng, C. Yang, Y. Shen, and B. Zhou, “3d-aware image synthesis via learning structural and textural representations”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 430–18 439.
- [199] E. R. Chan, C. Z. Lin, M. A. Chan, *et al.*, “Efficient geometry-aware 3d generative adversarial networks”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 123–16 133.
- [200] J. Sun, X. Wang, Y. Shi, L. Wang, J. Wang, and Y. Liu, “Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis”, *arXiv preprint arXiv:2205.15517*, 2022.
- [201] S. Galanakis, B. Gecer, A. Lattas, and S. Zafeiriou, “3dmm-rf: Convolutional radiance fields for 3d face modeling”, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3536–3547.
- [202] D. Rebain, M. Matthews, K. M. Yi, D. Lagun, and A. Tagliasacchi, “Lolnerf: Learn from one look”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1558–1567.
- [203] *ISO/IEC 30107-3:2017(E) Information technology – Biometric presentation attack detection – Part 3: Testing and reporting*, International Standard, Switzerland, Jun. 2017.
- [204] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-grained head pose estimation without keypoints”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018.
- [205] G. Mai, K. Cao, X. Lan, and P. C. Yuen, “Secureface: Face template protection”, *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 262–277, 2020.
- [206] Y. K. Jang and N. I. Cho, “Deep face image retrieval for cancelable biometric authentication”, in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 2019, pp. 1–8.
- [207] H. Lee, C. Y. Low, and A. B. J. Teoh, “Softmaxout transformation-permutation network for facial template protection”, in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 7558–7565.

- [208] P. Drozdowski, F. Struck, C. Rathgeb, and C. Busch, “Benchmarking binarisation schemes for deep face templates”, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 191–195.
- [209] K. H. Cheung, A. W.-K. Kong, J. You, D. Zhang, *et al.*, “An analysis on invertibility of cancelable biometrics based on biohashing”, in *CISST*, Citeseer, vol. 2005, 2005, pp. 40–45.
- [210] R. Belguechi, E. Cherrier, and C. Rosenberger, “How to evaluate transformation based cancelable biometric systems?”, in *Proceedings of the NIST International Biometric Performance Testing Conference (IBPC)*, 2012.
- [211] Y. Lee, Y. Chung, and K. Moon, “Inverse operation and preimage attack on biohashing”, in *Proceedings of the IEEE Workshop on Computational Intelligence in Biometrics: Theory, Algorithms, and Applications*, IEEE, 2009, pp. 92–97.
- [212] X. Dong, Z. Jin, and A. T. B. Jin, “A genetic algorithm enabled similarity-based attack on cancellable biometrics”, in *Proceedings of the IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2019, pp. 1–8.
- [213] C. E. Shannon, “A mathematical theory of communication”, *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [214] D. Keller, M. Osadchy, and O. Dunkelman, “Inverting binarizations of facial templates produced by deep learning (and its implications)”, *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4184–4196, 2021.
- [215] S. Sadrizadeh, H. Otroshi-Shahreza, and F. Marvasti, “Impulsive noise removal via a blind cnn enhanced by an iterative post-processing”, *Signal Processing*, vol. 192, p. 108378, 2022.
- [216] R. Sibson, “Information radius”, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.
- [217] S. Verdú, “ $\alpha$ -mutual information”, in *Proceedings of the Information Theory and Applications Workshop (ITA)*, IEEE, 2015, pp. 1–6.
- [218] A. Rényi, “On measures of entropy and information”, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, University of California Press, 1961, pp. 547–561.
- [219] Y. Wu and P. Yang, “Chebyshev polynomials, moment matching, and optimal estimation of the unseen”, *The Annals of Statistics*, vol. 47, no. 2, pp. 857–883, 2019. DOI: 10.1214/17-AOS1665. [Online]. Available: <https://doi.org/10.1214/17-AOS1665>.
- [220] J. Galbally, S. Marcel, and J. Fierrez, “Biometric antispoofing methods: A survey in face recognition”, *IEEE Access*, vol. 2, pp. 1530–1552, 2014.
- [221] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, “Biometrics systems under spoofing attack: An evaluation methodology and lessons learned”, *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

## Bibliography

---

- [222] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of biometric anti-spoofing: Presentation attack detection*. Springer, 2019, vol. 2.
- [223] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, “Bob: A free signal processing and machine learning toolbox for researchers”, in *Proceedings of the 20th ACM Conference on Multimedia Systems (ACMMM)*, Nara, Japan, Oct. 2012.
- [224] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel, “Continuously reproducing toolchains in pattern recognition and machine learning experiments”, in *Proc. of ICML 2017 Reproducibility in Machine Learning Workshop*, 2017, pp. 1–8. [Online]. Available: <https://openreview.net/forum?id=BJDDItGX->.
- [225] *Microsoft SEAL (release 3.6)*, <https://github.com/Microsoft/SEAL>, Microsoft Research, Redmond, WA., Nov. 2020.

# Hatef OTROSHI SHAHREZA

in hatef-otroshi •  Hatef Otroshi Shahreza •  otroshi

## Education

- **EPFL, Lausanne, Switzerland** 2020 – 2024
  - *Ph.D. in Electrical Engineering*
- **Sharif University of Technology, Tehran, Iran** 2016 – 2018
  - *M.Sc. in Electrical Engineering - Communication Systems*
- **University of Kashan, Kashan, Iran** 2012 – 2016
  - *B.Sc. (Hons.) in Electrical Engineering - Communications*

## Experience

### Research Assistantships

- **Idiap Research Institute, Switzerland (2020-2024)**
  - Research Assistant in Biometrics Security and Privacy group [**conducted Ph.D. thesis**]
  - **Research Topics:** Generative Models (GAN, NeRF, Diffusion Model), Face Reconstruction, Synthetic Data, Face Recognition, Security, Privacy, Information Leakage, Biometrics, Knowledge Distillation, Multi-modal Systems, Large Vision Language Models.
- **Sharif University of Technology, Iran (2017-2020)**
  - Research Assistant in Multimedia Lab [**conducted M.Sc. thesis**]
  - **Research Topics:** Image/Video Quality Assessments, Human Perception Modeling, Image Enhancement, Image Denoising, Image/Video Processing, Machine/Deep Learning for Signal Processing.

### Visiting Research/ Internships

- **ams OSRAM AG, Switzerland, 2023 (4 months)**
  - Research Intern in Innovation Office
  - **Research Topics:** Face Reconstruction, Lensless Imaging.
- **Hochschule Darmstadt (HDA), Darmstadt, Germany, 2022 (6 months)**
  - Visiting Researcher at Biometrics and Internet Security group da/sec.
  - **Research Topics:** Face/Speaker Recognition, Biometric Template Protection, Homomorphic Encryption.

### Teaching Assistantships

- **“Fundamentals in Statistical Pattern Recognition”** course, EE Dept, EPFL, 2023.
- **“Introduction to Biometrics”** course, Master of AI, Idiap Research Inst and UniDistance, 2021-2023.
- **“Deep Learning”** course, CE & EE Depts, Sharif University of Technology, 2018-2019.
- **“Machine Learning”** course, CE & Depts., Sharif University of Technology, 2018.

### Reviewing Activities

- **Conferences:** NeurIPS, ICML, ECCV, WACV, ICPR, IJCB, ICASSP.
- **Journals:** IEEE-TIFS, IEEE-TCSVT, IEEE-TIM, IEEE-TBIOM, Pattern Recognition, Scientific Reports

## Publications

The updated list of publications is available at my Google Scholar profile.

## Computer skills

**Programming Languages:** Python, C/C++, MATLAB

**Packages and Libraries:** PyTorch, PyTorch Geometric, TensorFlow, OpenCV

**General:** Linux, Git,  $\text{\LaTeX}$

## Honors and Awards

- Winner of the *European Association for Biometrics (EAB) Research Award 2023*
- Winner of the *Idiap Best PhD Student Award*, Idiap Research Institute, 2023
- Reception of the *H2020 Marie Skłodowska-Curie Fellowship* for doctoral program, 2020-2023
- Ranked 1st among 109 students of Bachelor of Electrical Engineering, University of Kashan, 2016