UNIL | Université de Lausanne

FACULTÉ DE DROIT, DES SCIENCES CRIMINELLES ET D'ADMINISTRATION PUBLIQUE

ÉCOLE DES SCIENCES CRIMINELLES

# Face morphing attacks in the era of deepfakes: risks, detection & source attribution

THÈSE DE DOCTORAT

présentée à la
Faculté de droit, des sciences criminelles et d'administration publique
de l'Université de Lausanne

pour l'obtention du grade de
Doctorat en science forensique

par
Laurent COLBOIS

Directeur de thèse
Prof. Sébastien Marcel

Co-directeur de thèse
Prof. Christophe Champod

Jury

Prof. Thomas Souvignet, président
Prof. Luisa Verdoliva, experte externe
Prof. Vitomir Štruc, expert externe,
Prof. David-Olivier Jaquet-Chiffelle, expert interne

LAUSANNE
2025

UNIL | Université de Lausanne
Ecole des Sciences criminelles

# IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise le dépôt de la thèse de Monsieur Laurent Colbois, candidat au doctorat en science forensique, intitulée :

**« Face morphing attacks in the era of deepfakes : risks, detection & source attribution »**

Professeur Thomas Souvignet
Président du jury

Lausanne, le 31 mars 2025

Unil

UNIL | Université de Lausanne

Ecole des sciences criminelles

## Thèse de doctorat en science forensique

## Présentation publique

Lundi 31 mars 2025, Amphipôle 210, 10h15 – 12h00

**Titre :** **Face morphing attacks in the era of deepfakes : risks, detection & source attribution**

**Candidat :** Laurent Colbois

**Jury :**

Prof. Thomas Souvignet, Ecole des sciences criminelles, Président du jury

Prof. Luisa Verdoliva, GRIP Lab, Université Federico II de Naples, Italie, experte externe

Prof. Vitomir Štruc, Laboratory for Machine Intelligence, Université de Ljubljana, Slovénie, expert externe

Prof. David-Olivier Jaquet-Chiffelle, Ecole des sciences criminelles, expert interne

Prof. Sébastien Marcel, Ecole des sciences criminelles, co-directeur de thèse

Prof. Christophe Champod, Ecole des sciences criminelles, co-directeur de thèse

A la suite des rapports des membres du jury et après avoir entendu le candidat exposer le sujet de sa recherche et répondre aux questions qui lui étaient posées, le jury a délibéré et, à l'unanimité, charge la Direction de l'Ecole des sciences criminelles de transmettre à la Direction de l'Université de Lausanne, un préavis invitant l'Université à délivrer le titre de Docteur ès Sciences en science forensique à Monsieur Laurent Colbois.

UNIL | Université de Lausanne

Ecole des sciences criminelles

Le jury relève l'étendue et diversité des travaux de recherche de M. Colbois ainsi que la qualité de rédaction de sa thèse de doctorat. La précision et la qualité des réponses aux différentes questions du jury souligne la grande maitrise de son sujet ainsi que ses qualités de synthèse et de vulgarisation.

M. Colbois réalise une étude rigoureuse des attaques par morphing de visage dans le contexte des « deepfakes », plus particulièrement dans la délivrance et la vérification des passeports et autres documents de légitimation.

Il fournit des contributions sous différentes perspectives et dans différents scénarios, des *Generative Adversarial Networks* (GANs) aux *Diffusion Models*, de la détection à l'attribution.

Son travail se distingue tout d'abord par la profondeur de l'analyse et l'originalité des solutions techniques proposées. Il est également remarquable par la dissémination de ses travaux tant au niveau de la littérature (conférences, 5 articles publiés et 2 soumis) que par le partage de code et de données de recherche.

La thèse de M. Colbois apporte une contribution unique, originale et significative permettant une application pratique, améliorant l'efficacité et la performance des systèmes de détection.

Le jury félicite le candidat pour la qualité de sa recherche.

Le Président du jury délivre l'imprimatur et autorise le candidat à déposer sa thèse.

Prof. Thomas Souvignet
Président du jury

Prof. Luisa Verdoliva
Experte externe

Prof. Vitomir Štruc
Expert externe

Prof. Sébastien Marcel
Directeur de Thèse

Prof. David-Olivier Jaquet-Chiffelle
Experte externe

Prof. Christophe Champod
Co-Directeur de Thèse

# Acknowledgements

This thesis was supported by the Swiss Center for Biometrics Research & Testing and the Idiap Research Institute. To enhance the style and flow of the text, language models were employed to suggest phrasing tweaks and improve vocabulary precision. These tools were only used as a means of stylistic refinement, the underlying ideas and content are entirely my own.

This thesis is the result of a four-year effort, during which I have been accompanied by an incredible support network, without which the PhD would certainly not have been such an enjoyable experience. I wish to thank all those involved:
**Sébastien**, for the quality of your supervision. You knew precisely when to provide guidance whenever I felt lost or needed inspiration, and when to trust my process when I felt confident in my direction and simply needed time. This balance is probably delicate and challenging for a supervisor, and you mastered it.
**Christophe**, for being much more than an administrative supervisor. You broadened my horizons by guiding me in discovering the fundamentals of forensic science, expanding my research perspective through your lens, and providing me with plenty of opportunities to engage with communities beyond machine learning.
**The members of my thesis jury**: Prof. Thomas Souvignet, for chairing the defense, and Professors Luisa Verdoliva, Vitomir Štruc, and David-Olivier Jaquet-Chiffelle for your valuable time and insightful feedback. Luisa, special thanks for hosting me at the GRIP Lab in Naples. That visit was enriching for my research as much as for my personal experiences.
**The Biometrics Security & Privacy group**: Amir, Alex, Tiago, Hatef, Sushil, Vedrana, Anjith, Alain, Luis, Pavel, Christophe, Vidit, Ibrahim, Parsa. I enjoyed our coffee break conversations either discussing research or various life things, as well as the collaborations I shared with some of you. You collectively created a comfortable work environment.
**All the friends I made at Idiap**: so many people have played a crucial role in making Idiap and Martigny feel like the greatest campus, and I truly hope I don't forget anyone![1] From the start, I felt kind of adopted into an amazing family of dear friends. To Amir, Zohreh, Radvin, Neha, François, Florian Piras, Florian Mai, Andrei, Chloé, Eklavya, Anshul, Tilak, Pablo, Julian: I struggled and procrastinated writing these acknowledgements precisely because of you all, because I don't know how I can ever adequately express my gratitude for all the wonderful moments we've shared from the beginning until now, and hopefully for many more to come. Thank you for the dinners, trips, hikes, skiing, board games, kicker games, running, swimming,

[1] Such as Fabio! Beyond everything else, special thanks for all the running sessions and the philosophical discussions!

cooking, baking, tasting your amazing culinary creations, dancing, music, concerts, silly memes and jokes... I extend these thanks to the many other Idiapers I had the pleasure of interacting with throughout my PhD journey: Karl, Laura, Carlos, Louise, Sarthak, Valentin, Mirko, Barbara, Michiel, Inga, Mathias, Luis, Victor, Enno, Olena, Darya, Alina-Elena, Roberto, Apoorv, Sargam, Alejandra, Maxime, Pierre, Arya, Alice, Haruki, and Shashi. Each of you contributed significantly to making life at Idiap and Martigny extraordinarily enjoyable.

**The Harmonie municipale de Martigny**, for providing an exciting and welcoming musical environment, which was a much needed source of relaxation. Special thanks to our conductor Dany Rossier, to our president and my trumpet companion Jonathan Bochatay, and to the committee, whose continued efforts allow us to enjoy playing music in a relaxed yet stimulating atmosphere. Thanks also to the close friends I made through the band : Flavia, Etienne, Julie, Aloïs, Vincent, Jean, Sofiia. I enjoy our time together, whether rehearsing, giving a concert, or outside of our music activities.

**My group of friends from the Master's program**: Sergio, Didier, Emily, Manuel. It was always a pleasure catching up for unexpected outings, hikes, or dinners. And to **Cécile** in particular: thank you so much for your consistent support during difficult moments (I will fondly remember our daily phone calls when both of us were deep in the thesis writing) and for taking me with you for incredible summer bike trips where I also got the chance to meet Alex and Emilie. You are a treasured friend, and I know I can always rely on you whenever life becomes intense.

**The Twitch co-working channel of BlackThunder1332**, for providing one of the most effective productivity tools during my thesis writing. Thanks to Nico and the whole PUB community for the focused Pomodoro sessions, accountability, encouragement, and fun breaks on marathon days.

Last but certainly not least, **my family**: Mom, Dad, Jeanne, Bérangère. Thank you for everything. Our Sunday evening family Zoom calls were an essential support when I felt overwhelmed or stuck. Special thanks to Mom for patiently listening to my spontaneous brainstorming sessions on the phone when I was working through a research problem. I feel extremely lucky to have you as my family !

*Lausanne, July 11, 2025* Laurent Colbois

# Abstract

Face morphing attacks exploit vulnerabilities in automated face recognition systems through the creation and submission of fake reference passport pictures. These images blend the faces of two individuals, allowing them to share the same passport and posing a significant security threat. Initially, face morphs were generated using relatively simple image processing techniques. However, recent advancements in generative artificial intelligence (AI) have enabled the creation of particularly realistic fake images, commonly known as deepfakes.

Deepfakes present an escalating challenge for society: generation techniques are rapidly advancing, diversifying, producing increasingly realistic results, and becoming more accessible to the public through widespread exposure and the availability of user-friendly commercial generators. The creation of synthetic faces or the alteration of real ones has been a primary application of deepfake technology since its inception. Consequently, the general issues posed by deepfakes are directly relevant to biometric security, particularly face morphing attacks. Deepfake technology has directly influenced the evolution of morphing attacks, leading to the emergence of novel 'deep' morphs. Traditional face morphs have been extensively studied, resulting in various methods that can detect them with reasonable accuracy. In contrast, deep morphs are a new phenomenon that requires further investigation, which is the central focus of this thesis.

We examine the risks posed by modern morphing attacks by assessing their ability to deceive existing face recognition systems and discussing their potential to fool human operators who process submitted passport pictures. Additionally, we introduce a novel deep morphing attack, which exploits models designed for reconstructing faces from face recognition embeddings. This demonstrates how techniques originally designed for face reconstruction can be repurposed to generate highly effective morphing attacks with minimal effort. Overall, our findings highlight that deep morphs can deceive face recognition systems as effectively as traditional morphs, underscoring the need for updated countermeasures.

To this end, we first focus on developing detection systems capable of identifying both traditional and deep morphs. We then extend our efforts to source attribution, which aims to identify the algorithm used to create a given fake media, thus going beyond mere detection. With the increasing diversity of morphing attack generators, source attribution provides valuable insights that can be integrated into investigations, serving as a partial indicator of the authorship of a fake media.

We conceptualize our studies within the framework of forensic science, framing them as a search for relevant traces in AI-generated images. These traces, whether semantic (visible anomalies, such as irregular facial features) or statistical (imperceptible patterns in noise or in frequency domains), are crucial for both detecting and attributing morphing attacks. In practice, we leverage representation learning techniques to discover salient traces in a data-driven manner. We consider not only end-to-end supervised learning approaches (dominant in the field), which train deep neural networks on examples of real and fake images for binary or multiclass classification, but also other methodologies that more clearly separate the trace extraction process from the actual modeling. Specifically, we evaluate three families of feature representations: (1) handcrafted features, manually designed using expert knowledge; (2) supervised features, learned through end-to-end supervised training on labeled real and fake images; and (3) foundational features, extracted using large vision models trained exclusively on real images, making them independent of any specific attack type. We evaluate and discuss the strengths and weaknesses of these three major families of representations in terms of discriminative power (for detection and source attribution), generality, robustness, and interpretability.

Overall, we find that approaches based on foundational features are particularly promising, especially in terms of generality and robustness, whereas supervised approaches tend to overspecialize to the types of deepfakes seen during training. Notably, foundational features enable the development of fully attack-agnostic detection models by providing a fine statistical description of natural images, allowing any attack to be detected as an out-of-distribution sample. Moreover, despite being developed using only real images, foundational features also provide a surprisingly well-structured space in which attacks from various generators are nicely clustered, enhancing their potential for source attribution tasks, even in open-set scenarios. Our findings demonstrate that deepfake analysis systems can greatly benefit from visual foundation models, which may surpass traditional supervised approaches. Given the rapid advancement of vision foundation models, these techniques offer a promising, forward-compatible strategy for deepfake forensic analysis, in an era where deepfake generators also keep progressing.

# Résumé

Les attaques par *morphing* de visages exploitent les vulnérabilités des systèmes automatisés de reconnaissance faciale par la création et la soumission de fausses photos d'identité de passeport. Ces images combinent les visages de deux individus, leur permettant de partager le même passeport et posant ainsi un risque de sécurité important. À l'origine, les *morphs* faciaux étaient générés à l'aide de techniques relativement simples de traitement d'image. Cependant, les récents progrès de l'intelligence artificielle (IA) générative ont permis la création d'images particulièrement réalistes, communément appelées hypertrucages (*deepfakes* en anglais).

Les hypertrucages représentent un défi croissant pour la société : les techniques de génération progressent rapidement, se diversifient, produisent des résultats toujours plus réalistes et deviennent accessibles au grand public grâce à la démocratisation des outils et à la disponibilité de générateurs commerciaux faciles d'accès. La création de visages synthétiques ou la modification de visages réels constitue une application centrale des hypertrucages depuis leur apparition. Par conséquent, les problématiques générales posées par ces technologies concernent directement la sécurité biométrique, et en particulier les attaques par *morphing* de visages. Les progrès en matière d'hypertrucage ont directement influencé l'évolution des attaques par *morphing*, donnant naissance à de nouveaux *morphs* "profonds" (*deep morphs*), qui se distinguent fondamentalement des précédents. Alors que les *morphs* faciaux traditionnels ont été largement étudiés et que des méthodes permettent de les détecter avec une précision raisonnable, les *morphs* profonds constituent un phénomène émergent nécessitant des recherches plus approfondies, qui sont au cœur de cette thèse.

Nous examinons les risques posés par les attaques modernes de *morphing* en évaluant leur capacité à tromper les systèmes de reconnaissance faciale existants et en discutant leur potentiel à induire en erreur les opérateurs humains traitant les photos d'identité soumises pour les passeports. De plus, nous introduisons une nouvelle attaque par *morphing* profond exploitant des modèles conçus pour reconstruire des visages à partir d'*embeddings* de reconnaissance faciale. Cela démontre comment des techniques initialement développées pour la reconstruction faciale peuvent être facilement réutilisées pour générer des *morphs* très efficaces. Globalement, nos résultats mettent en évidence que les *morphs* profonds peuvent tromper les systèmes de reconnaissance faciale aussi efficacement que les *morphs* traditionnels, soulignant ainsi la nécessité de contre-mesures adaptées.

Dans ce cadre, nous nous concentrons d'abord sur le développement de systèmes de détection capables d'identifier à la fois les *morphs* traditionnels et profonds. Nous étendons ensuite nos recherches à une tâche d'attribution de la source, qui vise à identifier l'algorithme ayant généré un média falsifié, allant ainsi au-delà de la simple détection. Avec la diversité croissante des générateurs d'attaques par *morphing*, l'attribution de la source fournit des informations précieuses qui peuvent être intégrées aux enquêtes judiciaires, constituant un indicateur partiel de l'origine d'un média falsifié.

Nous intégrons notre étude dans le cadre de la science forensique, en la concevant comme une recherche de traces pertinentes dans les images générées par l'IA. Ces traces, qu'elles soient sémantiques (anomalies visibles, telles que des caractéristiques faciales irrégulières) ou statistiques (signaux imperceptibles dans le bruit ou dans le domaine fréquentiel), sont essentielles tant pour la détection que pour l'attribution des attaques par *morphing*. En pratique, nous exploitons des techniques d'apprentissage de représentations pour extraire ces traces de manière guidée par les données. Nous considérons non seulement les approches d'apprentissage supervisé de bout en bout (dominantes dans le domaine), qui entraînent des réseaux de neurones profonds sur des exemples d'images réelles et falsifiées pour des classifications binaires ou multiclasses, mais aussi d'autres méthodologies qui séparent plus distinctement le processus d'extraction des traces du processus de classification. Plus précisément, nous évaluons trois familles de représentations : (1) les caractéristiques développées manuellement, conçues à partir de l'expertise humaine ; (2) les caractéristiques supervisées, apprises à travers un entraînement supervisé de bout en bout sur des images réelles et falsifiées annotées ; et (3) les caractéristiques fondationnelles, extraites de modèles de fondation de vision entraînés exclusivement sur des images réelles, les rendant indépendantes de tout type d'attaque spécifique. Nous analysons et comparons ces trois familles de représentations en fonction de leur pouvoir discriminant (pour la détection et l'attribution de la source), leur généralité, leur robustesse et leur interprétabilité.

Nos résultats montrent que les approches basées sur les caractéristiques fondationnelles sont particulièrement prometteuses, notamment en termes de généralité et de robustesse, tandis que les approches supervisées ont tendance à se spécialiser excessivement sur les types d'hypertrucages rencontrés lors de l'entraînement. En particulier, les caractéristiques fondationnelles permettent de concevoir des modèles de détection entièrement indépendants des attaques, en fournissant une description statistique fine des images naturelles, permettant ainsi de détecter toute attaque comme un échantillon hors-distribution. De plus, bien qu'elles soient développées uniquement à partir d'images réelles, ces représentations forment un espace vectoriel remarquablement structuré où les attaques issues de divers générateurs se regroupent naturellement, renforçant ainsi leur potentiel pour les tâches d'attribution de la source, y compris dans des scénarios *open-set*. Nos résultats démontrent que les systèmes d'analyse d'hypertrucages peuvent tirer un grand bénéfice des modèles de fondation visuels, qui pourraient surpasser les approches supervisées traditionnelles. Compte tenu de la progression rapide des modèles de fondation en vision, ces techniques offrent une approche évolutive et adaptable pour l'analyse forensique des hypertrucages, à une époque où les générateurs d'hypertrucages continuent également d'évoluer.

# Contents

# Introduction

In late 2017, a Reddit user under the pseudonym *u/deepfakes* starts posting fake videos where the face of the main actor is replaced by that of another celebrity. Shortly thereafter, they publicly release their tool, which utilizes recently developed techniques from the generative deep learning field to automate the face-swapping process. The era of **deepfakes** has started.

This event, while building on technology developed in preceding years, serves as a significant milestone marking the onset of the deepfake phenomenon. Not only did it coin the term *deepfake*, but it also encapsulated core aspects of the deepfake issue: the use of generative artificial intelligence for media manipulation, with applications ranging from artistic to malicious, notably in the realm of impersonation.

It was nevertheless just a precursor of what was to come. In the following years, generative AI progressed rapidly, enabling the increasingly simple and realistic creation of tampered media or even fully synthetic content. From faceswap videos requiring a relatively constrained source video for the editing to work and targeting only celebrities due to data requirements, visual deepfakes evolved towards increasingly flexible applications: generation of fully synthetic personas, precise editing of static faces to alter their age, haircut, expression, or virtually any aspect of them, face puppetry enabling the generation of a controllable animated face from a still picture, up to the generation of fake pictures matching a given caption of choice. In impersonation applications, the amount of reference data from the impersonated target has also become increasingly small, putting most of the population at risk of seeing themselves impersonated by a deepfake.

And this is just restricted to the case of face-related visual deepfakes. Modern generators can represent a virtually unending diversity of content, and other modalities such as audio (generated music, voice cloning) and text (spectacularly demonstrated by tools such as ChatGPT). Progress has not plateaued yet, as 2024, for instance, has seen major developments in the field of caption-controlled video generation.

In parallel with the explosion of deepfake types and applications, access to deepfake generators has become highly democratized. Since around 2022, the technology's increasing maturity has led to a rapidly growing market for commercial deepfake generators, which require minimal technical expertise to use. While these commercial generators can implement safety measures to limit malicious use, competitive open-source solutions also exist, providing the public with unrestricted access to state-of-the-art generative AI, without safeguards.

In summary, deepfakes are becoming increasingly diverse and realistic, and access to generators is becoming more widespread. This has led to a surge in deepfake content across the web and society at large. Deepfakes pose significant issues by facilitating both preexisting and novel forms of criminality, such as misinformation, fraud, and defamation, while also eroding trust in media evidence. Consequently, it is imperative to develop robust countermeasures that can improve the reliability of assessing origin and degree of authenticity[2] of media.

### Deepfakes under the spectrum of the forensic trace

To frame our research objectives, we situate our work within established principles of forensic science. The Sydney Declaration [1], which aims to formalize the foundations of forensic science, establishes as its first principle that "*Activity and presence produce traces that are fundamental vectors of information.*" These traces are the central subject of study in forensic science.

We adopt the formal definition of the Trace proposed by [2], which defines the Trace as the "full modification of a Scene, subsequently perceptible, resulting from an Event of interest." More specifically, it introduces the notion of the Tangible Trace, observable in the real world, as the *perceptible difference* between the current state of a Scene and a hypothetical state where an Event of interest did not occur. This difference is perceptible only up to a finite level of precision. What is commonly referred to as a "trace" in practice is thus an *observed facet* of this complete, underlying Trace, serving as the *fundamental vector of information* highlighted by the Sydney Declaration.

Under this framework, the relationship between deepfakes and the notion of Trace manifests in two distinct ways:

1. **Deepfakes *as* Traces**: Deepfake media itself (e.g., a synthetic image or manipulated video) represents an *observed facet* of a Tangible Trace. Its presence in a Scene (such as the Internet or a social media profile) constitutes a perceptible modification (an adjunction) resulting from an underlying Event (e.g., the creation and dissemination of that media). This trace acts as a fundamental vector of information, indicating broader activities (or suspected ones) such as misinformation campaigns or fraud attempts. In this sense, the deepfake is a trace of that activity.

2. **Deepfakes *containing* Traces**: Here, the process of creating a deepfake using a generative algorithm is the Event of interest. This Event modifies the digital Scene, which

[2] Note the notion of *degree of authenticity*: in many cases, authenticity is a continuous concept which does not enable the binary labeling of media as either authentic or inauthentic. We discuss this issue in more details in section 1.5

in this case can be restricted to the media itself (data and metadata). Following the first principle of the Sydney Declaration, we assume that the generation process with a deep generator necessarily leaves behind traces within the content itself, artifacts of the generation process. These intrinsic facets (e.g., statistical anomalies, inconsistencies) are *perceptible differences* between the current state of the Scene and the hypothetical state where the media would have been produced genuinely (i.e., through the use of an actual physical camera). These contained traces are the fundamental vectors of information that signal the synthetic origin of the media.

The second perspective, focusing on the traces *contained* in deepfakes, is the primary focus of this thesis. The objective of deepfake detection and source attribution is to detect, extract, and analyze these traces, interpreting the information they provide about the source and nature of the media's creation.

More precisely, what type of information can we hope to gain from observing traces contained in deepfakes? Considering the lens of the eight elementary functions of the trace as proposed by [3] provides elements of an answer. For deepfakes, three primary functions seem particularly relevant and define the different types of analysis tasks that can be considered:

1. **Gain information on the nature and profile of the source**: Assessing the likely origin of the analyzed media involves evaluating indicators that suggest whether it originates primarily from a physical capture device, or instead exhibits characteristics consistent with a synthetic generation process. This assessment aligns with the task of **detection**.
2. **Identify the source**: In cases where the media has been confirmed as a deepfake, identifying the exact tool or algorithm used to create it provides additional insights. In the simplest case, this process involves selecting among a finite list the generative algorithm most likely to have created the media of interest. This corresponds to a **closed-set attribution** task. In most cases however, one must expect the possibility that *none* of the considered algorithms is the true source. In that case, the analysis should also be able to detect the involvement of a novel unknown generative algorithm. This corresponds to a **novelty detection** task.
3. **Establish links with other traces**: When dealing with deepfakes generated by unknown algorithms, it might be possible to determine whether multiple deepfakes were produced by the same generation algorithm, even if the algorithm itself remains unidentified. For example, in misinformation networks utilizing fake social media profiles, this analysis can hint at which profiles might belong to the same network. This corresponds to a **source linking** task. As a whole, the joint tasks of novelty detection and source linking form a more general **open-set attribution** task, which aims to link together media with the same origin, while properly detecting novel sources.

Achieving these goals requires exploiting traces present within the suspected media itself. These traces can manifest either at the data level (e.g., artifacts from the generation process, visual inconsistencies compared to genuine media) or at the metadata level (e.g., absence of or inconsistencies in device capture information, signs of manipulation using specific software). Methods based on metadata traces will be discussed in Chapter 1, but the experimental part of this thesis will focus on methods based on data traces. The motivations for this choice are as follows: first, there are no fundamental differences between the types of exploitable metadata traces in the case of deepfakes and that of earlier media manipulation attempts. The goal of this thesis is to develop a better understanding of the specificities of deepfake manipulations, hence a focus on data traces makes more sense. Second, metadata traces are relatively straightforward to falsify by a malicious and informed actor. In the case of data traces, however, falsification methods are currently less obvious, making the exploitation of those traces potentially more robust. Finally, the main case study of this thesis will involve manipulations that transition from the digital domain into the physical one, as the manipulated samples are expected to be printed and re-scanned, as will be detailed later. This domain switch can naturally result in a complete loss or overwrite of metadata information, thus leaving only data traces as salient information.

Traces within the data can be categorized into *high-level* or *semantic* traces, which are irregularities in the media content that may be perceptible by humans (e.g., lack of blinking in face-swap videos, extra fingers in generated diffusion images depicting humans), and *low-level* or *statistical* traces, which are artifacts from the generation process that are not necessarily visible to humans (e.g., unrealistic frequency content, abnormal noise residuals). However, in practice, literature on deepfake detection and attribution often focuses on data-driven methodologies: because it is feasible to generate large datasets of deepfakes automatically, one can train a deep neural network for the task of interest based on a wide set of examples and obtain a relatively effective detector. After this process, one cannot describe **what** exactly the exploited traces are. By itself, the network has learned an effective representation of data. An ideal representation satisfies the following properties:

- **Discriminative power**: The representation should effectively distinguish between different classes for the task at hand (e.g., detection, attribution). It must encapsulate all relevant information from the traces in the data while ignoring irrelevant details.
- **Generality**: The representation should be applicable across a wide range of deepfake generation algorithms, maintaining its relevance and effectiveness.
- **Robustness**: The representation should withstand various data alterations, such as resampling, compression, or other post-processing operations, without losing its effectiveness.
- **Interpretability**: The representation should provide insights into the nature of the traces it captures, allowing us to understand **what** the relevant traces in deepfakes are.

The core focus of this thesis is the exploration and evaluation of such representations.

### Case study : face morphing attacks

A particular field in which deepfakes are introducing new challenges is biometric security. With the increasing use of automated face recognition systems, for instance, in automated border control at airports, there is a concern that deepfake technology can be used to attack and fool these systems. Throughout this thesis, we focus on a specific type of such attacks: **face morphing attacks**. A face morph is a fake face picture that combines the identities of two contributing subjects and is used as a reference picture during a passport application process. When the attack is successful, both contributing subjects can share the same passport, being successfully verified as the "rightful" owner by both human evaluators and automated face recognition systems.

While face morphs have existed before the era of deepfakes, advances in generative AI are leading to the emergence of a variety of novel morphing algorithms, which can challenge preexisting detection systems. It is thus crucial to update and improve existing security measures against morphing attacks to ensure proper handling of deepfake-based morphs.

Moreover, the development of countermeasures against face morphing attacks is also an interesting problem in the broader context of image forensic analysis. It presents two particular challenges that are still open problems. First, during the submission of a passport picture, the picture will usually be printed and rescanned, a post-processing operation that might drastically alter traces left in the image by potential digital alteration. Increasing robustness to post-processing operations is an important problem in deepfake forensic analysis, but it is usually approached by including post-processed data as part of the modeling phase, an approach that is not trivial for print-scan post-processing, due to the time necessary for creating the post-processed data. Second, face morphing algorithms predating the deep learning era are still highly relevant and among the most effective attacks, which means the developed security systems should handle both previous attacks and novel deepfake-like ones. Image manipulations not based on deepfake technology are nowadays called **cheapfakes**, and the development of systems handling both deepfakes and cheapfakes is another problem of interest.

We chose to adopt a narrower focus on morphing attacks rather than general deepfake forensic analysis. This specific focus does not limit the complexity of the issues encountered; instead, it introduces unique challenges, such as the significance of print-scan post-processing and the relevance of cheapfakes. Additionally, this approach allows us to generate all experimental data in-house, ensuring fully reliable ground truth for attack labeling. This reliability is crucial for studying the generalization of detection systems, understanding the limitations of attribution systems, and enabling systematic experimentation where all important factors are controlled.

We can thus summarize our central research question as follows:

> *What are the most effective representations of image data, satisfying requirements of discriminative power, generality, robustness, and interpretability, in the context*

*of face morphing attack detection and attribution?*

We will particularly focus on three major families of representation learning methodologies: expert-informed, manually designed representations (**handcrafted features**); autonomously learned representations through end-to-end training of a neural network on a large dataset of deepfake examples (**supervised features**); and autonomously learned representations through training of a foundation model using web-scale amounts of real data only (**foundational features**). The strengths and weaknesses of each approach will be highlighted and discussed.

The thesis is structured in two parts. The first part aims to provide important contextual information related to deepfakes (Chapter 1) and morphing attacks (Chapter 2). Those chapters provide an overview of those technologies, both from a generation and detection point of view, and discuss the types of challenges posed by them. The second part presents four research contributions, focusing on morphing attacks specifically. In details:

- **Chapter 1** provides an overview of deepfakes, their history, and the challenges they pose. It also reviews existing work in deepfake detection and attribution and introduces fundamental concepts in machine learning and deep learning essential for understanding the thesis.
- **Chapter 2** focuses on introducing morphing attacks and the unique challenges they present. This chapter also details the generation methods and the creation process of the experimental datasets used in this thesis, which was a crucial part of the technical work. Finally, it introduces relevant metrics for evaluating the effectiveness of morphing attacks and the performance of morphing attack detectors.
- **Chapter 3** experimentally motivates the necessity of morphing attack detection and attribution systems by evaluating the risks posed by existing attacks. It also demonstrates the applicability of face template inversion to morphing attack generation, proposing a novel family of attacks and illustrating that new attacks are still to be expected in the current research landscape. This chapter is based on a published conference article [4] and a journal extension currently under review.
- **Chapter 4** presents an evaluation of representations for face morphing attack detection. It compares two main families of representations: **handcrafted texture-based features** and **attack-aware supervised features**, developed by training deep neural networks on examples of morphing attacks. This chapter is based on a published conference article [5].
- **Chapter 5** builds on the insights from the previous chapter to enhance evaluation protocols for morphing attack detection systems. It also experiments with a new family of representations: **attack-agnostic foundational features**, developed by training deep neural networks on bona fide data using a distinct pretext task, and compares them to supervised features. This chapter is based on a published conference article [6].

- **Chapter 6** assesses representations for face morphing attack attribution, focusing on the attack-agnostic features developed in the previous chapter and evaluating their effectiveness for the closed-set attribution task.
- **Chapter 7** concludes the work by a discussion of the overarching takeaways of the experimental work, notably by discussing the strengths and weaknesses of the three considered representation learning methodologies in terms of discriminative power, generality, robustness, and interpretability. Future research directions resulting from this thesis are then highlighted. Finally, we conclude with a broader discussion on effective mitigation strategies against deepfakes given the current state of the art in deepfake forensic analysis, taking notably in account both technical and non-technical measures.

# Contextualization Part I

# 1 The era of deepfakes

In this chapter, we provide the necessary context to understand the field of deepfake and how they are approached in forensic science. First, we introduce fundamental notions in deep learning, which are essential for understanding both the creation and detection of deepfakes. We focus on the concept of **representation learning** and its various forms. Next, we give an overview of existing deepfake generation methods, highlighting the rapid growth of the field and the need for an increasingly broad definition of deepfakes. We then discuss the main societal challenges posed by deepfakes. Following this, we present approaches for deepfake forensic analysis, emphasizing methods for extracting salient traces in deepfake media through representation learning. We also describe the main open challenges in the area. Finally, we motivate our later focus on face morphing attacks, which are a significant issue in face biometric security, and also present interesting challenges from a forensic point of view.

## 1.1 Introductory notions

Deepfake generation, detection, and attribution systems all rely on Deep Neural Networks (DNNs). Conceptually, a DNN is a **parametric function** $f_\Theta$ whose parameters $\Theta$ are optimized through a process called *training* or *learning*. This optimization aims to satisfy a specific objective using a dataset of examples, known as the training data. The objective is characterized by a loss function $\mathcal{L}$, which measures how much the behavior of the DNN deviates from the desired outcome, using the training samples as reference. The training process involves updating the parameters $\Theta$ to minimize the average value of the loss $\mathcal{L}$ over the training dataset.

For a simple analogy, consider the common task of fitting a polynomial to a set of 2D data points $(x, y)$ using the least squares method. Here, $f_\Theta$ represents the polynomial, $\Theta$ denotes the polynomial coefficients, and $\mathcal{L}$ is the squared error:

$$\mathcal{L}(\Theta, x, y) := \big(y - f_\Theta(x)\big)^2$$

DNNs have several distinctive characteristics:

- The function $f_\Theta$ is constructed as a complex composition of many simpler subfunctions, i.e., $f_\Theta := f^1_{\Theta_1} \circ f^2_{\Theta_2} \circ \cdots \circ f^N_{\Theta_N}$. These subfunctions are referred to as **layers** of the neural network. The term **deep** in deep neural networks (and by extension, deep learning) originates from the use of many layers, in contrast to early neural networks which had only a few layers and are now sometimes called **shallow** networks.

- The layers are designed to be differentiable, which ensures that the overall function $f_\Theta$ is differentiable. This property allows for parameter optimization through a **gradient descent** algorithm, known as **backpropagation**.

- Due to the depth of modern neural networks, they contain a huge number of parameters (millions a few years ago, and billions nowadays), making them computationally expensive to optimize and requiring access to massive amounts of training data. Advances in computing hardware (particularly Graphical Processing Units (GPUs)) and the increasing availability of large-scale datasets over the past decade have facilitated significant breakthroughs in deep learning.

- As a result, DNNs are now capable of modeling highly sophisticated functions, demonstrating surprisingly strong robustness and generalization capabilities (i.e., the ability to perform well, to some extent, even on data that is somewhat distinct or distant from the training dataset).

Applications of DNNs are diverse, depending on the choice of input data and the function to be modeled. We are particularly interested in their use for the following tasks:

- **Classification**
  - **Input**: Image, audio, text, or signal.
  - **Output**: Probability of belonging to each class in a predefined list.
  - **Example**: Differentiating pictures of cats from pictures of dogs.
- **Representation Learning**
  - **Input**: Image, audio, text, or signal.
  - **Output**: Low-dimensional vector encapsulating all the relevant information in the input while ignoring irrelevant information.
  - **Example**: From a face picture, extracting a vector that depends on the identity of the person, but not on unrelated factors such as facial expression or background.
- **Generative Modeling**
  - **Input**: “Simple” probability distribution.
  - **Output**: “Complex” probability distribution.

  - **Example**: Mapping a multidimensional normal distribution (simple) to a distribution of face images (complex). This enables the random sampling of face pictures by sampling an input vector from the multidimensional normal distribution and then feeding this input to the trained DNN.

It is important to note that these high-level examples mask much of the complexity involved in training an effective DNN. Specifically, the precise design of the layers and their composition (referred to as the network's **architecture**), as well as the formulation of an appropriate loss function, are challenging issues that lie at the heart of fundamental deep learning research.

### 1.1.1 Representations of data

There is an important connection between representation learning and classification, as finding an appropriate representation of data facilitates the development of effective classifiers. We want to highlight the three main categories of approaches for finding good representations of data, which we will consider throughout this thesis.

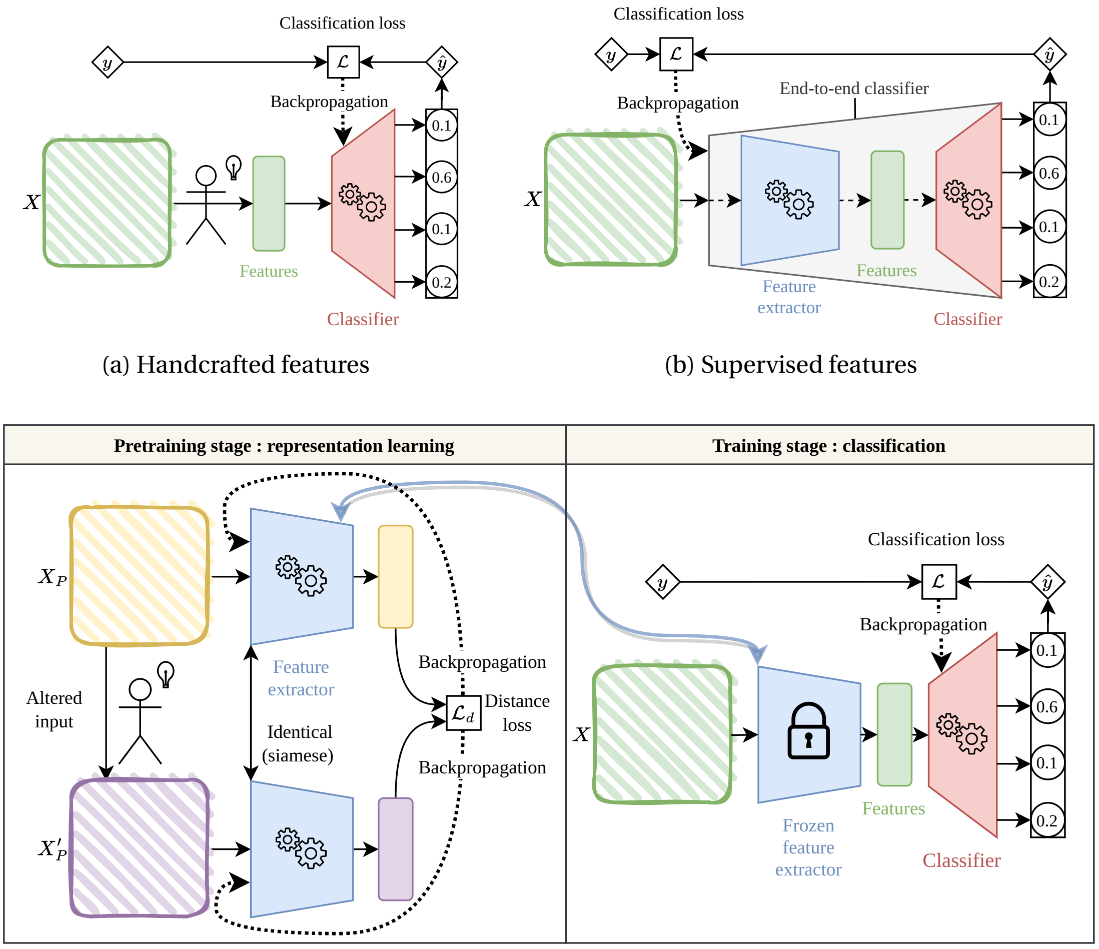
#### Handcrafted features

Before the era of neural networks, machine learning decomposed classification systems into two components. First, features of interest were extracted from the raw input data. In the case of image recognition, for instance, these features could be edge maps computed by applying edge detection filters to the input images. A shallow classifier such as a Support Vector Machine (SVM) [7] was then trained on these feature representations of the data. This process is illustrated in Figure 1.1a.

Importantly, the choice of which features to consider depended on the expertise of a human, based on their domain knowledge and experiments. Hence, we use the term **handcrafted features**. A good set of features is one that summarizes all the relevant information in the input data while rejecting task-irrelevant information. In other words, a good set of features for a classification task should typically be compressive and especially discriminative, meaning that representations from data samples belonging to different classes should be distinct enough to enable the final classifier to model proper decision boundaries. In this context, the "representation learning" itself was achieved by the human experimenter through a mixture of domain knowledge and empirical experimentation.

#### Supervised features

A paradigm shift began in 2012 with the introduction of AlexNet [8], a Convolutional Neural Network (CNN) that significantly outperformed previous machine learning techniques in image recognition, marking the beginning of the current era of deep learning. In this paradigm, the DNN is conceived as a deep classifier that learns to perform classification directly on raw



(a) Handcrafted features

(b) Supervised features

(c) Foundational features

Figure 1.1: Different approaches for finding effective representations of data, and application of these representations in a classification context. **Handcrafted features**: Relevant data representations are manually designed by an expert based on domain knowledge and experiments. Classification relies on simple classifiers trained on those representations. **Supervised features**: A complex DNN classifier is trained end-to-end on raw data, with internal representations serving as feature representations. **Foundational features**: An extractor is pretrained for representation learning using a pretext task (often self-supervised learning). The extractor is then frozen, and a task-specific classifier is trained on top.

data, without the need for any form of manual preprocessing. The process is illustrated in Figure 1.1b.

This type of learning is called *end-to-end*, illustrating that no human-designed intermediate representation is involved between the raw input data and the final classification output. Nevertheless, because of the layered structure of the network, it is still possible to conceive it as a two-component model, with early layers acting as feature extractors and late layers acting as a shallow classifier. Commonly, the entirety of the network until its penultimate layer is considered as the feature extractor, while the last layer is considered as a simple linear classifier. This allows the trained model to be used as a fixed feature extractor for other tasks by dropping the last layer. The output space of the feature extractor component is typically quite structured when considering data that is part of the training distribution.

As these representations are learned through the supervision of the training process by class labels, we call them **supervised features**.

### Foundational features

A third methodology has become prevalent over the past few years, deriving partly from research focused on pure representation learning. This approach aims to find data representations that are useful across a wide variety of tasks, contrasting with end-to-end supervised learning, which necessitate a full retraining of the DNN when the task or data domain are altered.

A common algorithm for pure representation learning is **self-supervised learning**, as presented in [9] and illustrated in Figure 1.1c. This process involves a training dataset of unlabeled samples, such as images. These images are subjected to various distortions (e.g., mirroring, blurring, cropping). A feature extractor DNN is then trained with the objective that two distorted versions of the same image should be close to each other in the feature space, while distorted versions of different images should be far apart. This implies that the network learns to ignore "meaningless" alterations caused by the distortions.

To evaluate the relevance of these representations for various tasks, one can freeze the parameters of the extractor, use it to extract representations of the training dataset for the task of interest, and then train a shallow classifier on top of these features.

The initial representation learning stage is named **pretraining**. The second stage, where the shallow classifier is trained, is called **transfer learning**. The classification task of interest is named the **downstream task**, and the shallow classifier is referred to as the **downstream classifier**, or as a **probe** in recent literature. In some variants, the parameters of the feature extractor are not frozen during transfer learning, but rather updated in order to slightly adapt the representations, a process known as **fine-tuning**.

Empirically, this methodology becomes extremely effective when the size of the pretraining dataset and the pretrained model scales up. This has led to works like [10] and [11], which, by pretraining models on web-scale amounts of image data, learn representations that enable very good downstream performance on a wide variety of tasks. In other words, the learned representations are very general.

These models are costly to train in terms of data and computation, but this training only has to happen once. Then, they can be shared publicly (including architecture and weights) and adapted by other researchers for their own purposes with much fewer data and compute. Due to the generality of these pretrained models, they are now called **foundation models**, and we will refer to the data representations they produce as **foundational features**.

### 1.1.2 Generative models

DNNs have also been successful for generative modeling. Two approaches in particular are of high relevance for the type of synthetic media considered in the experimental part of this thesis. They are respectively **Generative Adversarial Networks (GANs)** and **diffusion models**. We present here the fundamentals behind those generative approaches.

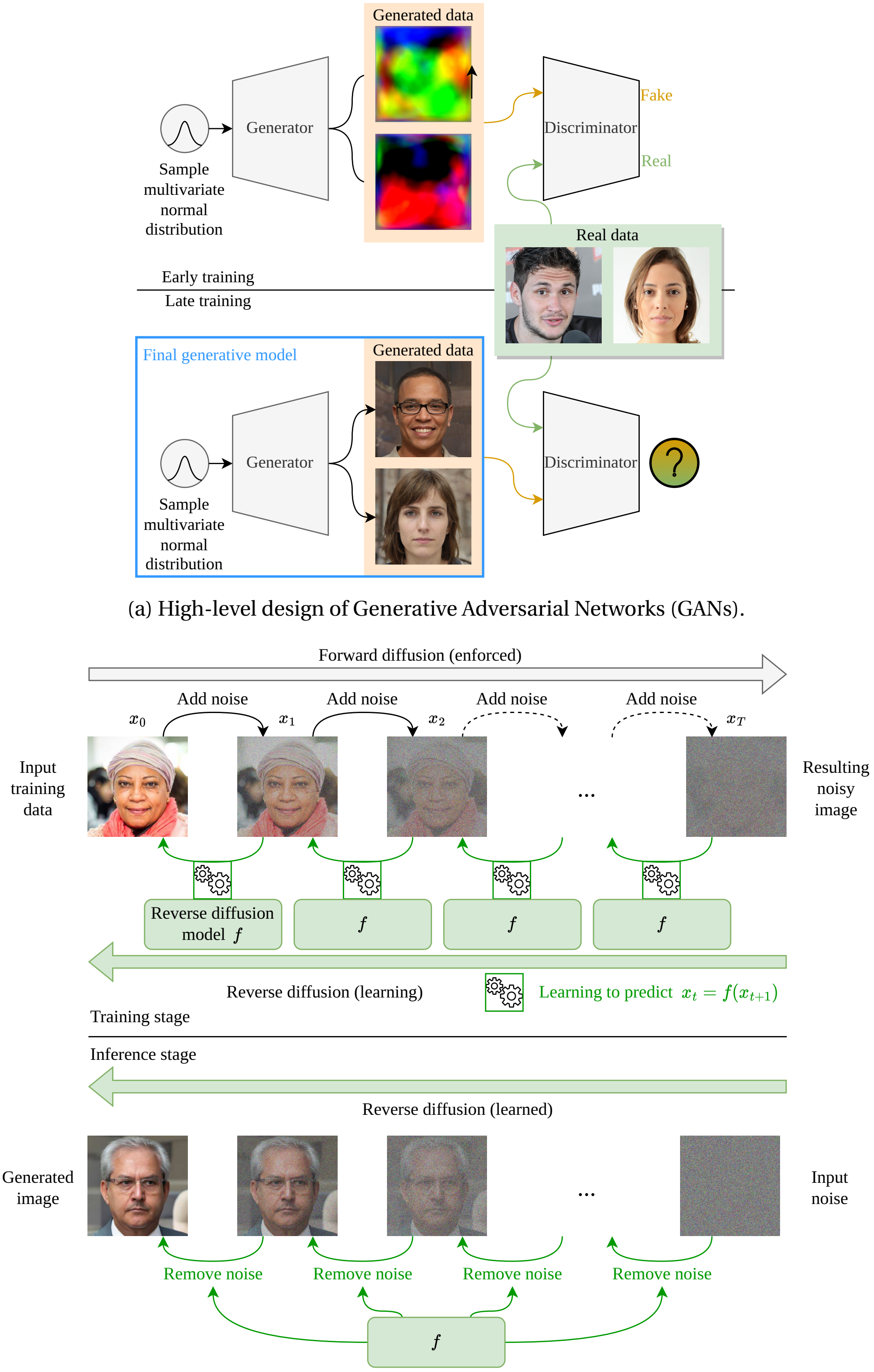
#### Generative Adversarial Networks

GANs derive their name from the concept of *adversarial* training, which involves the simultaneous training of two DNNs with opposing objectives. These two networks are known as the **generator** and the **discriminator**.

The generator's goal is to learn to map a simple distribution (e.g., a multivariate normal distribution), from which sampling is straightforward, to a complex and more "abstract" distribution, such as the distribution of "images of faces". As illustrated in Figure 1.2a, before training, the generator produces abstract-looking images, whose style is solely dependent on the architecture and initial weights of the model.

The discriminator's objective is to distinguish between synthetic images produced by the generator and real images from a dataset. This is achieved using a classification loss that penalizes the discriminator for misclassifying fake images as real and real images as fake. While the discriminator is optimized to minimize this loss, the generator is trained in contrary to *increase* it, hence the term "adversarial".

At the beginning of the training, the discriminator's task is relatively easy due to the significant difference between real and generated images. However, as the discriminator improves, the generator must produce increasingly realistic outputs to continue fooling the discriminator. This creates a cat-and-mouse dynamic where both the generator and the discriminator progressively get better at their respective objectives. At the end of the training, it is expected that the generator produces content that is nearly indistinguishable from real images, and the discriminator struggles to differentiate between the two.



(a) High-level design of Generative Adversarial Networks (GANs).

(b) Diffusion

Figure 1.2: High-level design of diffusion models.

A significant strength of this method is that it does not require explicit specification of "what makes an image realistic". The discriminator learns this autonomously, and the type of content the generator can create is influenced only by the choice of the training data.

### Diffusion Models

Diffusion models are based on training a model to learn to reverse a process known as **forward diffusion**, which can be summarized as a Gaussian random walk in pixel space. As illustrated in Figure 1.2b, the idea is to take images from the training set and add a slight amount of Gaussian noise to them. This process is repeated multiple times, resulting in a sequence of images $x_0, x_1, \ldots, x_T$ that become progressively noisier. In practice, it is possible to compute $x_T$ directly from $x_0$ without calculating intermediate steps.

The main objective of the model is to enable reverse diffusion, i.e., to invert this diffusion process. This is achieved by training the model to predict the noise contained in an image. During a single training step:

1. An image $x_0$ is taken from the training dataset.
2. A noisy version $x_t$ at a randomly picked time $t$ is computed by adding $t$ steps of Gaussian noise. The total added noise is denoted by $\epsilon$.
3. The reverse diffusion network $f$ produces an output $f(x_t, t) := \hat{\epsilon}$, which aims to be an accurate prediction of $\epsilon$.
4. The loss is formulated as the L2 pixel-wise norm:

$$\mathcal{L} := ||\hat{\epsilon} - \epsilon||_2 = ||f(x_t, t) - \epsilon||_2.$$

After training, the reverse diffusion model can accurately predict the noise in the image for any input time step. This allows starting from completely random noise $x^{\mathcal{N}}$ (not generated by adding noise to a real image), using the model to "predict the noise in the image" and removing it to obtain a *fully novel* synthetic image $x^S$:

$$x^S := x^{\mathcal{N}} - f\left(x^{\mathcal{N}}, T_S\right)$$

Note that this still requires selecting an arbitrary step $T_S$ and pretending that the initial noisy image is the noisy version of a fictional image, after $T_S$ steps of forward diffusion. In practice:

- The denoising process itself is done iteratively: only a portion of the predicted noise is removed from the initial noisy image, and this denoising step is repeated many times, resulting in an increasingly pristine output. This process empirically leads to better outcomes than performing a single denoising step.

- The total number of denoising steps, $T_S$, can be flexibly chosen. Low values lead to low-quality images, while high values result in high-quality images but at the cost of increased computation time.

## 1.2 Deepfakes: a widening definition

The breakthroughs in generative modeling over the past decade have enabled the creation of hyperrealistic fake media in a highly automated manner. This type of fake media is now commonly referred to as **deepfakes**. However, the definition of the term has evolved since its inception. In this section, we present a brief history of deepfakes, from the early days of deep generative modeling to the current era of democratized commercial deepfake generation tools.

The term "deepfakes" originates from the pseudonym of a Reddit user in 2017, who first proposed the implementation of a face-swapping algorithm using a DNN. This algorithm could replace the face of a person in a video with that of another person. Initially, the term was specifically restricted to face-swapping videos. However, over the last five years, its increasingly common use to describe a wide range of AI-based generative algorithms has led to a much broader accepted definition. The following definitions from the literature illustrate this evolution, listed in increasing order of generality:

1. "*The very popular term 'DeepFake' is referred to a deep learning based technique able to create fake videos by swapping the face of a person by the face of another person.*" [12]
2. "*A deepfake is content, generated by an artificial intelligence, that is authentic in the eyes of a human being. The word deepfake is a combination of the words 'deep learning' and 'fake' and primarily relates to content generated by an artificial neural network, a branch of machine learning.*" [13]
3. "*Believable media generated by a deep neural network.*" [13]

In this thesis, we adopt the broader definitions 2 and 3. This broad definition includes three main categories: visual deepfakes, audio deepfakes, and text deepfakes. This thesis will focus on visual deepfakes solely, and in particular face-related ones. However, we will provide an overview of all categories, given they all are a cause for concern for society.

### 1.2.1 Origins of visual deepfakes

Modern visual deepfakes are based on three major families of generative models: variational autoencoders [14], generative adversarial networks [15], and diffusion models [16].

These original works established the theoretical foundations behind these models, and the training processes required to approximate a given distribution of images. However, they also had several practical limitations:

- **Limited image realism**: Early deep generative models produced images that were clearly artificial. Generating high-resolution images with realistic levels of sharpness was particularly challenging.
- **Lack of controllability**: Early models allowed for the random sampling of synthetic images but lacked control over specific semantic details of the output. For instance, in the context of face images, there was no control over demographic attributes, age, or expression.
- **Limited generation diversity**: Early models were relatively specialized, typically trained on a single category of objects (e.g., faces, cats, churches). As a result, distinct models were needed to generate different types of content.

Subsequent advancements have mitigated these issues through the discovery of various engineering tricks and the increased availability of larger, higher-resolution, and better-annotated image datasets.

GANs were the first generative models to achieve significant breakthroughs in image realism and resolution. Notable examples include the Progressive Growing GAN (ProGAN [17]) and the subsequent StyleGAN [18], which were capable of generating highly realistic synthetic images at resolutions up to 1024×1024. Progress in these aspects for Variational Autoencoders (VAEs) came later through several parallel works, such as VQ-VAE2 [19] and NVAE [20], although these models did not achieve resolutions higher than 256×256. The first major iteration on diffusion models [21] directly enabled good realism at a 256×256 resolution, and was followed by further improvements that facilitated even higher resolutions. For example, Latent Diffusion [22] combines the diffusion process with an autoencoder-like image downsampler/upsampler, performing the diffusion in a lower, less computationally expensive resolution. Another example is [23], which explores model design choices necessary to facilitate higher resolution diffusion directly in the original pixel space.

In order to improve controllability of the generation, two families of approaches exist. First, in some cases, the input space of the generator showcases some sort of structure which can be exploited : instead of sampling images completely at random, a more careful selection of the input enables some level on control of the generation. When such structure of the input space is available, it is common to rename it a **latent space**, and the inputs **latent vectors** or simply **latents**.

For instance, it is common that sampling in a small neighborhood of an initial latent will generate images which are visually similar to each other. An even stronger property is that of **perceptual continuity**, which is satisfied when moving regularly in one direction of the latent space leads to regular perceptual change in the output image. This is notably the case with the StyleGAN model [18], which enables this property through the introduction of a learned intermediate latent space, with better structural properties emerging as byproduct of the training. The model also benefits from another property which the authors call **linear separability of attributes**. It means that one can find directions in the latent space along which it is possible to move while altering a *single* semantically interpretable attribute of the output image. In the case of a generator trained solely on faces, those directions can for example correspond to only alter the age or the hair color of the output face. In this context, some necessary work is needed to actually "cartography" the latent space in order to use it for controlled generation.

Another orthogonal approach for enabling controllability is by introducing a so-called **conditioning** of the generator, through the form of an additional input to the generator. The generator then has two inputs : a latent input which is sampled randomly, and a conditioning input which must influence the generated outcome. The resulting **conditional generator** is trained to ensure not only the realism of the generated image, but also that the generated image satisfies the input condition.

The generator can be conditioned on a class label [24], which enables the user to generate any of the classes considered during training reliably, just by switching the value of the class label.

The generator can also be conditioned on an image, which can be useful when developing a generator intended to only alter portions of a real image instead of generating a fully synthetic one. For instance, the StarGAN model [25] enables face attribute editing through a GAN model conditioned on both face images and class labels. Another notable example is ControlNet[26], a method enabling very strong controllability of the output of diffusion models, through conditioning with a wide variety of inputs such as edge maps, depth maps, segmentation maps, or pose skeletons. Finally, maybe the most popular approach available to the large public is conditioning through text, where the content of the generated image will match a provided input caption [27].

### 1.2.2 The commercial era

All the aforementioned generation methods have been developed by the research community. The training and inference code, as well as the pretrained weights of the models, were typically open-sourced, enabling the community to reuse and build upon them. This process still required a relative amount of technical know-how. However, as the field of generative AI reached maturity, there was a significant shift with the appearance of commercial models. This greatly facilitated the penetration of the technology into the public in two distinct ways: first, by greatly lowering the barrier of entry for any individual to produce their own deepfake content, and second, by increasing the visibility of the technology to a public that might have been unaware of it beforehand.

This trend was primarily led by OpenAI with the release of the text-to-image generation model

DALL-E 2[1] [28]. Initially published in March 2022, the model was not open-sourced but was made accessible to a selected group of researchers and artists to explore its potential applications and discover creative use cases for showcasing the technology. In June 2022, DALL-E 2 was made available to the public through a subscription-based closed beta, and by September 2022, it was accessible without a waitlist. The model was also integrated into Bing Search in early 2023, further increasing its visibility.

The success of DALL-E 2 sparked a surge of interest, leading to the rapid development of other proprietary text-to-image models by various industry actors, such as MidJourney, Adobe Firefly, and Google's Imagen. In contrast, the company Stability AI, advocating for the continuation of open-source practices, open-sourced their own text-to-image diffusion model, Stable Diffusion. This model was widely adopted by online tech communities, who greatly enhanced its usability (e.g., by creating a user-friendly web interface[2]), expanded its functionalities (identity consistency, localized editing of real images, superresolution, etc.), and released alternative sets of weights to specialize the model for specific types of content or styles (e.g., through the sharing platform CivitAI[3]). While these tools still require some technical skills to use compared to fully commercial models, they offer significantly more functionalities and flexibility than their commercial counterparts.

Since then, major actors have kept upgrading their models following progresses in generative AI research, improvements in data availability and quality, and investment in computing hardware to train increasingly large models, the size often being correlated with achievable realism and quality of the generated content. As an outcome, the year 2024 has seen the first very compelling attempts at text-to-video generation, such as the SORA model from OpenAI (still private), Stable Diffusion Video from StabilityAI, or the Gen1, Gen2 and Gen3-Alpha models from RunwayML.

### 1.2.3 Face deepfakes

We also aim to provide an overview of the main categories of face-related deepfakes, one of which will be our main focus, and clearly distinguish between these categories. Let us consider $X_s$ and $X_t$, facial data (image or video) of a source and target identity, respectively, and $X_g$ a deepfake generated from $X_s$ and $X_t$. We can identify the following subtypes of face-related deepfakes. For each subtype, we provide references to some of the most significant works having contributed to developing or increasing the realism of this type of face manipulations. However, we note that the field of face deepfake generation is very large, and state-of-the-art methods are often the result of combining many tricks developed both in research works or as contributions from the open-source community. Those references are thus not intended to be exhaustive, as several extensive review works already exist for that purpose [12], [29].

[1] https://openai.com/dall-e-2/

[2] https://github.com/AUTOMATIC1111/stable-diffusion-webui

[3] https://civitai.com/

**Identity swap / Face replacement [30], [31], see Figure 1.3**

It consists in replacing the face of the target identity by that of the source. Optionally, only a portion of the face is replaced, such as the mouth or the eyes (gaze). This can be done in two slightly different ways:

- Face Transfer : $X_g$ is obtained by fully replacing the content of $X_t$ by the content of $X_s$. The resulting $X_g$ will have the head, body, and background of $X_t$, but with the face of $X_s$.
- Face Swap : $X_g$ is obtained by replacing the content of $X_t$ by the content of $X_s$, but let it still be driven by $X_t$. This means that the resulting $X_g$ will have the head, body, background, **and expressions** of $X_t$, but with the face of $X_s$.

**Face reenactment [32], [33], see Figure 1.4**

It consists in maintaining the face of the target identity but alter its expressions to imitate the expression of the source identity. I.e, $X_g$ is an altered version of $X_t$ with expressions driven by $X_s$. Optionally, only a portion of the face is altered, such as the mouth or the eyes (gaze).

**Attribute editing [25], [34], see Figure 1.5**

It consists in changing specific semantic attributes of $X_t$, such as the facial hair, the glasses, or the age.

**Entire face synthesis [18], see Figure 1.6**

It consists in creating a fully synthetic face $X_g$ with no target as basis (i.e., a completely imaginary identity). Possibly, several face images of the same synthetic identity can be created with several factors of variation such as pose or expression, as in [35].

**Morphing attack [36], [37], see Figure 1.7**

It consists in creating a synthetic face $X_g$ that is composed of a mixture of face attributes from both $X_t$ and $X_s$, such that this $X_g$ would match both $X_t$ and $X_s$ in a face recognition setting. Morphing attacks will be explained in more details in Chapter 2.

### 1.2.4 Other modalities

Deepfakes are not restricted to visual content but also extend to other modalities, mainly audio and text. While these are outside the scope of this thesis, they pose societal and security challenges that are at least as significant, if not more so, than visual deepfakes. Therefore, we provide a brief overview of existing methodologies for sake of completeness.

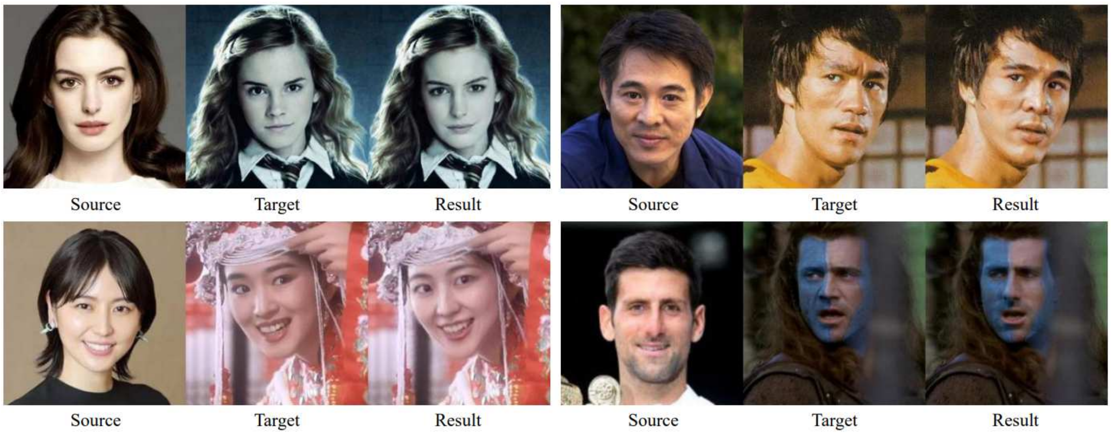


Figure 1.3: Face swapping (from [31])

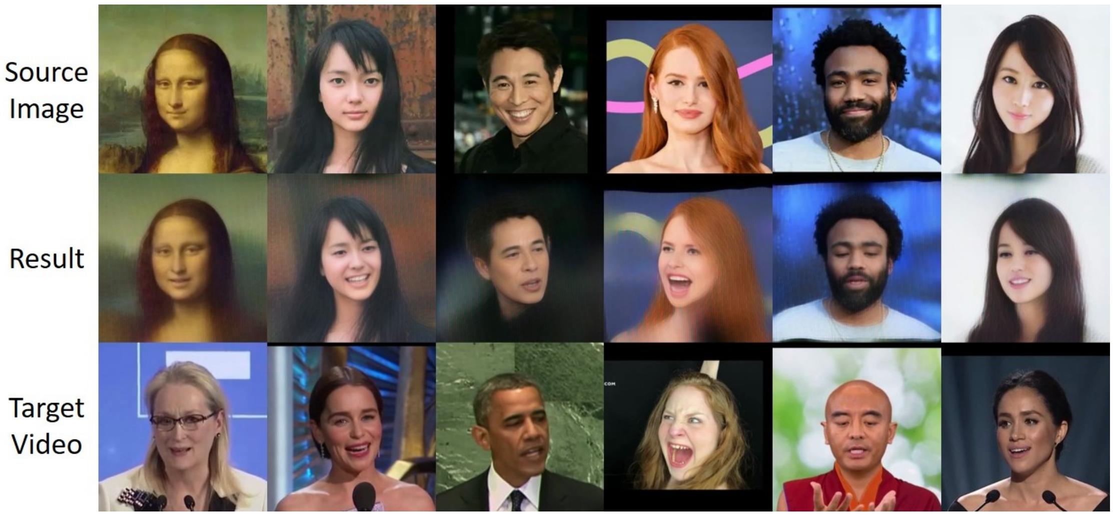


Figure 1.4: Face reenactment (from [33])

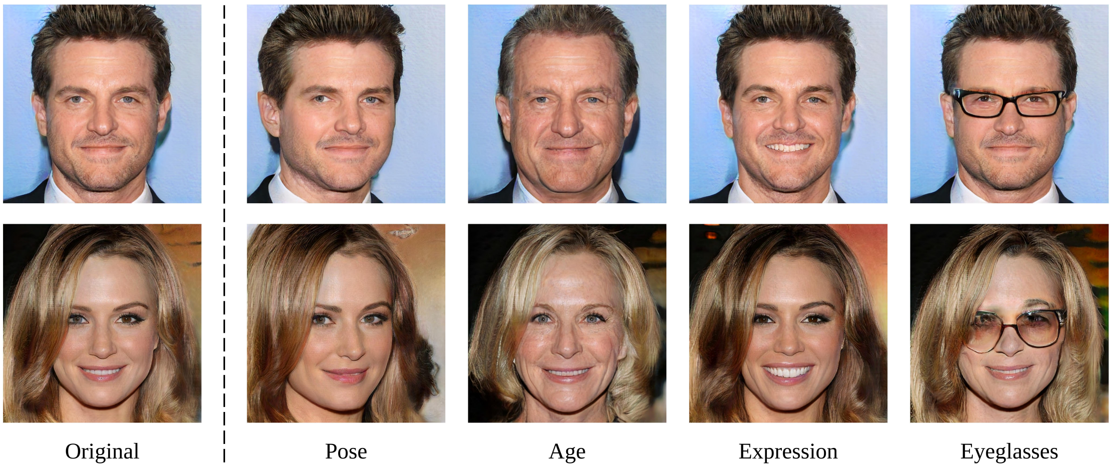


Figure 1.5: Attribute editing (from [34])

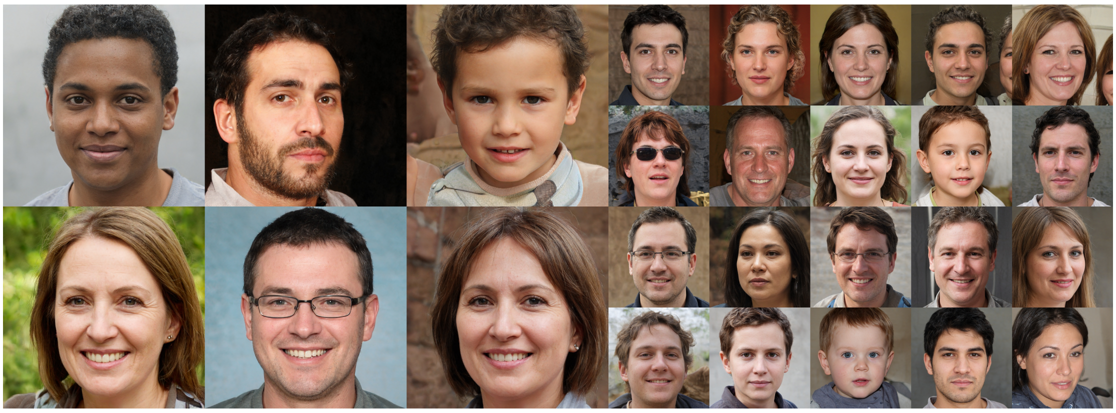


Figure 1.6: Entire face synthesis (from [18])

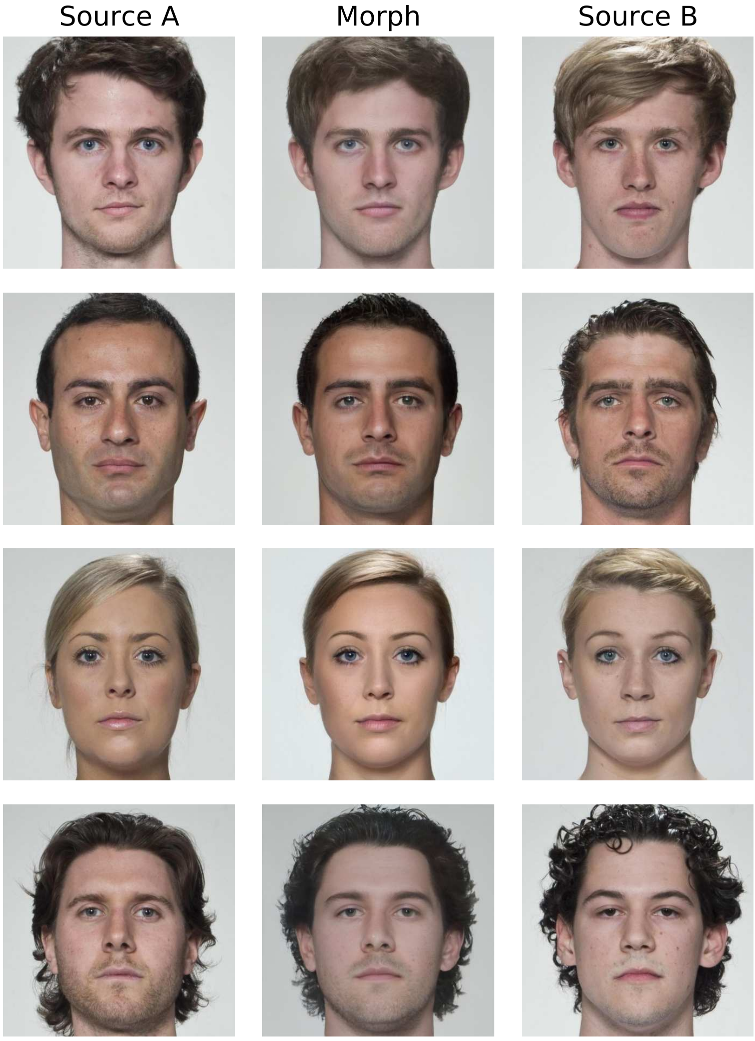


Figure 1.7: Morphing attacks

Typical audio deepfakes consist of either synthetic speech or synthetic music. Commercial solutions for synthetic speech, such as ElevenLabs, offer not only text-to-speech generation but also voice cloning, which enables the synthetic recreation of an individual's voice, including timbre, tone, and accent. This technology can be combined with video face-swapping techniques to generate audiovisual deepfakes, which impersonate both someone's face and voice in a forged video [38].

Commercial solutions for synthetic music generation (e.g., SUNO, UDIO) enable users to specify lyrics and style for a song, which is then generated by the model. While not necessarily as problematic as speech deepfakes, such methods are likely to drastically alter the landscape of commercial music, and raise important legal questions about plagiarism and copyright infringement, given the need to use existing songs for training data.

Finally, text deepfakes have become the most omnipresent, especially following the release of ChatGPT by OpenAI in late 2022. This event marked a significant milestone in the era of Generative AI, with quick and widespread adoption, and applications across a huge variety of domains involving text or coding. Similar to the impact of text-to-image models like DALL-E 2, the success of ChatGPT triggered industrial competition, leading to the development of other Large Language Models (LLMs) by both established and emerging companies. Notable examples include Gemini by Google, Llama by Meta, and Claude by Anthropic. Meta, in particular, has adopted a more open approach by releasing the weights of all iterations of their Llama model. Although the training code and data remain proprietary, this policy allows the community to fine-tune the initial model for specialized applications.

A similar approach has been taken by the Chinese company DeepSeek, which, in early January, released DeepSeek-R1, the first reasoning LLM with open weights. In addition to achieving competitive performance compared to OpenAI's models, it is also distributed under a more permissive license than Llama's iterations. This has allowed open-source communities, such as HuggingFace, to collaborate on fully replicating the model and its training process, thereby further accelerating the narrowing of the performance gap between open-source and commercial models.

In conclusion, the phenomenon of deepfakes has been steadily increasing in magnitude since its inception in 2017. With the shift to commercial products leading advancements in the field around 2022, this growth is not only steady but seemingly accelerating. The number and diversity of available models are expanding, technical breakthroughs are occurring more frequently, adoption is becoming more widespread, and applications are multiplying. It is unclear when this wave of generative AI will slow down, and it seems likely that at least several more years of progress are ahead, particularly in video generation models, which are still in their infancy. The main limitation the field might face is the increasing demand for computational power: most current improvements are strongly correlated with increasing model size, requiring more compute for training and inference. Manufacturing sufficient hardware and providing the necessary power could become a significant bottleneck. However, this does not rule out the possibility of further advancements through theoretical breakthroughs, by improving models' design or efficiency. In any case, deepfakes are becoming commonplace and are here to stay. Significant efforts are needed to manage their impact on the job market, on creative arts, and on various forms of criminality facilitated or even enabled by the democratized access to powerful generative AI models.

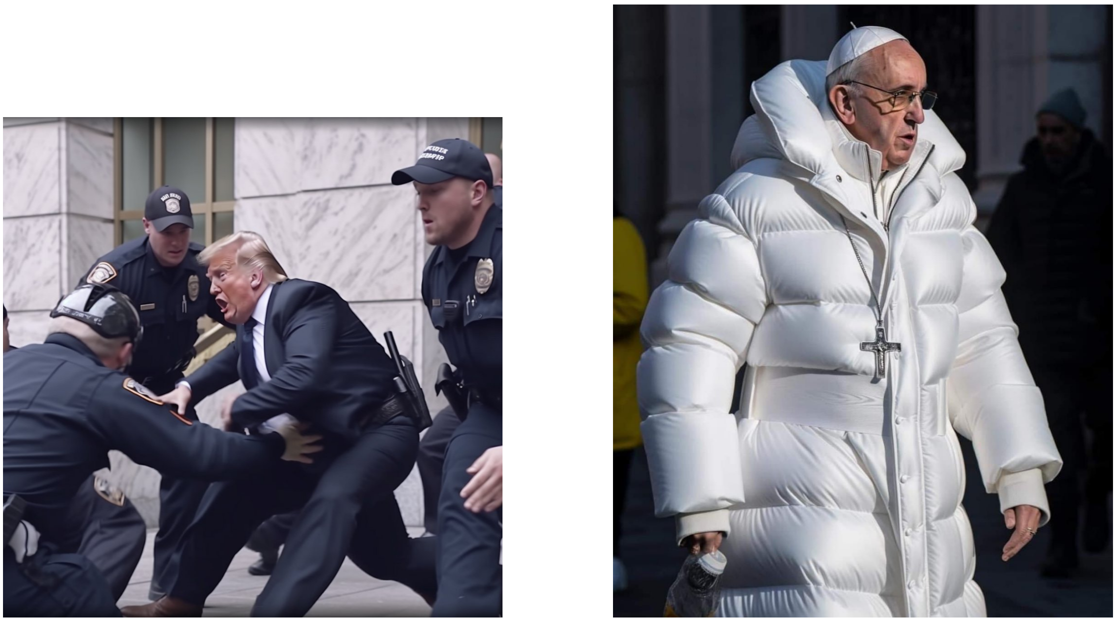
## 1.3 Challenges posed by deepfakes

### Non-consensual pornography

Very early on, the development of deepfake generators has strongly been driven by Internet communities, which has quickly led to unethical and unlawful applications. During the inception of the term *deepfakes* itself, it referred to fake pornographic videos obtained by swapping the faces of celebrities onto those of adult film actresses [39]. The face swapping script was made public on Reddit, enabling anyone to create similar videos with very little work or computational resources. While the Reddit administration eventually banned the sharing of this type of content on their website, it was too late to stop the propagation of non-consensual deepfake pornography on the Internet. This specific model could only perform one-to-one face swapping between two specific identities, and thus could only target celebrities as a large dataset of images from the target was needed to train the model. However, the continuous progress in many-to-many deepfake generation (that can swap between any pair of identities) can nothing but eventually lead to similar applications able to target virtually anyone, thus greatly increasing its harmful potential.

Two years later, another anonymous programmer shared an app called DeepNude, that when fed an image of a clothed woman would output an undressed version of the image [40]. The creator shut the app down a few days later, and did not share the source code. However, it still inspired the creation of similar open-source projects, some of which are still available to anyone to this day. Note also that this application is already able to target virtually anyone. A notable case happened in 2024 in a school in Spain, when AI-generated naked pictures of 20 girl students were circulated on WhatsApp group [41]. Those pictures had been generated using a commercial mobile application.

More recently, text-to-image generation models have been extended with capability to finely control identities in the generated images to match a real person (e.g., the DreamBooth algorithm [42]), enabling to represent them in a variety of contexts, and requiring only few example images from the target identity. This technology uses only open-source tools and can also be used for generating fake pornographic images.



(a) Fake arrest of Donald Trump

(b) Pope Francois' fake coat

Figure 1.8: Viral fake images of famous people generated using identity-conditioned text-to-diffusion images

### Impersonation

Deepfakes can be used to impersonate public figures such as politicians (Barack Obama [43], Donald Trump [44], Volodymr Zelensky and Vladimir Putin [45]). Until now, very realistic looking impersonation videos that have circulated on social media have never actually aimed to fool the public, as they have been accompanied by a disclaimer stating the content was fake. Often, those videos have been developed by researchers, media networks or FX artists, with the aim of showcasing the progress in deepfake generation realism, and raise awareness about the topic.

To be successful, this type of attack requires a particularly high realism of the deepfake, given the amount of scrutiny it will be exposed to. Therefore, some particular care and manual editing might be required, thus limiting the ease of creation for a single person. However, it is quite reasonable nowadays to imagine a small team producing a realistic enough fake video in order to smear a political figure. Initially, a core limitation was that the audio content of the video had to be recorded by a human imitator, but this limitation has disappeared following progress in audiovisual deepfakes generation. Finally, this type of attack being very public by definition, it can be quickly refuted by the impersonated target. However, it could still cause massive damage in the meantime if it gets diffused at a critical time, such as right before an election.

Here as well, identity-conditioning of text-to-diffusion models also enables the creation of very diverse fake pictures stealing someone's likeness (either a public or private persona). Figure

1.8 showcases some examples of such images which went viral after their creation.

Deepfake-based impersonation can also be used for fraud purposes, rather than for political defamation or misinformation. During a meeting in late 2023 at UNIL, representatives from the cantonal police of Argovia explained that fake phone calls impersonating voices are the main type of deepfakes they have encountered in real criminal cases. Generally, these are used for scamming the call receiver: by impersonating a close family member or a work supervisor, the scammer attempts to urgently request money or ask for a fake work-related financial transaction. Text deepfakes also facilitate such fraudulent activities as they enable the generation of personalized and highly realistic phishing emails on a large scale.

### Pushing misinformation with fake profiles

Fully synthetic faces are sometimes used to serve as profile picture for fake accounts on social media platforms. This has recently been observed on various platforms (Facebook, Twitter, Instagram, YouTube) where the approach was used to build several pro-China misinformation or propaganda networks [46], [47]. The process is called **astroturfing**, where a certain message is pushed while keeping the true original source hidden, aiming instead to give the impression that the message is originating from grassroots movements and is spread by popular opinion. In [47] for example, the account of a completely fake biology expert called Wilson Edwards had been denouncing pressure applied by the US on the World Health Organization to orient blame towards China regarding the exact origin of the COVID-19 pandemic. Those claims were then reshared by a network of both fake and (seemingly) real other users, and finally in Chinese media. This phenomenon is accentuated with the emergence of text deepfakes, which enable the creation of fake automated social media profiles that can convincingly mimic real users, interact with each other, and engage with genuine users. This can create the illusion of grassroots movements, which can be used to manipulate public opinion for political or other purposes.

### Fooling border control with morphing attacks

A common preoccupation linked to Automatic Border Control systems based on face recognition (FR) is the eventuality of so-called morphing attacks. The attack scenario (see also Chapter 2) consists in an individual who want to conceal his facial identity and an accomplice to first create a morph, which is a fake picture mixing their two faces. The accomplice then submits the results as its passport photo when registering into the FR system. In the case of a successful attack, *both* the accomplice and the criminal can successfully authenticate against this altered reference, enabling the individual to freely navigate through border control. While morphing attacks can be prevented by acquiring the reference image in a controlled context, several countries still allow providing one's own picture. A group of activists already demonstrated the feasibility of the attack in 2018, when applying for a passport in Germany [48]. Recent models have emerged recently to generate morphs with the help of a deep neural network [37], [49], thus adding deep morphing attacks to the list of possible nefarious applications of deepfake technology. Morphing attacks will be introduced more extensively in Chapter 2.

#### Reducing the trust in genuine media

As the exposure of the public to highly realistic fake media continues to increase, it is likely that overall trust in image and video evidence will decline. In the United Kingdom, a study has shown that while people are not typically misled by deepfakes, they become very uncertain when evaluating the authenticity of some content [50]. This uncertainty leads to increased cynicism and skepticism towards media that were once considered trustworthy.

This phenomenon can sometimes lead to significant real-life impacts. For example, in 2018, after Gabon's President Ali Bongo had not been seen in public for several months, the population began to suspect he was ill or even dead. The government released a real video of the president for his New Year's address, but suspicions arose that it was actually a deepfake used as a cover-up. This led to an unsuccessful coup by the military, with the video being cited as one of the motivations for the attempt.

The loss of trust in information circulating online might be partially mitigated by public media organizations taking a central role in vetting this information. Given that this vetting process demands significant resources and organization, it is unlikely to be achievable by individuals or automated algorithms alone. However, official fact-checking organizations might struggle to earn and maintain the trust of the public, as any single error could severely damage their reputation as a trustworthy source. Additionally, the vetting process could be time-consuming and challenging to implement without causing major delays incompatibley with the required speed and fluidity of news communications.

Doubt about the genuineness of a video has also appeared in a legal context. In 2021, during a defamation trial against French comedian and political activist Dieudonné, the defense claimed that the core video evidence containing the problematic statements was actually a deepfake [51]. Further investigation was necessary to assess the validity of this claim. With the increasing realism of deepfakes, it is becoming easier to use this “deepfake defense” as a way to discard visual and audio evidence. Yet, the invalidation of such claims is also getting increasingly challenging.

In this context, it is crucial to ensure legal professionals have a clear understanding of the capabilities and limitations of media-altering technology and are guided through the process of assessing the credibility and value of media produced or collected as part of investigations. Initiatives focusing on these goals have already been initiated, such as the TRUE Project in the UK, which notably proposes a guide for judges on the evaluation of digital open-source imagery [52]. A main challenge in this area will be the speed at which the field of generative AI is evolving; the training of legal professionals should ideally be a continuous effort, ensuring they are regularly updated on developments in the field.

### Legal challenges

The rise of deepfakes has facilitated certain types of offenses, raising questions about whether existing legislation is sufficient to address these issues or if new laws are needed. In Switzerland, several parliamentary motions have been proposed, urging the Federal Council to establish new legislation specifically targeting deepfake-related problems. These include the creation of general regulations in civil, penal, and administrative domains [53], protection against pornographic deepfakes [54], and possibly the prohibition of certain apps or online services, such as nudification technology [55].

So far, the Federal Council's stance is that existing legislation already covers all problematic uses of deepfakes. For example, actions involving fraud, impersonation, or the creation of fake pornographic content are already addressed through technology-agnostic laws. Additionally, they anticipate the future establishment of more general AI regulations that could help govern deepfakes within a broader framework of problematic applications.

Organizations like the Foundation of Technology Assessment (TA-SWISS) are working to identify potential gaps and issues in existing legislation that might not adequately cover critical applications of deepfakes. For instance, TA-SWISS [56] points out that deepfake offenses predominantly occur and spread on major international platforms, making it difficult to enforce nationally set regulations and principles aimed at either removing identified fake content or at least transparently signaling it. They notably suggest the implementation of advisory centers aimed at supporting victims of cybercrime, for which individual confrontation to online platforms is unlikely to be effective.

As another example, [57] examines whether the production of pornographic deepfakes is punishable under the current Swiss criminal code. At the time of writing, they point out that the law primarily penalizes the exposure of pornographic content to unwilling individuals, which does not typically extend to the distribution of fake pornographic content on online platforms. However, they note that offenses against personal honor could potentially be used to address the production and dissemination of fake pornography.

Fortunately, recent legislation has been introduced that should help make pornographic deepfakes a punishable offense. The first is the article on identity theft (Art. $179^{\text{decies}}$ CP), effective as of September 2023. The second is the article on the unauthorized sharing of private sexual content (Art. 197a CP), which came into effect in July 2024. Although primarily aimed at addressing revenge porn, the latter law is likely applicable to deepfake pornography, as it penalizes 'the sharing of private sexual content [...] without the consent of the person recognizable therein."

Legal questions surrounding deepfakes extend beyond their punishability. Another significant issue is copyright law. The widespread use of commercial, closed-source generative models trained on undisclosed but extensive datasets raises concerns about potential copyright infringement involving samples included in the training data. For instance, this issue has led to an ongoing copyright infringement lawsuit between the New York Times and OpenAI [58]. Additionally, there is the challenge of determining whether content generated by artificial intelligence in a fully automated manner (without human creative intent or intervention) qualifies for copyright protection.

As deepfakes continue to diversify and proliferate, these legal issues are likely to grow in significance, necessitating clear and specific legislation.

## 1.4 Countermeasures

### 1.4.1 General approaches

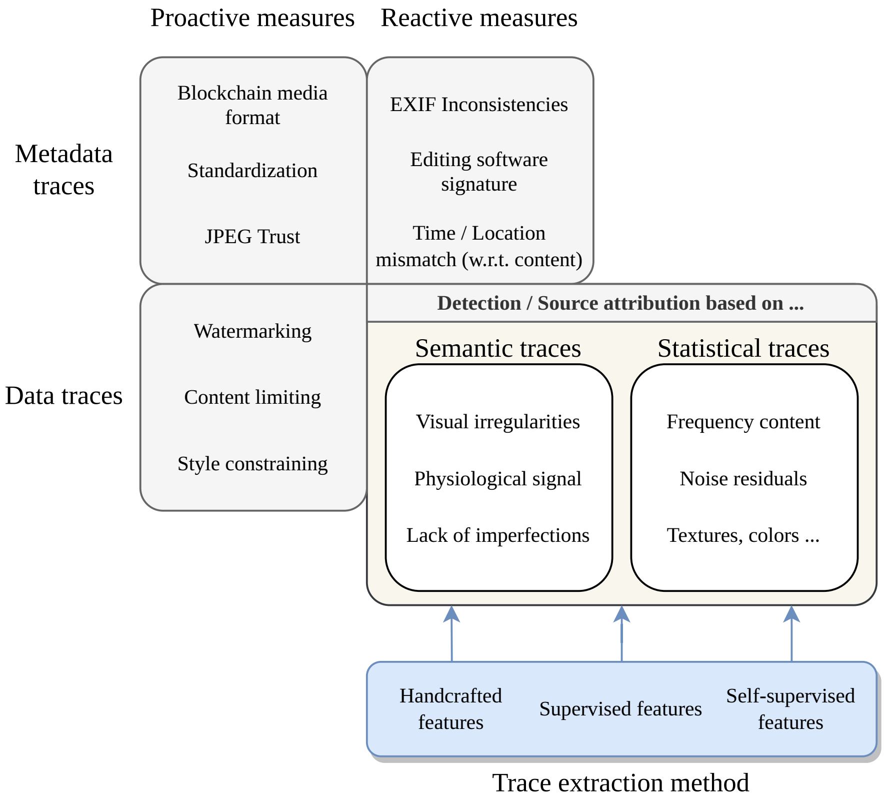
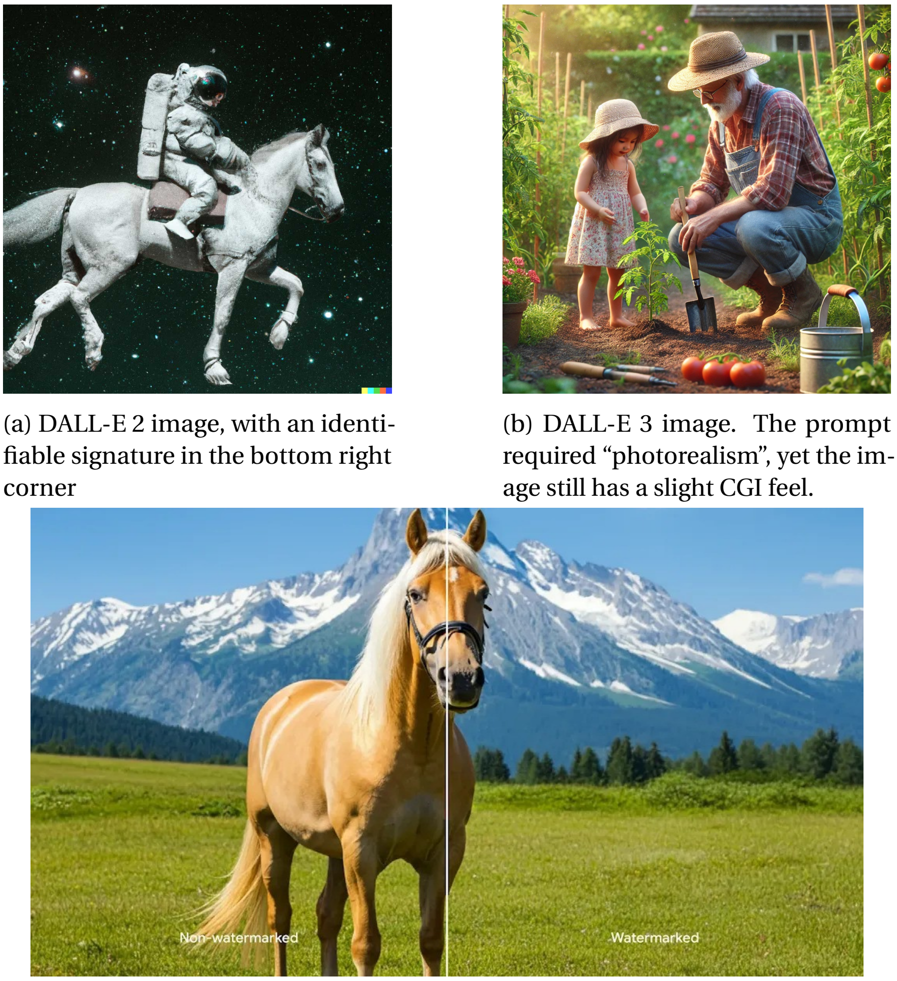


Figure 1.9: Overview of the different categories of countermeasures against deepfakes.

Effective countermeasures against deepfakes require methods for reliably evaluating the likely authenticity of a given media (task of **detection**) and potentially identifying the likely algorithm used to generate a fake media (task of **source attribution**). This second step can provide complementary investigative information regarding the activities involving deepfakes. For example, identifying that an image was produced by a common commercial generator could direct the investigation towards examining the server logs of the generator. These logs might be linked to a specific user account or credit card information, providing additional



(a) DALL-E 2 image, with an identifiable signature in the bottom right corner

(b) DALL-E 3 image. The prompt required "photorealism", yet the image still has a slight CGI feel.

(c) Image watermarked with SynthID (non-watermarked on the left half, watermarked on the right half) : the watermark is imperceptible

Figure 1.10: Proactive methods for enabling deepfake identification, through the inclusion of (perceptible or imperceptible) signature in the generated images, or by restricting the style or content.

leads for identifying the perpetrator. Even with private generators, source attribution can still be useful for linking series of deepfakes together or excluding certain suspects from the investigation if it is clear they could not have had access to the considered generator.

We first provide an overview of possible approaches for detecting and attributing deepfakes, summarized in Figure 1.9. At a high level, methods can be categorized along two axes: whether they focus on metadata or on the data itself (e.g., pixel values in an image), and whether they are proactive (requiring active measures before and during the creation of the deepfake generator or media) or reactive (relying on currently exploitable properties of both genuine and generated media). We then focus in more detail on reactive methods that exploit data traces.

Methods exploiting metadata traces aim to identify missing or inconsistent information within the media file. A robust proactive methodology involves tracking all stages an image goes through from its creation to its final form. This requires the media format to carry information about the initial source (camera, generator, or artist) and all subsequent edits or post-processing steps. Ideally, this information should be difficult to falsify. In [59], the authors propose the use of blockchain technology, which enables secure tracking of "transactions" (i.e., all operations an image undergoes). More recently, the Coalition for Content Provenance and Authenticity (C2PA) project proposes the development of technical standards for the certification of the source and history of media content [60], in a way that prevents falsification by malicious actors. Similarly, the JPEG Trust initiative[4] is working towards developing a new media format standard that incorporates similar features and is notably aligned with the C2PA specifications. However, relying on new standards for ensuring media legitimacy has the significant drawback of requiring a long timespan for widespread adoption.

Reactive methods based on metadata are already achievable, given that existing media formats carry a large amount of potentially useful information. For instance, [61] study how video manipulations can leave traces in the metadata contained in the MP4 video wrapper, which can be detected. [62] focus on image provenance analysis, aiming to understand the path an image or video has taken from its creation to the time of analysis. They consider both content and metadata information, showcasing that using metadata improves provenance analysis over the sole use of content. Alternatively, inconsistencies between the metadata and the content of the image itself can hint at forgeries. For instance, [63] develop a method that analyzes whether visual information present in an image (notably the illumination and weather conditions) is consistent with the timestamp and geolocalization present in the metadata, thus enabling the detection of some manipulations.

The main weaknesses of reactive metadata-based approaches include the ease of falsifying metadata and the potential loss of important information through the many resharing steps of the media on social networks. For this reason, most of the research literature in deepfake detection has focused on exploring the use of data traces instead.

Methods exploiting data traces aim to either embed or discover discriminative signals directly in the data itself, which can provide reliable indicators regarding its likely authenticity or origin.

Proactive methods involve embedding specific signals (**watermarks**) in any generated media, as a signature of the generator. Examples of these approaches are shown in Figure 1.10. For example, the DALL-E 2 generator includes a small set of colored squares in the bottom right corner to clearly identify the source of the image. However, this method has low robustness as a simple tighter crop of the generated image can defeat it. An alternative proposed by Google DeepMind, called SynthID, embeds an imperceptible signal in the image data that is robust to various perturbations (including cropping) and enable the identification of the image's source.

[4] https://jpeg.org/jpegtrust/

As an additional safety measure, commercial generators can constrain the generated content to non-problematic use cases. For instance, DALL-E 3, available through ChatGPT, applies tight restrictions on content (e.g., preventing the user from generating representations of existing people) and on the style of the generated images, which always have a slightly "cartoonish" appearance, making it obvious that they are not real. While these approaches can effectively ensure some traceability of media generated using commercial systems, they can easily be bypassed by a malicious actor using open-source solutions, which are numerous.

A remaining solution is to seek core imperfections or artifacts contained "natively" in generated deepfakes and exploit them for detection and source attribution. This type of reactive solutions is the core focus of this thesis, with the main questions being *what these imperfections could be* and *how we can exploit them.* In forensic science terms, the question is to determine which useful traces are left in the generated media by the generator. In machine learning terminology, the question is to find appropriate representations of data in the context of deepfake image forensic investigations.

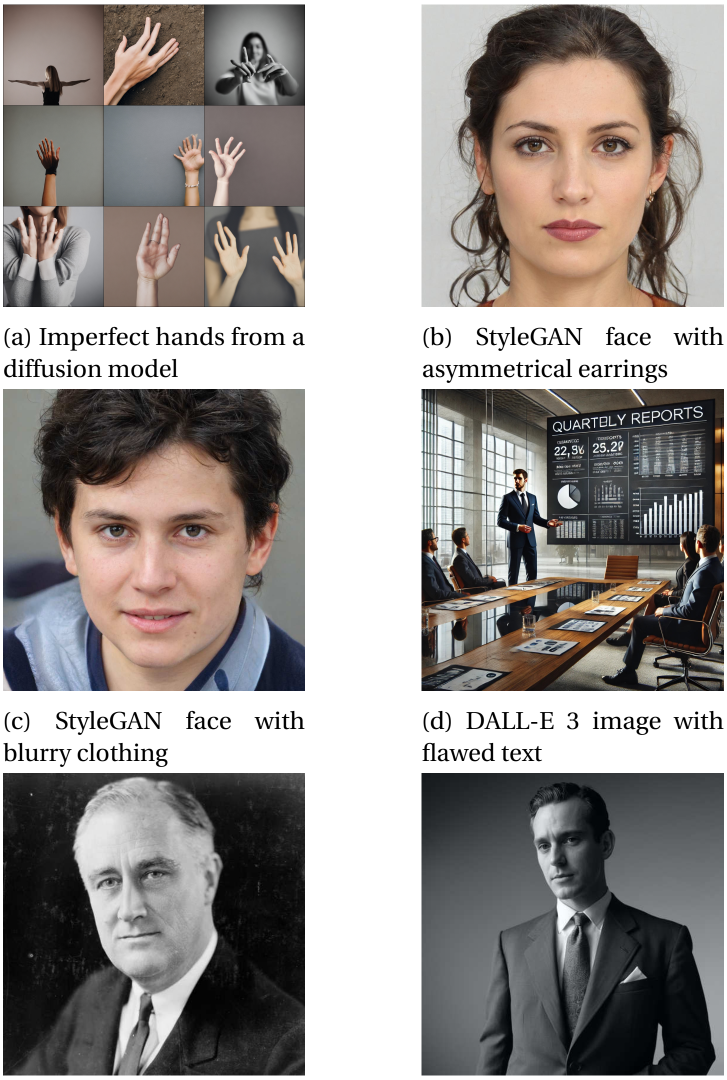
If appropriate representations can be found, the principal deepfake forensic analysis tasks can be conceived as follows:

- For practical modeling, **detection** is often approached as a binary classification problem, aiming to distinguish media exhibiting characteristics primarily associated with genuine capture from those exhibiting characteristics indicative of synthetic generation. Alternatively, it can be framed as an **anomaly detection** problem, where the goal is to model the feature distribution of genuinely captured media (one-class modeling) and identify synthetic media as out-of-distribution samples.
- **Closed-set attribution** can be approached as a multiclass classification problem, where the objective is to identify the specific generator used to create the synthetic media among a known set of generators.
- **Open-set attribution** can be divided into two tasks: novel deepfake detection and source linking. Novel deepfake detection can be framed as an **anomaly detection** problem, aiming to model the feature distribution of synthetic images from known generators and identify media from novel generators as out-of-distribution samples. Source linking can be approached as a **metric learning** problem, aiming to find a representational space where samples from the same generator are close to each other, while samples from different generators are far apart.

Closed-set attribution is useful in controlled scenarios, but its reliance on a fixed set of generators limits its applicability in broader, real-world contexts, where the occurrence of new, unknown generators cannot be excluded. However, in specific cases, closed-set attribution can be sufficient. For instance, consider a license infringement conflict: a provider produces a publicly available model with specific license restrictions (e.g., for non-commercial applications only) and then observes a company using synthetic images very similar to those produced by the generator as part of a marketing campaign. Assessing the potential breach of license would necessitate verifying whether the synthetic images are produced by the provider's generator or by another generator claimed by the company to be the actual source.

In this thesis, our experiments will be restricted to closed-set attribution scenarios. While this is a limitation, it is also an essential building block towards the development of open-set methods. By first validating the feasibility of source attribution in a controlled environment, we establish a basis on which future research can build, adding the complexities and uncertainties of handling unknown generators.

### 1.4.2 Traces in deepfake images



(a) Imperfect hands from a diffusion model

(b) StyleGAN face with asymmetrical earrings

(c) StyleGAN face with blurry clothing

(d) DALL-E 3 image with flawed text

(e) Real historical picture, with flaws

(f) Synthetic historical picture, overly perfect

Figure 1.11: Examples of semantic traces for deepfake detection

### Semantic traces

The detection of fake images based on data can exploit two broad categories of traces. **Semantic traces** correspond to common high-level flaws in the content of the image. While not always easily noticeable for an untrained human evaluator, it is feasible to establish a list of these common flaws, which can be used as a “check-list” for a human evaluator to assess a media of interest. Examples of such traces are presented in Figure 1.11. We can mention:

- **Body irregularities**: many diffusion-based generators struggle to represent certain complex body parts such as hands correctly. Taking a more detailed look at hands in pictures of people can be a surprisingly effective way to detect if the image is synthetic. Sometimes, unexpected asymmetries can also be a telltale sign, such as asymmetrical earrings which are common in StyleGAN-generated faces.
- **Flaws in secondary subjects**: generators that target a relatively narrow type of generated content can produce important flaws in the contextual portions of the images, which are not the main target of the modeling process. For StyleGAN faces, for example, while the faces are usually pristine, it is common for the clothing to lack realism.
- **Text**: diffusion-based generators struggle with representing textual content in their output, instead producing gibberish which is a very easy signal to notice.
- **Inconsistent physiological signals**: in deepfake videos in particular, subtle physiological signals are not always well reproduced. Inconsistent eye blinking can be exploited, as in [64] where cropped out eye sequences are extracted from the video and fed as input data to a convolutional-recurrent neural network. Alternatively, [65] and [66] aim to detect absent or inconsistent heartbeat patterns, which can be perceived through small variations of the color of the skin of the face in real videos.
- **Lack of imperfections**: perhaps counterintuitively, synthetic images often look a little too “perfect”. GAN-generated faces are absolutely pristine, with no “human” flaws or skin imperfections. Diffusion-images often have a cinematic feel to them, seemingly with perfect studio lighting and depth of field, instead of following more raw visuals that could be captured in the wild with a smartphone, for example. These aspects enable a trained eye to develop a reasonably good intuitive detection ability, at least in a differential context (i.e., having to distinguish among two images which one is real and which one is synthetic). This can be experimented with on public websites, for faces [5] and general text-to-image[6].
- **Post-processing artifacts**: Some forgery algorithms combine the use of a deep generator with additional post-processing operations. For example, in faceswap video generation, deepfake generators produce only a tight crop of the tampered face, which then has to be

[5] https://www.whichfaceisreal.com/

[6] https://whichisai.raphaelreynouard.com/

blended back into the target video. Detection systems can focus on artifacts introduced by this post-processing rather than by the deep generator itself. For instance, [67] trains a DNN to localize the boundary between the blended face and the rest of the frame in faceswap videos.

Semantic traces are particularly useful for human evaluators. However, due to the adversarial nature of deepfake generation, identifying effective semantic features for deepfake detection often highlights weaknesses in the generation process. This, in turn, leads to the development of more realistic deepfakes as generators are improved to address these weaknesses. For instance, depth-map conditioning of text-to-image generators using ControlNet [26] helps them produce better-looking hands. For another example, composite algorithms that combine raw image generation with post-processing steps to embed text can mitigate issues with textual content.

### Statistical traces

The other family of traces that can be exploited for detection are **statistical traces**. These correspond to differences in the typical statistical distribution of pixel values in synthetically generated images compared to those found in genuinely captured images. Importantly, these are low-level traces that are not dependent on the content of the image itself but are introduced as a byproduct of the generation process. Such traces are typically not visible, making them more difficult to exploit by a human, but they have been very useful for developing automated detection systems, probably more than semantic traces.

Indeed, it has been observed that synthetic images produced by deep generators typically present spectral distributions that do not accurately match those observed in real data. This is the case for both GANs [68]–[70] and diffusion models [71]. The specific distribution of a particular network forms what is called a **fingerprint** or **signature** of the network, which has the potential to be detected (we will use the second terminology in this thesis). While the presence of signatures in synthetic data is not fully explained, the consensus is that it is caused by the convolutional architecture of the generators themselves (particularly the upsampling steps according to [72]). As all common image generators exploit upsampling layers, such signatures are strong candidate representations for developing robust detection systems encompassing a wide variety of generators. Figure 1.12 showcases visualizations of signature-like signal in synthetic images.

However, it is important to note that the distinctive spectral distributions observed in the aforementioned works are derived from averaging the spectra of many synthetic images. Consequently, it is not guaranteed that the signal present in *individual images* is sufficiently salient to enable robust detection.

Another type of statistical trace involves noise patterns that are either present or absent in synthetic data. [73] propose the use of Photo Response Non-Uniformity (PRNU) features for deepfake detection. PRNU is a noise pattern introduced in real images during the capture process by digital image sensors, and can be characteristic of individual camera devices. Conversely, synthetically generated images are not expected to exhibit similar noise patterns, making PRNU an interesting feature for detection.

### 1.4.3 Trace extraction for deepfake images

While we have described conceptually the traces that might be present in deepfake images, it remains to explain how to extract those traces in practice.

Extracting relevant traces for deepfake detection and source attribution is a representation learning problem. The relevant information in the analyzed media is any that is helpful for distinguishing real images from synthetic ones, i.e., the portion of the semantic content that tends to be flawed in generated media, possible statistical defects, and possible signatures. The irrelevant information is the rest of the semantic content, and statistical signal introduced by processes other than the synthetic image generation itself (for example, through JPEG compression of the image). In practice, all three families of data representations presented in section 1.1.1 can be considered : handcrafted features, supervised features and foundational features.

#### Handcrafted features

Approaches based on handcrafted features usually target statistical traces. [74] and [75] aim to extract signature information through the use of simple spectral analysis features, and [76] through an analysis in color spaces. [73] focuses instead on noise patterns through PRNU features. Handcrafted features can also be used in the context of attribution, as shown in [68]. They proceed by measuring the correlation between an image-signature to a model-signature, and attributing the image to the model which maximizes correlation. The model-signature is computed as an average of image-signature from its output. [72] builds on those results and improves attribution performance by working in the frequency domain (using a Discrete Cosine Transform of the images).

#### Supervised features

However, most of the literature focused on supervised features, approaching detection and attribution as end-to-end classification problems. Due to the generative nature of deepfakes, large datasets of examples can be created relatively easily in an automated manner, enabling a focus on data-driven approaches. This is a fundamental difference compared to physical identification in forensic science, which require a resource- and time-consuming collection process of traces before considering data-driven approaches.

The community has experimented with end-to-end training of the main architecture families

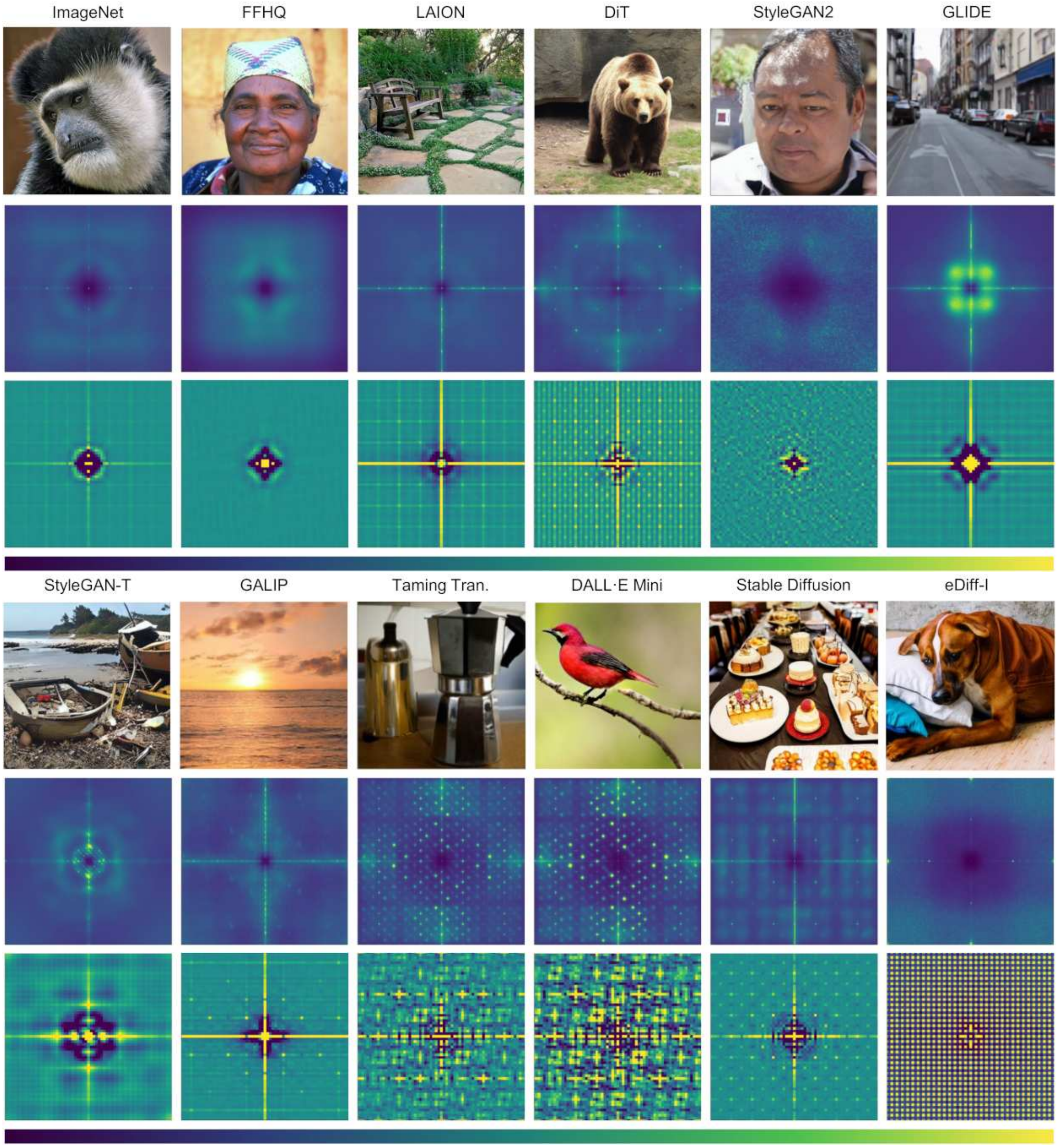


Figure 1.12: Figure taken from [71]. Examples of signature-like traces present in deep synthetic images. Those are obtained by 1) computing noise residuals from a set of images with the same source; 2) computing the average power spectra of those residuals (rows 2 and 5) or their average autocorrelation (rows 3 and 6). The first 3 sources (ImageNet, FFHQ, LAION) are datasets of real images, while other columns correspond to images produced by deep generators. Synthetic images showcase peculiar patterns, observable as peaks in the spectral domain, or as periodic patterns in the autocorrelation. However, it is important to point out that patterns can also appear for real images (e.g., LAION), linked for instance to JPEG compression artifacts.

of image-processing DNNs, mainly CNNs [69] and Vision Transformers (ViTs) [77]. Research has focused on increasing the generalization capability of the detectors, and their robustness to common image degradations such as resizing or JPEG compression. As proposed in [69], improvements in this regard can be achieved through the use of training-time data augmentations: the training images are subjected to various blurring, resizing, or cropping operations, which helps the detector become more insensitive to these transformations. Further extensions also focus on minor but critical architectural changes aimed at preserving salient signature signals. For this purpose, past works propose to drastically limit careless downsampling operations in the models, by redesigning early layers in [70], or by performing detection on cropped image patches rather than full downsampled images in [78]. These works, having focused mainly on GAN images (as those predate the main breakthroughs in diffusion models), have been followed by works applying similar ideas to the case of diffusion images [79].

Supervised learning approaches have also been proposed for deepfake attribution. [80] propose to learn GAN-signatures in a data-driven manner and perform closed-set attribution by training attribution DNNs for multiclass classification of GAN images. While effective, these methods are influenced not only by the architecture of the generative model but also by its training data, meaning that they will be affected by fine-tuning or retraining of the generator. To address this issue, [81] propose an attribution system at the architecture level, capable of associating images generated using the same generative architecture retrained on different data. They propose a network trained in two stages: first pretrained in a self-supervised manner to learn strong representational features, then fine-tuned for GAN-image attribution. Another notable work is the HiFi-IFDL model [82], which focuses on forgery detection and localization. During training, they include a proxy objective of forgery attribution to learn multi-resolution feature maps that can then be post-processed for forgery localization.

Due to their black-box nature, it is not clear which types of traces (semantic or statistical) are exploited by the trained detectors. Nevertheless, given the observation that the detectors lack robustness to minor image degradations altering pixel statistics without changing the content, it appears that statistical traces are the primary features learned by the detector.

#### Foundational features

Finally, a more recent research line proposes a shift towards foundational features.

[83] study video deepfake detection, and note that past works using supervised learning were usually starting from an image recognition model pretrained on the ImageNet dataset [8]. They challenge this practice and propose to consider a variety of initial pretrained models instead, trained not only for image recognition but also for other tasks such as face recognition and age detection, and also models trained in a self-supervised manner. They evaluate whether the learned feature spaces are appropriate for unsupervised clustering of real and fake videos in separate clusters. They also train binary classifiers for detection starting from these pretrained

models, either by only training a final fully connected layer with cross-entropy, or by fully fine-tuning the model. They observe that models pretrained for ImageNet recognition are rarely performing best under those settings, and in particular that self-supervised models are the most effective, suggesting their learned representation space is much more relevant to the downstream detection task.

Similarly, [84] address the generalization issue of fake image detectors. They demonstrate that robust systems capable of accurately detecting various GAN-generated images fail significantly when exposed to images from diffusion models. They hypothesize that the highly salient signatures present in synthetic data are “too easy” for supervised detectors to identify, resulting in internal feature representations that are effective for GAN images but lack robustness. To circumvent this issue, they utilize CLIP, a pretrained large vision-language model [11] (which can be considered foundational in nature), as the feature extractor. Detection is then performed by applying a simple nearest-neighbor classifier on top of the learned feature space, achieving surprisingly strong detection accuracy for both GAN and diffusion images, despite the pretrained model having *never* been exposed to any GAN images during pretraining. Similar findings are reported in a parallel study [85].

These works suggest that effective generalizing detectors (and probably as well strong open-set attribution system) should rely on feature spaces *not biased towards a given set of attacks*, and prefer feature spaces emerging from vision models trained with more generic tasks, delegating the rest of the modeling to a classifier applied on top of this space. A possible intuition would be that these features are a good representation of “what makes a natural image”, and hence:

- Non-natural images can be detected as outliers in this feature space

- Non-natural images coming from different synthetic generators are unlikely to differ *in the same way* from the distribution of natural images, thus enabling some attribution capability.

The development of powerful vision foundation models being also currently a very active research line, there exists other models than CLIP which could also be interesting to explore as they specifically target the learning of general representations. We can mention for instance DINO [86] or AIM [87], respective efforts from Meta and from Apple.

Table 1.1 provides a comparative overview of the strengths and limitations of each trace extraction method, focusing on the required data and the four key axes of analysis: discriminative power, generality, robustness, and interpretability. Notably, it highlights two critical open questions regarding the use of foundational features: their discriminative power for deepfake detection and their robustness to common media post-processing operations.

Table 1.1: Comparison of the three primary trace extraction methods, highlighting their characteristics and ability to derive meaningful representations from input data.

| | Handcrafted features | Supervised features | Foundational features |
|---|---|---|---|
| Data-driven | No | Yes | Yes |
| Data for representation learning | None | Real + Fake images | Real images |
| Dataset size for downstream classifier | Small | Large | Small |
| Discriminative power | Medium | High | To be evaluated |
| Generality | Medium (Adaptation to new generators might be challenging) | Limited (risk of over-specialization to seen classes) | High (by design) |
| Robustness | High (if integrated in the design) | Medium (can be improved through augmentation) | To be evaluated |
| Interpretability | High | Low | Low, but less at risk to exploit spurious correlation |
| Examples | Spectral features [74], PRNU [73] | Most detection literature, exploring various architecture and training processes. GAN-detection CNNs [69], [70], Diffusion-detection CNNs [79], HiFi-IFDL [82] | CLIP [11], [84], [85], DINO [86], AIM [87] |

## 1.5 Challenges in detection and attribution

### Generalization

The main current challenge in the field is **generalization**: a detection model trained on many different types of deepfakes will still typically perform poorly when encountering deepfakes produced by a previously unseen method. Given the prevalence of open-source generators, this is problematic as little effort might be required for a malicious actor to tweak their own deepfake generation model, and subsequently bypass implemented detection algorithms with relative ease.

Generalization has been a major focus of research on deepfake detection, and breakthroughs have been achieved that improve generalization from detectors trained on a single GAN to other unseen GANs, through data augmentation [69] or architectural tweaks [70].

However, the field of deepfake generation has been evolving rapidly and is likely to continue doing so, leading to a high diversification of generative models, including novel approaches. For instance, [79] illustrate that following a similar methodology as the best GAN-image detectors, but in the context of diffusion models, does not provide as strong generalization capability from a single diffusion model to other unseen ones. Therefore, even within a given family, generalization is still not guaranteed. Moreover and for illustration, we point to [88] which propose an autoregressive image generator based on a new family of architectures called Mamba. As both the learning process and the architecture differ drastically from what is used in GANs and diffusion generators, generalization to this new family of deepfakes is improbable.

A solution orthogonal to generalization is that of continuous learning, which aims to develop detectors that are regularly updated with new data from novel generators. [89] propose a framework for this approach and evaluate its effectiveness. While their method improves over zero-shot generalization performance, they also observe that the outcomes are strongly dependent on the order in which the novel classes are integrated into the system.

Such approaches might be more practical in the long run, as developing a universal detector is somewhat idealistic. However, continuous learning also presents challenges, including the difficulty of avoiding catastrophic forgetting (where performance on previously seen examples decreases after learning from new samples), the need for explicit access to actual samples from all targeted generators, and the risk that performance might saturate after a certain number of generators.

#### Robustness

Another challenge in deepfake detection is that generating the raw synthetic image using a deep generator is only one step in the lifecycle of the deepfake media. Commonly, post-processing operations are applied to the media. These operations can be forgery-related (part of the forgery creation process) or forgery-unrelated (part of the normal lifespan of a typical visual media). An example of forgery-related post-processing is in the context of generating a faceswap video, where the generated deepfake is just a tight crop of the face, which has then to be blended back into the target video. The impact of this post-processing on the detectability of the forged media is twofold: on the one hand, the blending process could alter the statistical properties of the raw generated images, potentially degrading the signature and making detection more challenging. On the other hand, blending could create an exploitable signal for detection. For example, low-resolution deepfakes might need to be upscaled to match the original video resolution, generating warping artifacts.

Examples of forgery-unrelated post-processing operations are numerous, and more problematic. These correspond to typical operations a visual media might undergo after its creation, such as resizing, cropping, or compression applied to media shared online. These processes can significantly alter the statistical properties of the media, making detection more challenging.

In practice, research shows that detectors trained on raw generated deepfakes tend to perform poorly on post-processed version of those deepfakes [69]. To improve the robustness of detectors to forgery-unrelated post-processing, this work proposes the use of training data augmentation. Specifically, images from the training set are subjected to various degradations (resizing, blurring, compression) without altering their labeling as real or synthetic. This process forces the detector to learn that these degradations are irrelevant for predicting the label of test images. While this approach enhances robustness and generalization capability to unseen generators, [90] illustrates that most published detectors nevertheless still suffer from a significant performance drop when evaluated on post-processed images.

Post-processing operations are less likely to degrade semantic traces, as these operations are typically expected not to alter the content of the media. Therefore, developing detectors that focus more on semantic cues is a promising direction for improving robustness. However, training detectors in a supervised manner often results in them seemingly focusing on statistical traces. Significant regularization of the training process might be necessary to ensure sole reliance on semantic traces. Alternatively, the use of foundational features could be a promising direction. As hypothesized by [84], statistical traces are almost “too salient” and easy for supervised models to pick up on, yet they lack robustness. In contrast, foundational features naturally enable the development of models that are not biased towards the specific signatures of generators seen during training.

### Cheapfakes

While the focus on deepfake detection is increasing, it is crucial not to overlook other traditional types of image manipulations (e.g., manual Photoshop edits and computer graphics), which can still be highly effective. These types of manipulations are now referred to as **cheapfakes** (or sometimes **shallow fakes**) to distinguish them from deepfakes.

Due to their fundamentally different nature, the semantic and statistical traces of cheapfakes are likely to be completely distinct from those present in deepfakes. As a result, current deepfake detectors are unlikely to be effective on cheapfakes.

Unfortunately, datasets used in deepfakes forensic analysis literature almost never include cheapfakes, which limits both the training of joint detection systems and the evaluation of cheapfake detection performance. A notable exception is the FaceForensics++ dataset [91], which includes both computer graphics and deep generative-based algorithms for video face swapping and face expression editing. However, such mixed datasets are not the norm.

### Interpretability

Both for detection and attribution of deepfakes, most of the current best algorithms rely on DNNs, especially through supervised learning approaches. Such models are inherently black-boxes, meaning their internal processing, which results from the training process, is completely opaque. The network’s prediction is not accompanied by any justification, and the cause-effect relationship between the input and the output is not explicit. This makes it challenging to allow for some degree of appreciation of the prediction by a human expert or a jury. Moreover, there is always a risk that the prediction is actually based on irrelevant components of the input.

From a forensic science perspective, this limits the applicability of such detectors in evaluative contexts (e.g., analysis of a specific image in a criminal case). However, they might remain adequate for investigative contexts and large-scale applications where occasional errors can be tolerated (e.g., automated flagging of dubious images uploaded to a social network).

A straightforward workaround is to renounce using DNNs and focus on algorithms that exploit handcrafted features, which can be better explained given they are designed by a human expert. However, the use of such features typically comes at the cost of a decrease in detection or attribution accuracy.

Alternatively, DNN diagnostic tools and post-hoc analysis techniques can enhance the understanding of model decisions. For instance, the Grad-CAM algorithm [92] provides insights into which areas (pixel-wise) of the input contribute most to the final decision, generating what is known as a saliency map. However, saliency maps can be misleading. They often produce visually compelling results for human evaluators but fail to meet basic sanity checks. For example, [93] demonstrates that models trained on data with random labels and models with partially randomized weights still produce relatively convincing saliency maps.

Finally, score calibration techniques can help to give a likelihood-ratio interpretation to the models' output, making them more fit for forensic applications. Indeed, a deepfake detector's output is simply a score which is low for real images and high for synthetic images. However, the value of this score is not calibrated: it is not expressive of the detector's confidence level in the prediction. Through the use of a calibration dataset, a transformation of the scores can be computed to ensure the transformed score is actually a proper confidence level, i.e., it is aligned with the practical error rates of the detector. This calibration process is presented in more detail, for instance, in [94] in the context of automated face recognition.

While the process still does not really add to the interpretation of the model's decision, it enables at least the *quantification* of the confidence level of the models, in particular highlighting their potential shortcomings. However, the relevance of the calibration is directly dependent on the choice of calibration data, which should be representative of the actual case of interest. In face recognition, there is a general understanding of which the important characteristics from the case of interest should be present in the calibration data (similar demographics, pose, image quality, . . . ). Unfortunately, this understanding is much more limited in the context of deepfake detection, especially because many important characteristics might be statistical rather than semantic, and thus hard to judge by the expert.

### Authenticity is a continuous notion

Evaluating media authenticity is particularly challenging due to the continuous nature of the concept. While it is generally possible to distinguish between media materialized by a physical camera and media generated by a deep generator, this distinction often becomes ambiguous. For instance, composite media (where a genuine image or video has been locally edited for falsification, such as altering only the mouth region in a deepfake video) can be difficult to classify holistically. In such cases, establishing a ground truth for authenticity is meaningful only at a local level. Even then, ambiguity may persist, as individual pixel values can result from a weighted influence of both the original genuine image and the manipulation process.

Another challenge arises from the fact that modern digital cameras often apply image enhancement algorithms natively during capture. Should the resulting media be considered authentic? A reasonable approach is to align the notion of authenticity with the impact of these algorithmic edits on the exact *message* conveyed by the media. For example, post-processing operations like compression, which do not alter the media's message, can be considered authenticity-preserving. Conversely, edits that significantly change the semantic content of the media, such as altering *what* the subject is saying or doing, clearly suppress its authenticity.

Importantly, this classification depends on the media's context, not solely on the algorithmic edits. For instance, a slight beautification filter might be considered to preserve authenticity if applied to a video of a politician delivering a policy speech, but could suppress authenticity if used in a cosmetic advertisement featuring a model.

In this thesis, our experimental work focuses on fully generated synthetic face images depicting non-existent identities (more precisely, non-existent mixtures of existing identities). In this context, the ground truth for authenticity is relatively clear: if the face image is produced by a camera, it is considered authentic. Any other scenario should make it classified as fake.

Note that it is essential to distinguish between two types of uncertainty: uncertainty in the ground truth authenticity classification of an image and uncertainty in the authenticity classification of an image with unknown ground truth. While the former is inherent to the concept of authenticity, the latter arises from the limitations of the analysis system.

### Attribution levels

A similar, and potentially even more challenging issue arises in the context of deepfake attribution: determining which generative models should be classified as "same" or "different." For example, should a pretrained generator, and a distinct fine-tuned version of it, be treated as the same source or as a distinct one? To address this, existing literature introduces the concept of **levels of attribution**. For instance, [95] outlines the following distinct tasks:

- **Architectural- or model-level attribution**: Identifying a family of generators based on generic design aspects, such as the architecture, loss function, or training procedure.

- **Instance-level attribution**: Identifying a specific pretrained generator, characterized not only by its design but also by its unique set of weights, which can be altered through any modification in training data, random initialization, or training stochasticity.

This hierarchy is defined by analogy with source camera identification. The attribution level framework provides a useful lens for analyzing existing datasets, particularly in the domain of video deepfakes, for instance:

- *Korean Deepfakes* dataset [96]: This dataset includes five distinct deep generators, thus representing a model-level attribution problem.
- *FaceForensics++* dataset [91]: This dataset includes three deep learning generators and two computer graphics methods, presenting a model-level attribution problem that combines deepfakes and cheapfakes. Given the fundamental differences in nature between those two families, one could argue this is even a higher methodology-level attribution problem.
- *Deepfakes from Different Models* dataset [97]: This dataset uses a single autoencoder for faceswapping but creates distinct classes by altering hyperparameters like the number of layers and resolution. While these variations might classify as different architectures under strict model-level attribution, practical scenarios might warrant treating them as a single class.

  For example, consider an open-source generative model published under a restrictive license. If a commercial derivative appears with minor architectural tweaks, attribution efforts would focus on linking generated outputs to the original codebase, requiring an approach resilient to small hyperparameter changes.

Moreover, many published datasets reuse the same source code for some of their deepfake classes, potentially with minor architectural or pipeline modifications. As most datasets have originally been released with deepfake detection in mind, details about the exact generation might sometimes be missing. Since these datasets are typically designed for detection rather than attribution, precise details about generation methods are often missing. This ambiguity complicates cross-dataset evaluations, as classes from different datasets might either be identical or exhibit slight differences—often without clear documentation. Consequently, while the notion of attribution levels is theoretically sound, its practical application is challenging. Accurately labeling classes requires fine-grained definitions, which may vary depending on the application context.

Addressing those challenges ideally requires:

- Full control on the generation process: this ensures that attribution classes can be well-defined, whatever the considered attribution level.
- Application-specific focus: this clarifies which attribution level is relevant.

## 1.6 Face morphing attacks as a case study

Based on the challenges highlighted in the previous sections, we chose to focus the experimental work on the specific problem of face morphing attacks. Face morphing attacks are a significant issue in face biometrics and present interesting aspects from an image forensic analysis perspective:

- **Mixture of deepfakes and cheapfakes**: Many recent methods for face morph generation use deep generators, but traditional image processing methods are still among the most effective. Therefore, face morphing attack forensic analysis requires handling both deepfakes and cheapfakes.
- **Print-scan post-processing**: As part of the attack, face images go through a printing and rescanning process, which can drastically alter statistical traces. Unlike other common degradations like compression or resizing, robustness to print-scan post-processing is difficult to achieve through data augmentation, as automated generation of print-scan training data is challenging.
- **Generation control**: Morphing attack generation is less complex than video deepfakes, making it more reasonable for us to generate our experimental datasets in-house. This gives us full control over the specifics of the generation and enables precise definition of classes for attribution.

In the next chapter, we will thus present the problematic of face morphing attacks in more details.

# 2 Face morphing attacks

In this chapter, we will present a specific type of fake media that is relevant in the field of biometric security: face morphs. These are images created by blending the facial features of at least two individuals, called the sources. Malicious actors can exploit face morphs to carry out a **face morphing attack**, which involves submitting the morphed image during a passport application process. This tactic allows both contributing subjects to share the same passport, thereby violating a fundamental principle of "one person ↔ one passport" and posing a security threat, notably at Automated Border Control (ABC) gates.

The development of face morphs predates the era of deepfakes, but recent advancements have introduced novel generation algorithms that produce "deep" morphs, notably by exploiting GANs and diffusion models. This positions face morphing attack analysis in an interesting niche of fake media countermeasures research, as the problem encompasses both cheap- and deepfakes, along with an uncommon type of image degradation: the print-scan postprocessing. Typically, a generated face morph must be printed before being physically submitted to the passport issuing office, where it is subsequently scanned for enrollment into a face recognition system. This process can significantly alter the statistical properties of the image, making detection or attribution more challenging.

The aim of this chapter is to properly introduce the phenomenon of morphing attacks. First, we will describe the threat model, discuss the feasibility and real-world occurrences of such attacks. We will then explain the basic principles behind face morph generation, and describe the process of evaluating the effectiveness of the attack. We will present the face morphing data that will be used in the experimental portion of this thesis, and finally provide an overview of relevant works and methodologies for morphing attack detection.

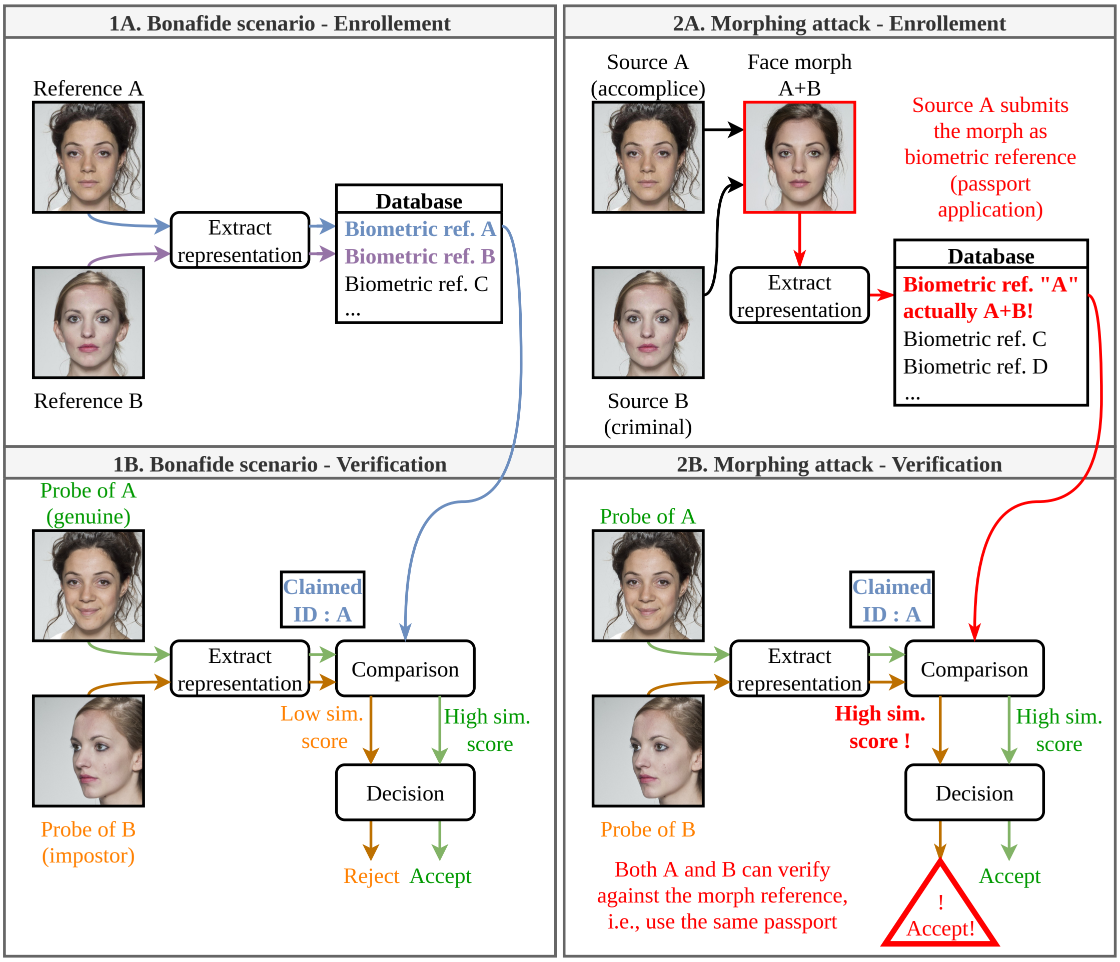


Figure 2.1: Comparison between the bona fide scenario of biometric verification, and the scenario of a morphing attack. The attack is performed during the enrollment phase (2A) and leads to verification errors in the verification phase (2B).

## 2.1 Threat model

### 2.1.1 Biometric verification

Before explaining the process of a morphing attack, it is necessary to provide a high-level view of the principles of a biometric recognition system.

A biometric recognition system's primary function is to compare pairs of biometric samples. A biometric sample is an analog or digital representation of a biometric characteristic of a human subject. Examples of biometric characteristics include a fingerprint, an iris, or a face, which is the focus of Face Recognition Systems (FRSs). A biometric sample typically contains both identity-related information (e.g., unique facial features) and unrelated information (e.g., facial expressions, lighting conditions). Before comparison, a biometric system processes the biometric sample to extract relevant information (biometric features) while discarding irrelevant data. Extracted features from two biometric samples are then numerically compared to measure their similarity. Ideally, biometric features from two samples of the same subject should be highly similar, whereas features from samples of different subjects should be highly dissimilar.

A biometric recognition system can be used to perform biometric verification. In biometric verification, a subject presents themselves to the biometric system while claiming a given identity, and the system must assess whether this claim is valid or not. The process is decomposed into two stages.

In the **enrollment** (stage 1A in Figure 2.1) stage, a biometric reference of the subject is captured and stored in a database for later retrieval. This stage corresponds, for example, to setting up face recognition on a smartphone by storing an initial capture of the user's face, or to the creation of an identity document based on a reference picture.

In the **verification** stage (stage 1B in Figure 2.1), the subject presents themselves again to the system while claiming an identity. The live biometric sample is captured, features are extracted, and a comparison is performed with the stored biometric reference of the claimed identity. The decision to validate or not the comparison is made by comparing the resulting similarity score with a predefined decision threshold. This stage corresponds to an attempt to unlock a smartphone using one's face, or to passing through an ABC gate using one's passport.

### 2.1.2 Attack scenario

A face morphing attack targets FRSs during the *enrollment* stage, in the context of passport applications. It starts with the creation of a face morph of two contributing subjects, i.e., a fake face image mixing both their facial features. This face morph is then submitted as reference for enrollment (stage 2A in Figure 2.1). Typically, the individuals involved are a wanted criminal and an unknown accomplice who bear some resemblance to each other. The accomplice applies for a passport using the face morph as the reference photo. Since the morph resembles the accomplice, it does not raise suspicion from the office worker processing the application.

Moreover, when the attack is successful, the target FRS will erroneously accept verification attempts from *both* contributing subjects against the morph reference. As a result, both the accomplice and the criminal can use the same passport, effectively claiming the same identity, which allows the criminal to pass through ABC gates without raising alarms during the verification stage (stage 2B in Figure 2.1). The morph's resemblance to the criminal also enables them to deceive the border control personnel, on top of the FRS. In summary, when successful, this type of attack undermines the fundamental principle of "one person ↔ one passport," leading to significant security risks.

### 2.1.3 Feasibility and occurrences

The concept of a morphing attack was first introduced by [98]. They not only described the threat model but also demonstrated its theoretical feasibility by showcasing the effectiveness of manually edited face morphs in compromising contemporary FRSs. An actual implementation of the attack was successfully carried out in Germany in 2018 [48] by a group of activists aiming to disrupt mass surveillance programs by contaminating governmental databases. Through a standard application process, they managed to obtain an authentic passport containing a face morph between one of the activists and Federica Mogherini, a politician who was then the High Representative of the European Union for Foreign Affairs and Security Policy.

In 2021, the Slovenian Police reported having detected more than 40 cases of morphing at Ljubljana Airport over the previous year [99]. These cases were seemingly part of an organized operation providing Slovenian passports to Albanian citizens, enabling them to travel to Canada.

A natural defense against face morphing attacks is to ensure that the reference passport picture is captured in a controlled environment, such as a governmental office, where it can be directly processed to prevent any intermediate manipulations. Alternatively, the picture can be taken at a government-recognized photo office, which digitally signs the document and sends it to the authorities. However, only a few countries have implemented such systems (e.g., Sweden and Norway for the former, Finland for the latter).

In most other countries (e.g., Austria, France, Germany), applicants can bring a physical passport picture to the passport office, enabling the submission of a manipulated image, provided it does not raise the suspicion of the employee collecting the picture. Moreover, some countries (e.g., Ireland, New Zealand) even allow the upload of digital photos on an online application platform, further increasing the potential for manipulations.

In Switzerland, the requirements vary between cantons. Most cantons require passport applicants to have their reference face picture taken at a governmental biometric office. However, some cantons still allow the submission of a self-taken photo, as listed by [100], specifically Aargau, Jura, Lucerne, Obwald, Schaffhausen, Solothurn, and Zurich.

Therefore, while standardized controlled capture remains the most robust defense against morphing attacks, its widespread implementation is still a distant goal. In the meantime, it is necessary to develop and deploy Morphing Attack Detection (MAD) systems to detect and mitigate the risk posed by passports containing morphing attacks, both those already in circulation and those that may be issued in the near future. These observations have led to several research programs focused on both assessing the vulnerability of existing FRSs to morphing attacks and developing robust detection methods. A notable example is the FATE Morph benchmark from NIST [101], which has underscored the extensive susceptibility of various FRSs to morphing attacks.

## 2.2 Morph generation

### 2.2.1 Overview

The initial morphing approach introduced by [98] is now referred to as **landmark-based** morphing. The core idea involves localizing specific facial landmarks in both source images, warping the images to align these landmarks, and then averaging the pixels. The original process required significant manual effort and was causing visible *ghosting* artifacts at the non-aligned boundaries of the face. Subsequent works proposed an automated approach [102] and addressed the ghosting artifacts issue by blending the morphed face back into one of the source images [103].

More recently, a new family of morphing generators has emerged, leveraging deep learning techniques (**deep morphs**). The first proposed deep morphing method was introduced in [104]. They trained a GAN for face generation, along with an encoder into the latent space of the GAN. This latent space was used for face representations, and the synthesis network of the GAN served as the decoder. A significant limitation of their approach was the low resolution of the resulting images.

Subsequent works achieved higher resolution morphs by leveraging the StyleGAN model [18], a powerful high-resolution face generator. Faces were encoded in the latent space of the GAN, specifically in the $\mathcal{W}$ space [36] or the $\mathcal{W}+$ space [105]. In the absence of an image-to-latent encoder, the encoding was performed by optimization, searching for a latent vector that could accurately decode into the target face. Building on this, [37] introduced the latent morph fine-tuning process, optimizing the input of the generator using a biometric loss to ensure the resulting image was an effective morph. This process significantly improved the morph's effectiveness.

More recently, approaches leveraging diffusion probabilistic models (DPMs) as the primary generative process have been proposed. Although DPMs are generally not associated with a structured latent space, [106] introduces a diffusion autoencoder that enables the encoding of real images into a semantically structured latent space. Exploiting this new latent space for performing latent morph interpolation, [49] and [107] independently propose similar diffusion-based deep morphing attacks.

### 2.2.2 Formalization

At a high level, all morphing algorithms considered in this thesis can be described as a combination of four main steps: encoding, latent manipulation, decoding, and optionally post-processing, as illustrated in Figure 2.2.

1. **Encoding**: A latent representation is extracted from each of the contributing source images. This latent representation should enable arithmetic operations, particularly

averaging two latent representations, which is used in most algorithms.

2. **Latent Manipulation**: From the latent representations of the two sources, a latent representation for the morph is computed. Generally, this is done by averaging the latent representations of the sources. However, the latent morph can potentially be refined, i.e., a latent morph "better" than just a simple average can be found by exploring the latent space. This usually requires introducing a biometric loss into the process, to evaluate how effective the candidate morph would be.
3. **Decoding**: The latent representation of the morph is decoded back into an image.
4. **Post-processing**: In some cases, further post-processing is applied. This can include blending the morphed face back into one of the source images for landmark-based morphing. It can also involve a print-scanning process or other forms of image degradation.

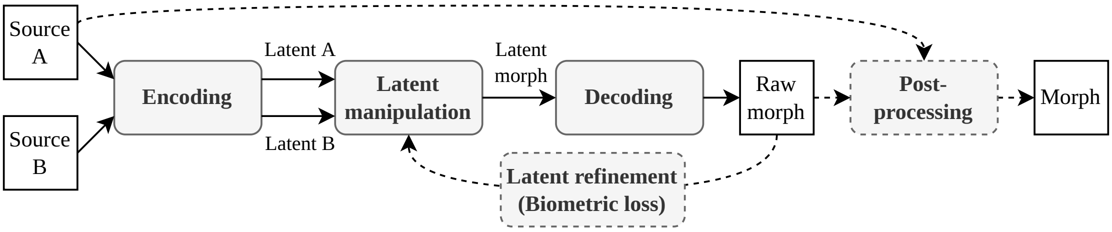


Figure 2.2: Overview of the morphing algorithms discussed in this thesis. The dashed lines indicate optional steps. Specific algorithms are characterized by the details of the encoding, latent manipulation, decoding, and post-processing steps.

We now provide additional details on the two main families of attacks : landmark-based morphs, and deep morphs.

### 2.2.3 Landmark-based morphs

Landmark-based morphing involves manipulating the source images at the pixel level, guided by the facial landmarks of the contributing identities. The latent representation in this context is a combination of the image itself and the extracted facial landmarks. This process is illustrated in Figure 2.3a.

Given two source faces, a naive approach would be to average the pixels of both images to create the morph. However, due to differences in facial structure between the subjects, this results in visible artifacts caused by misaligned features such as the mouth, nose, or eyebrows. To avoid this, an initial step involves deforming the source images to align all key facial landmarks (e.g., the centers of the eyes, the corners of the mouth, the base of the nose).

First, a specific set of facial landmarks is detected for both source images. For a set of $N_L$ landmarks, the landmarks for each source are represented by matrices $\boldsymbol{L}_A$ and $\boldsymbol{L}_B$, each of dimensions $(N_L, 2)$, describing the x- and y- positions of each landmark. The target position $\boldsymbol{L}_M$ of the landmarks of the morph is then computed as a weighted average:

$$\boldsymbol{L}_M := (1 - \alpha_M)\boldsymbol{L}_A + \alpha_M \boldsymbol{L}_B,$$

where $\alpha_M \in [0, 1]$ is called the **morphing factor** and controls the influence of each source face on the final morph. In most cases, it is set to $\frac{1}{2}$, balancing the contributions from both subjects.

Next, 2D meshes associated with each set of landmarks are computed through triangulation of the source faces, using the detected landmarks as vertices. All past works on landmark-based morphing use Delaunay triangulation for this purpose. Each constituting triangle is then deformed using an affine transformation (affine warping) to be mapped to the corresponding triangle of the morph triangulation. This operation results in warped versions $W_A$ and $W_B$ of the source images, where the facial landmarks are aligned to $\boldsymbol{L}_M$.

Performing pixel averaging between the two warped images results in the so-called **complete morph** $M_{\text{complete}}$, as proposed in [102]. The pixel values of the complete morph are given by:

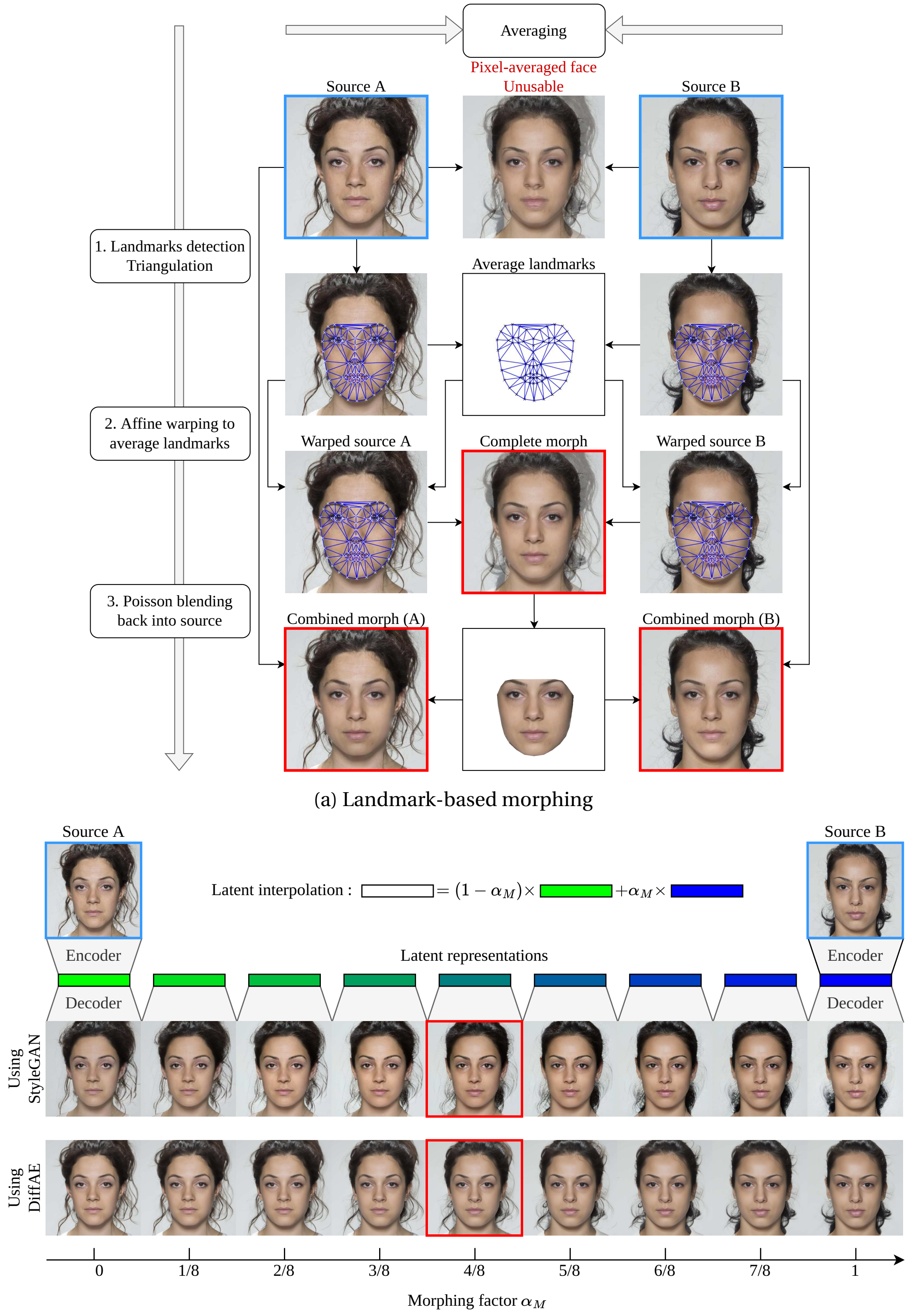
$$M_{\text{complete}}[i, j] := (1 - \alpha_M) W_A[i, j] + \alpha_M W_B[i, j]$$

It is worth noting that the morphing factor $\alpha_M$ used for landmark averaging does not necessarily have to be the same as the one used for pixel averaging. However, for simplicity, we assume they are identical in this thesis.

A remaining issue with complete morphs is that, while the inner portion of the face looks relatively neat thanks to the landmark alignment phase, the contours of the face usually exhibit noticeable superposition of the original images, known as **ghosting artifacts**.

To address this issue, an additional post-processing step can be applied. It involves cropping the region of the face in the complete morph where the landmarks have been aligned, and then blending it back into one of the source images. This process results in a **combined morph**, as proposed by [103]. The authors recommend using Poisson blending [108] to ensure high visual quality and smooth transitions between the morphed face region and the original source image.

This type of morphing algorithm was the first to be introduced and remained prevalent until the more recent development of deep morphing algorithms.



(a) Landmark-based morphing

(b) Deep morphing

Figure 2.3: Overview of the two main families of morphing algorithms (morphs in red).

### 2.2.4 Deep morphing

Deep morphing emerged following breakthroughs in generative AI, particularly with GANs and diffusion models introduced in the previous chapter. To develop a face morph generation algorithm using a deep generative model, this model requires the following elements:

1. **Synthetic Face Generation**: The ability to generate synthetic faces based on a conditioning latent vector. This is akin to the decoder component from Figure 2.2.
2. **Perceptual Continuity**: The property that a regular change in the latent input should lead to a regular change in the resulting image, as perceived by humans.
3. **Latent Space Encoding**: The ability to encode real face images into the latent space, i.e., finding a corresponding latent vector for an input target face that enables accurate reconstruction of the target face. This is akin to the encoder component from Figure 2.2.

We will now present the generators and associated techniques for both the GAN and diffusion families that enable these three properties.

#### Properties of GANs

By design, GANs trained on a dataset of faces inherently possess the first property, with the generator component functioning as the decoder. GANs are designed to map a simple distribution (e.g., a multivariate normal distribution) to the distribution of images represented by the training data. In this context, the random vectors sampled from the multivariate normal distribution and fed into the generator serve as the latent vectors.

However, the property of perceptual continuity is not inherently guaranteed. Fortunately, the StyleGAN model [18], particularly the StyleGAN2 variant [109], addresses this issue. This is achieved through the introduction of an intermediate latent space. As illustrated in Figure 2.4, random vectors sampled from a multivariate normal distribution, considered to belong to a space called $\mathcal{Z}$, are first mapped to another vector of the same dimensionality using a learned transformation based on a simple multilayer perceptron (the mapping network $g_m$). The resulting space is called $\mathcal{W}$, and it is this latent space that is used to condition the synthesis network $g_s$. It has been observed that the $\mathcal{W}$ space provides the desired properties of perceptual continuity, particularly better than the $\mathcal{Z}$ space.

Finally, StyleGAN2 does not include an encoder to project real images into its latent space. However, since the synthesis network is differentiable, it is possible to “invert” it through an optimization process.

Given a target image $I$, its latent representation $\boldsymbol{w}_I^*$ can be found by the following optimization:

$$\boldsymbol{w}_I^* := \min_{\boldsymbol{w}\in\mathcal{W}} d_{\text{perceptual}}\left(I, g_s(\boldsymbol{w})\right)$$

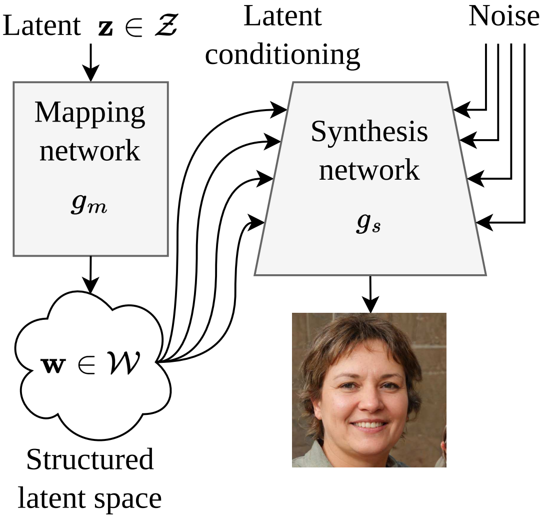


Figure 2.4: High-level illustration of the StyleGAN generator structure. Initial latent vectors $\boldsymbol{z}$ are sampled from a multidimensional normal distribution and mapped to an intermediate latent space $\mathcal{W}$ using a learned transformation. The $\mathcal{W}$ space offers better interpolation properties. The latent vector $\boldsymbol{w}$ from the $\mathcal{W}$ space conditions the synthesis process. The synthesis network $g_s$ processes images at progressively increasing resolutions, applying the same latent vector $\boldsymbol{w}$ at each resolution. Conceptually, the latent vectors are aimed to condition the semantic content of the resulting image, while the noise inputs are aimed to control the stochastic variation in the images (e.g., the detail of the hair or of freckles).

where $g_s(\boldsymbol{w})$ is the output of the synthesis network for the input $\boldsymbol{w}$, and $d_{\text{perceptual}}$ is a distance function that measures the perceptual difference between $I$ and $g_s(\boldsymbol{w})$. This objective can be minimized using gradient descent, backpropagating through the synthesis network. In practice, the distance function $d$ is often chosen to be the perceptual loss proposed in [110], which has been shown to be an effective proxy for human perception of image differences. Thanks to this encoding process, a StyleGAN2 network can be utilized for morphing attack generation through the encoding-interpolation-decoding loop illustrated in Figure 2.2. The morph image $M$ from source $I_A$ and $I_B$ is obtained as:

$$M := g_s\left((1-\alpha_M)\boldsymbol{w}^*_{I_A} + \alpha_M \boldsymbol{w}^*_{I_B}\right),$$

where, like for landmark-based morphing, we see the appearance of a morphing factor $\alpha_M$ weighing the importance of each source face in the final morph. In practice, it is common to use $\alpha_M = \frac{1}{2}$. Three variants of this methodology are considered in this thesis. The base method, as presented, was proposed in [36]. A first alternative involves exploiting a slightly different latent space for encoding real images. As shown in Figure 2.4, the $\boldsymbol{w}$ latent from StyleGAN is used repeatedly to condition the synthesis process. This can be denoted as follows:

$$g_s(\boldsymbol{w}) = g_s^+(\boldsymbol{w}, \boldsymbol{w}, \ldots, \boldsymbol{w})$$

where $g_s^+$ provides a finer description of each input to the synthesis network. [105] propose to improve the generator inversion process by relaxing the constraint of using the same single latent vector for each input of the synthesizer. This approach involves considering a richer latent space $\mathcal{W}^+$ by defining latent vectors $\boldsymbol{w}_+ \in \mathcal{W}^+$ and associated synthetic images $g_s(\boldsymbol{w}_+)$:

$$\begin{aligned} \boldsymbol{w}_+ &:= [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_R] \qquad\qquad \text{s.t.} \quad \boldsymbol{w}_i \in \mathcal{W} \quad \forall i \in [1, R] \\ g_s(\boldsymbol{w}_+) &:= g_s^+(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_R) \end{aligned}$$

with $R$ being the total number of inputs to the synthesis network. As before, morphs can be generated through the encoding-interpolation-decoding loop, but using the $\mathcal{W}^+$ space this time. The latent representation $\boldsymbol{w}^*_{+,I}$ of a real image $I$ is obtained as:

$$\boldsymbol{w}^*_{+,I} := \min_{\boldsymbol{w}_+ \in \mathcal{W}^+} d_{\text{perceptual}}\left(I, g_s(\boldsymbol{w}_+)\right)$$

In practice, working in the $\mathcal{W}^+$ latent space results in better reconstruction of the source images, which in turn produces more effective morphs.

Finally, [37] introduces a process of latent refinement using a biometric loss, as illustrated in Figure 2.2. The idea is to directly explore the $\mathcal{W}^+$ latent space to find a latent vector $\boldsymbol{w}_{+,M}$ such that the associated synthetic image $M = g_s(\boldsymbol{w}_{+,M})$ is an effective morph. This is achieved by minimizing a combined loss function:

$$\begin{aligned} \boldsymbol{w}_{+,M} := \min_{\boldsymbol{w}_+ \in \mathcal{W}^+} & \frac{\lambda_1}{2}\left[d_{\text{perceptual}}\left(I_A, g_s(\boldsymbol{w}_+)\right) + d_{\text{perceptual}}\left(I_B, g_s(\boldsymbol{w}_+)\right)\right] \\ & + \frac{\lambda_2}{2}\left[d_{\text{biometric}}\left(I_A, g_s(\boldsymbol{w}_+)\right) + d_{\text{biometric}}\left(I_B, g_s(\boldsymbol{w}_+)\right)\right] \end{aligned}$$

where $d_{\text{biometric}}$ measures an identity-based distance between two images, which can be computed using a preexisting FRS. The factors $\lambda_1$ and $\lambda_2$ weight the contribution of each term of the loss, and optimal values have to be determined experimentally. This generation algorithm is named MIPGAN (*Morphing through Identity Prior driven GAN*).[1]

### Properties of diffusion models

There is a large variety of publicly available diffusion models able to generate synthetic faces. However, their synthesis outcome is determined by the noisy input which is image-like, and will be slowly denoised by the reverse-diffusion process to reach the final image. In that context, the closest thing to a conditioning latent vector can only be the noisy input, which is

[1] The full MIPGAN system introduces additional loss terms which are not detailed here. For those details, we refer the reader to the original work [37].

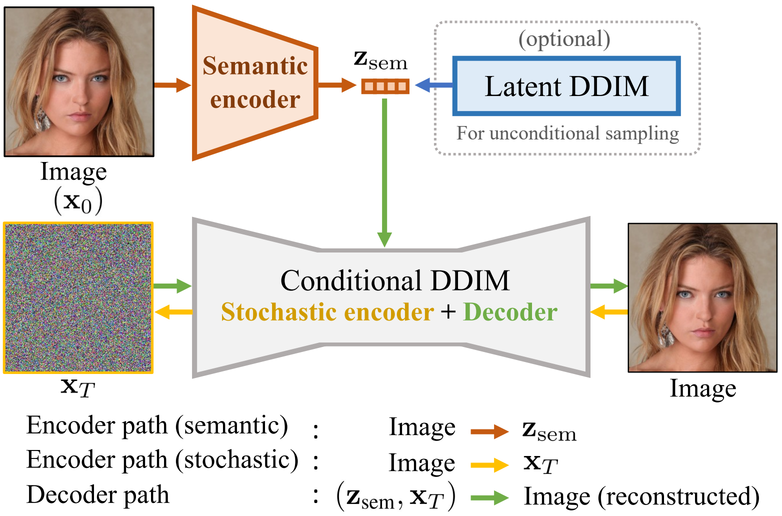


Image
$(\mathbf{x}_0)$
Semantic encoder
$\mathbf{z}_{\text{sem}}$
(optional)
Latent DDIM
For unconditional sampling
Conditional DDIM
Stochastic encoder + Decoder
$\mathbf{x}_T$
Image
Encoder path (semantic) : Image → $\mathbf{z}_{\text{sem}}$
Encoder path (stochastic) : Image → $\mathbf{x}_T$
Decoder path : $(\mathbf{z}_{\text{sem}}, \mathbf{x}_T)$ → Image (reconstructed)

Figure 2.5: High-level illustration of the Diffusion Autoencoder (DiffAE) structure. A standard diffusion model includes only the bottom component, without conditioning. During the forward diffusion process (stochastic encoding), input images are progressively degraded by adding noise until reaching a noise-only image $\boldsymbol{x}_T$. The diffusion generator learns the reverse diffusion process, enabling the generation of pristine images by successively refining an initial fully noisy image (unconditional decoding). The main change for the diffusion autoencoder is the joint learning of a semantic encoder for real images, whose output is used as conditioning to the diffusion model during both the forward and reverse diffusion processes. This enables separate encoding of the semantic and stochastic components of the image ($\boldsymbol{z}_{\text{sem}}$ is the semantic latent, $\boldsymbol{x}_T$ is the stochastic latent). Unconditional generation remains feasible by randomly sampling from the semantic latent space. Figure taken from [106].

empirically difficult to work with and in particular does not provide convenient perceptual continuity. However, an improved diffusion model which uses a more convenient latent conditioning has been developed in [106], and is presented in Figure 2.5. The idea is to extend the design of the diffusion model with an additional encoder, which is learns to encode images into a convenient latent representation $\boldsymbol{z}_{\text{sem}}$ which is further used as conditioning to the actual diffusion model.

With this update, the resulting DiffAE is much more convenient to work with. In the original work, the authors showcase that the $\boldsymbol{z}_{\text{sem}}$ latents appropriately encode the semantic content of the images (with the stochastic content handled by the $\boldsymbol{x}_T$ noisy input still provided to the diffusion model). Moreover, the semantic latent space has a nice structure and good properties of perceptual continuity. Finally, encoding real images is straightforward : the semantic latent of a real image is obtained by inference through the semantic encoder, while the stochastic latent is obtained by forward diffusion process. The complete latent representation of an image is then the combination of its semantic and stochastic latents. With all the requirements satisfied, the DiffAE can effectively be used for face morphing, again using the encoding-interpolation-decoding process. Examples are presented in Figure 2.3b.

## 2.3 Morph vulnerability evaluation

We have seen that by using either landmark-based image manipulations or exploiting latent spaces inherent to preexisting deep face generators, one can generate faces that, to human perception, reasonably seem to mix the identity features of two contributing subjects. This is crucial for an effective morphing attack, as it must deceive a human operator (the office worker receiving the passport picture). However, it also needs to fool an automated FRS. The process of *morph vulnerability evaluation* aims to assess the extent to which a given face morph, or a morph generation algorithm, can lead to verification errors when attacking existing FRSs. This is achieved by simulating attacks on the system and measuring the percentage of those attacks that are considered successful.

Let us consider a set $\mathbb{S}$ of subjects, for each of which one reference source image and at least one probe image is available. We denote by

$$\begin{aligned}
&\mathbb{R} && := \{R_s \,:\, s = 1\ldots|\mathbb{S}|\} && \text{the set of all reference images,}\\
&\mathbb{P}_s && := \left\{P_s^k \,:\, s = 1\ldots|\mathbb{S}|,\ k = 1\ldots|\mathbb{P}_s|\right\} && \text{the set of probes of an individual subject, and}\\
&\mathbb{P} && := \bigcup_{s=1.....|\mathbb{S}|} \mathbb{P}_s && \text{the set of all probes.}
\end{aligned}$$

A set of face morphs $\mathbb{M} = \{M_{ij}\}$ can be created using any generative algorithm $\mathcal{A}$ of choice, by selecting pairs of source images:

$$M_{ij} := \mathcal{A}(R_i, R_j) \qquad \text{s.t. } i \neq j, R_i \in \mathbb{R}, R_j \in \mathbb{R}$$

Not all possible pairs need to be considered. To simulate a realistic scenario, pairs should be selected such that both contributing subjects already bear some resemblance to each other.

Attacks are simulated by measuring the similarity scores between the generated morphs (assumed to be enrolled into a FRS) and the probes from the corresponding contributing subjects (assumed to be verification attempts) using a face recognition system $\mathcal{F}$. For each comparison between a morph $M$ and a probe $P$, the system produces a similarity score $s_{\mathcal{F}}(M,P)$. A verification attempt is considered successful if this score is higher than a predefined decision threshold $\tau(\mathcal{F})$. We can define the indicator function of a successful verification attempt:

$$\chi_v(M,P,F) := \begin{cases} 1 & \text{if } s_{\mathcal{F}}(M,P) > \tau(\mathcal{F}) \\ 0 & \text{otherwise} \end{cases}$$

Several metrics then arise, which depend on the exact criterion to satisfy for a morphing attack to be considered successful:

The Mated Morph Presentation Match Rate (MMPMR) metric [111] assumes only one probe per subject, and considers an individual attack to be successful **if and only if both subjects**

**are successful in their verification attempt.**

$$\text{MMPMR} := \frac{1}{|\mathbb{M}|} \sum_{M_{ij} \in \mathbb{M}} \left[ \chi_v \left( M_{ij}, P_i^1, \mathcal{F} \right) \cdot \chi_v \left( M_{ij}, P_j^1, \mathcal{F} \right) \right]$$

Several extensions of the MMPMR handling multiple probes per subjects (simulating multiple verification attempts) were proposed in the literature [111],[112]. The MinMax-MMPMR considers an attack to be successful **if at least one attempt per subject is successful.**

$$\text{MinMax-MMPMR} := \frac{1}{|\mathbb{M}|} \sum_{M_{ij} \in \mathbb{M}} \left[ \bigvee_{P_i^k \in \mathbb{P}_i} \chi_v \left( M_{ij}, P_i^k, \mathcal{F} \right) \cdot \bigvee_{P_j^k \in \mathbb{P}_j} \chi_v \left( M_{ij}, P_j^k, \mathcal{F} \right) \right]$$

A drawback of the MinMax-MMPMR is that it can somewhat artificially inflate the estimated vulnerability when the number of attempts becomes large (it is likely, in the long run, that at least one will be successful). A proposed alternative is ProdAvg-MMPMR, which first compute the ratio of successful attempts per each subject, then multiplies those ratios to estimate the rate of success of each morph.

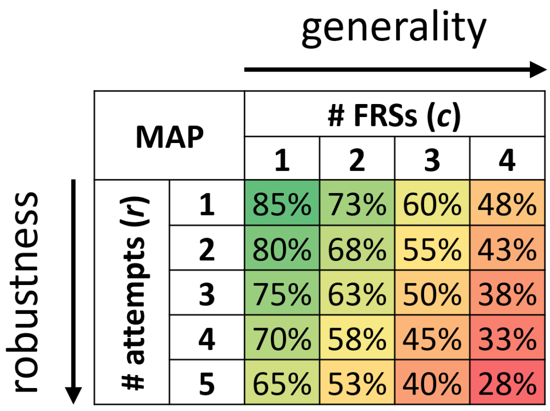
$$\text{ProdAvg-MMPMR} := \frac{1}{|\mathbb{M}|} \sum_{M_{ij} \in \mathbb{M}} \left[ \frac{1}{|\mathbb{P}_i|} \sum_{P_i^k \in \mathbb{P}_i} \chi_v \left( M_{ij}, P_i^k, \mathcal{F} \right) \cdot \frac{1}{|\mathbb{P}_j|} \sum_{P_j^k \in \mathbb{P}_j} \chi_v \left( M_{ij}, P_j^k, \mathcal{F} \right) \right]$$

Finally, the Fully Mated Morph Presentation Match Rate (FMMPMR) [112] aims to establish a much tougher criterion, which is that the morphing attack should be considered successful only **if *all* verification attempts from both subjects are successful.**

$$\text{FMMPMR} := \frac{1}{|\mathbb{M}|} \sum_{M_{ij} \in \mathbb{M}} \left[ \bigwedge_{P_i^k \in \mathbb{P}_i} \chi_v \left( M_{ij}, P_i^k, \mathcal{F} \right) \cdot \bigwedge_{P_j^k \in \mathbb{P}_j} \chi_v \left( M_{ij}, P_j^k, \mathcal{F} \right) \right]$$

More recently, [113] proposed the introduction of the Morphing Attack Potential (MAP) as an enhanced methodology for evaluating morphing attack vulnerability. The MAP is not just a metric but an evaluation protocol with two primary aims: first, to assess the *robustness* of the attack (i.e., its effectiveness across multiple verification attempts from the contributing subjects), and second, to evaluate its *generality* (i.e., its ability to deceive multiple FRSs). Considering multiple verification attempts and multiple FRSs, the MAP is defined as a matrix (cf. Figure 2.6) where the element MAP[$r$,$c$] indicates the percentage of attacks for which at least $r$ verification attempts of both contributing subjects were successful using at least $c$ of the considered FRSs.

In practice, the element MAP[1,1] corresponds to the MinMax-MMPMR, and the element MAP[R,1] corresponds to the FMMPMR in the case where $R$ is the total number of considered verification attempts. Elements for which $c > 1$ indicate how well the attack generalizes to multiple FRSs.


generality
robustness

| MAP | | # FRSs (*c*) | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| # attempts (*r*) | 1 | 85% | 73% | 60% | 48% |
| | 2 | 80% | 68% | 55% | 43% |
| | 3 | 75% | 63% | 50% | 38% |
| | 4 | 70% | 58% | 45% | 33% |
| | 5 | 65% | 53% | 40% | 28% |

Figure 2.6: Example of a MAP matrix. Each element $(r, c)$ of the matrix indicates the rate of attacks for which at least $r$ attempts from both contributing subjects reach a match decision using at least $c$ of the considered FRSs. Figure Taken from [113].

An advantage of the MAP is that it provides a more holistic view of the attack's effectiveness compared to earlier metrics, given its consideration for generalization across FRSs. However, its computation is relatively complex, the results can be quite sensitive to the choice of considered FRSs, and it does not enable easy comparison between different types of morphing attacks due to its matrix form as opposed to a scalar form. Moreover, in operational scenarios, it is expected that a single FRS would typically be used. This FRS could, of course, be designed internally as a combination of several subsystems (which, according to [113], is a sound way to increase its robustness against morphing attacks); but from a vulnerability assessment point of view, we only care about the behavior of the FRS in its entirety. For all these reasons, we chose in this thesis to focus on the MinMax-MMPMR metric (MAP[1,1]) and the FMMPMR metric (MAP[R,1] with R the total number of verification attempts) when performing vulnerability analysis.

## 2.4 Morph datasets

A significant challenge in morphing attack research is the limited availability of morphing datasets. Creating such a dataset requires leveraging an existing face dataset from which the contributing subjects are selected. This face dataset should have clear identity labels and images captured in a controlled environment, ideally meeting quality and constraints requirements expected from passport pictures. Consequently, datasets originally developed for face recognition research are typically used. However, due to the sensitive nature of biometric data, the licenses and consent forms associated with these datasets often prohibit the publication of derivatives, notably encompassing morphing attacks in this restriction.

Two notable public datasets are worth mentioning: the first is the FRLL-Morphs dataset introduced by [36]. This dataset is based on the Face Research Lab London (FRLL) dataset, which contains high-resolution images of 102 subjects with passport-like qualities (good illumination, neutral expression, and frontal pose). The morphing attack data includes three landmark-based morphing algorithms and one GAN-based algorithm. Unfortunately, due to the very limited amount of bona fide samples in FRLL, FRLL-Morphs can only be used for testing purposes, and the resulting metrics can only be relatively coarse.

The second dataset is the Synthetic Morphing Attack Developement (SMDD) dataset introduced by [114]. This dataset addresses privacy concerns by utilizing fully synthetic faces generated by a GAN as the morphing sources. Due to its larger scale, the SMDD dataset is convenient for training MAD systems. However, it has two main drawbacks. First, the released version includes only a single morphing algorithm, which is relatively limiting. Second, the fact that the source images themselves are synthetic poses conceptual challenges in the context of this thesis. As we aim to draw inspiration from deep synthetic image analysis literature, it is problematic to work with data where the "bona fide" class is actually synthetic.

Finally, some datasets have been created specifically for the development of the two main public MAD benchmarks. The first benchmark, FATE-Morph by NIST [101], evaluates both morphing attack vulnerability and detection. The second, FVC-Ongoing ny the University of Bologna [115], focuses primarily on detection. However, the associated datasets are kept sequestered on the benchmarking platforms to prevent their use as training data, which would compromise the fairness of comparisons among submitted solutions. This restriction means they cannot be used for diagnosing failure cases of detection methods in development, as there is no way to examine misclassified examples looking for patterns. Additionally, these datasets are heavily biased towards landmark-based morphs, and typically contain a limited number of bona fide samples.

Therefore, we opted to generate our own experimental datasets using established morphing attack algorithms. This approach provides precise control over the experimental data, allows the inclusion of the latest morphing techniques, and ensures a well-defined ground truth regarding the source algorithm for each morph. This latter point is particularly important for attribution tasks.

Table 2.1: Number of samples in each dataset and split. We indicate the number of attack samples *per morphing algorithm*, i.e. the total number of attack samples used in experiments should be obtained by multiplying the provided value by the number of considered morphing algorithms.

| | # bona fide | | # per attack | |
|---|---|---|---|---|
| Src. dataset | Train | Test | Train | Test |
| FRGC | 9228 | 2304 | 2014 | 507 |
| FRLL | - | 204 | - | 1140 |
| FFHQ | 8000 | 2000 | 4000 | 1000 |

Table 2.2: Available image sets. Attacks are grouped into higher level families indicated in the first row. Most attacks are available in digital format (○), some of them have their test set in print-scan format as well (•). The FRGC-MIPGAN attack is used only for testing purpose in the print-scan domain (⋆).

| | LB | | GAN | | Diff | | |
|---|---|---|---|---|---|---|---|
| | **LB-Complete**[102] | **LB-Combined**[103] | **SG2-W**[36] | **SG2-W+**[105] | **MorDIFF**[49] | **MIPGAN**[37] | **Bona fide** |
| FRGC | ○ | • | ○ | ○ | ○ | ⋆ | • |
| FRLL | ○ | ○ | ○ | ○ | ○ | | ○ |
| FFHQ | ○ | ○ | ○ | ○ | ○ | | ○ |

We create morphs using three distinct source datasets: the FRLL dataset [116], the FRGC dataset [117], and the FFHQ dataset [18]. The FRLL and FRGC datasets are extensively utilized in prior research on face morphing due to their constrained facial images (frontal pose, neutral expression) with consistent backgrounds and illumination. These characteristics render them suitable for morph generation. In contrast, the FFHQ dataset, collected from Flickr, exhibits greater diversity. We hypothesize that employing a more diverse source dataset is advantageous for our research objectives, particularly in studying cross-source dataset generalization and one-class modeling of the bona fide class.

For the FRLL and FRGC datasets, we select identity pairs for morph creation following previous research works [103] and [37], respectively. This results in 1'140 pairings for FRLL and 2'521 pairings for FRGC. For the FFHQ dataset, we initially select 10'000 images from the original dataset of 70'000 images, focusing on those with the most frontal poses, which are then randomly paired to form 5'000 morphing pairs. While this process might yield some unrealistic morphs (e.g., morphs between different genders), it allows the creation of a large set of samples containing the relevant attack artifacts and showcasing high diversity. Hence, this set remains valuable for training MAD systems.

Using consistent pairings, we generate morphs from these source datasets employing five different attack algorithms. These include two landmark-based algorithms (**LB-Complete** [102] and **LB-Combined** [103]), two GAN-based algorithms using StyleGAN2 (**SG2-W** [18] and **SG2-W+** [105]), and one diffusion-based algorithm using DiffAE (**MorDIFF** [49]). Examples of the generated morphs are presented in Figure 2.7.

For the bona fide sets, we use original images from the source datasets. For FRLL, we use the only available 204 frontal images, some of which have also been used as sources for morphing.

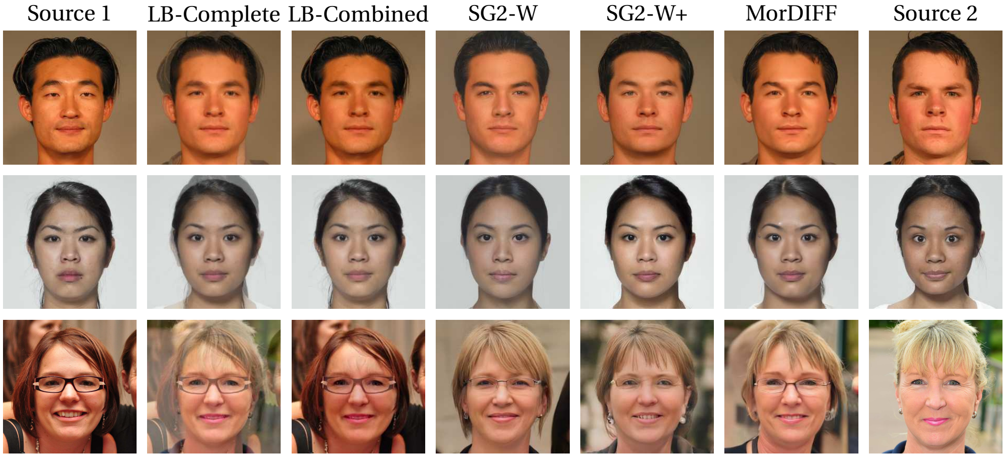


Figure 2.7: Examples of generated morphs using as source dataset respectively Face Recognition Grand Challenge (FRGC) (first row), FRLL (second row) and Flicker-Faces HQ (FFHQ) (third row). The first and last column show the two real sources for which a morph must be created, and other columns show the results using each considered morphing algorithm.

Due to this low amount of bona fide images, we restrict the usage of FRLL to test purposes. For FRGC and FFHQ, we select bona fide images containing identities never used for morphing, with 11,532 and 10,000 images, respectively. We split both the bona fide sets and attack sets into training and test sets using an 80-20 ratio, ensuring that identities are disjoint between the training and test sets for bona fide images, and that pairs of identities are disjoint between the training and test sets for the attacks. The exact number of samples in each dataset is detailed in Table 2.1.

Additionally, we create a “real-world” test dataset by printing and scanning a subset of images. Specifically, the bona fide test samples from FRGC, morph test samples created using FRGC with the LB-Combined and MorDIFF algorithms, and an additional set of FRGC morphs created using another unseen algorithm, MIPGAN [37]. This simulates a challenging scenario where we must generalize from the digital to the print-scan domain, and towards unseen attacks. The morphs are printed at a size of 35 mm ×35 mm then rescanned at a resolution of 300 DPI, using a *Kyocera TASKalfa 2554ci* (laser printer + scanner). As preprocessing, all images are cropped to 256×256 pixels while ensuring consistent landmark alignement.

Available image sets are summarized in Table 2.2. For the experiments, we sometimes regroup attacks into higher level families, respectively landmark-based (LB), GAN-based (GAN), and diffusion-based (Diff).

## 2.5 Morphing attack detection

### 2.5.1 Prior work

Morphing Attack Detection (MAD) systems can be broadly categorized into single MAD and differential MAD. Single MAD aims to assess the authenticity of a single image, such as a registered passport picture, while differential MAD also exploits probe information, such as the live-captured image of the passport holder at the Automated Border Control (ABC) gate. Our work will deal only with single MAD.

MAD systems can be categorized into those using handcrafted features and those using deep features [118]. Handcrafted features typically rely on texture cues (e.g, Local Binary Patterns (LBPs) [119]) or cues regularly used in image forensic analysis (e.g., frequency analysis, or noise patterns such as the PRNU [120]). Deep features, on the other hand, are learned in a data-driven manner by training a neural network (usually a CNN) on examples of bona fide and morphed images.

A significant portion of research has focused on landmark-based morphs, with somewhat more limited attention given to GAN-based morphs and almost none to the more recent diffusion-based morphs, such as those introduced in [49]. This is notably illustrated by the rarity of deep morphing algorithms in the NIST FATE MORPH [101] benchmark, and the SOTAMD dataset from the benchmark of University of Bologna [115]. In practice, handcrafted features developed for landmark-based MAD are not particularly effective for deep morphs, as demonstrated in [121]. The effectiveness of deep features is strongly dependent on the training data, and generalization from a training dataset containing only landmark-based morphs to one containing deep morphs is not guaranteed, as observed in [49].

Notable exceptions include two works that approach MAD as an anomaly detection problem. Both design an image-reconstruction network that aims to degrade then reconstruct bona fide input images. This process is done through an autoencoder in [122], and by a noise-denoise process in [123] using diffusion models. They then observe that the reconstruction error differs between bona fide images and morphs, although it is *lower* for morphs in [122] but *higher* in [123]. The reconstruction error is thus discriminative for detection purposes. One main advantage of such approaches is that they are one-class, relying only on bona fide data and not on specific attacks in the training set, making them less prone to bias towards a specific family of morphing methods. However, evaluation on diffusion morphs, for example, is not provided in these works.

Overall, the detection of deep morphs has often been approached by adapting existing methods designed for landmark-based morphs, such as incorporating deep morphs into the training datasets for data-driven detectors. A core rationale of this thesis is to consider the reverse perspective: treating MAD as a *deepfake* detection problem, specifically focusing on detecting deep synthetic images. With respect to typical deepfake detection, one must then address two additional challenges: keeping the ability to handle the fundamentally different nature of landmark-based morphs, and ensuring robustness against print-scan post-processing—a degradation not typically considered in deepfake detection literature. This reasoning enables us to take inspiration from a wide set of literature on deep synthetic image detection, cf. the methods presented in the previous chapter.

### 2.5.2 Metrics

MAD is approached as a binary classification problem, which involves developing a detector $\mathcal{D}$ that, given an input image $I$, produces a score $s$. There is a slight mismatch between the formalisms of the biometrics community and that of the deepfake detection community regarding the interpretation of this score. We aim to clarify this nuance:

- In the biometrics community, the score $s$ is typically expected to be high for a bona fide face and low for an attack. This is because the system is expected to accept verification attempts with bona fide probes (positive result) and reject attacks (negative result).
- In the deepfake detection community, the score $s$ is expected to be high for attacks and low for bona fide samples. This is because the system is expected to raise alarms in the presence of forgeries (positive result) and not react if no forgery is observed (negative result).

In this work, we adopt the second formalism. For the final decision, the produced score $s$ is compared to a decision threshold $\tau$; the image is classified as an attack if the score is above the threshold, and as bona fide otherwise. Systems are then compared by computing predictions on a reference test dataset containing both bona fide and attack samples, and measuring the following error rates:

- Attack Presentation Classification Error Rate (APCER): the proportion of attack samples incorrectly classified as bona fide.
- Bona Fide Presentation Classification Error Rate (BPCER): the proportion of bona fide samples incorrectly classified as attacks.

These metrics are conceptually equivalent to the false accept and false reject rates in the biometrics community, or to the false negative and false positive rates in the deepfake detection community. However, the specific terminology ensures clarity and avoids confusion. Table 2.3 summarizes the different conventions.

Since these metrics depend on the choice of threshold, it is common practice to select a target value for the BPCER, compute the operating threshold that satisfies this target value, and then report the corresponding APCER at this threshold. In other words, this approach involves setting a convenience criterion (how often we tolerate a bona fide face raising alarms) and evaluating the corresponding level of security (how often we miss an attack).

To obtain a threshold-independent analysis, it is common to visualize Detection Error Tradeoff (DET) curves, which plot the BPCER values against the corresponding APCER values across the full range of possible thresholds. Additionally, works in the literature often report the Detection Equal Error Rate (D-EER) value, which is the error rate at the threshold where APCER and BPCER are equal.

| Community | Ambiguity | Fake misclassification rate | Bona fide misclassification rate |
|---|---|---|---|
| Biometrics (verification) | What is *accept* ? →Bona fide (valid attempt) | False accept rate (FAR) | False reject rate (FRR) |
| Deepfake detection | What is *positive*? →Fake media | False Negative Rate (FNR) | False Positive Rate (FPR) |
| Biometrics (Presentation Attack Detection) | Unambiguous | Attack Presentation Classification Error Rate (APCER) | Bona Fide Presentation Attack Error Rate (BPCER) |

Table 2.3: Overview of the naming conventions present in biometrics and deepfake detection. Literature in morphing attack detection reuses the convention from presentation attack detection, which relieves ambiguity through the use of more specific terminology. In rare cases in MAD, the term APCER is replaced by the Morphing Attack Classification Error Rate (MACER).

With this, we are ready to delve into the research contributions of this thesis, which we start in chapter 3 by providing a comprehensive analysis of the vulnerability of FRSs to morphing attacks, with a particular focus on comparing the risks posed by deep morphs versus landmark-based morphs. Additionally, we introduce a novel family of deep morphing attacks that exploit **face template inversion systems**, which are models capable of reconstructing face images from face embeddings.

# Research contributions Part II

# 3 Risks: inversion-based morphing and vulnerability assessment

In this chapter, we present two experimental contributions aimed at assessing and quantifying the practical risks posed by morphing attack algorithms. Specifically:

- We observe the development of **face embedding inversion** models, which enable the reconstruction of face images from embedding representations extracted by face recognition networks. We hypothesize that these models have direct applications in face morphing generation, by using them to invert a theoretical optimal morph embedding.
- We design and implement a novel family of morphing algorithms based on face embedding inversion, evaluating whether this new category of **inversion morphs** poses significant risks.
- We conduct a comprehensive evaluation of face morphing attack vulnerability, considering all existing families of face morphing algorithms (landmark-based, GAN-based, and diffusion-based) in addition to our proposed inversion-based morphs. By having access to all considered generation algorithms, we ensure completely fair comparisons between algorithms by using identical vulnerability assessment protocols in all cases, both for morphing pair selection and probe selection.
- Additionally, we evaluate the studied attacks in terms of realism (both qualitatively and quantitatively) and detectability using publicly available detectors.

## 3.1 Methodology

### 3.1.1 Inversion morphing

The concept of inversion-based morphing seamlessly integrates into the overarching morphing framework presented in the previous chapter (Figure 2.2).

All the deep morphing methods discussed earlier (particularly those not involving a latent refinement step) heavily rely on the assumption of linear perceptual continuity within the

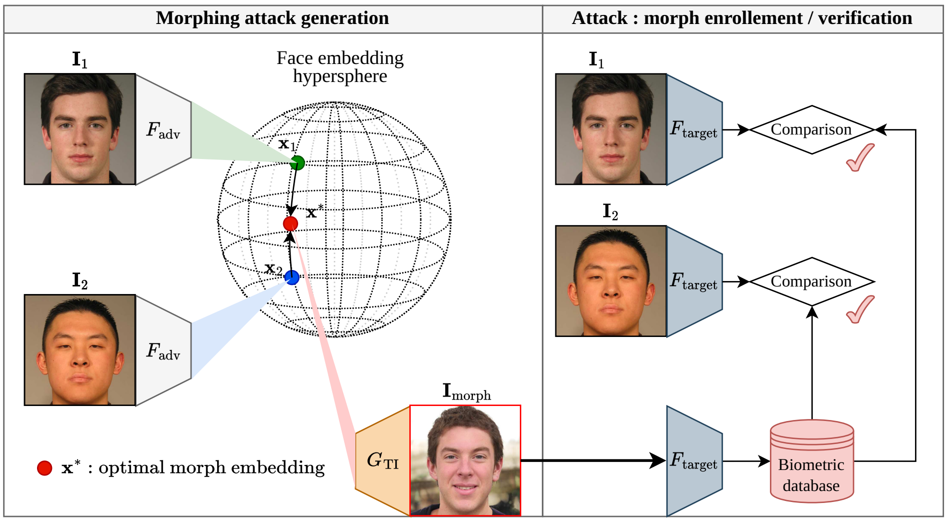


Figure 3.1: Morphing attack based on the optimal morph embedding: 1) Morph Generation: face embeddings are extracted from the source face images using a face recognition network $F_{adv}$ available to the adversary. Then, an optimal morph embedding is computed in the embedding space. Finally, the optimal morph embedding is fed to a template inversion model $G_{\text{TI}}$ to generate a candidate morph image. 2) Attack: the generated morph is registered as biometric reference in a database using a distinct face recognition network $F_{\text{target}}$, the target of the attack. In successful attacks, both contributing subjects can authenticate against the stored reference e.g., share a passport.

latent space. This assumption ensures that the midpoint between two source latents will decode into a satisfactory morph, exhibiting roughly equal perceptual similarity to both source images. However, this perceptual similarity relates to generic image features rather than specific identity attributes. In other words, while the morph may "look similar *overall*" to both sources, it does not necessarily "look similar *in identity*". Although these two notions often align (as demonstrated by the success of previous deep morphing works), they remain conceptually distinct.

This raises the question of identifying a latent space with stronger conceptual guarantees that a gradual transition along a segment will correspond to a gradual change in the identity of the generated image. Fortunately, such a space exists: it is the embedding space of a pretrained face recognition network. Encoding faces in this embedding space and performing interpolation there provides the strongest theoretical guarantee that the resulting morph will be highly effective. By definition, the identity distance between two face images is measured as the distance between their face embeddings. Therefore, the midpoint between two face embeddings will mathematically be the **optimal morph embedding**, i.e., the face embedding that minimizes the identity-distance to both sources (at least, in the metric space derived from the considered face recognition network).

The theoretical notion of optimal morph embedding is proposed in [124]. Let us consider a

FRS $f : X \to Z$ mapping face images $\boldsymbol{x} \in X$ to their face embedding $\boldsymbol{z} \in Z$, and $d$ a distance function on $Z$ used to evaluate the identity distance between pairs of embeddings in $Z$. The optimal morph embedding $\boldsymbol{z}^*$ for two face images $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ is defined as:

$$\boldsymbol{z}^* = \arg\min_{\boldsymbol{z} \in Z} \left[ \max\left( d(\boldsymbol{z}, f(\boldsymbol{x}_1)), d(\boldsymbol{z}, f(\boldsymbol{x}_2)) \right) \right]$$

In the case where $d$ is the cosine distance, which is the case for all FRSs considered in this work, the following is a possible solution for the optimal morph embedding:

$$\boldsymbol{z}^* = \frac{\boldsymbol{z}_1 + \boldsymbol{z}_2}{||\boldsymbol{z}_1 + \boldsymbol{z}_2||}$$

where $\boldsymbol{z}_1 = f(\boldsymbol{x}_1)$, $\boldsymbol{z}2 = f(\boldsymbol{x}_2)$.

Building on the diagram from Figure 2.2, we can conceptualize inversion morphing as the process of

- utilizing a *pretrained face recognition network* as the encoder,
- performing latent manipulation (optimal morph embedding computation) within the face embedding space, which serves as the latent space,
- and finally decoding face embeddings back into images.

The process is illustrated in Figure 3.1. A crucial component is the decoder that converts face embeddings back into face images. Interestingly, there is a research line in biometrics security that specifically focuses on reconstructing faces from face embeddings, a process known as **template inversion**.

Originally, template inversion research aimed to assess the security of biometric template databases, particularly evaluating the sensitive information that could be reconstructed from stolen templates. Recent studies [125], [126] have demonstrated the feasibility of reconstructing face images from stolen face templates and using these reconstructed images to breach face verification systems. A byproduct of this research is the development of decoders that convert face embeddings back into face images, which are essential for our theoretical inversion morphing method.

A brief terminology clarification is necessary to distinguish between a face embedding and a face template. In machine learning, “embedding” refers to numerical representations of information (here, a face image) in a continuous space that captures semantic meaning (here, the identity). In biometrics, “template” refers to the reference of a given biometric subject which is stored in a biometric database. A face template might be based on multiple face embeddings combined in a certain way, or on other data than face embeddings. Nevertheless, for our application to morph inversion, we are dealing with a subcase where the two notions align:

the face templates targeted by template inversion literature *are* individual face embeddings extracted by a FRS.

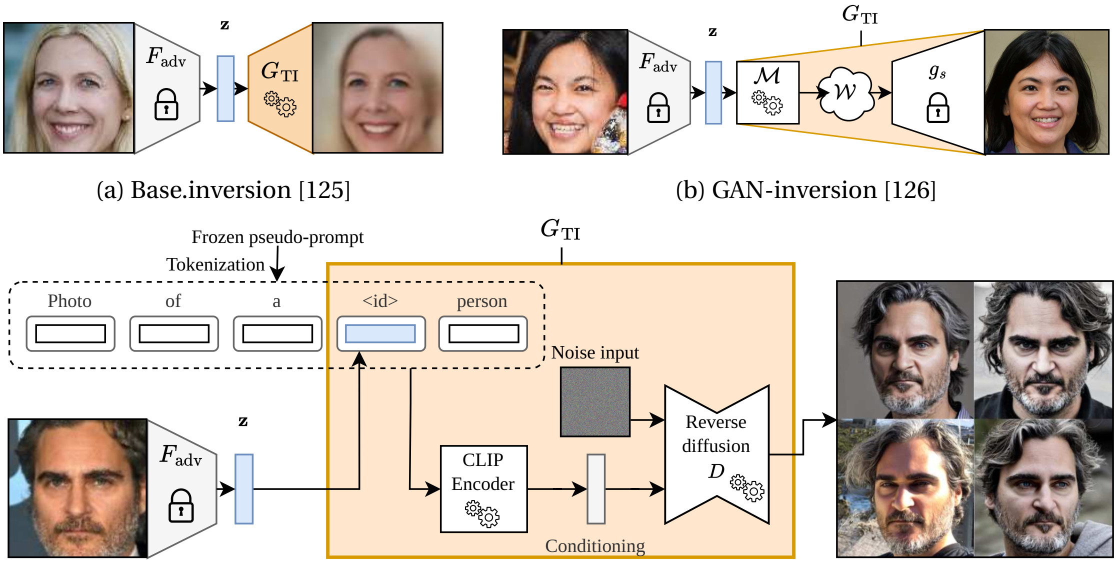
### 3.1.2 Threat model

Morphing attacks based on inversion morphing involve a face recognition system as part of the generation process, and another face recognition system that is the target of the attack. Those two systems may or may not be identical, depending on the exact threat model, which we clarify hereafter:

- *Adversary's goal*: The adversary aims to create a face morph image $I_{\text{morph}}$, mixing the identities from two source images $I_1$, $I_2$, which are for two different subjects; then enroll $I_{\text{morph}}$ into a face recognition database (e.g., passport creation). Afterhand, the goal is for both contributing subjects to successfully authenticate against the stored reference (e.g., enabling them to share the passport to go through an automated border control gate).
- *Adversary's knowledge:* The adversary is assumed to have the following information:
  - The adversary has access to a face recognition network $F_{\text{adv}}$ and a template inversion network $G_{\text{TI}}$, which is able to invert face embeddings extracted by $F_{\text{adv}}$.
  - The adversary may also have a *white-box* knowledge of the target face recognition system $F_{\text{target}}$, and can use it during the morph generation process (i.e., $F_{\text{adv}} = F_{\text{target}}$). Otherwise, the target face recognition system $F_{\text{target}}$ is not used by the adversary, and therefore $F_{\text{adv}} \neq F_{\text{target}}$. We refer to this scenario (i.e., $F_{\text{target}}$ is not white-box) as the *black-box* scenario in our experiments.
- *Adversary's capability:* The adversary can submit the generated morph for enrollment into a target face recognition system $F_{\text{target}}$. We consider two scenarios for the enrolment process:
  1. The adversary can submit $I_{\text{morph}}$ as a digital image for enrollment.
  2. The adversary needs to print the image $I_{\text{morph}}$, which will be then scanned for enrollment (print-scan).
- *Adversary's strategy:* The adversary's strategy is to compute face embeddings from both contributing subjects using $F_{\text{adv}}$ and average them to obtain an optimal morph embedding, then invert this embedding back into $I_{\text{morph}}$ using the template inverter $G_{\text{TI}}$.

### 3.1.3 Implementation

We thus intend to take preexisting face template inversion models and use them to invert optimal morph embeddings, before evaluating to what extent the resulting face images can be used for morphing attacks. We consider three distinct pretrained inversion models from the literature, denoted by $G_{\text{TI}}$, which are elaborated in Figure 3.2. The first model [125], referred to as **base-inversion**, constructs $G_{TI}$ as an independent CNN trained from scratch for the purpose of face reconstruction. The second model [126], called **GAN-inversion**, learns a mapping from the face embedding space to the $\mathcal{W}$ space of a StyleGAN3 model trained for face generation. This approach leverages the generation realism of StyleGAN3 and simplifies the inversion learning process by mapping between two structured spaces. The third model [127], called **diffusion inversion**, adapts a text-to-image diffusion model (Stable Diffusion) through fine-tuning to replace text conditioning with conditioning based on a face embedding. An additional advantage of this last generator is the ability to generate multiple images representing the same reconstructed identity by sampling the diffusion model with various noise inputs while using the same face embedding for conditioning.



(a) Base.inversion [125]

(b) GAN-inversion [126]

(c) Diffusion-inversion [127]

Figure 3.2: Overview of the models considered for face template inversion. **Base-inversion**: an independent CNN is trained from scratch for face reconstruction. **GAN-inversion**: a mapping network $\mathcal{M}$ is trained to map face embeddings to a latent $\boldsymbol{w}$ in the $\mathcal{W}$ space of a StyleGAN2 network, ensuring that the resulting synthetic image $g_s(w)$ (with the synthesis network $g_s$ frozen) accurately reconstructs the identity. **Diffusion-inversion**: utilizes a text-to-image diffusion model (Stable Diffusion). Face embeddings are injected into a fixed pseudo-prompt, then processed by the CLIP encoder to generate the conditioning vectors. Both the CLIP encoder and the conditioned diffusion model are fine-tuned to ensure precise identity reconstruction. After fine-tuning, the CLIP encoder relies solely on the face embedding, and the diffusion model accurately reconstructs the identity. Synthetic variations of the same image can be generated by altering the noise input to the reverse diffusion process.

### Latent refinement

Any considered inversion template $G_{\text{TI}}$ can be used as is to generate a candidate morph $I_{\text{morph}}$ from the optimal morph embedding $z^*$, i.e.,

$$I_{\text{morph}} := G_{\text{TI}}\left(z^*\right)$$

Ideally, we then have $F_{\text{adv}}(I_{\text{morph}}) = \boldsymbol{z}^*$, meaning that extracting a face embedding from the morph image $I_{\text{morph}}$ directly recovers the optimal embedding $z^*$. However, due to potential imperfections in the template inverter $G_{\text{TI}}$, this property is not guaranteed. To address this, we can enforce it through a latent refinement process, which computes a slightly adjusted input $\boldsymbol{z}^*_{\text{opt}}$ to $G_{\text{TI}}$ that satisfies:

$$\boldsymbol{z}^*_{\text{opt}} = \underset{z \in Z}{\arg\min} \left|\left|\boldsymbol{z}^* - [F_{\text{adv}} \circ G_{\text{TI}}](\boldsymbol{z})\right|\right|_2$$

In other words, we use a morph embedding that deviates from the optimal $\boldsymbol{z}^*$, but such that its corresponding reconstructed face with $G_{\text{TI}}$ actually maps back to the optimal embedding $\boldsymbol{z}^*$. This corrects potential reconstruction errors or biases in the template inverter, leading to an even stronger morphing attack candidate.

In the case of GAN-based inversion, since we have $G_{\text{TI}}(\boldsymbol{z}^*) = g_s\left(\mathcal{M}\left(z^*\right)\right)$, we can also perform the latent refinement not in the face embedding space, but instead refine the StyleGAN latent $\boldsymbol{w}^* = \mathcal{M}(\boldsymbol{z}^*)$. This optimization can be done either in the $\mathcal{W}$ or $\mathcal{W}+$ space, leading to the respective objectives:

$$\boldsymbol{w}_{\text{opt}} = \underset{\boldsymbol{w}}{\arg\min} \left|\left|\boldsymbol{z}^* - [F_{\text{adv}} \circ g_s](\boldsymbol{w})\right|\right|_2, \quad \boldsymbol{w} \in \mathcal{W}$$

$$\boldsymbol{w}_{+,\text{opt}} = \underset{\boldsymbol{w}_+}{\arg\min} \left|\left|\boldsymbol{z}^* - [F_{\text{adv}} \circ g_s](\boldsymbol{w}_+)\right|\right|_2, \quad \boldsymbol{w}_+ \in \mathcal{W}+$$

In practice, we always perform this optimization by gradient descent, using $\boldsymbol{z}^*$ as the initialization, or $\boldsymbol{w}^* = \mathcal{M}(\boldsymbol{z}^*)$ when refining in $\mathcal{W}$ or $\mathcal{W}+$. We use the Adam [128] optimizer for 100 iterations with a learning rate of $2.5 \times 10^{-3}$.

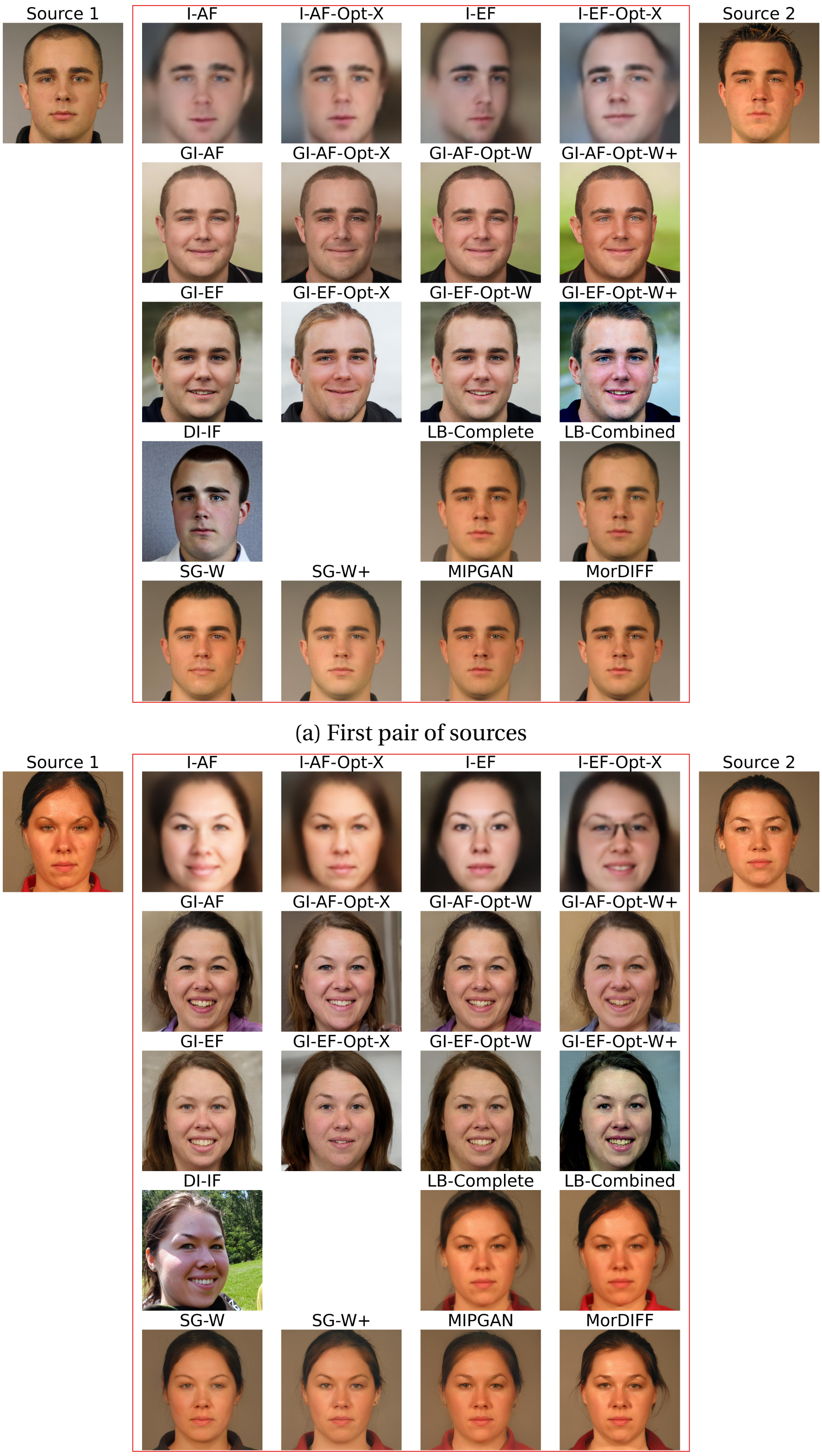
We note that refining in $\mathcal{W}+$ has similarities with MIPGAN, which creates morphs by an optimization-based exploration of the $\mathcal{W}+$ space in search of effective morph latents. The main differences are that we initialize the optimization process using the output of the template inverter (which can lead to a different optimization result depending on the presence of local minimas in the loss landscape), and we only use a biometric loss, whereas they use a composite loss also considering pixel-level aspects.

Finally, in the case of diffusion-inversion, performing this optimization process is not feasible due to the heavy computational cost of backpropagating through the reverse diffusion process. However, we can generate multiple face reconstructions by using different noise inputs. Exploiting this, we perform a greedy selection of the best face morph by using $k$ different noise inputs, computing candidate morphs $I = D(\boldsymbol{z}^*, \boldsymbol{n}, t)$ with each noise $\boldsymbol{n}$ and number of reverse diffusion steps $t$, and then selecting the generated image whose embedding $\boldsymbol{z} = F_{\text{adv}}(\boldsymbol{I})$ is the most similar to the optimal morph embedding $\boldsymbol{z}^*$. In our experiments, we generate each image with $t = 25$ iterations and generate $k = 10$ images for each given optimal morph embedding.

We summarize in Table 3.1 the resulting inversion morphing attack variants that are considered for our experiments. We use the same pairing protocols as explicited in Chapter 2. For both base-inversion and GAN-inversion, we consider two distinct template inversion models that have been trained to respectively invert the ArcFace model [129] and the ElasticFace model [130]. The diffusion-inversion model is trained to invert ArcFace embeddings only. Additionally, we will also consider previously developed attacks, which have been already introduced in Chapter 2. This will provide us with a complete picture of the vulnerability of face recognition systems to deep morphs, including but not limited to our proposed inversion morphs. We present in Fig. 3.3 examples of the resulting morphs using each of the considered methods.

Table 3.1: Overview of the considered morphing methods. The first block contains all considered inversion attacks, which are all our contribution. The second block lists previously proposed methods which we include in our analyses.

| Name | Approach |
|---|---|
| I-AF/EF | Base-inversion of optimal embedding (ArcFace or ElasticFace) |
| + Opt-X | + optimization of the inverter's input |
| GI-AF/EF | GAN-inversion ($\mathcal{W}$) of optimal embedding (ArcFace or ElasticFace) |
| + Opt-Z/W/W+ | + optimization of the inverter's input (Z) or the synthesis network input (W, W+) |
| DI-AF | Diffusion-inversion of optimal embedding (ArcFace) with greedy selection of the best morph among 10 candidates |
| [36]SG2-W | Encoding and interpolation in StyleGAN2 $\mathcal{W}$ space |
| [105]SG2-W+ | Encoding and interpolation in StyleGAN2 $\mathcal{W}+$ space |
| [37] MIPGAN | Optimization in StyleGAN2 $\mathcal{W}+$ space |
| [49] MorDIFF | Encoding and interpolation in the latent space of a Diffusion Autoencoder |
| [102]LB-Complete | Landmark-based complete morph |
| [103]LB-Combined | Landmark-based combined morph |



(a) First pair of sources

(b) Second pair of sources

Figure 3.3: All types of considered deep morphs for two different pairs of source identities.

## 3.2 Experiments

We aim to evaluate the effectiveness of our proposed morphing attacks. The effectiveness of a morphing attack depends on its ability to fool both a face recognition system (e.g., at an automated border control gate), and a human operator (e.g., the administrative employee receiving and processing the image submitted for a passport application). We compare the performance of our attack across those two aspects against previous attacks proposed in the literature.

The ability of the attack to fool FR systems is evaluated through the mean of a vulnerability analysis, which simulates attacks by enrolling the morphs into a biometric verification system, then evaluating what percentage of them allow both contributing subjects to successfully authenticate (cf. Chapter 2) While in some cases, passport application photos can be submitted in digital format, they might sometimes have to rather be printed and physically sent to the processing office, where they will be scanned for redigitalization. Independent vulnerability analyses should thus be performed using the morphs either in their digital form, or after performing this print-scan operation for a more real-world setting. These analyses are performed in Section 3.2.1. We first compare several variants of our proposed attack to evaluate the importance of the input optimization step for the effectiveness of the attack. We then compare our proposed attack to previously proposed ones, both considering deep morphing and landmark-based morphing. Finally, we select a subset of attacks which are subjected to a print-scan process, and evaluate whether our conclusions change in this more realistic setting.

The ability of the attack to fool humans has two components. First, the morphed image should be sufficiently realistic looking to not raise suspicion. Secondly, it might also be necessary that the morph is good enough of a lookalike to both source identities (or at least of the accomplice's identity). While this is in principle already evaluated in the vulnerability analysis, there might be cases where human perception of identity differs from that of the automated system. Those aspects are discussed in Section 3.2.2, first with a qualitative discussion of the morphed faces, then with a quantitative evaluation of the morphs realism using the Fréchet Inception Distance (FID) metric.

### 3.2.1 Effectiveness on face recognition systems

We evaluate the ability of our proposed morphing attacks to fool face recognition systems through a **vulnerability analysis** study, whose point is to simulate morphing attacks on a FR system and evaluate the rate of successful attacks. The morphs are enrolled as reference in a FR system, simulating a passport application. A specific operating threshold for the FR system is calibrated on a bona fide evaluation protocol, with a tolerance for a false match rate (FMR) of $10^{-3}$, following the FRONTEX guideline [131]. For each morph, probes from both contributing subjects are then presented to the system, trying to authenticate under the same registered identity represented by the morph. The morph is considered successful if both subjects manage to authenticate. We evaluate this through the Mated Morph Presentation Match Rate, described in [111], for which we consider the MinMax variant, which considers an attack successful if at least one probe from each subject leads to a succesful verification attempt. We also measure the Fully Mated Morph Presentation Match Rate [112], which considers an attack successful if *all* probes from each subject lead to successful verification attempts. Mathematical details of the metrics have been provided in the previous chapter (section 2.3).

We restrict the analysis to morphs created from the FRLL and FRGC datasets (cf. Table 2.2). For the vulnerability evaluation with FRLL morphs, we probe the system with all available frontal poses of the contributing subjects. When working with FRGC, we reuse the probes from [37], de facto replicating their vulnerability evaluation protocol.

We make use of two FR systems for the analysis: ArcFace (AF) [129], and ElasticFace (EF) [130]. Given inversion morphs necessitate a FR system for the generation itself, we generate independent morph sets using each of those models, except for diffusion-inversion for which we only have access to a model conditioned on ArcFace embeddings. We then use the same two models for simulating the considered attacks and evaluating their effectiveness. This approach in particular enables comparison between white-box attack scenarios (FR system available at morph generation time, i.e., when the same network is used for generation and evaluation), and black-box attack scenarios (when the attacked network differs from the one used for generation). On that aspect we also note that MIPGAN makes uses of the ArcFace network for computing the biometric component of the loss when fine-tuning the latent morph; for this reason, we classify a MIPGAN attack on the ArcFace system as *white-box*. Finally, for calibrating the operating threshold we use the FRGC Experiment 2 protocol [117] and operate at a tolerated FMR of $10^{-3}$.

#### Importance of fine-tuning the input to the inverter

Using this analytical approach, we first assess the impact of fine-tuning the latent morph representation on the attack's effectiveness. As detailed in Section 3.1.3, this involves optimizing the input to the inverter to better ensure that the generated morph accurately maps back to the optimal morph embedding. The results of the vulnerability evaluation across these different methodologies are presented in Table 3.2.

We observe that in all instances, the optimization process enhances the attack's effectiveness. For base inversion methods, while the effectiveness is already notably high without optimization, incorporating it further elevates the MMPMR close to its maximum potential. For GAN-Inversion, the impact is even more pronounced, with approximately three to six times more successful attacks when optimization is applied. Additionally, we note that performing the optimization in the $\mathcal{W}+$ space yields the best results.

One might be concerned that incorporating the optimization step could lead to overfitting the attack to the specific FR system used, especially given that the same system is employed to train the inverter. However, our observations indicate otherwise: even in black-box attack scenarios, the optimization step significantly increases the attack's effectiveness. This implies that the image adjustments introduced by the optimization are robust and generalize well across different FR systems.

Table 3.2: Effect of optimizing the input to the inverter on the attack effectiveness. We distinguish white-box □ and black-box ■ attacks.

| | | MinMax-MMPMR (%) | | FMMPMR (%) | |
|---|---|---|---|---|---|
| FRS | Attack | FRLL | FRGC | FRLL | FRGC |
| AF | □ I-AF | 97.54 | 89.88 | 91.49 | 54.90 |
| | □ I-AF-Opt-X | **100.00** | **99.80** | **99.82** | **94.25** |
| | □ GI-AF | 52.02 | 41.81 | 33.86 | 8.41 |
| | □ GI-AF-Opt-X | 91.75 | 80.72 | 81.40 | 44.23 |
| | □ GI-AF-Opt-W | 88.77 | 79.06 | 77.54 | 40.66 |
| | □ GI-AF-Opt-W+ | **99.91** | **98.53** | **98.16** | **82.63** |
| | ■ I-EF | 90.70 | 79.65 | 80.61 | 42.48 |
| | ■ I-EF-Opt-X | **98.86** | **94.01** | **97.11** | **74.77** |
| | ■ GI-EF | 17.19 | 12.77 | 7.46 | 1.11 |
| | ■ GI-EF-Opt-X | 62.46 | 47.32 | 44.74 | 15.47 |
| | ■ GI-EF-Opt-W | 56.32 | 44.39 | 41.14 | 13.65 |
| | ■ GI-EF-Opt-W+ | **95.35** | **83.74** | **89.82** | **56.84** |
| EF | □ I-EF | 96.67 | 87.78 | 89.39 | 57.75 |
| | □ I-EF-Opt-X | **100.00** | **99.52** | **99.74** | **89.77** |
| | □ GI-EF | 33.68 | 27.45 | 20.26 | 4.76 |
| | □ GI-EF-Opt-X | 80.35 | 61.13 | 66.93 | 30.54 |
| | □ GI-EF-Opt-W | 75.79 | 61.96 | 62.72 | 29.91 |
| | □ GI-EF-Opt-W+ | **99.21** | **93.57** | **96.67** | **73.94** |
| | ■ I-AF | 91.75 | 75.09 | 81.93 | 40.18 |
| | ■ I-AF-Opt-X | **97.19** | **83.22** | **93.16** | **57.08** |
| | ■ GI-AF | 39.74 | 27.37 | 24.39 | 5.28 |
| | ■ GI-AF-Opt-X | 68.51 | 47.56 | 58.77 | 22.05 |
| | ■ GI-AF-Opt-W | 71.49 | 52.76 | 60.70 | 21.18 |
| | ■ GI-AF-Opt-W+ | **92.46** | **77.07** | **87.81** | **46.97** |

**Comparison with other methods**

We select the best-performing configurations of inversion morphs (both with and without optimization) and compare their effectiveness to previous methods from the literature. The results are presented in Table 3.3. Our primary focus is on the performance of our methods relative to MIPGAN and MorDIFF, which are considered state-of-the-art for deep morphing using GANs and Diffusion models, respectively, and against landmark-based methods.

First, focusing on the base inversion morphs, we observe that they perform exceptionally well against FR systems. In the white-box scenario, the ArcFace-based inversion, with or without fine-tuning, surpasses MIPGAN. In the black-box scenario, base inversion without

Table 3.3: Comparison of the effectiveness of the best morph generation methods. We distinguish white-box □ and black-box ■ attack scenarios.

| | | MinMax-MMPMR (%) | | FMMPMR (%) | |
|---|---|---|---|---|---|
| FRS | Attack | FRLL | FRGC | FRLL | FRGC |
| AF | □ I-AF | 97.54 | 89.88 | 91.49 | 54.90 |
| | □ I-AF-Opt-X | **100.00** | **99.80** | **99.82** | **94.25** |
| | □ GI-AF | 52.02 | 41.81 | 33.86 | 8.41 |
| | □ GI-AF-Opt-W+ | 99.91 | 98.53 | 98.16 | 82.63 |
| | □ DI-AF | 97.11 | 84.69 | 92.02 | 54.07 |
| | □ MIPGAN | - | 73.22 | - | 35.74 |
| | ■ I-EF | 90.70 | 79.65 | 80.61 | 42.48 |
| | ■ I-EF-Opt-X | **98.86** | **94.01** | **97.11** | **74.77** |
| | ■ GI-EF | 17.19 | 12.77 | 7.46 | 1.11 |
| | ■ GI-EF-Opt-W+ | 95.35 | 83.74 | 89.82 | 56.84 |
| | ■ SG2-W | 1.05 | 4.32 | 0.44 | 0.20 |
| | ■ SG2-W+ | 62.63 | 60.10 | 45.79 | 21.42 |
| | ■ MorDIFF | 90.09 | 74.81 | 78.95 | 41.13 |
| | ■ LB-Complete | **99.21** | **95.48** | **97.11** | **75.41** |
| | ■ LB-Combined | 93.95 | 84.97 | 88.51 | 53.51 |
| EF | □ I-EF | 96.67 | 87.78 | 89.39 | 57.75 |
| | □ I-EF-Opt-X | **100.00** | **99.52** | **99.74** | **89.77** |
| | □ GI-EF | 33.68 | 27.45 | 20.26 | 4.76 |
| | □ GI-EF-Opt-W+ | 99.21 | 93.57 | 96.67 | 73.94 |
| | ■ I-AF | 91.75 | 75.09 | 81.93 | 40.18 |
| | ■ I-AF-Opt-X | **97.19** | **83.22** | **93.16** | **57.08** |
| | ■ GI-AF | 39.74 | 27.37 | 24.39 | 5.28 |
| | ■ GI-AF-Opt-W+ | 92.46 | 77.07 | 87.81 | 46.97 |
| | ■ DI-AF | 92.11 | 72.19 | 83.68 | 38.60 |
| | ■ SG2-W | 3.07 | 10.19 | 1.32 | 0.79 |
| | ■ SG2-W+ | 72.28 | 67.63 | 53.95 | 31.30 |
| | ■ MIPGAN | - | 75.80 | - | 42.76 |
| | ■ MorDIFF | 89.74 | 76.76 | 77.37 | 44.51 |
| | ■ LB-Complete | **99.47** | **96.63** | **97.54** | **77.95** |
| | ■ LB-Combined | 95.96 | 87.58 | 90.70 | 59.58 |

fine-tuning is already competitive with MIPGAN and becomes significantly stronger after fine-tuning, making it a leading method for deep morphing effectiveness. Moreover, base inversion achieves performance levels between LB-Complete and LB-Combined morphs. To the best of our knowledge, this is the first deep-learning-based morphing method to achieve effectiveness comparable to that of landmark-based morphing.

Now, focusing on GAN-Inversion, we observe that without fine-tuning, the method is only moderately effective in both white-box and black-box scenarios, with MMPMR values lying between SG2-W and SG2-W+ morphing. However, applying the input optimization steps (specifically optimizing in the $\mathcal{W}+$ space) brings the performance to competitive levels. In the white-box scenario, it provides performance close to base-inversion but with the advantage of much higher realism, as will be demonstrated in Section 3.2.2, notably surpassing MIPGAN and MorDIFF. In the black-box scenario, it achieves slightly better performance than MIPGAN and MorDIFF but remains slightly below the landmark-based methods.

Finally, the diffusion-inversion method also performs competitively, notably more effective than base-inversion and GAN-inversion without the input optimization step in the black-box scenario, and only marginally weaker than the same method with input optimization. We can expect that performing actual gradient-based input optimization for the diffusion-inversion method would further enhance the morph effectiveness; however, as mentioned earlier, this process is non-trivial due to computational requirements.

Examining the previously proposed attacks, we observe that landmark-based attacks generally remain the most effective. However, the disparity between these and deep morphing techniques is gradually diminishing. While SG2-W morphs exhibit limited effectiveness, their variant utilizing the $\mathcal{W}+$ space demonstrates significantly improved performance. MIPGAN, which incorporates further latent refinement in $\mathcal{W}+$, achieves even better results. Finally, MorDIFF showcases very high effectiveness, making is the closest non-inversion-based method to landmark-based morphs in terms of cause vulnerability.

Comparing the visual aspects of the morphs (discussed in more depth in Section 3.2.2), we notice that the attack effectiveness seems somewhat uncorrelated with the visual realism of the morphs. Given the nature of inversion morphing (using an FR system at generation time), one might wonder whether it is more akin to an adversarial attack. In other words, the high effectiveness of the morph might not be linked to actual facial semantic attributes in the generated image but rather to humanly imperceptible noise patterns introduced in the image that somehow manage to fool the FR system. This phenomenon might explain the mismatch between the effectiveness of these morphs on humans versus systems.

#### Print-scan analysis

The fact that the attack generalizes well in black-box scenarios is a first hint against this adversarial hypothesis, showing the morphs actually contain some meaningful high-level

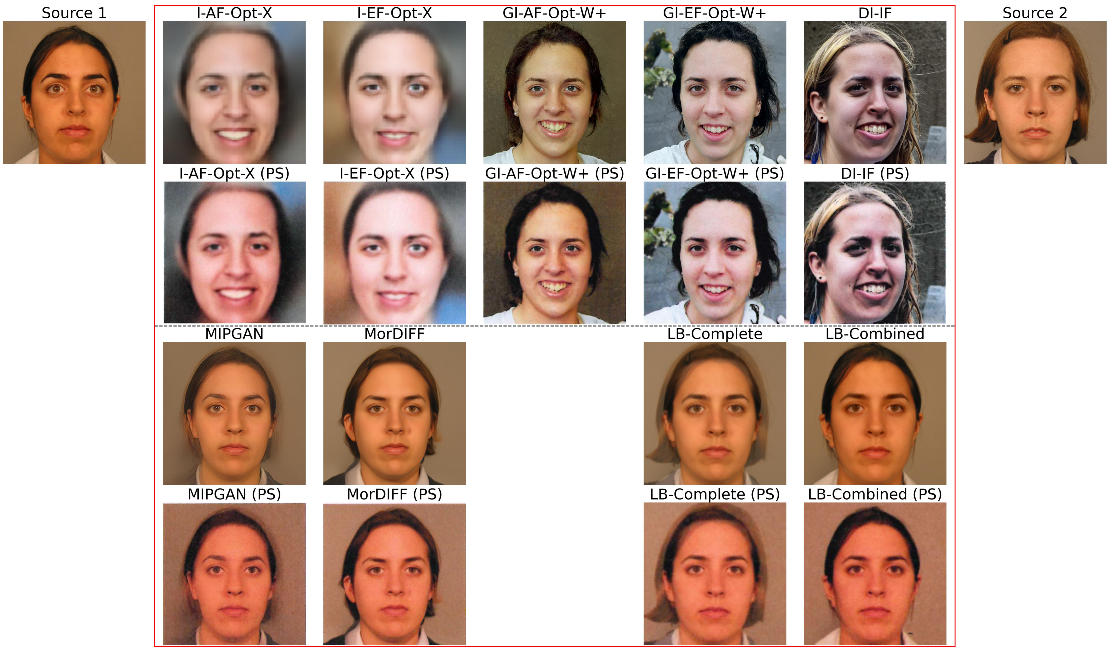


Figure 3.4: Examples of the selected subset of morphing attacks before and after print-scan (PS).

Table 3.4: Vulnerability analysis on a subset of attacks using the FRGC source dataset, in a print-scan setting. We distinguish between white-box □ and black-box ■ attack scenarios.

| FRS | Attack | MinMax-MMPMR (%) | FMMPMR (%) |
|---|---|---|---|
| AF | □ MIPGAN | 62.16 | 26.10 |
| | □ I-AF-Opt-X | **99.09** | **87.94** |
| | □ GI-AF-Opt-W+ | 97.26 | 78.62 |
| | □ DI-AF | 80.56 | 48.35 |
| | ■ MorDIFF | 66.56 | 31.85 |
| | ■ I-EF-Opt-X | **91.51** | **67.83** |
| | ■ GI-EF-Opt-W+ | 81.63 | 53.23 |
| | ■ LB-Complete | 88.22 | 58.07 |
| | ■ LB-Combined | 72.03 | 36.89 |
| EF | □ I-EF-Opt-X | **97.82** | **84.17** |
| | □ GI-EF-Opt-W+ | 91.79 | 69.93 |
| | ■ MorDIFF | 69.97 | 35.86 |
| | ■ MIPGAN | 69.46 | 33.32 |
| | ■ I-AF-Opt-X | 78.70 | 51.65 |
| | ■ GI-AF-Opt-W+ | 74.22 | 43.47 |
| | ■ DI-AF | 67.27 | 34.11 |
| | ■ LB-Complete | **91.19** | **64.02** |
| | ■ LB-Combined | 77.91 | 44.59 |

identity information able to fool a variety of FR networks. However, we can further test this hypothesis by conducting a print-scan vulnerability analysis. In this context, the morphs are printed and redigitalized before being enrolled in the system. This study has two main interests: first, it is a better simulation of a real-world scenario, in which a submitted passport picture can typically be subject to such a print-scan process before enrollment. Secondly, the print-scan process has the potential to degrade certain features of the image, causing a decrease in its vulnerability. With our inversion morphs in particular, if their effectiveness was due to some fine adversarial signal, a print-scan process could likely degrade such subtle patterns: hence, it is important to check how the attack performance varies in this setting, to evaluate how robust it is to minor degradations. We consider all available print-scan morphs presented in Chapter 2 (Table 2.2), and apply the same print-scan process to the best-performing inversion morphs. Figure 3.4 presents examples of the printed morphs. The probes for the vulnerability study are kept digitalized, to simulate a live capture and comparison at an ABC gate. The results of the vulnerability analysis are presented in Table 3.4.

We observe that while degraded, the performance of our morphing attacks is still preserved in this print-scan setting. The I-EF-Opt-X system performance degradation is even low enough that the method actually becomes stronger than landmark-based morphing in this setting. It is non-trivial to understand how this phenomenon can occur, but it at least suggests that the relevant face patterns in the morph are robust to print-scan degradation. In parallel, the GAN-Inversion morphs with fine-tuning remain more effective than MIPGAN and MorDIFF ones in this setting, and the GI-EF-Opt-W+ configuration in particular actually reaches a better performance than LB-Combined morphs in the black-box scenario.

### 3.2.2 Perceptual analysis

We now aim to discuss the perceptual aspects of the generated morphs, focusing on three key points: whether the morphs can be perceived as good lookalikes to both source identities, their potential for use in real-world attack scenarios, and their overall realism.

Regarding the morphs' ability to resemble both source identities, inversion morphs exhibit distinct behavior compared to other methods. Traditional methods typically ensure global image similarity to the source identities, providing intermediate facial features, face shape, facial expression, and hair. In contrast, inversion morphing primarily blends the central facial features of both sources, while peripheral elements like face shape, hair, and certain covariates such as expression or pose can differ significantly. For instance, in Fig. 3.3a, GAN-inversion morphs display a rounder face shape and a smiling expression, unlike the source identities. This discrepancy arises from the morph generation process, where the inverted face recognition system focuses on a tight crop of the face, rendering boundary features like cheek shape and hair irrelevant. Additionally, an ideal face recognition system's independence from facial expressions means there is no inherent reason to invert an embedding into a neutral expression.

Considering the practical application of inversion morphs in real-world attack scenarios, two main limitations emerge. Firstly, the non-uniform background, especially in GAN-inversion morphs, may not meet passport photo standards in many countries. This issue is minor, as background post-processing is relatively straightforward. Secondly, the non-neutral expression may also conflict with passport photo standards. Addressing this requires further work, such as constraining the template inversion network to produce neutral expressions. For GAN-Inversion, another approach could involve editing the $\mathcal{W}$ representation of the latent morph to neutralize the expression, as proposed in [35]. Additionally, fine-tuning the GAN generator on the source dataset, a process used notably in MIPGAN, could help GAN-Inversion morphs better match visual aspects from the source dataset. However, this fine-tuning might lead to catastrophic forgetting of the original weights, potentially decreasing the generated image's realism. This is illustrated in the case of MIPGAN, where many images exhibit blurry artifacts around the face, potentially due to degradation of the generator's ability during fine-tuning.

In terms of realism, GAN-inversion morphs are sharp and realistic, comparable to other StyleGAN-based approaches. They may even surpass MIPGAN morphs, which sometimes display blurry contours around the face, and LB-Complete morphs, which contain ghosting artifacts. However, base-inversion morphs are of lower quality overall; while the center of the face is sharp, the contours are excessively blurry. Post-processing operations, such as blending the face morph onto one of the source images, are likely necessary for these morphs to convincingly fool a human evaluator.

Table 3.5: Fréchet Inception Distance (FID). A lower value indicates an estimated stronger perceptual realism, i.e., lower is better.

| Attack | FRLL | FRGC |
|---|---|---|
| I-AF | 270.14 | 264.60 |
| I-EF | 269.87 | 259.52 |
| I-AF-Opt-X | 271.19 | 265.14 |
| I-EF-Opt-X | 268.28 | 265.24 |
| GI-AF | 96.33 | 64.05 |
| GI-EF | 92.02 | 76.73 |
| GI-AF-Opt-W+ | 87.99 | 62.99 |
| GI-EF-Opt-W+ | 82.56 | 70.79 |
| DI-AF | 124.57 | 116.56 |
| SG2-W | 25.99 | 21.31 |
| SG2-W+ | **23.75** | **17.76** |
| MIPGAN | - | 35.52 |
| MorDIFF | 58.79 | 81.65 |
| LB-Complete | 42.86 | 30.48 |
| LB-Combined | **27.87** | **28.36** |

An exact quantitative evaluation of the morphs' realism would require to run a perceptual study on human subjects, however we can get a first insight on the topic using the Fréchet Inception Distance metric (FID) [132] which maps human judgement relatively well and has been used in the literature for realism evaluation (e.g., [18], [107]). It estimates the perceptual distance between two sets of images by extracting their feature representations using a pretrained Inception network, fitting a Gaussian to both distributions, then computing the Fréchet distance between those. To compute the FID score, we use for each dataset the full set of real data (all source images used for morphing plus all the probes used in the vulnerability study, see Section 3.2.1) as the bona fide sets, and the generated morphs with each individual attack as the fake set. The results are presented in Table 3.5.

The FID results confirm the relatively clear-cut observation that base inversion morphs are not of the greatest realism (high FID), but we actually also observe that the GAN-Inversion and diffusion-inversion morphs are not performing as well as StyleGAN or landmark-based morphs. One possible explanation is that while often used for that purpose, the FID technically does not measure exactly "realism", but rather differences between the bona fide and fake set: as we commented before, inversion morphs being less constrained, they do not preserve some image elements like background, expression or pose, which sets the morph distribution further away from the source distribution, in contrary to other morphing methods which constrain both the identities and other image factors. However, those aspects are actually crucial in a real world scenario, as it is necessary for the submitted morphs to presents the appropriate constraints expected from passport pictures. In that regard, we see that the SG2-W+ morphs are the most appropriate in terms of general visuals, followed by the SG2-W morphs and then the landmark-based morphs. In contrast, MIPGAN, MorDIFF and inversion morphs are not as realistic.

We must emphasize that our perceptual analysis is incomplete without conducting an actual perceptual study with human evaluators. Implementing such a study is complex and was considered beyond the scope of this thesis. Nevertheless, we outline the structure such a study could follow.

First, subjects should be exposed to a series of passport pictures containing both bona fide samples and morphing attacks, and asked to determine whether the samples are legitimate or fake. Second, a simulation of a passport application process should be implemented as follows: subjects would act as governmental office workers responsible for receiving and processing submitted passport pictures. Subjects should be presented with two images: the submitted reference picture and an alternative probe image (or ideally, a video) of the person applying for the passport. Subjects should then assess whether the applicant represented by the probe matches the submitted passport picture. This study should include both bona fide scenarios (where the submitted picture is legitimate and matches the applicant) and attack scenarios (where the submitted picture is a morph, and the probe represents one of the contributing identities). This study also simulates another setting in which human evaluators might be involved in the security process, which is during a verification attempt at an ABC gate, where security personnel might double-check the automated analysis from the face recognition system.

These studies could compare the abilities of different categories of evaluators, such as police officers, researchers familiar with face manipulations, or super-recognizers with enhanced face recognition abilities [133] (although research has not shown their higher capability to detect deepfakes [134]). The studies could also evaluate whether the detection of attacks by human evaluators can be improved through specific training.

We note that the original article introducing the MIPGAN algorithm [37] proposes such a perceptual study and considers attacks partially overlapping with those in our work, though it misses more recent diffusion morphs and inversion-based morphs. They focus solely on realism assessment, observing that over 30% of morphs are typically undetected by evaluators, a phenomenon consistent across morphing algorithms. However, they do not study the perceived identity match between the morphing attack and the contributing subjects. Moreover, their study is restricted to the digital domain, and a study focusing on printed morphs could better simulate a real-world scenario, while also making the detection of morphs more challenging by hiding salient artifacts during the print-scanning process.

Alternatively, [135] proposes a study of identity-match between morphing attacks and contributing subjects for human perception, and observes that human evaluators are regularly fooled by morphing attacks in that regard as well. However, this study predates the emergence of deep morphs and focuses solely on landmark-based ones, and an update including newer generative algorithms could be valuable.

### 3.2.3 Detectability

We also evaluate the detectability of the generated morphs using publicly available morphing attack detectors. Specifically, we consider MixFaceNet-SMDD [114], a CNN detector trained as a binary classifier (supervised learning), and SPL-MAD [122], a one-class detector that models the distribution of bona fide images and detects morphing attacks as out-of-distribution samples (anomaly detection). The MixFaceNet-SMDD model is trained on the SMDD dataset, which contains landmark-based morphs created from a set of synthetic source identities generated using a StyleGAN model. The SPL-MAD model is trained on Casia-WebFace, originally a face recognition dataset. Table 3.6 presents the detection performance of these two models on each type of considered attack (using the same set of bona fide samples). We specifically report the Detection Equal Error Rate (D-EER), which corresponds to the error rate at an operating threshold where the proportion of attacks classified as bona fide and the proportion of bona fide classified as attacks are equal.

We first observe that base-inversion morphs are very effectively detected, which is likely related to their low resolution and subpar realism. However, both GAN-Inversion morphs and Diffusion-inversion morphs pose a high challenge for both detectors. This is notably despite the detectors performing decently on other GAN-based morphs (e.g., MIPGAN), and other

Table 3.6: Detection Equal Error Rate (D-EER), in %, using publicly available morphing attack detectors. For each attack, the same set of bona fide is used and correspond to frontal samples extracted from the source dataset which is also used to create the morphs.

| Attack | Model<br>Dataset | MixFaceNet-SMDD<br>[114] | SPL-MAD[122] |
|---|---|---|---|
| I-AF-Opt-X | FRLL | 0.09 | 0.00 |
| | FRGC | 0.92 | 0.08 |
| I-EF-Opt-X | FRLL | 0.49 | 0.00 |
| | FRGC | 0.86 | 0.08 |
| GI-AF-Opt-W+ | FRLL | 49.57 | 70.63 |
| | FRGC | 53.57 | 67.76 |
| GI-EF-Opt-W+ | FRLL | 55.46 | 75.53 |
| | FRGC | 67.88 | 78.18 |
| DI-IF | FRLL | 70.14 | 74.40 |
| | FRGC | 84.72 | 70.31 |
| SG2-W | FRLL | 11.90 | 13.06 |
| | FRGC | 8.88 | 3.73 |
| SG2-W+ | FRLL | 22.77 | 25.00 |
| | FRGC | 15.07 | 6.35 |
| MIPGAN | FRGC | 24.54 | 15.48 |
| MorDIFF | FRLL | 2.18 | 0.53 |
| | FRGC | 8.88 | 1.19 |
| LB-Complete | FRLL | 0.45 | 0.00 |
| | FRGC | 2.37 | 0.16 |
| LB-Combined | FRLL | 1.65 | 0.00 |
| | FRGC | 10.81 | 0.51 |

diffusion-based morphs (e.g., MorDIFF). We also observe that despite their high effectiveness, landmark-based morphs are typically well detected by available detectors. This is partially due to them appearing first historically, and thus being systematically considered in the development of morphing detectors : the SMDD dataset used to train the MixFaceNet detector, which is one of the largest publicly available dataset, is focused on landmark-based morphs only. However, it is interesting to observe that the SLP-MAD model is performing well on landmark-based morphs, despite its one-class design.

As several inversion-based morphs have been showcased to be simultaneously effective at fooling face recognition systems and not easily detected by existing detectors, it suggests that future improvements on the detectors might need to consider inversion morphs as an additional possible attack.

## 3.3 Discussion

We have introduced a novel deep morphing method that approximates the optimal face morph by leveraging template inversion techniques to reconstruct an image from the optimal morph embedding. Specifically, we have showcased three distinct template inversion systems: one based on a fully trained embedding-to-image synthesis network, another utilizing the latent space of a pretrained face GAN to enhance the realism of the morphs, and a third employing an identity-conditioned diffusion model. We have proposed two generation settings: one using the default output of the template inverter, and another where the output is optimized to better align with the optimal morph embedding. Our experiments have demonstrated that this optimization significantly enhances the morphs' ability to deceive face recognition systems.

Following our comprehensive evaluation of the effectiveness, realism, and detectability of all major families of morphing attacks, we can summarize our key findings as follows:

- **Morphing attack vulnerability**: Landmark-based morphing attacks remain highly effective against face recognition systems. However, advancements in deep morphing are closing the gap. Notably, GAN-based algorithms like MIPGAN and diffusion-based algorithms such as MorDIFF produce attacks with a high success rate, raising significant concerns. Our proposed inversion-based morphing methods further increase attack effectiveness compared to previous deep morphing approaches, even in black-box scenarios. Some inversion-based methods, such as base-inversion and GAN-inversion, achieve effectiveness comparable to landmark-based methods, outperforming LB-Combined attacks in several evaluation scenarios. Additionally, our results indicate that the effectiveness of these attacks is largely preserved after a print-scan process.
- **Morphing attack realism**: The primary limitation of inversion morphs is their lack of visual realism and control over facial covariates, making them unsuitable for use in official documents like passports. This suggests that a straightforward application of

template inversion methods to morphing attack generation may not pose an immediate real-world threat. However, as the field of template inversion evolves, and particularly might improve conditional control over output face semantics, it is crucial to monitor the potential of template inversion for morphing attack generation. Orthogonally to improvements in the inverter, post-processing techniques could also enhance the visual realism of inversion morphs, such as editing the background to be uniform or neutralizing facial expressions in GAN-Inversion morphs by appropriate $\mathcal{W}$ latent editing.

For previously proposed attacks, GAN-based methods (SG2-W+, SG2-W) exhibit the highest realism in terms of FID, closely followed by landmark-based approaches. In contrast, MIPGAN shows somewhat lower realism, which is also observed through visual inspection. However, although MorDIFF has a worse FID score, it qualitatively appears more realistic than MIPGAN attacks. This discrepancy may be due to the FID metric's unfairness towards diffusion images, as discussed in [136].

- **Morphing attack detectability**: Publicly available detection models perform relatively well on landmark-based morphs and base-inversion morphs, but are less accurate on MorDIFF morphs and struggle significantly with other attacks. Note that this evaluation does not account for the challenges introduced by print-scan post-processing, which will be analyzed in Chapter 5. This highlights the challenge posed by the diversity of existing attacks, as a robust detector must perform consistently well across various attack types.

This study highlights the existence of both cheapfakes (landmark-based morphs) and deepfakes (deep morphs) that can effectively deceive automated face recognition systems. Their ability to also deceive human operators is yet to be fully confirmed through the implementation of a perceptual study on human subjects, the structure of which we also outlined. While our work provided initial insights into the estimated realism of the generated morphs, it lacks confirmation from an actual attack implementation and an assessment of the identity-match between generated morphs and their contributing subjects for human perception. Implementing these studies is the last missing piece for demonstrating the real-world feasibility of modern morphing attacks.

While the lack of a perceptual study is a main limitation of our work, we note that due to the pace at which novel morph generation algorithms are emerging due to breakthroughs in generative AI, keeping track of them through continuously updated perceptual studies might be challenging, as such studies can be costly and time-consuming.

For instance, we demonstrated how template inversion models can be easily adapted to create highly effective morphing attacks, despite their current visual limitations. This highlights the rapid development and diversification of deep generative models, which introduce new challenges in biometric security. Instead of waiting for comprehensive risk assessments of novel attacks through perceptual studies, a more prudent approach is to assume that these attacks could be problematic, notably as they are already deceiving automated face recognition systems. Therefore, it is crucial to develop morphing attack detectors that are both general enough to handle novel, unseen attacks (typically deepfakes) and capable of addressing the ongoing threat posed by landmark-based morphs (cheapfakes). In the next chapter, we thus propose a first research contribution on the topic of detection, which focuses on GAN-based morphs in particular.

# 4 Handcrafted and supervised features for detection of GAN-based morphing attacks

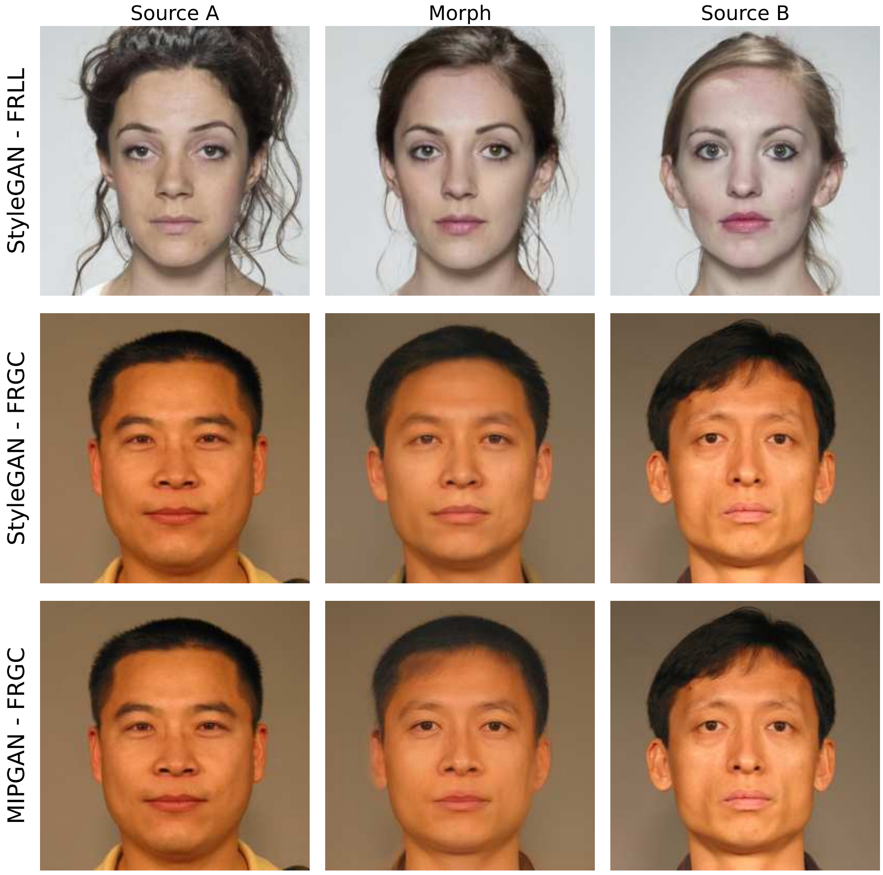


Figure 4.1: Example of generated morphs, using the SG2-W+ method (first and second row) and the MIPGAN method (third row)

This chapter presents a study conducted in 2022, focusing on the detection of morphing attacks generated by GANs. This research reflects the state of the field before the emergence of diffusion-based morphs and foundational feature-based detection methods, making its relevance somewhat limited in the current landscape. It addresses a narrower subset of the attacks introduced in Chapter 2 and does not tackle challenges posed by print-scan post-processing. Nevertheless, it was a crucial stepping stone that provided valuable insights, informing the experimental design of other contributions in this thesis, particularly Chapter 5. Presenting this earlier work highlights its contribution to subsequent advancements. It also illustrates the rapid evolving landscape in this area which in itself is a security challenge. For any authority, keeping a watching brief on the threat situation is very challenging.

The study was conducted during the period following the development of initial GAN-based morphing techniques, specifically SG2-W+ [105] and MIPGAN [37]. At that time, the detection of this novel category of morphing attacks had not yet received significant attention. Detection methodologies could draw inspiration from two primary sources: existing research on MAD, which focused on landmark-based attacks, and the field of deep synthetic image detection.

The core objectives of this study are as follows:

- Establish baseline performance for the detection of GAN-based morphing attacks.
- Consider methodologies from both landmark-based morph detection literature and deep synthetic image detection literature to verify their applicability for this task.
- Compare the performance of methods using handcrafted features to those using supervised features, and investigate the potential trade-off between detection accuracy and interpretability.

For the handcrafted features, we specifically consider:

- **Fourier features**: A simple summary of the images' frequency content through the azimuthal average of their Fourier transform. This representation was proposed in [74] for faceswap deepfake detection.
- **LBPs**: A texture descriptor effectively used previously for the detection of landmark-based morphing attacks in [137].

These representations are motivated not only by their previous usage in the cited works but also by qualitative observations of GAN-based morphs. We notice that these morphs tend to appear smoother than real images. For example, in the first row of figure 4.1, skin irregularities such as moles vanish, thin hair details in the eyebrows disappear, and the skin looks overall more “plastic”. We hypothesize that GAN-based morphs tend to contain reduced high-frequency content compared to real images (hence the use of Fourier features) and that texture-related features might also be successful (hence the use of LBPs). Additionally, Fourier features also relate to the notion of GAN-signature, whose initial conceptualization was done through observations of the frequency response of generated images [69].

For the supervised features, we consider two distinct approaches:

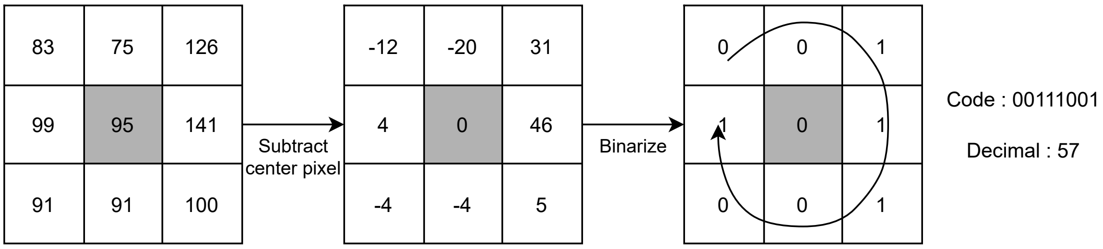


Figure 4.2: Illustration of Local Binary Pattern extraction in a single 3×3 pixel neighborhood. Each pixel value is compared to the value of the center pixel, then the result is binarized. The resulting bits are read in a fixed order, thus generating an 8-bit code describing the specific local texture. For such a shape of neighborhood, this results in 256 possible different codes.

- Training a CNN from scratch for morphing attack detection as a binary classification problem.
- Reusing a preexisting CNN model trained for generalized GAN-Image detection: RN50-PGGAN [69].

The study is restricted to three groups of morphing attacks selected from the complete list presented in Chapter 2. These are the SG2-W+ attack (both for the FRLL and FRGC datasets) and the MIPGAN attack (only for the FRGC dataset). These three attacks are abbreviated as FRLL-SG2, FRGC-SG2, and FRGC-MIPGAN in this chapter, and they are considered as three independent datasets.

We first present in more detail the considered representations. We then establish baseline performance in an intra-dataset setting (i.e., when the test set contains attacks seen during training) before focusing on cross-dataset performance (i.e., when the test set contains attacks not seen during training). Finally, we will discuss the interpretability of the proposed approaches, notably through the use of saliency maps.

## 4.1 Methodology

For all subsequent experiments, face images are cropped for face landmark alignment, following the same cropping used for images fed into the ArcFace model [129], a commonly used open-source FR system. We specifically use the ResNet100 version trained on MSCeleb-1M that is provided by InsightFace[1].

We decide to experiment with three types of features for MAD:

- **Fourier features**: As a very simple baseline, we convert the images to grayscale and analyze their frequency content by computing their Fourier transform, then integrating

[1] https://github.com/deepinsight/insightface

it over radial frequency bands. This results in a 1D, angle-independent signal. Such features were originally proposed in [74] for Face Swap detection.

- **Local Binary Patterns (LBP) histogram features**: Local Binary Patterns (LBP) work as a local texture extractor. They have been successfully used in the past for face anti-spoofing [138] and MAD [137]. The principle is to process pixel neighborhoods to summarize them as a single binary code. A limited number of different codes are possible, each of which describes a specific local texture. These LBP codes are extracted over the whole image, then a histogram is computed describing the occurrence of each of the possible codes, thus describing the overall texture content of the image. These histograms are then used as the extracted features for MAD. We perform experiments using the proposed LBP implementation in the Bob toolkit[2]. The images are converted to grayscale beforehand. We test several configuration variants, differing by the size of the neighborhood ($(n, r) = (4, 1)$ or $(8, 2)$ with $n$ the number of neighbors and $r$ the radius of the neighborhood), the shape of the neighborhood (square or circular), and the optional regrouping of similar codes in the same histogram bins (*RIU2* regrouping where local textures equivalent up to a rotation are grouped in the same bin). This leads to a total of eight different LBP configurations.

- **Neural-net features (CNN)**: Finally, we also evaluate the efficiency of neural-net based features (supervised features) by training various CNN architectures for binary classification (bona fide *vs* morph). In general deepfake detection, this approach has shown the most success compared to using more hand-engineered features, therefore it is important to evaluate it as well. However, it comes at the cost of interpretability: there is no straightforward way to understand what components of the input image cause the model to produce one classification or the other, which is impractical if focusing on a specific case in a forensic context. We explore five different architectures: two of them are the XCeption and EfficientB4 models that have been successful for video deepfake detection in [139]. In this previous work, those models were trained on frames extracted from videos, which represented a much larger amount of data (at least 10x more images) than our morphs datasets, which are composed of individual images. For this reason, and aiming to limit the risk of overfitting the training data, we also experiment with two smaller scale architectures, specifically MobileNet and MobileNetV2 [140]. The models are loaded with pretrained ImageNet weights, then trained as binary classifiers for MAD using a final fully connected layer with a single output neuron, jointly with the cross-entropy loss and the Adam optimizer with a learning rate of 2E-4, following the setup from [139]. We independently train on each of the morphs datasets, use the two remaining datasets as validation sets, and select the best model based on the validation loss.

Finally, we also experiment with RN50-PGGAN, the pretrained GAN-image detection model proposed in [69]. We specifically use the Blur+JPEG(0.1) version that is available for download

[2] https://gitlab.idiap.ch/bob/bob.ip.base/

on GitHub[3]. This model has shown strong generalization capabilities and is likely to perform well in our context, given that our morphs are generated by GANs. However, it remains to be seen whether the detector's generality holds, as our morph generation methodology differs significantly from simple GAN sampling. Specifically, our approach involves computing morph latents through a projection-interpolation process or an optimization-based exploration of the $\mathcal{W}+$ space. This means the input latents provided to the StyleGAN synthesis network could be out-of-distribution compared to the $\mathbf{w}$ latents typically seen during normal sampling. It is uncertain whether these differences in the synthesis network's input will significantly affect the detectable traces in the generated images.

## 4.2 Results

### 4.2.1 Detection

We aim to evaluate the effectiveness of the MAD systems in both intra-dataset and cross-dataset scenarios. We approach the task as a binary classification problem, where images belonging to the 'morph' class are taken from FRLL-SG2, FRGC-SG2, and FRGC-MIPGAN, and images belonging to the 'bona fide' class are taken from the corresponding source datasets.

In the intra-dataset case, we create separate train and test sets by splitting the identities of the source datasets into two groups: 3/4 for training and 1/4 for test. This split is straightforward for bona fide examples; for morph examples, we include in each set only morphs for which both source identities are also part of the set. Due to the small scale of our datasets, further reduced by the train/dev split, evaluation of CNN approaches is impractical in the intra-dataset case and is thus reserved for the cross-dataset case.

In the cross-dataset case, which evaluates generalization capability, we systematically train on each full dataset and evaluate on the other two. We note that FRLL-SG2 and FRGC-SG2 use the exact same GAN (architecture and weights) and the same morph generation algorithm but use different source datasets. FRGC-SG2 and FRGC-MIPGAN use the same source dataset and the same GAN architecture but different sets of weights (for FRGC-MIPGAN, the GAN is fine-tuned on the FRGC dataset) and a different morph generation algorithm. We also report results of RN50-PGGAN, which we reuse out-of-the-box.

We focus on summary metrics rather than specific operating point performances, mainly the area under the AUC ($\in [0,1]$, higher is better), which is popular in deepfake detection literature (e.g., [139]), and the D-EER % ($\in [0,100]$, lower is better). We report only the best-performing configuration for each type of system. Intra-dataset and cross-dataset results are presented in Tables 4.1 and 4.2, respectively.

We observe that despite the simplicity of the classifiers, both types of features enable significantly better than random detection capability, with LBP-LDA systems consistently out-

[3] https://github.com/peterwang512/CNNDetection

| Dataset | Extractor | Specs | AUC | D-EER (%) |
|---|---|---|---|---|
| FRLL- | Fourier | | 0.92 | 17.31 |
| SG2 | LBP | (8,2) ○ | **0.99** | **5.77** |
| FRGC- | Fourier | | 0.92 | 16.21 |
| SG2 | LBP | (8,2) ○ | **1.00** | **2.86** |
| FRGC- | Fourier | | 1.00 | 1.43 |
| MIPGAN | LBP | (8,1) □ | **1.00** | **0.00** |

Table 4.1: MAD performance on the test sets of our three morphing attack datasets. For LBP features, we only report the LBP configuration with the maximum AUC. ○ indicates a circular LBP shape while □ indicates a square LBP shape.

performing Fourier-based ones. Notably, MIPGAN morphs appear much easier to detect, with the Fourier-LDA system performing significantly better on these morphs compared to SG2-morphs. The (8,1) □ LBP-LDA system even achieves perfect separation. This contrasts with the fact that MIPGAN morphs usually cause higher vulnerability in common FR systems than SG2 morphs. Visually, MIPGAN morphs often contain noticeable artifacts, particularly around the hair, which sometimes appears blurry. This blurriness might result in a more distinctive spectral distribution, thus facilitating detection.

Moving to the cross-dataset setting, we observe a significant decrease in detection performance of the LDA classifiers. LBP-LDA classifiers still showcase better performance than Fourier-LDA ones in almost every case (often significantly). However, the cross-dataset generalization error they exhibit is often more pronounced. For example, alternating between FRLL-SG2 and FRGC-SG2 causes very limited performance decrease with Fourier-LDA, whereas the D-EER is multiplied by around 3x-6x with LBP-LDA. A similar phenomenon is observed when alternating between FRGC-SG2 and FRGC-MIPGAN, except for the Fourier-LDA generalization error being multiplied by approximately 20 when going from FRGC-MIPGAN to FRGC-SG2. This is likely an additional indication of the surprisingly high effectiveness of Fourier features on FRGC-MIPGAN observed in the intra-dataset experiment. Finally, alternating between FRLL-SG2 and FRGC-MIPGAN, Fourier-LDA classifiers perform very poorly, while LBP-LDA showcases similar generalization error as in the other cases. Overall, it seems the LBP-LDA features are more consistently robust than Fourier-LDA ones, which are relatively robust when keeping the same morphing algorithm, but not otherwise (and perform worse than LBP features overall).

We note, however, that no LBP configuration emerges as the systematic best choice. Overall, rotation-invariant regrouping and 8-neighbors shapes generally correlate with good performance, but we do observe that no bins regrouping is the best setup for FRGC-SG2 / FRGC-MIPGAN generalization. In this latter case, we hypothesize that this might be an effect of the shared bona fide set between both datasets: not using bins regrouping decreases regularization and could enable the classifier to learn more patterns specific to this bona fide set, thus helping with generalization.

| Train on | Test on | Extractor | Specs | AUC | D-EER (%) |
|---|---|---|---|---|---|
| FRLL-SG2 | FRGC-SG2 | Fourier | | 0.88 | 20.46 |
| | | LBP | RIU2 (8,1) □ | 0.93 | 15.16 |
| | | CNN | Xception | **0.99** | **5.72** |
| | FRGC-MIPGAN | Fourier | | 0.45 | 53.25 |
| | | LBP | RIU2 (8,2) □ | 0.96 | 11.39 |
| | | CNN | Xception | **0.99** | **5.72** |
| FRGC-SG2 | FRLL-SG2 | Fourier | | 0.92 | 16.04 |
| | | LBP | RIU2 (8,1) □ | 0.91 | 19.97 |
| | | CNN | EfficientNet | **1.00** | **1.20** |
| | FRGC-MIPGAN | Fourier | | 0.85 | 23.24 |
| | | LBP | (8,2) ○ | 0.99 | 3.91 |
| | | CNN | EfficientNet | **1.00** | **0.00** |
| FRGC-MIPGAN | FRLL-SG2 | Fourier | | 0.62 | 41.8 |
| | | LBP | RIU2 (8,2) ○ | **0.92** | **16.98** |
| | | CNN | MobileNet | 0.89 | 19.2 |
| | FRGC-SG2 | Fourier | | 0.79 | 28.51 |
| | | LBP | (8,2) ○ | 0.97 | 8.76 |
| | | CNN | MobileNet | **0.99** | **5.41** |
| PGGAN | FRLL-SG2 | CNN | RN50-PGGAN | **1.00** | **0.00** |
| | FRGC-SG2 | CNN | RN50-PGGAN | **1.00** | **0.56** |
| | FRGC-MIPGAN | CNN | RN50-PGGAN | **1.00** | **0.00** |

Table 4.2: Cross-dataset MAD performance on our three morphing attack datasets. For LBP features and CNNs, we pick the best performing configuration on the test set and report performance on the evaluation set. □ indicates a square LBP shape, while ○ indicates a circular shape.

The CNN classifiers showcase drastically better cross-dataset performance, significantly outperforming LBP-LDA in all cases except when training on FRGC-MIPGAN and testing on FRLL-SG2. In this specific case, it seems more about the CNN underperforming rather than LBP-LDA performing very well. Interestingly, this is not true in the reverse case (training on FRLL-SG2 and testing on FRGC-MIPGAN), suggesting that features learned by the CNN on SG2-based datasets might generalize better to the MIPGAN-based dataset, rather than the reverse.

Last but not least, the RN50-PGGAN system outperforms every other classifier by a large margin, despite being used out-of-the-box with no fine-tuning. This demonstrates the generalization power of this GAN-image detector is preserved despite replacing a simple image sampling procedure with a much more complex $\mathcal{W}+$ latent space exploration process. Again, the preservation of this generalization capacity was not guaranteed a priori.

It thus appears that reusing a GAN-image detector is currently the most effective approach for MAD. However, we remain interested in exploring ways to enhance the performance of

| Calib. on | Test on | Specs | Fused D-EER (%) | Sys. 0 D-EER (%) | Sys. 1 D-EER (%) |
|---|---|---|---:|---:|---:|
| FRLL-SG2 | FRGC-SG2 | All LDA | 10.00 | 15.16 | 20.46 |
| | | LBP-LDA+RN50-PGGAN | **0.15** | 17.94 | 0.56 |
| | FRGC-MIPGAN | All LDA | 10.98 | 11.39 | 13.09 |
| | | LBP-LDA+RN50-PGGAN | **0.00** | 39.23 | 0.00 |
| FRGC-SG2 | FRLL-SG2 | All LDA | 12.03 | 19.97 | 16.04 |
| | | LBP-LDA+RN50-PGGAN | **0.00** | 20.99 | 0.00 |
| | FRGC-MIPGAN | All LDA | 3.65 | 3.91 | 7.90 |
| | | LBP-LDA+RN50-PGGAN | **0.00** | 26.03 | 0.00 |
| FRGC-MIPGAN | FRLL-SG2 | All LDA | 15.02 | 16.98 | 20.99 |
| | | LBP-LDA+RN50-PGGAN | 0.76 | 19.03 | **0.00** |
| | FRGC-SG2 | All LDA | 8.09 | 8.76 | 21.44 |
| | | LBP-LDA+RN50-PGGAN | **0.15** | 10.98 | 0.56 |

Table 4.3: Cross dataset MAD performance using linear logistic regression score-level fusion of 2 systems. Among all considered combinations, we pick the best performing one on the test set and report performance on the evaluation set. For LDA systems, the same data is used for training & score calibration. In the LDA+RN50-PGGAN case, system 1 always refers to the RN50-PGGAN system.

classifiers based on handcrafted features, which offer advantages such as lower computational cost and higher interpretability. Given access to many classifiers potentially focusing on complementary aspects of the data, we further experiment with score-level fusion of these classifiers. This is achieved by pairing independent systems and training a linear regression classifier on one of the datasets (the calibration set), using the two sets of scores as features. Performance on the remaining datasets is then evaluated. We consider two scenarios: fusion between two LDA systems (*can we improve performance while still only considering simple systems?*) and fusion between an LBP-LDA system and the RN50-PGGAN system (*can an LBP-LDA system still help to enhance the performance of RN50-PGGAN?*). Results are reported in Table 4.3.

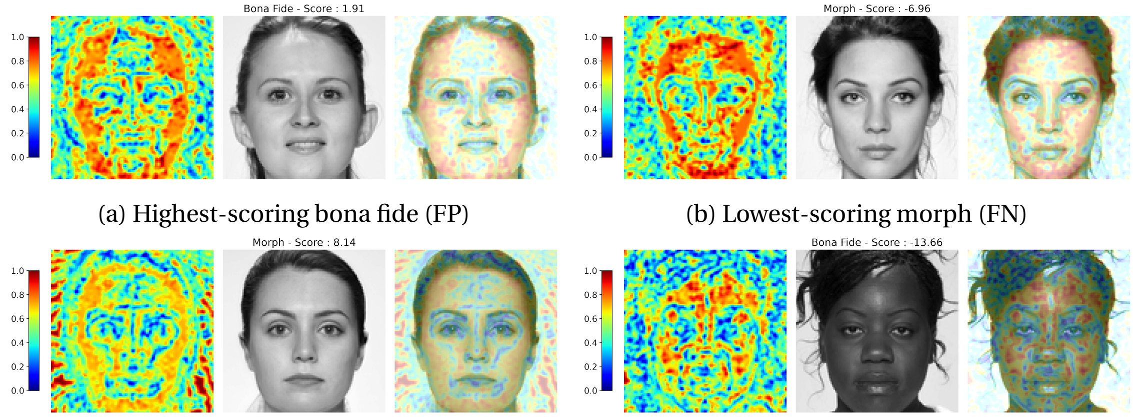
We observe that this fusion process almost systematically leads to performance improvements, sometimes significant for the LDA systems. More interestingly, it also improves performance in some cases (e.g., testing on FRGC-SG2) over using the standalone RN50-PGGAN. This improvement is small (D-EER decrease of around 0.4%) but non-negligible, corresponding to a decrease from 11 to 3 erroneous classifications over 1940 queries. This suggests that the LDA systems still provide complementarity to the RN50-PGGAN, despite the significantly better single-system performance of the latter. On more challenging datasets especially, simple handcrafted features could thus still be useful to improve performance beyond that of a single CNN.

### 4.2.2 Interpretability

The performance analysis in the previous section highlights the effectiveness of different systems for automated detection. In a forensic context, this applies to an *investigative* phase, where the goal is to flag suspicious content. However, in an *evaluative* phase, where the aim is to analyze an individual case and build an argument on the genuineness of the data, the binary output provided by the classifiers is not sufficient. We propose to qualitatively evaluate the relevance of saliency maps in this analysis, envisioning them as a potential tool to support a human expert by highlighting relevant areas in the suspect image. We focus the analysis on extreme examples in the test set, i.e., the lowest- and highest-scoring bona fides and morphs. Figures 4.3 and 4.4 present some of these maps for an LBP-LDA classifier and the RN50-PGGAN model, respectively. The CNN saliency maps are obtained using the GradCAM method [92], while the LBP-LDA saliency maps are obtained using the approach described below.

Consider a pretrained LDA classifier with coefficients $\mathbf{w}$, an input image $\mathbf{x}$, an LBP operator $LBP(\cdot)$, and the histogram of LBP codes $\mathbf{h}_{LBP(\mathbf{x})}$. The elementwise product $\mathbf{w} \odot \mathbf{h}$ specifies the individual contribution of each possible code to the final classification, with negative values indicating a push towards classification as a real image, and positive values indicating the opposite. We can rank the codes in increasing order of contribution, creating a mapping from each possible code to its rank. Applying this mapping to the LBP image $LBP(\mathbf{x})$ and rescaling it to values in [0, 1], we obtain a heatmap describing the contribution of each LBP neighborhood to the final classification, with higher values for neighborhood types that push towards classification as a morph. Optionally, we can apply a light Gaussian smoothing to this heatmap to mitigate the pixelated aspect of the visualization. Superimposing this heatmap over the original image allows us to visualize which regions are particularly salient for detection.

One main benefit of LBP-LDA saliency maps over Grad-CAM visualization is their more human-friendly structure, as they tend to highlight regions that make semantic sense. This allows for certain explanations, such as '*The eyes, mouth, and ears regions do not typically matter for detection*', or '*In example 4.3c, the most salient signal is actually coming from the background.*' This level of structure is inherent to the classifier's design, which naturally defines large groups of pixels sharing a role in the analysis (all pixels with an LBP code lying in the same bin). In comparison, Grad-CAM visualization is harder to interpret; while there might be an argument that the hair area seems important in cases 4.4a and 4.4c, the most salient areas tend to be tiny blobs not confined to specific regions. We also remind that the article introducing the RN50-PGGAN model claims it relies mostly on GAN-signature information, understood as differences in the spectral properties of GAN-generated images compared to real images. This suggests the salient features might be local in the spectral domain only, and thus highly non-local in the image domain, making saliency maps inappropriate for analysis. In other words, while the learned approach to detection is robust, it makes it difficult to build compelling explanations, especially for non-technical experts such as a jury.



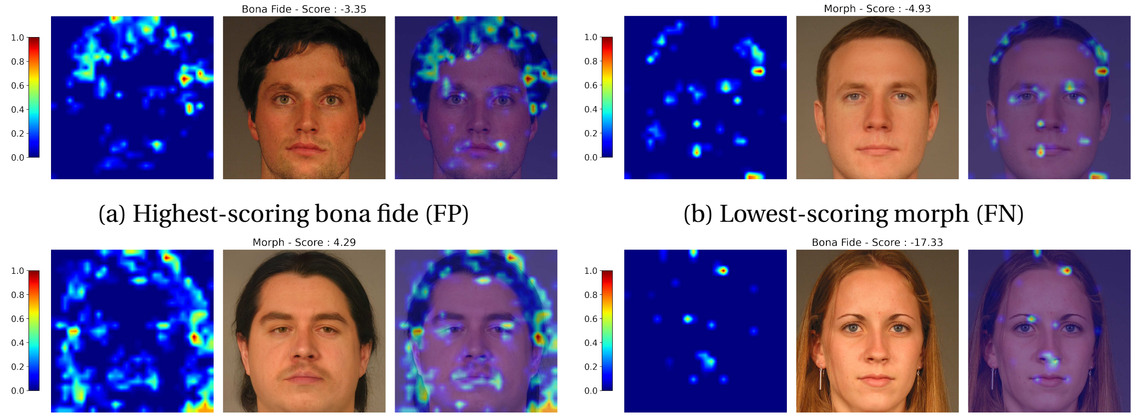
(a) Highest-scoring bona fide (FP)

(b) Lowest-scoring morph (FN)

(c) Highest-scoring morph (TP)

(d) Lowest-scoring bona fide (TN)

Figure 4.3: Examples of best and worst classified FRLL-SG2 images using the best LBP-LDA classifier trained on FRGC-MIPGAN (RIU2 (8,2) ○). The left column is the computed heatmap using a smoothing of 1.0, higher values corresponds to pixels pushing towards a classification as *morph*, which is the positive class. The center column is the classified image, and the right column overlays it with the heatmap. The D-EER threshold is at $\tau \simeq -2.49$.



(a) Highest-scoring bona fide (FP)

(b) Lowest-scoring morph (FN)

(c) Highest-scoring morph (TP)

(d) Lowest-scoring bona fide (TN)

Figure 4.4: Examples of best and worst classified FRGC-SG2 images using the RN50-PGGAN model. The left column is the computed heatmap using Grad-CAM visualization, higher values corresponds to pixels pushing towards a classification as *morph*, which is the positive class. The center column is the classified image, and the right column overlays it with the heatmap. The D-EER threshold is at $\tau \simeq -3.97$.

However, LBP-LDA maps provide limited explanatory power. For example, examples 4.3a and 4.3b, corresponding to the most erroneous classifications, do not showcase elements in the saliency maps that would explain the misclassification. In example 4.3c, the highlighted background areas do not seem to contain human-perceptible signals. Thus, the imperfect alignment between algorithmic and human vision limits the explanation power.

## 4.3 Discussion

We identify the following main takeaways:

- **Data representations**: Approaches based on supervised features are typically much more effective than those based on handcrafted features. This includes retraining a CNN end-to-end for MAD or utilizing the pretrained RN50-PGGAN detector. The latter suggests that integrating methods from deep synthetic image detection literature into the MAD problem might be a promising direction. This is especially relevant as generative AI trends means that deep morph generation methods are expected to diversify, whereas landmark-based morphs might remain limited to existing techniques. While LBP-based approaches show reasonable merit in intra-dataset scenarios, they suffer from significant generalization errors in cross-dataset settings. Although they can lead to minor improvements when fused with the RN50-PGGAN model, the pretrained detector still does most of the heavy lifting.

- **Generalization**: The challenges in generalization are not only due to the presence of unseen attacks in the test set but also, and perhaps *more significantly*, due to shifts in the *source dataset* used for testing compared to the training set. For instance, Table 5.2 shows better generalization from FRGC-SG2 to FRGC-MIPGAN than from FRGC-SG2 to FRLL-SG2. This underscores the importance of carefully analyzing different generalization challenges and isolating the various factors that contribute to these issues.

- **Interpretability**: One advantage of LBP-LDA classifiers is that they produce saliency maps with a stronger semantic structure compared to Grad-CAM visualizations on CNNs, enabling more explicit human-level explanations for predictions and mitigate the fear of black-box systems used in forensic decision-making. However, these maps are still insufficient as they do not clearly explain failure cases. The gain in interpretability from using handcrafted features is limited and does not compensate for the loss in detection accuracy compared to supervised features. Further work on interpretability approaches could be of interest not only in this context but also in general for synthetic image detection. The saliency maps approach might have limited applicability as it typically relies on some form of locality of salient features to provide easy-to-analyze visualizations, which might not be the case in our context, in particular if the salient information is actually localized in the frequential domain (and thus non local in the image domain). Other explanation methods exist and could be explored in future work, such as providing the expert with a few “most similar examples” to the analyzed input, taken from a set of known reference examples. One could also imagine developing a two-stage approach, where a very opaque but accurate classifier is used for flagging suspicious content (investigation phase), but a set of more transparent models are used to further analyze a case of interest with the aim of providing additional explanations (evaluation phase).

Additionally, we must acknowledge the following limitations of our study:

- **Lack of print-scan study**: Our work has been restricted to the digital domain, which limits its real-world applicability since morphing attacks are often expected to undergo print-scan post-processing. It is particularly important to evaluate the robustness of the RN50-PGGAN model to such post-processing. Although the model has been trained to withstand other degradations like resizing and compression through data augmentation, it remains uncertain whether this robustness extends to print-scan post-processing.
- **Absence of landmark-based morphs**: Given the high effectiveness of landmark-based morphing attacks, as highlighted in Chapter 3, a comprehensive detection system should also be capable of handling these attacks. While a robust GAN-image detector like RN50-PGGAN performs well for GAN-based morph detection, it is unlikely to be effective against landmark-based morphs, which are not expected to exhibit the same type of traces.

From these takeaways and limitations, we establish the following guidelines for orienting our subsequent work on MAD:

1. Approach MAD as a deepfake detection problem and draw from methods in deep synthetic image detection. Importantly, *evaluate the performance of such methods on landmark-based morphs* (or in other words, cheapfakes). Ensure that all major families of morphing attacks are represented in our experimental data.
2. Properly isolate different generalization challenges: generalization to unseen attacks and generalization to new source datasets.
3. Thoroughly evaluate the robustness to print-scan degradations.
4. Focus on data-driven approaches (supervised features or foundational features) over handcrafted approaches. While the latter might be interesting from an interpretability point of view, it also makes it difficult to establish strong baselines, especially with the diversification of morphing attacks. Overall, a less interpretable but more accurate detection system is likely to be preferred over a more interpretable but less accurate one.

# 5 Attack-agnostic features and generalization in morphing attack detection

This chapter presents a second research contribution on the topic of morphing attack detection, following the lessons learned and recommendations from Chapter 4. We focus on methods inspired by deepfake detection research, consider all major families of morphing attacks (landmark-based, GAN-based, and diffusion-based), isolate individual generalization challenges, and systematically evaluate the robustness of the developed methods to print-scan post-processing.

In particular, we aim to explore methods based on **foundational features** and evaluate their advantages over supervised features. As presented in Chapter 1, recent advancements in deepfake detection have demonstrated the effectiveness of using internal features from pretrained large vision models. These features can be used in conjunction with simple downstream classifiers to perform remarkably accurate detection. Notably, features extracted using a pretrained CLIP model, originally trained for image-caption alignment, have shown promise in previous studies [84], [85].

A crucial aspect of such foundation models is that they are solely trained on real images, and not on any form of synthetic ones. This means that the learned representation *cannot* be biased towards a specific subfamily of synthetic images by design, yet it appears the representation can still be useful for deep synthetic image detection. This is particularly interesting in our case because it also means the representation is not going to overly focus on some synthetic signature information which would be absent from cheapfake images. It is thus a promising direction for developing models that can jointly detect deepfakes and cheapfakes.

To further emphasize the fact that the considered foundation models are not trained on any form of synthetic data (and in particular no morphing attacks), we call the representation they produce **attack-agnostic**, which we contrast with **attack-aware** features when produced from models trained on some amount of synthetic data. The use of task-specific data (i.e., morphing attacks) is relegated to the training of a downstream classifier on top of the learned representation.

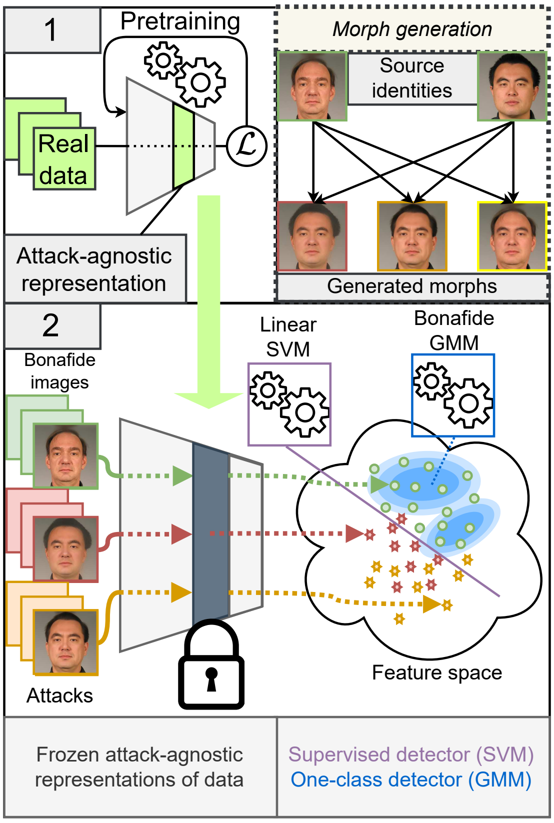


Figure 5.1: We tackle the problem of MAD using pretrained attack-agnostic extractors. Morph generation: we generate morphs using a variety of algorithms (landmark-based, GAN-based, and diffusion-based). Stage 1: the attack-agnostic extractor is a large vision model trained on real images for a pretext task. We reuse it to summarize any image by extracting an internal representation as the feature vector. Stage 2: features are extracted for bona fide images and face morphs. We train a supervised morphing attack detector as a linear SVM on top of this features space. We train a one-class detector by modeling the distribution of bona fide features with a Gaussian Mixture Model (GMM), then using the likelihood of incoming samples as the discriminative score.

This study thus focuses on evaluating the applicability of attack-agnostic features for MAD. Specifically:

- We develop and evaluate MAD systems using simple probe classifiers trained on attack-agnostic feature representations. In particular, we extend the idea of using CLIP features proposed in [84], and consider a wider set of publicly available foundation models.
- We also develop and evaluate MAD systems based on one-class modeling of the bona fide class, detecting morphs as out-of-distribution samples, an approach enabled by the use of attack-agnostic representations.
- We compare our methodology against traditional supervised CNN training through

extensive experiments involving three different datasets and five types of morphing attacks spanning three categories: landmark-based, GAN-based, and diffusion-based. Our evaluation includes a variety of scenarios, focusing on the generalization capabilities across different families of attacks, source datasets, and domains (digital to print-scan).

For completeness, we mention the existence of a parallel work with methodological similarities [141], using a pretrained Vision Transformer as the main extractor. Compared to this work, our study considers a broader set of foundation models and integrates diffusion-based morphs in the analysis, which are absent from their analysis.

## 5.1 Methodology

We aim to evaluate the performance of MAD systems in various settings, with a focus on assessing generalization capability across unseen attack families (LB, GAN, or Diff), unseen source datasets, and different domains (digital to print-scan). The considered attacks are all the ones presented in Table 2.2 in Chapter 2, which also details which attacks are part of each family. Additionally, we seek to evaluate the performance of one-class detectors trained solely on bona fide data. The following evaluation scenarios are considered:

### 5.1.1 Evaluation scenarios

1. **Baseline** : the detector is trained and tested on digital bona fide and morph samples from the same source dataset (FRGC or FFHQ), with all families of attacks seen during training.

2. **Generalization to unseen attacks** : unlike the baseline, the detector is trained using only a single family of attacks (LB, GAN, or Diff.) and tested on the other two families.

3. **Generalization to different source datasets** : unlike the baseline, the detector is tested on bona fide and morph samples from an unseen source dataset, specifically FRLL.

4. **Generalization to print-scan data** : unlike the baseline, the detector is tested on print-scanned bona fide and morph samples. This scenario is evaluated only using FRGC, for which print-scanned data is available.

5. **One-Class Detection** : the detector is trained solely on bona fide samples and then tested on all attacks. For this setting, we restrict ourselves to a single source dataset and to the digital domain.

The first four scenarios involve training the detector in a supervised manner as a binary classifier. In the last scenario, one-class detectors are achieved by modeling the statistical distribution of the features of bona fide samples, then using the likelihood score of incoming samples under the learned distribution as the discriminative score. For both types of systems, performance is evaluated by reporting the D-EER on the respective test sets.

### 5.1.2 Models

We consider two types of detection models. The first type, which is the focus of our study, involves training a simple downstream classifier on top of pretrained features extracted from an attack-agnostic vision model, i.e., a network trained solely on bona fide data for some auxiliary task (cf. Figure 5.1). The second type is used for comparison purposes, and consist in fully training a convolutional neural network directly on image samples, either as a binary classifier (in the supervised setting) or as an autoencoder (in the one-class setting).

#### Probed Attack-Agnostic Models

We consider the following attack-agnostic feature extractors:

- **RN50-IN** [142] : this baseline extractor is a ResNet50 network trained for image classification on ImageNet. We use the output of the penultimate layer before the image classification layer as the feature representation of images.

- **DINOv2** [10] : this extractor is trained in a self-supervised manner with the goal of learning general image representations. It has demonstrated effectiveness for a broad variety of downstream classification tasks and serves as a more sophisticated baseline compared to RN50-IN. We specifically use the ‘giant‘ variant, and use the learned general representation as feature vector.

- **CLIP** [11] : this vision-language model is trained to represent matched image-caption pairs jointly in the same feature space. Despite being trained for a seemingly unrelated task, previous research [84], [85] has shown that CLIP-extracted features showcase strong discriminative power to differentiate between bona fide and synthetic images. We use the L/14 variant as suggested by [84], and use the output of the vision encoder as feature vector.

- **AIM** [87] : this extractor is pretrained for autoregressive image modeling, which involves decomposing images into ordered sequences of patches and predicting subsequent patches using only the context of previous patches. This autoregressive objective is theoretically equivalent to learning the true underlying image distribution. Trained on a massive dataset of 12.8 billion images, AIM has the potential to approximate the distribution of “natural” images. Given that deep synthetic images typically exhibit salient statistical differences from bona fide ones [85], we hypothesize that they might lie outside the distribution learned by AIM. We use the 600M variant, and use the pool-averaged output of the trunk as the feature representation.

- **DNADet** [81] : this extractor is originally designed to improve the accuracy of source attribution for GAN-generated images. It is pretrained using real images for a task of patchwise contrastive learning of image transformations, where images undergo various degradations (e.g., blurring, JPEG compression) and are decomposed into patches. The model learns to represent patches subject to the same degradations close to each other, and patches subject to different degradations far apart, and additionally has to classify incoming patches based on their applied degradation. Given the already demonstrated efficacy of this pretraining in learning salient features to differentiate various GAN models, DNADet is a strong candidate for synthetic image detection, particularly as its pretraining data includes face images, making it also content-specific for our case. We use the output of the penultimate layer, right before the fully connected layer used for the classification, as the feature representation.

For supervised modeling, we train a downstream linear probe on top of the extracted features, specifically a binary linear SVM, preceded by a Principal Component Analysis (PCA) decomposition achieving 99% of explained variance. This initial projection mitigates the challenges posed by the high dimensionality of certain feature spaces.

For one-class modeling, we fit the distribution of bona fide features using a GMM, also preceded by PCA decomposition achieving 99% of explained variance. The log-likelihood of incoming samples under this statistical model is then used to distinguish between bona fide samples, which are expected to have high log-likelihood values, and attacks, which are expected to have low log-likelihood values. To determine the optimal number of components for the GMM (ranging from 1 to 256), as well as the type of covariance matrix (diagonal or spherical), we perform 4-fold cross-validation on the training set. The validation set includes attack samples, and the D-EER on the validation set is used as the selection criterion.

#### Reference MAD Models

For the supervised detection setting, we use as comparative reference the MixFaceNet architecture, which has been employed in prior work as a backbone for training MAD systems. The model is a CNN trained as a binary classifier directly on image examples. We reproduce the backbone setup and training process as described in [114], and reuse their provided code.[1] For the one-class setting, we compare our models to the SPL-MAD model from [122]. The SPL-MAD model is trained as a convolutional autoencoder on the Casia-WebFace dataset (bona fide face images). At inference time, the authors observe that the reconstruction error is smaller for morphs than for bona fide images, and thus can be used as a discriminative score for detection, despite using only bona fide data at training time. We use the code and pretrained model provided by the authors[2].

[1] https://github.com/naserdamer/SMDD-Synthetic-Face-Morphing-Attack-Detection-Development-dataset

[2] https://github.com/meilfang/SPL-MAD

Table 5.1: **Baseline.** D-EER (%) on the test split when all attacks are seen at training time. **Bold** values indicate setups where probed attack-agnostic models perform better than the MixFaceNet MAD reference. Underlined values are the best performing models.

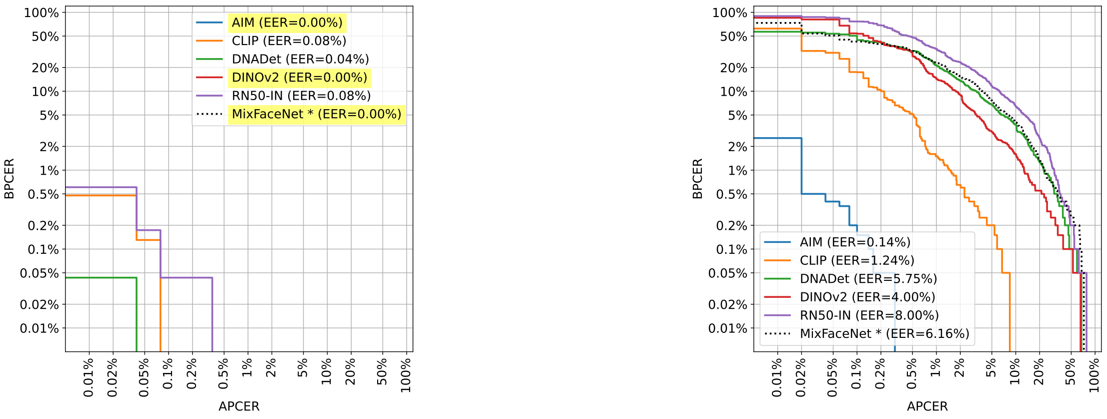
| Src dataset | FRGC | | | FFHQ | | |
|---|---|---|---|---|---|---|
| Test on | LB | GAN | DIFF | LB | GAN | DIFF |
| Model | | | | | | |
| AIM | 0.00 | 0.00 | 0.00 | **0.20** | **0.05** | **0.05** |
| CLIP | 0.00 | 0.00 | 0.13 | **1.45** | **0.40** | **1.65** |
| DNADet | 0.00 | 0.04 | 0.00 | **5.70** | 5.75 | **6.10** |
| DINOv2 | 0.00 | 0.00 | 0.00 | **5.50** | **2.45** | **3.25** |
| RN50-IN | 0.04 | 0.17 | 0.00 | 9.70 | 7.25 | **6.35** |
| MixFaceNet * | 0.00 | 0.00 | 0.00 | 6.70 | 5.30 | 7.05 |

## 5.2 Results

We present the results in two formats. First, we use D-EER tables to summarize performance with a single metric, facilitating comparison between approaches and breaking down results by the main families of test attacks. Second, we provide BPCER@APCER DET curves for each scenario, grouping performance across all families of test attacks. This visual comparison allows us to evaluate specific operating point performances.

However, in several evaluation scenarios, no model performs satisfactorily enough to achieve reasonable operating performance. Therefore, the D-EER remains the most valuable metric, as it effectively highlights meaningful performance differences across various models.

### Baseline performance



(a) Src. Dataset: FRGC

(b) Src. Dataset: FFHQ

Figure 5.2: **Baseline**: Train on all digital attacks from 1 source dataset, test on all digital attacks from the same source dataset. Systems highlighted in yellow actually reach a D-EER of 0.0% in this setting (hence the curve is not visible).

Table 5.1 and Figure 5.2 show results for the baseline scenario where both training and testing data come from the same source dataset in the digital domain, and all attacks are known during training. For attacks from a constrained dataset (FRGC), all methods perform well, achieving nearly perfect separation between bona fide and attack samples regardless of the attack family. However, with a more diverse dataset (FFHQ), performance declines for most methods, and differences become more evident. Here, our linear probes generally outperform the MixFaceNet detectors, with AIM and CLIP achieving the best results across all attack families.

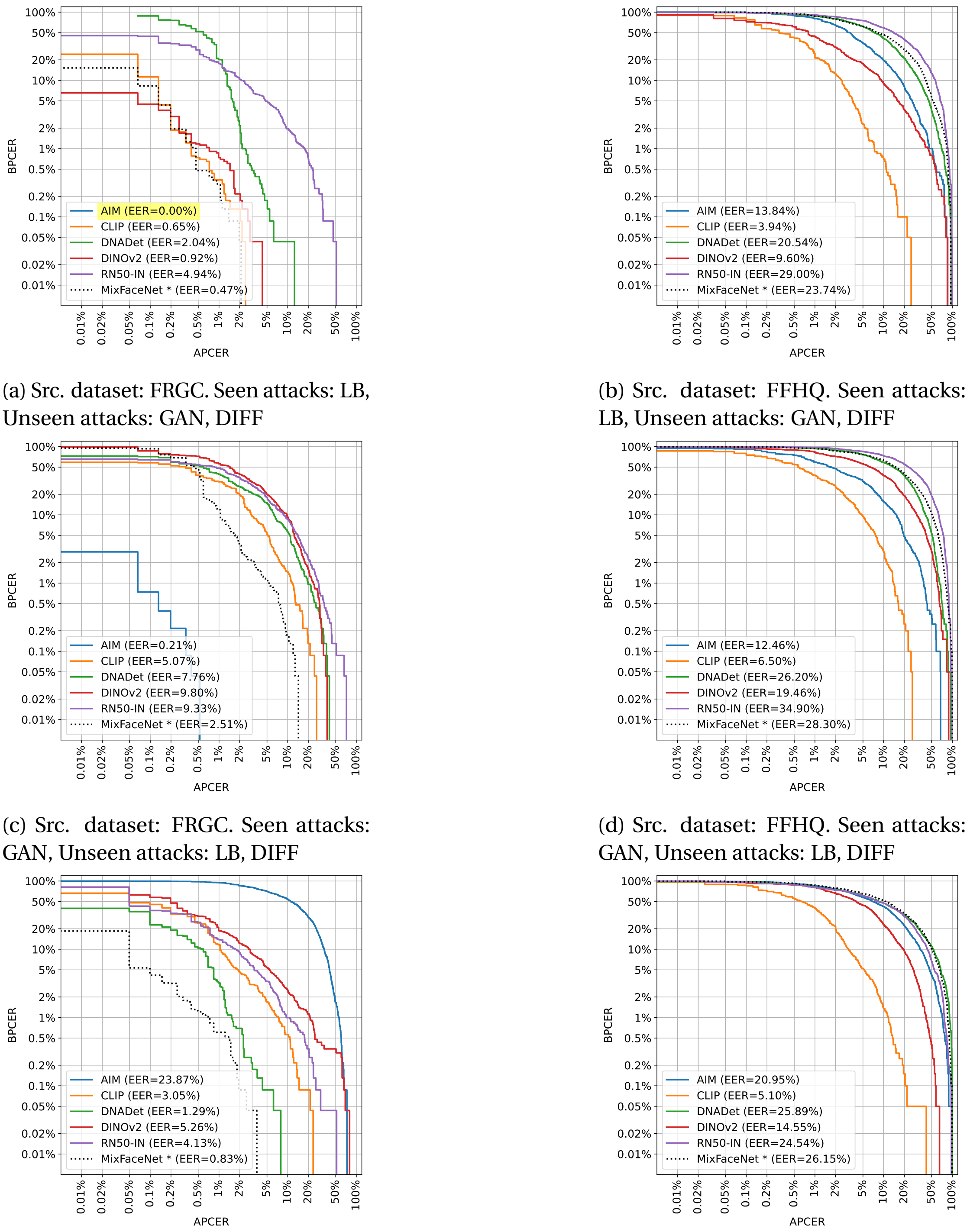
### Generalization to unseen attacks

Table 5.2 and Figure 5.3 present results where only one attack family is known during training. For FRGC attacks, AIM features perform best except when only diffusion attacks are known; in this case, MixFaceNet is superior for diffusion to landmark-based generalization, and CLIP is best for diffusion to GAN generalization. The DNADet probe shows comparable generalization from diffusion to both landmark-based and GAN attacks, though it performs slightly worse than MixFaceNet overall.

For FFHQ attacks, CLIP probes consistently outperform AIM probes and MixFaceNet across all generalization scenarios. Our linear probing approach also frequently surpasses MixFaceNet.

Table 5.2: **Unseen attacks generalization.** D-EER (%) on the test split when a single family of attacks is seen at training time. **Bold** values indicate setups where probed attack-agnostic models perform better than the MixFaceNet MAD reference. Underlined values are the best performing models.

| | Train attacks | LB | | | GAN | | | Diff | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Test attacks | LB | GAN | Diff | LB | GAN | Diff | LB | GAN | Diff |
| Src. dataset | Model | | | | | | | | | |
| FRGC | AIM | 0.00 | **0.00** | **0.00** | **0.22** | 0.00 | **0.39** | 33.81 | 8.68 | 0.00 |
| | CLIP | 0.00 | **0.00** | 1.22 | 5.21 | 0.00 | 5.03 | 4.34 | **0.22** | 0.00 |
| | DNADet | 0.00 | 2.91 | **0.00** | 9.77 | 0.00 | **0.65** | 1.39 | 1.09 | 0.00 |
| | DINOv2 | 0.00 | 0.69 | 1.13 | 10.81 | 0.00 | 5.90 | 7.34 | 2.78 | 0.00 |
| | RN50-IN | 0.09 | 6.03 | **0.00** | 11.50 | 0.09 | **0.74** | 2.86 | 4.86 | 0.00 |
| | MixFaceNet * | 0.00 | 0.48 | 0.13 | 2.73 | 0.00 | 1.87 | 0.95 | 0.61 | 0.00 |
| FFHQ | AIM | **0.00** | **11.20** | **19.60** | **12.90** | **0.00** | **11.90** | 27.90 | **13.00** | **0.00** |
| | CLIP | **1.25** | **0.90** | **8.30** | **5.70** | **0.00** | **7.90** | **7.75** | **1.20** | **0.30** |
| | DNADet | **2.30** | **17.05** | **27.20** | **16.15** | 1.95 | 41.45 | **25.30** | 26.35 | **0.90** |
| | DINOv2 | 5.10 | **8.25** | **12.20** | **20.75** | **0.30** | **17.85** | **19.05** | **10.15** | **0.55** |
| | RN50-IN | 7.75 | 33.00 | **17.65** | 33.90 | 2.45 | 36.60 | **21.90** | 26.95 | **2.25** |
| | MixFaceNet * | 5.00 | 18.10 | 33.75 | 24.05 | 0.85 | 34.55 | 26.15 | 26.15 | 2.40 |



(a) Src. dataset: FRGC. Seen attacks: LB, Unseen attacks: GAN, DIFF

(b) Src. dataset: FFHQ. Seen attacks: LB, Unseen attacks: GAN, DIFF

(c) Src. dataset: FRGC. Seen attacks: GAN, Unseen attacks: LB, DIFF

(d) Src. dataset: FFHQ. Seen attacks: GAN, Unseen attacks: LB, DIFF

(e) Src. dataset: FRGC. Seen attacks: DIFF, Unseen attacks: LB, GAN

(f) Src. dataset: FFHQ. Seen attacks: DIFF, Unseen attacks: LB, GAN

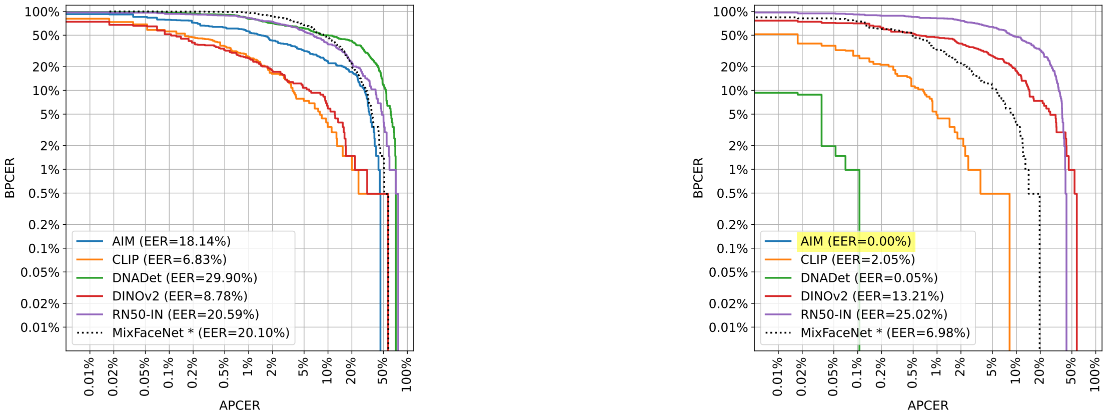
Figure 5.3: **Generalization to unseen attacks**: train on **one family of attacks** (digital) from one source dataset, test on **unseen digital attacks** from the same source dataset. Systems highlighted in yellow actually reach a D-EER of 0.0% in this setting (hence the curve is not visible).

### Generalization to different source datasets

Table 5.3 and Figure 5.4 reports results when the source dataset differs between training and testing, focusing on the FRLL dataset. The experiment highlights in particular the importance of source dataset diversity for effective generalization. Indeed, detectors trained on FFHQ attacks generally perform better on FRLL attacks than those trained on FRGC. This trend holds for both our linear probes and the MixFaceNet detector, with DINOv2 being a notable exception. Overall, linear probes typically outperform MixFaceNet. When trained on FFHQ attacks, AIM, DNADet and CLIP-based models achieve almost perfect separation between FRLL morphs and bona fide samples but perform poorly when trained on FRGC attacks. In this latter case, CLIP probes provide the most balanced performance for generalization to unseen datasets across all attack types.

Table 5.3: **Source dataset generalization.** D-EER (%) on FRLL bona fide & morph images when all attacks based on a *different* source dataset are seen at training time. **Bold** values indicate setups where probed attack-agnostic models perform better than the MixFaceNet MAD reference. Underlined values are the best performing models.

| Train src. dataset | FRGC | | | FFHQ | | |
|---|---|---|---|---|---|---|
| Test attacks | LB | GAN | DIFF | LB | GAN | DIFF |
| Model | | | | | | |
| AIM | **1.47** | **23.53** | **11.76** | **0.00** | **0.00** | **0.00** |
| CLIP | **6.86** | **4.90** | **7.84** | 3.43 | **0.49** | **0.98** |
| DNADet | **10.29** | 35.78 | 42.65 | **0.00** | **0.00** | **0.00** |
| DINOv2 | **9.80** | **8.33** | **3.92** | 15.20 | 13.24 | 5.88 |
| RN50-IN | 13.24 | 29.41 | **19.61** | **1.96** | 38.73 | **0.98** |
| MixFaceNet * | 12.75 | 28.92 | 20.10 | 2.94 | 11.76 | 1.47 |



(a) Train on FRGC attacks, test on FRLL attacks

(b) Train on FFHQ attacks, test on FRLL attacks

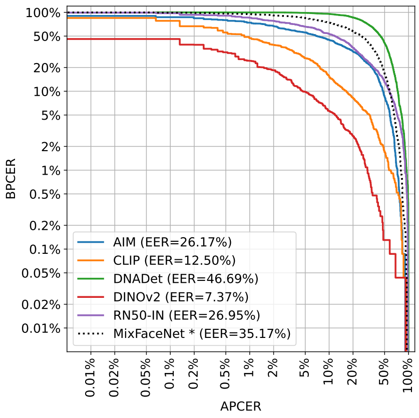
Figure 5.4: **Generalization to different source datasets**: train on all digital attacks from one source dataset, test on all digital attacks from another source dataset. Systems highlighted in yellow actually reach a D-EER of 0.0% in this setting (hence the curve is not visible).

### Generalization to print-scan data

Table 5.4 and Figure 5.5 show results when detectors trained on digital data are evaluated on print-scan data. This scenario is challenging because artifacts left by deep morph generators on generated samples are likely degraded during the print-scan process. Most detectors, which achieved perfect separation in the baseline protocol, show significant performance drops on print-scan data (notably LB-PS and Diff-PS, even though they contain attacks whose digital counterpart has been seen during training). The MIPGAN-PS attacks are particularly challenging due to being totally unseen during training. Nevertheless, our linear probes still generally outperform MixFaceNet, with DINOv2 features being the most effective, followed by CLIP.

Table 5.4: **Print-scan generalization.** D-EER (%) on test split when all digital attacks are seen at training time, but test attacks are in the print-scan domain. **Bold** values indicate setups where probed attack-agnostic models perform better than the MixFaceNet MAD reference. Underlined values are the best performing models.

| Src dataset | FRGC | | |
|---|---|---|---|
| Test on | LB-PS | MIPGAN-PS | DIFF-PS |
| Model | | | |
| AIM | **4.77** | **32.47** | **30.12** |
| CLIP | **3.99** | **15.02** | **14.97** |
| DNADet | **16.19** | 60.50 | 56.55 |
| DINOv2 | **8.85** | **5.60** | **7.51** |
| RN50-IN | **20.57** | **35.24** | **26.78** |
| MixFaceNet * | 22.05 | 50.22 | 32.34 |



(a) Src. dataset: FRGC

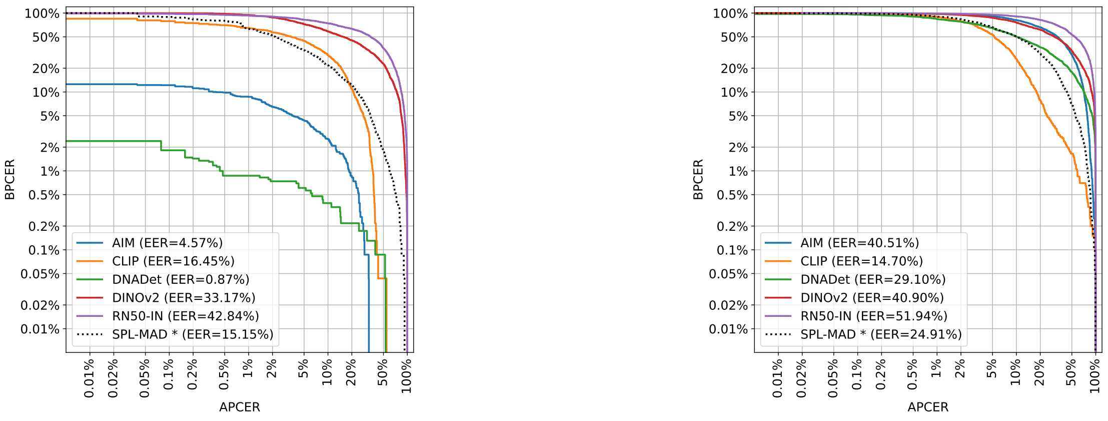
Figure 5.5: **Robustness to print-scan**: train on all digital attacks from one source dataset, test on all **print-scan** attacks from the same source dataset.

### One-class detector

Finally, Table 5.5 and Figure 5.6 present the performance of one-class detectors trained only on bona fide data. It is important to note that the comparison to SPL-MAD is not entirely fair, as SPL-MAD is trained on Casia-Webface data, while our detectors are specifically tuned to the considered source dataset, providing an advantage. Nonetheless, for FRGC attacks, AIM and DNADet probes show quite strong performance, and significantly better than SPL-MAD. DNADet probes in particular lead to an impressive D-EER of under 1% for all considered families of attacks, even though the detector is never exposed to any attack for its development. For FFHQ attacks however, the overall detection performance is unsatisfactory, with CLIP features proving to be the most effective in this scenario.

Table 5.5: **One-class model.** D-EER (%) on the test split when only bona fide sample are seen at training time. We compare to the SPL-MAD model from [122]. **Bold** values indicate setups where probed attack-agnostic models perform better than the SPL-MAD reference. Underlined values are the best performing true one-class models.

| Src. dataset | FRGC | | | FFHQ | | |
|---|---|---|---|---|---|---|
| Test attacks | LB | GAN | DIFF | LB | GAN | DIFF |
| Model | | | | | | |
| AIM | **6.08** | **0.39** | **0.00** | 34.40 | 56.10 | **7.20** |
| CLIP | 23.87 | **1.52** | 20.92 | **14.50** | **4.75** | **27.70** |
| DNADet | **0.87** | **0.82** | **0.48** | **27.10** | 29.10 | **32.80** |
| DINOv2 | 35.72 | 32.86 | 30.16 | 35.80 | 48.90 | **34.00** |
| RN50-IN | 51.56 | 43.23 | **18.75** | 46.75 | 61.25 | 46.10 |
| SPL-MAD * | 16.28 | 11.02 | 20.23 | 28.15 | 14.10 | 34.20 |



(a) Src. dataset: FRGC

(b) Src. dataset: FFHQ

Figure 5.6: **One-class detection**: train on bona fide samples only, test on all digital attacks.

## 5.3 Discussion

### 5.3.1 Data representations

The results demonstrate that the considered attack-agnostic feature representations are highly effective for morphing attack detection. Training simple probes on these features consistently outperforms a CNN detector trained end-to-end on image samples across all generalization scenarios. They also lead to improved performance over an out-of-the-box one-class detector from the recent literature. However, *which* representation is the most effective is scenario-dependent.

The key outcomes can be summarized as follows:

- **DNADet features** are particularly effective for one-class modeling in the digital domain and when targeting a single passport standard. The DNADet one-class detector achieves a D-EER under 1% for all attack families on FRGC attacks. However, these features exhibit poor performance in print-scan generalization. This limitation is likely due to DNADet's pretraining task of contrastive learning of image transformations, which may result in a different representation manifold for print-scan images compared to digital ones. Incorporating print-scan data into the bona fide training set may resolve this issue, which we plan to explore in future work.

- **AIM features** excel for generalizing to unseen attacks but show inconsistencies in other generalization scenarios. While AIM features behave overall similarly to DNADet features, their more irregular performance across different attack families may limit their practicality in real-world applications.

- **DINOv2 features** are particularly suitable for print-scan generalization. In scenarios where we assume a limited known set of possible attacks (i.e., all attacks can be seen during training), these features are valuable when generating actual print-scan data for training is impractical or too time-consuming. Future work should in particular verify if this print-scan generalization performance holds across a wider variety of physical devices.

- **CLIP features**, even though they are rarely the best, consistently perform well across all generalization scenarios, making them interesting for scenarios where multiple generalization challenges are simultaneous. By enabling robust generalization to unseen attacks, strong source dataset generalization, and decent print-scan generalization, they become a strong candidate for training detectors in a supervised way on a small set of attacks. In the one-class setting, CLIP features, while less effective than DNADet on FRGC attacks, are the most effective for FFHQ attacks. Coupled with their strong source dataset generalization capability, this fact makes them potentially well-suited for developing more general-purpose one-class MAD systems that target *multiple* passport standards.

### 5.3.2 Generalization

In terms of generalization, we reinforce the observation from Chapter 4 that generalization to novel source datasets is at least as challenging as generalization to unseen attacks. However, we demonstrate a relatively effective methodology through the use of a source dataset with increased diversity.

Despite the FFHQ-based morphs being created with somewhat less care in the selection of morphing pairs, using it as a training set leads to strong performance transfer on FRLL morphs, much more than when using FRGC morphs as the training set. This does not seem to be simply due to differences in dataset sizes, given that the orders of magnitude are similar (we have 8000 training set bona fide images for FFHQ versus 9228 for FRGC, and 4000 training set morphs per attack for FFHQ versus 2014 for FRGC). It suggests that diversity is also significantly contributing to the phenomenon.

This is a promising result as the FFHQ dataset has a relatively more permissive license compared to typical passport-quality face datasets such as FRGC, notably regarding the publication of derivative for non-commercial purposes. It might thus enable the public release of a large-scale morphing dataset based on real source images, which is still not yet available (the main large-scale public dataset, SMDD, being based on synthetic source images). Given the generalizability of detection models trained on FFHQ morphs to other source datasets, this dataset might facilitate MAD research, which is often slowed down by the lack of easily accessible training data. However, it is essential to first refine the dataset creation process by selecting FFHQ samples with the highest face image quality and carefully choosing morphing pairs that use lookalikes as the two sources. This refinement allows us to evaluate whether a more meticulous process can lead to additional generalization benefits.

### 5.3.3 Future work

Future work should focus on several key areas to further enhance the robustness and generalizability of our proposed approach. First, a more systematic evaluation of one-class detection performance is necessary, particularly to ensure fairer comparisons with existing methods, notably by making sure equivalent bona fide sets are seen at training time. Second, an evaluation of the one-class performance of DNADet in the print-scan domain is needed, likely requiring the inclusion of bona fide print-scan data in the training set. Third, there is a broad research space exploring alternative pretraining methods for developing attack-agnostic extractors with particularly strong downstream performance for morphing attack detection. One could notably consider that training on content-specific data (i.e., faces exclusively) might be an advantage; it is actually the approach that is adopted notably for the SPL-MAD model. Lastly, the print-scan generalization capabilities of DINOv2 should be evaluated using additional print-scan devices to verify its effectiveness across a broader range of physical conditions.

In conclusion, the study validates the effectiveness of attack-agnostic representations for

MAD, with DNADet and CLIP feature representations standing out in one-class and generalist performances, respectively, and DINOv2 in print-scan generalization. The outlined future work aims to address current limitations and further optimize these models for practical deployment in diverse real-world scenarios.

# 6 Closed-set source attribution of morphing attacks using attack-agnostic features

This chapter presents a contribution to the task of source attribution of morphing attacks. While crucial, morph detection capability is only a portion of the problem. In a broader forensic investigation framework, it is essential not only to determine the fundamental status of the analyzed data (bona fide or fake) but also to gain insight into the involved criminal actor. For instance, in the series of morphing cases observed in Slovenia mentioned in Chapter 2, one might ask whether several observed morphs are produced by the same entity or by distinct organizations. A first step towards this goal is to achieve **source attribution** at an algorithm level: *help to identify* ***which*** *generative algorithm has been used to create an observed morph*, as illustrated in Figure 6.1. This question is becoming increasingly relevant with the diversification of morphing algorithms. While not directly providing information on the actual perpetrator, source attribution offers insights that can be integrated into a larger investigation. It is reasonable to assume that a given organization would produce morphs using a single method or closely related methods, while different organizations might use lesser related methods.

Ideally, one would like to perform source linking, i.e., assess whether two fake media are produced by the same algorithm, even if the algorithm is unknown. However, given the current state of research, it is necessary to tackle easier problem settings such as closed-set attribution (where all possible generative algorithms are assumed to be known beforehand) and out-of-distribution detection, which aims to identify whether an incoming fake data sample is generated using one of the known algorithms or a novel, unknown one. This study focuses on opening this line of research by tackling closed-set attribution of morphing attacks.

Given the observed success of attack-agnostic features for morphing attack detection, as shown in the previous chapter, we aim to evaluate the effectiveness of similar features for the task of attribution. This has several potential benefits. First, it would be ideal to find a representation of data that is effective both for detection and attribution, i.e., one that not only separates bona fide and attack samples neatly but also distinguishes different attacks. Hence, it makes sense to explore the effectiveness of representations that have been shown to be useful for detection. Moreover, attack-agnostic representations might be necessary for extensions to open-set attribution, as it is important that new attacks, which cannot be used as part of the representation learning process, appear clearly distinct from known ones in the representational space.

However, the focus of this chapter is limited to closed-set attribution, excluding the bona fide class. In other words, we assume a scenario where a given data sample has already been identified as fake, and we then try to associate it with a known algorithm.

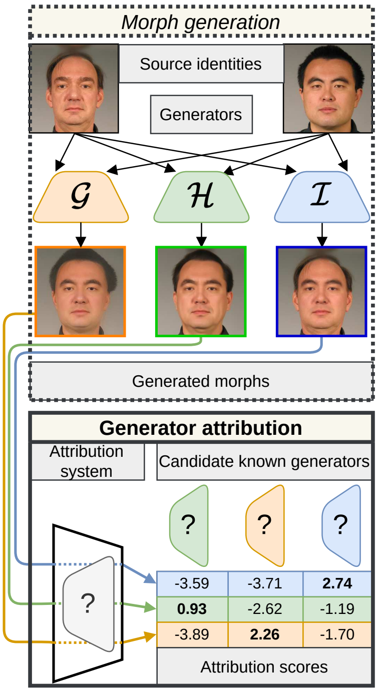
The main contributions are as follows:

- We develop and evaluate systems for closed-set attribution of morphing attacks. To the best of our knowledge, this is the first work proposing to tackle the problem of attribution for morphs.
- We evaluate the benefit of attack-agnostic pretrained feature extractors for morphing attribution, particularly because they are more practical in a scarce-data setting, more relevant for eventual extension to open-set scenarios, and have the potential to mitigate generalization issues. We include comparisons with attack-aware pretrained extractors to evaluate the advantages and drawbacks of attack-agnosticism.
- We assess the attribution performance in various settings, considering specific challenges posed by print-scan morphs and generalization across source datasets.
- We develop partial insights into the structure of the attack-agnostic feature spaces, particularly how they separate or cluster attacks, both qualitatively through T-SNE visualizations and quantitatively through K-Means clustering.

## 6.1 Methodology

Table 6.1: Considered experimental protocols. D indicates all available digital attacks, PS indicates all available print-scan attacks. For protocols involving both FRGC and another source dataset, the MIPGAN attack is ignored. For protocols involving digital and print-scan attacks, it is assumed both versions of an attack should be considered to be the same.

| | Training | | Evaluation | |
|---|---|---|---|---|
| Protocol | Src. dataset | Attacks | Src. dataset | Attacks |
| FRLL | FRLL | D | FRLL | D |
| FRGC | FRGC | D | FRGC | D |
| FRGC→FRLL | FRGC | D | FRLL | D |
| FRLL→FRGC | FRLL | D | FRGC | D |
| FRGC D+PS | FRGC | D + PS | FRGC | D + PS |
| FRGC D→PS | FRGC | D | FRGC | PS |



| | | |
|---|---|---|
| -3.59 | -3.71 | **2.74** |
| **0.93** | -2.62 | -1.19 |
| -3.89 | **2.26** | -1.70 |

Figure 6.1: We tackle the problem of morphing attack generator attribution. Morph generation phase: starting from the same pairs of source identities, morphs are created using a variety of morph generators. Generator attribution phase : the developed attribution system, aware of all possible generators (closed-set scenario), has to assign attribution scores for each "morph/candidate generator" pair. This score should be maximal when the morph is matched against its true generator.

We start by an important point of terminology : confusion should be avoided between the notion of the *source dataset* (set of bona fide samples used as base images to create morphs) and *source attribution* (the commonly used general term for the identification of the origin of a data sample). In our particular context, we are interested in *generator* attribution, meaning that we care to identify the exact morphing algorithm that is used to generate a given morph data sample.

In terms of data, we consider all attacks presented in Table 2.2 from Chapter 2, and restrict ourselves to FRGC and FRLL source datasets. Final experimental data is obtained by assembling train splits from several morph sets together for training, and similarly several test splits together for evaluation, according to specific experimental protocols which specify exactly which sets are included in this process.

The main types of protocols are

- baselines (FRGC, FRLL) which use only digital attacks and for which the same source dataset and attacks are used at training and evaluation time,
- source generalization protocols (FRGC→FRLL, FRLL→FRGC) which use only digital

attacks but a different source dataset between training and evaluation,

- a print-scan protocol (FRGC D+PS) in which both digital and print-scan attacks are seen at training and evaluation,
- and a print-scan generalization protocol (FRGC D→PS) in which only digital attacks are seen at training, but print-scan attacks are presented at evaluation.

The considered protocols are summarized in Table 6.1.

### 6.1.1 Feature representations

The central goal of our work is to study the relevance of **attack-agnostic pretrained feature extractors** in the context of morphing attack attribution. By attack-agnostic, we mean that the pretraining stage only involves the use of real data. This notion is opposed to that of attack-aware extractors, for which some amount of synthetic data is used at pretraining time to learn the feature representation, as illustrated in stage 1 in Figure 6.2.

As mentioned in the previous section, this approach has two important benefits. First, there is a concern that end-to-end data-driven feature learning will cause features to overly specialize to the specific subsets of morphs that we are considered (this type of phenomenon has been illustrated in past works, for instance in [84] for the task of deepfake detection). This is problematic in particular if the end goal is to aim towards open-set systems, which need a sufficiently generic feature representation. Second, learning data-driven features in an end-to-end setting is also tricky given the scarce available data. By transferring features from some attack-agnostic pretrained model, and introducing attacks only to train a downstream classifier, the amount of required data is significantly lesser.

The considered feature extractors are presented in Table 6.2, and are all open-sourced by their respective authors. In our experiments, we reuse the provided pretrained weights and inference code.

We consider 5 attack-agnostic extractors pretrained using real data only, and for comparison, 5 attack-aware extractors which are pretrained using some amount of synthetic data (GAN or Diffusion images). Attack-agnostic extractors are pretrained either for image recognition, for text-guided pretraining (which consists in aligning image representations with the CLIP embedding of their available caption), or with a self-supervised learning (SSL) objective. DNADet-Pretrain in particular has been developed with deepfake attribution in mind, and is pretrained by augmenting input real images with a large set of compression, blurring, resampling and noising transformations, then training with a patch-wise contrastive learning objective which forces randomly cropped image patches with the same transformations to be represented close-to each other, and patches with different transformations to be far apart. Attack-aware extractors are either trained for GAN or Diffusion image detection, GAN image attribution in the case of DNADet-CelebA, or for forgery detection and localization in the case of HiFi-IFDL.

Our hypothesis is that attack-aware extractors are at risk to overspecialize to the specific family of synthetic data that they see at pretraining time, as illustrated in [84], leading the representation to be unfit when facing other types of attacks (diffusion attacks for GAN-aware extractors, GAN-attacks for diffusion-aware extractors, and landmark-based or print-scan attacks for both types). In contrast, attack-agnostic extractors are not subject to this risk, but it remains to assess whether their learned feature space shows any utility for morphing attack attribution.

The morphs, which are already cropped to align face landmarks as a byproduct of the generation process, are uniformly scaled to a resolution of 256×256, before being fed into the extractors to produce feature representations. For the models that can be straightforwardly decomposed as a feature extractor component and a classifier component (RN50-IN, ViT-B/16, Wang2020, Grag2021-SG2, Corvi2023), we use the output to the final classifier as the chosen representations. For models designed with a specific goal of representation learning (CLIP-L/14, DINOv2-L, both variants of DNADet), we simply use the learned representation. Finally, for HiFi-IFDL, we drop the forgery segmentation component, and consider only the multi-resolution feature maps from the multi-branch feature extractor component; we average-pool them and concatenate the resulting feature vectors to obtain the final representation.

### 6.1.2 Experiments

To evaluate the relevance of any given feature extractor for morphing attack attribution, we consider three distinct experiments:

1. **T-SNE visualization**: we select a given set of attacks (training split) and visualize their T-SNE projection in the considered space.
2. **Clustering**: we select a given set of attacks (training split), then perform unsupervised clustering of their representations using K-Means, setup to have as many clusters as there are different attacks. A test set is then labeled using the learned clustering, and we assess how well this labeling matches the ground-truth attack labels. Intuitively, this experiment measures how well the feature space "naturally" separates different morphing attacks into distinct clusters. This is quantified using the V-Measure [143] metric, which is defined as the harmonic mean between *homogeneity* (measures if the learned clusters tend to contain only data points coming from a single ground-truth class) and *completeness* (measures if data points coming from a single ground-truth class tend to belong to the same single cluster). The V-Measure is independent of the permutation of the cluster or ground-truth labels.
3. **Attribution**: we select a given set of attacks (training split), then train a multiclass linear SVM on their representations to learn to classify the attacks (in a one-vs-rest fashion). We then evaluate the attribution accuracy of the resulting classifier on a test set.

For the clustering and attribution experiment, we always precede the fitted model (KMeans, resp. SVM) by a dimensionality reduction step using PCA to a fixed dimension of 256, which is necessary given the wide variety in dimensionality of the feature spaces, ranging between 512 (for DNADet) to 2048 (for the ResNet extractors) as described in Table 6.2. The complete methodology for the attribution task is illustrated in stage 2 in Figure 6.2.

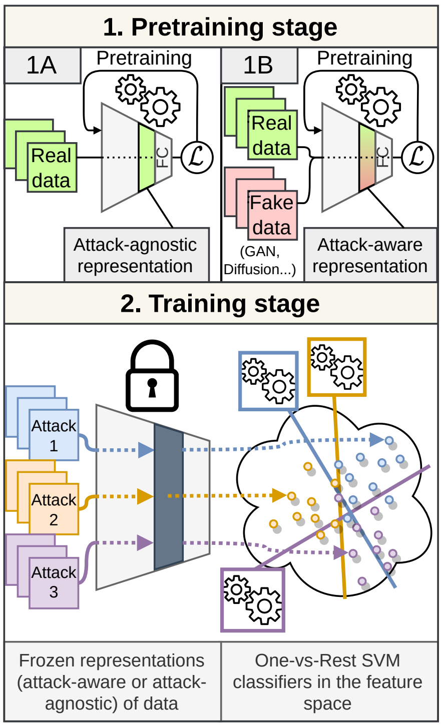

1. Pretraining stage
1A
Pretraining
Real data
FC
$\mathcal{L}$
Attack-agnostic representation
1B
Pretraining
Real data
Fake data
(GAN, Diffusion...)
FC
$\mathcal{L}$
Attack-aware representation
2. Training stage
Attack 1
Attack 2
Attack 3
Frozen representations (attack-aware or attack-agnostic) of data
One-vs-Rest SVM classifiers in the feature space

Figure 6.2: Considered design for our attribution systems. For feature extraction, we consider models pretrained for some auxiliary task and use an internal representation (the output of the last layer before the final fully connected layer) as the feature vector. Those extractors are pretrained either only on real data (1A), leading to attack-agnostic representations, or on a mixture of real and fake data (1B), leading to attack-aware representations. To train the attribution system (2), we freeze the pretrained model, extract features for each considered morphing attack, then fit one-vs-rest SVMs in to perform multiclass classification of the attacks in the considered feature space.

### 6.1.3 Representation fine-tuning

The five considered attack-agnostic extractors were initially trained on diverse datasets. In contrast, our focus is exclusively on face images. We thus propose to fine-tune these attack-agnostic representations using domain-specific data. We utilize only bona fide face images for fine-tuning to maintain the attack-agnostic nature of the resulting representations. We use for this purpose a set of 10’000 bona fide face images sampled from the FFHQ dataset.

The pretraining tasks of the considered extractors (e.g., image recognition, text-guided pre-training) are not directly applicable to this face dataset. Consequently, we fine-tune all the

Table 6.2: List of considered pretrained extractors. Extractors in bold are trained using only real data.

| | Architecture | Pretraining data | Pretraining task | Feature dim. |
|---|---|---|---|---|
| [142] **RN50-IN** | CNN | Real (ImageNet-1K) | Image recognition | 2048 |
| [144] **ViT-B/16** | ViT | Real (ImageNet-21K) | Image recognition | 768 |
| [11] **CLIP-L/14** | ViT | Real (400M images) | Text-guided pretraining | 768 |
| [10] **DINOv2-L** | ViT | Real (142M images) | SSL (with self-knowledge distillation) | 1024 |
| [81] **DNADet-Pretrained** | CNN | Real images (LSUN + CelebA) | SSL (contrastive learning of image transformations) | 512 |
| [69] Wang2020 | CNN | Real, GAN (ProGAN) | GAN image detection | 2048 |
| [70] Grag2021-SG2 | CNN | Real, GAN (StyleGAN2) | GAN image detection | 2048 |
| [71] Corvi2023 | CNN | Real, Diffusion | Diffusion image detection | 2048 |
| [82] HiFi-IFDL | CNN | Real, GAN, Diffusion, Cheapfakes | Forgery detection and localisation | 270 |
| [81] DNADet-CelebA | CNN | Real, GAN trained on CelebA | Fine-tune of DNADet-Pretrained for attribution | 512 |

extractors using the same self-supervised task, inspired by the DNADet pretraining process. This task involves augmenting the input images with various transformations (compression, blurring, resampling) and employing a contrastive learning objective. This objective ensures that images subject to identical transformations are represented close to each other, while those subject to different transformations are represented far apart. We adhere to the DNADet training setup [81], including hyperparameters, augmentations, and loss design. The key difference is that our contrastive objective is applied to the entire image rather than to patches. This adjustment is necessary because the pretrained extractors are designed to process full images. Moreover, given our objective of specializing the extractor to normalized face data, it makes sense to allow it to leverage spatial localization cues, which are otherwise lost with randomly cropped patches.

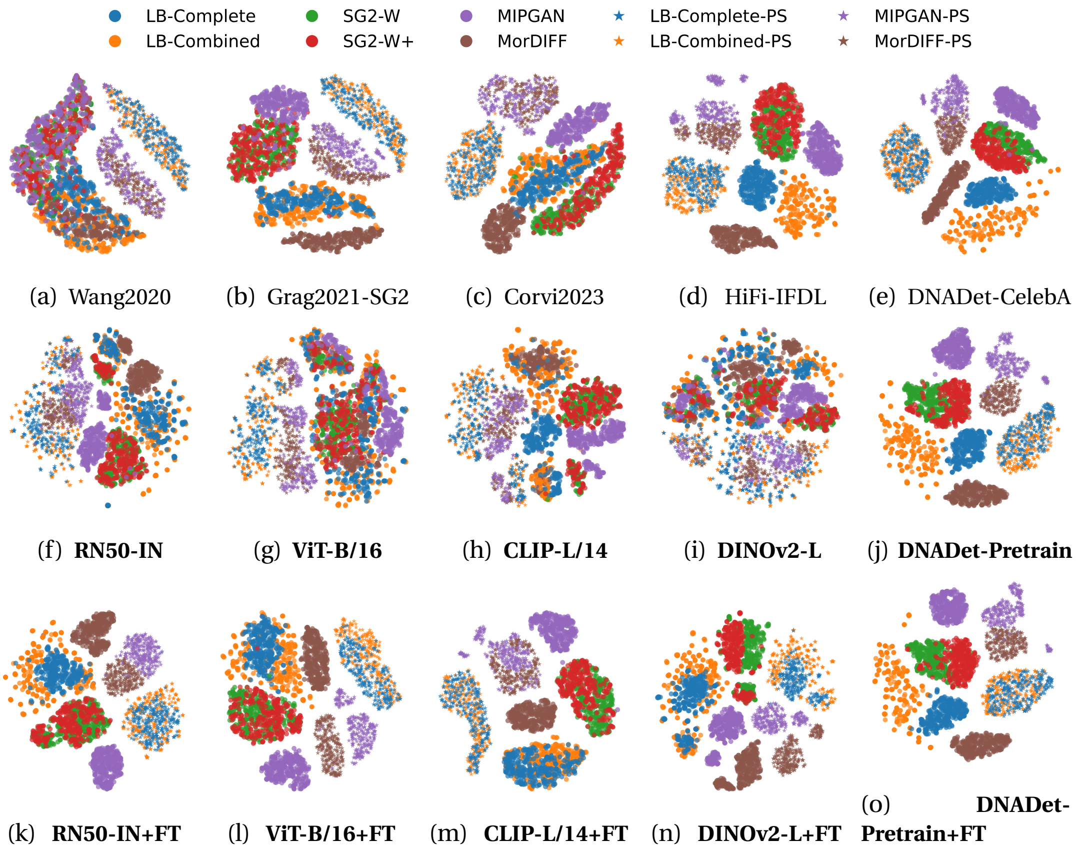
We perform a learning rate sweep for the feature extractor parameters, ranging from $10^{-7}$ to $10^{-5}$, and select the best model based on downstream attribution performance using a Linear SVM on a validation set of FFHQ-based morphs. Subsequently, we replicate the evaluation experiments described in Section 6.1.2 using the fine-tuned representations.

Overall, we insist that at no point morphing attacks are used to update or fine-tune the feature representations. The training and test sets described in Table 6.1 only describe which morphs are used for training the K-Means clusterings and SVMs, and which morphs are used for evaluating the performance using the V-Measure and accuracy metrics.

## 6.2 Results & discussion

### 6.2.1 T-SNE visualization

We focus on the T-SNE visualization of all FRGC attacks, as this dataset includes the largest set of attacks, notably including print-scan morphs. Visualizations for all considered feature spaces are presented in Figure 6.3.



(a) Wang2020 (b) Grag2021-SG2 (c) Corvi2023 (d) HiFi-IFDL (e) DNADet-CelebA

(f) **RN50-IN** (g) **ViT-B/16** (h) **CLIP-L/14** (i) **DINOv2-L** (j) **DNADet-Pretrain**

(k) **RN50-IN+FT** (l) **ViT-B/16+FT** (m) **CLIP-L/14+FT** (n) **DINOv2-L+FT** (o) **DNADet-Pretrain+FT**

Figure 6.3: TSNE visualization of FRGC attacks in the different considered feature spaces. Feature spaces in bold (bottom row) are attack-agnostic, i.e., learned without ever seeing synthetic data. We indicate fine-tuned extractors with the suffix "+FT"

#### Pre-fine-tuning analysis

We first examine the extractors before fine-tuning, highlighting their strengths and weaknesses in separating different attack classes.

Both attack-aware and attack-agnostic features provide some separability of different attacks, as shown in Figure 6.3. In particular, print-scan attacks are always clearly distinct from digital ones (see also Figure 6.6). However, attack-aware features tend to group individual attacks into single clusters, whereas most attack-agnostic features (except DNADet-Pretrain) spread

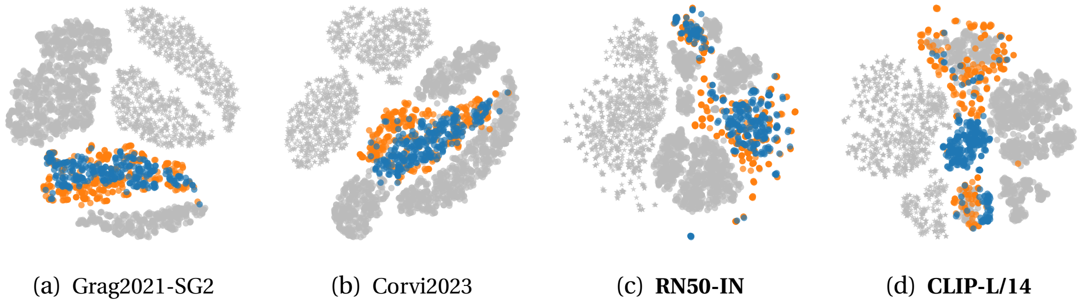


Figure 6.4: T-SNE projection of FRGC attacks highlighting the 2 landmark-based attacks (LB-Complete in blue, LB-Combined in orange). See also the legend in Figure 6.3.

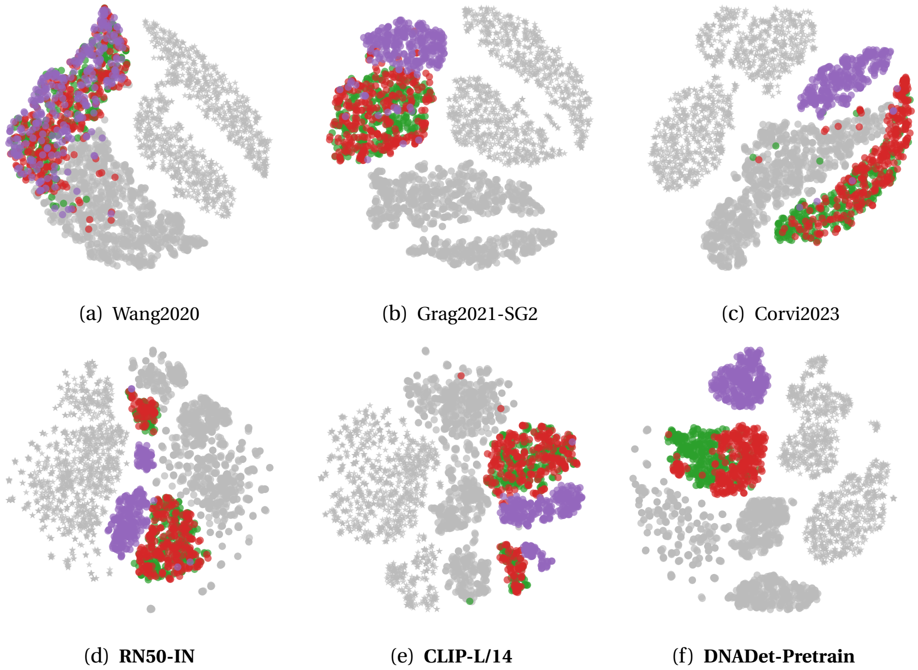
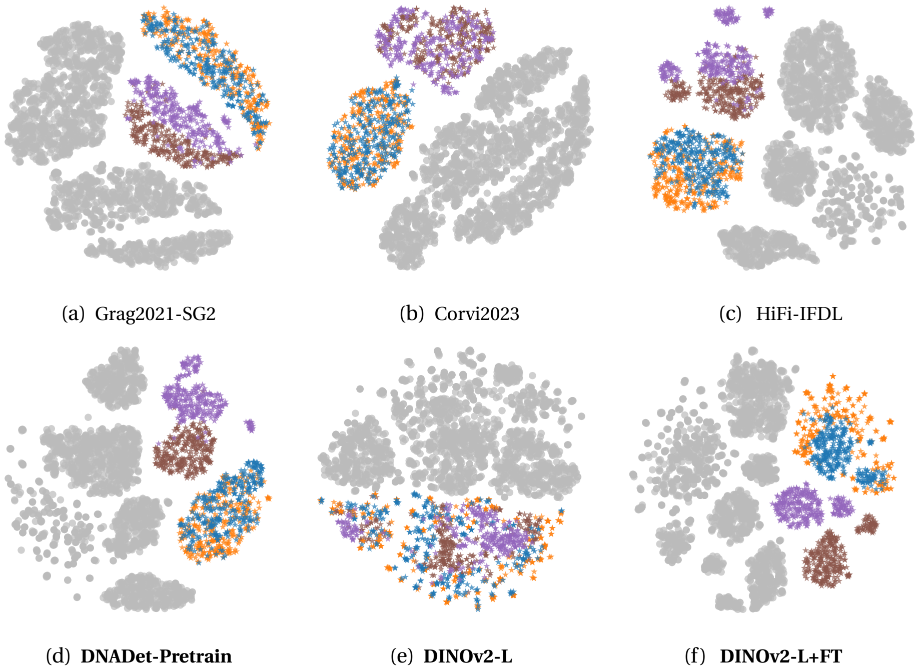


Figure 6.5: T-SNE projection of FRGC attacks highlighting the 3 GAN-based attacks (SG2-W in green, SG2-W+ in red and MIPGAN in purple). See also the legend in Figure 6.3.



(a) Grag2021-SG2 (b) Corvi2023 (c) HiFi-IFDL

(d) **DNADet-Pretrain** (e) **DINOv2-L** (f) **DINOv2-L+FT**

Figure 6.6: T-SNE projection of FRGC attacks highlighting the print-scan attacks (LB-Complete-PS in blue, LB-Combined-PS in orange, MIPGAN-PS in purple and MorDIFF-PS in brown). See also the legend in Figure 6.3.

each attack across several clusters. This is likely due to the need for these features to encode content-related information for their pretraining tasks, leading to additional separations based on content.

A consistent separation is observed between general families of attacks (landmark-based, GAN-based, diffusion-based), though finer separation within families is not always achieved. For GAN-based attacks (Figure 6.5), some attack-aware extractors such as Wang2020 (*6.5a*) struggle to separate the SG2-W, SG2-W+ and MIPGAN attacks, which are all based on the same StyleGAN2 generator. This could be because their pretraining task focuses on robust GAN-image detection, leading to a feature space that represents any GAN-image on the same manifold. This could suggest that pretrained deepfake detectors are a bad choice for an attribution task, however we see that the phenomenon is more complex, as Grag2021-SG2 (specifically trained for detection on StyleGAN2 images) actually manages to separate MIPGAN from other StyleGAN2-based morphs (*6.5b*), and so does Corvi2023 (*6.5c*). We note that as this latter is trained for detection solely on diffusion images, it is technically agnostic to GAN-based attacks which seems to actually *improve* its ability to separate MIPGAN attacks from other GAN-based ones. Interestingly, attack-agnostic extractors such as RN50-IN (*6.5d*) or CLIP-L/14 (*6.5e*) are also quite effective. Nevertheless, separation between SG2 and SG2-W+ is only achieved with DNADet and to some extent with HiFi-IFDL. This can be expected as those attacks are conceptually very close: they simply correspond to different ways of sampling the latent vectors that are fed into the synthesis network, meaning relevant artifacts linked to the synthesis itself should be similar in both cases. MIPGAN, in contrast, while using the same generator architecture, uses a different set of weights as it is fine-tuned on the source dataset, which makes separation from other GAN-morphs potentially easier.

For landmark-based attacks, attack-aware extractors do not seem to provide very good separation (e.g., Grag2021-SG2, *6.4a* and Corvi2023, *6.4b*). In contrast, attack-agnostic extractors again show better separated clusters (RN50-IN, *6.4c*), sometimes even very well (CLIP-L/14, *6.4d*).

In the print-scan domain, separation between deep and cheapfakes is relatively clear (Figure 6.6). Finer separation is however not always achieved, for instance with HiFi-IFDL (*6.6c*) and DNADet (*6.6d*) mixing together the two landmark-based attacks much more than in the digital domain.

### Post-fine-tuning analysis

We observe that DNADet extractors (Figures *6.3e* and *6.3j*) provide typically vastly better natural separation than most other attack-aware approaches. The clusters look very well localized and separated, even when considering attacks from the same family. In the print-scan domain, the MIPGAN and MorDIFF attacks are better separated than with other extractors, and the only remaining limitation is a struggle to separate the 2 landmark-based attack in this domain. This motivates the choice of reusing the DNADet contrastive learning objective for fine-tuning the extractors.

And indeed, we observe in the last row of Figure 6.3 that after fine-tuning, the extracted representations share similar properties to that of DNADet-Pretrain. In particular, they seem to provide the same quality of separability as HiFi-IFDL 6.3d, despite the latter having been trained on a wide variety of forged data. Yet, we still observe differences across the different fine-tuned extractors, for example with DINOv2-L+FT (Figure *6.3n*) showcasing better separation of the GAN attacks, as well as of the print-scan attacks, than other fine-tuned extractors. As seen in Figures 6.6f and 6.6d, DINOv2-L even compares favorably to DNADet for separating the print-scan version of landmark-based morphs. This suggests that while the fine-tuning process can still bring some additional discriminative power to the feature space, the initial separation capability of attack-agnostic extractors remains crucial even after fine-tuning.

All those observations are however only qualitative and limited to a very low-dimensional space. It is possible that natural feature separations between apparently confounded attacks are still present but not observed in a 2D T-SNE. To perform this analysis in a more quantitative way, we evaluate in the next section how well a higher dimensional unsupervised K-Means clustering of data samples actually matches a ground truth of attack labels.

### 6.2.2 Clustering

Table 6.3: V-Measure of the matching between unsupervised clusters learned using K-Means and the ground-truth attack labels, in function of the feature representation. Higher is better. A V-Measure of 0.0 means the cluster labeling and ground-truth are completely unrelated, a V-Measure of 1.0 means the cluster labels exactly match the ground truth labels. The number of clusters is set to the true number of clustered attacks, as the goal is to evaluate whether how well the unsupervised predicted clusters match the labeled ones. Results are reported on the evaluation set on each protocol. Attack-agnostic extractors are in bold, and fine-tuned extractors are indicated with the suffix "+FT".

| | FRLL | FRGC | FRGC D+PS | FRGC D→PS | FRGC→FRLL | FRLL→FRGC |
|---|---|---|---|---|---|---|
| Wang2020 | 0.25 | 0.25 | 0.37 | 0.24 | 0.26 | 0.21 |
| Grag2021-SG2 | 0.67 | 0.47 | 0.53 | 0.01 | 0.42 | 0.33 |
| Corvi2023 | 0.33 | 0.41 | 0.58 | 0.29 | 0.20 | 0.32 |
| DNADet-CelebA | 0.68 | 0.68 | 0.74 | 0.00 | 0.58 | **0.61** |
| HiFi-IFDL | **0.68** | 0.69 | 0.79 | 0.04 | **0.68** | 0.44 |
| **RN50-IN** | 0.43 | 0.43 | 0.46 | 0.02 | 0.24 | 0.23 |
| **RN50-IN+FT** | 0.62 | 0.69 | 0.84 | 0.00 | 0.60 | 0.57 |
| **ViT-B/16** | 0.05 | 0.06 | 0.35 | 0.04 | 0.02 | 0.01 |
| **ViT-B/16+FT** | 0.63 | 0.62 | 0.61 | 0.13 | 0.32 | 0.05 |
| **CLIP-L/14** | 0.41 | 0.54 | 0.53 | 0.00 | 0.35 | 0.38 |
| **CLIP-L/14+FT** | 0.66 | 0.74 | 0.82 | 0.00 | 0.55 | 0.57 |
| **DINOv2-L** | 0.36 | 0.05 | 0.31 | 0.03 | 0.03 | 0.05 |
| **DINOv2-L+FT** | 0.66 | 0.71 | 0.72 | 0.27 | 0.50 | 0.28 |
| **DNADet-Pretrain** | 0.68 | **0.81** | 0.83 | 0.16 | 0.39 | 0.21 |
| **DNADet-Pretrain+FT** | 0.65 | 0.76 | **0.84** | **0.40** | 0.47 | 0.34 |

We present in Table 6.3 the clustering performance results, following the experiment described in section 6.1.2 and the protocols listed in Table 6.1.

We see that DNADet-Pretrain is the best performing extractor on the baseline protocols, being either better than other approaches or equivalent (for example when compared to HiFi-IFDL for the FRLL protocol). Somewhat more surprisingly, it also shines on the FRGC D+PS protocol. Good performance on this latter protocol means that the unsupervised clustering process reasonably groups together print-scan and digital versions of the same attack, which is not trivial as we have seen they tend to be represented in distinct area of the feature space. In particular, under those settings, the fine-tuning of DNADet on CelebA-derived GAN images actually leads to *decreased* clustering performance (as seen in the DNADet-CelebA row). We hypothesize that during this fine-tuning, the extractor's features have to specialize for a finer representation of GAN images, which comes at the cost of discriminative power for non-

GAN attacks (landmark-based and diffusion-based). Other attack-agnostic extractors such as DINOv2-L+FT also perform well on those protocols, with performance comparable to the best attack-aware extractor (HiFi-IFDL). We note that is relatively remarkable, as HiFi-IFDL is aware of all types of considered broad family of attacks (GAN, Diffusion, manual editing), yet is not performing better than the best attack-agnostic extractors.

We also observe that attack-agnostic features seem much more sensitive to the source dataset. Indeed, in the source generalization protocols (FRGC→FRLL and FRLL→FRGC) the decrease in V-Measure is much stronger for attack-agnostic extractors than attack-aware ones, to the point where DNADet-CelebA ends up beating its Pretrain counterpart. This could be explained by the attack-agnostic feature being more sensitive to image content (which attack-aware extractors better learn to ignore), thus causing attacks from the unseen source dataset to be out-of-distribution with respect to the learned clustering.

On the print-scan generalization protocol (FRGC D→PS), performance is drastically degraded in almost all cases. As already observed on the T-SNE visualization, print-scan attacks tend to be clearly separated from digital attacks in all feature spaces, which means they would be out-of-distribution with respect to the learned clusters in this setting. We consider the variations in performance of the extractors in this case to be mostly random artifacts of the exact cluster structure, and assume they do not have much meaning. Finally, considering the effect of fine-tuning the feature extractor, we observe it leads almost always to an improvement in clustering performance. This matches our observations from the T-SNE visualization: before fine-tuning, the different attacks are separated, but dominant clusters do not relate directly to attacks. This is likely because the pretraining tasks typically will care more about the content of the image rather than to low level statistics which might be more salient for morph attribution. After fine-tuning however, we obtain a feature space which behaves more like DNADet's one, and this is reflected in the clustering performance. However, we do not go up to the performance of the base DNADet-Pretrain extractor, which suggests that the diversity of its pretraining data (not only faces) still ends up being beneficial overall. In practice, after fine-tuning, RN50-IN+FT and CLIP-L/14+FT are performing competitively with attack-aware extractors except HiFi-IFDL.

### 6.2.3 Attribution

We report in Table 6.4 the attribution accuracy of one-vs-rest linear SVMs multiclass classifiers trained on each feature space, according to the protocols specified in Table 6.1.

We focus first on baseline protocols (FRLL; FRGC), which are considered easiest as only digital attacks are considered, and the same source dataset is used for training and evaluation. Those baselines establish that strong attribution performance is definitely achievable in a constrained setting. Almost perfect accuracy is achieved with all extractors. This shows in particular that, despite most of the attack-agnostic feature spaces not providing a very "natural" separation of attacks (as showcased by the T-SNE and clustering experiments), *they*

Table 6.4: Attribution accuracy (in %) of a one-vs-rest multiclass linear SVM, for each protocol, and in function of the chosen feature representation. Results are reported on the evaluation set on each protocol. Attack-agnostic extractors are in bold, and fine-tuned extractors are indicated with the suffix "+FT".

| | FRLL | FRGC | FRGC D+PS | FRGC D→PS | FRGC→FRLL | FRLL→FRGC |
|---|---|---|---|---|---|---|
| *Random chance* | 20.00 | 16.67 | 16.67 | 25.00 | 20.00 | 20.00 |
| Wang2020 | 97.81 | 96.55 | 94.95 | 25.00 | 72.81 | 51.52 |
| Grag2021-SG2 | 99.74 | 99.54 | 99.25 | 25.00 | 54.56 | 55.31 |
| Corvi2023 | 97.54 | 98.32 | 97.59 | 25.00 | 65.18 | 57.40 |
| DNADet-CelebA | 99.91 | 99.84 | 98.44 | 23.32 | 40.00 | 24.10 |
| HiFi-IFDL | 99.30 | 99.01 | 97.04 | 25.00 | 70.00 | 35.27 |
| **RN50-IN** | 99.21 | 98.52 | 97.99 | 24.51 | 57.63 | 41.03 |
| **RN50-IN+FT** | 99.21 | 99.21 | 98.24 | 4.54 | 77.72 | **68.40** |
| **ViT-B/16** | 99.12 | 99.18 | 98.93 | 46.70 | 58.25 | 29.27 |
| **ViT-B/16+FT** | 99.82 | 99.97 | 99.66 | 29.04 | 55.70 | 36.21 |
| **CLIP-L/14** | 99.74 | 99.34 | 99.35 | 25.39 | 56.23 | 63.94 |
| **CLIP-L/14+FT** | 99.82 | 99.93 | 98.86 | 31.71 | 50.79 | 53.89 |
| **DINOv2-L** | 99.74 | 99.51 | 99.41 | **47.63** | 68.68 | 60.95 |
| **DINOv2-L+FT** | **100.00** | **100.00** | **99.90** | 29.49 | **77.89** | 58.50 |
| **DNADet-Pretrain** | 99.56 | 99.97 | 98.58 | 37.77 | 40.18 | 36.17 |
| **DNADet-Pretrain+FT** | 99.47 | 99.90 | 98.28 | 20.12 | 53.07 | 30.85 |

*all actually still enable linear separability of digital attacks.* This observation still holds when introducing print-scan attacks in both the training and evaluation data, as showcased by the FRGC D+PS protocol results. In this context, attack-agnostic extractors actually seem to perform marginally better to attack-aware ones. Moreover, in most cases, fine-tuning the extractors on face data does provide marginal improvements. The main exception is for DNADet, for which the fine-tuning leads to slight degradation in performance. As before, it suggests that the diversity of the pretraining data is beneficial for the generalization capability of the feature space.

One situation where attack-agnostic extractors really shine is for print-scan generalization, as showcased in the FRGC D→PS column. The protocol is challenging : it requires representations of print-scan versions of each attack to lie on the same side of the decision boundary learned using the digital version of the attacks. This is non-trival, given the significant alterations that the print-scanning can introduce into the images low-level statistics. For this reason, we observe overall relatively low accuracy. Nevertheless, a clear distinction can be made between attack-agnostic and attack-aware features. Linear SVMs trained on digital attacks using attack-aware features are absolutely incapable of generalizing to the print-scan counterpart of those attacks, performing no better than random chance. In contrast, while SVMs using attack-agnostic feature also struggle to generalize, they still showcase significantly better print-scan generalization capability overall. One most striking example is the degradation of print-scan generalization from DNADet-Pretrain to DNADet-CelebA : the fine-tuning of the extractor on a set of GAN-generated data seems to destroy any form of ability to relate digital and print-scan version of a similar attack. The possibility of this phenomenon occurring, likely because of an overspecialization of the feature space to brittle features in the synthetic data, was one main motivation for our approach, and really showcases the potential of attack-agnostic features for more robust morphing attack attribution. For this protocol, fine-tuning on face data actually leads to significant decrease of the performance. Intuitively, we have observed that the DNADet pretraining task also tends to separate digital and print-scan versions of the same attack, which is likely a drawback for this protocol. The refining of pretraining task aiming to correct for this issue should be investigated in future work.

Focusing now on the source generalization experiments (protocols FRGC→FRLL and FRLL →FRGC), we observe that the performance is somewhat more balanced between attack-aware and attack-agnostic features. Nevertheless, attack-agnostic extractors are often still performing better, notably DINOv2 and RN50-IN. Fine-tuning on face data leads to somewhat mitigated outcomes, with both improvements and degradations in performance depending on the situation. Even the best obtained attribution accuracy is not ideal. As mentioned in the discussion of the clustering experiment, we hypothesize that most attack-agnostic extractors are expected to be overly sensitive to image content, given the pretraining tasks, which could specifically challenge them for source dataset generalization. Indeed, FRLL and FRGC datasets are both quite constrained, but also quite different looking as showcased in Figure 2.7. Therefore, morphs coming from the unseen source dataset might lie out-of-distribution in the feature space with respect to the features seen at training time by the classifier. However, we note that for real-world application, such generalization is not fundamentally necessary. As passport pictures are typically very standardized, it should be feasible to simply assemble a source dataset that matches the particular considered standard.

Finally, as a general observation, we emphasize the strong performance of DINOv2-L in all protocols, in particular the base version (not fine-tuned on face data), which provides the most balanced performance overall, including on more challenging protocols.

## 6.3 Discussion

In this work, we have proposed an approach for closed-set attribution of morphing attacks that relies on the use of pretrained attack-agnostic features for data representation, and a simple linear classifier for the actual attribution task. We have showcased the potential of this approach as one providing better robustness guarantees in comparison to using attack-aware features, notably its stronger ability to associate print-scan morphs with their corresponding digital counterpart. While attack-agnostic features do not always provide a very natural regrouping of distinct attacks into separate clusters, they enable a linear separability of attacks which can be easily exploited by the classifier. Moreover, we observed that with the right pretraining task (specifically contrastive learning of real image transformations, proposed by [81] with DNADet), a fully attack-agnostic feature extractor showcases very good ability to represent distinct attacks in distinct clusters. While DNADet is originally developed for GAN image attribution, we see that their methodology still holds strong when including other types of attacks such as landmark-based and diffusion-based ones. This is further highlighted by using the same training objective to fine-tune the attack agnostic extractors on face data, leading in many cases to further performance improvements. In practice, we observed that a DINOv2-L extractor fine-tuned on bona fide face data with the DNADet contrastive learning task provide features which always outperform those extracted by a model like HiFi-IFDL, despite the latter having seen a wide variety of attack at pretraining time. Our approach is thus promising for future extension of this work to the open-set case, even accounting for the eventuality of yet another completely different morphing attack algorithm to appear, for example based on some novel deep generative modeling method.

We identified that digital to print-scan generalization of an attribution system remains a challenging task, even for our best performing models. However, we also showed that including print-scan morphs in the classifier's training data is enough to mitigate this issue. This is actually feasible in a realistic scenario, but typically not for *all* considered morphs, and using only a limited set of printer and scanners. Studying the print-scan generalization capability when only part of the training attacks are seen in print-scan format, and when facing unseen printers and scanners at evaluation stage, is an important research direction which is left for future work. Moreover, we observed that attack-agnostic features, in particular *not* fine-tuned, are more fit for print-scan generalization than attack-aware features. A future research direction could be to investigate whether improvements on the DNADet contrastive learning task could enable better generalization to the print-scan domain.

We also observed that source dataset generalization is somewhat challenging. We hypothesize that this is linked to the very constrained nature of the source datasets selected for morph creation in our experiments, and that the use of a more diverse source dataset for creating the training set attacks might significantly improve generalization capability. The verification of this hypothesis is left for future work. We do note that in practice, the ability of attribution systems to generalize to different source datasets might not necessarily be needed in a real-world scenario : passport pictures are usually quite constrained, and it should be feasible to just build a training dataset that follows those constraints. Nevertheless, this study stills helps to gain insight on the possible strengths and limitations of the proposed approach.

In parallel to a future extension to open-set cases, we also consider extending this work to perform joint detection and attribution of morphing attacks. We have indeed shown in Chapter 5 the effectiveness of attack-agnostic features for morphing attack detection, which hints that detection and attribution could be performed jointly in the same representation space.

Finally, a natural extension of this work is to not only train a linear probe on top of pre-extracted features, but to even fine-tune the full extractor specifically for the morph attribution task. Our experiments (notably the degradation in performance switching from DNADet-Pretrain to DNADet-CelebA) hint that this might lead to an overspecialization of the learned representation to the attacks seen during fine-tuning, thus potentially reducing generalization capabilities. A thorough study, considering all permutations of fine-tuning and probing attacks is necessary, and is left for future work as it might in particular necessitate the gathering of a wider set of attacks.

# 7 Conclusion

## 7.1 Takeaways

### 7.1.1 Summary of findings

**On the impact of deepfake technology on the issue of morphing attacks**

In Chapter 3, we observed that morphing attacks based on deep generative models are becoming increasingly effective, closing the gap with landmark-based attacks, which remain the most problematic. We demonstrated the direct application of template inversion literature to morphing attack generation through the inversion of the optimal morph embedding. While these generated attacks are highly effective at deceiving face recognition systems, they lack realism and are too unconstrained to be used as actual passport photos, limiting the real-world feasibility of inversion-based morphing attacks. However, we demonstrated that grounding the morph creation process on more solid theoretical guarantees, by performing latent interpolation directly in the embedding space of a face recognition network, leads to morphing attacks that are particularly effective on automated FRSs. Notably, these generated attacks remain highly effective even when targeting FRSs different from the one used to model the identity space.

Face template inversion models can be systematically trained by learning a mapping between the embedding space of a face recognition network and the latent space of a deep synthetic face generator. Each time a novel face generator is developed, applying it to inversion morphing should be relatively straightforward. The only missing ingredient to generate passport-like inversion morphs is increased control over the face generator, enabling conditioning on pose and expression to ensure a neutral face. This type of control is generally of interest in synthetic face generation literature, and it is likely that such capability will emerge in the near future, increasing the risk posed by inversion-based morphing. This illustrates a general trickle-down phenomenon in the field of generative AI: by nature, deep generators are relatively straightforward to tune or adapt for new objectives. Therefore, each improvement in face generative models can directly lead to an application in face template inversion, subsequently leading to an application in morphing attack generation through optimal embedding inversion. We can thus expect the landscape of morphing attack generation algorithms to evolve at a similar pace to that of generative AI, which is currently extremely fast.

This chapter also more generally demonstrates the growing importance of deep morphing in the context of FRSs security, by showing novel morphs can deceive FRSs to a problematic extent. It underscores the need to improve morphing attack detectors to handle the new reality of deep morphs. Additionally, it highlights the increasing diversity of morphing attacks, which increases the challenges in detection, but also adds value to attribution systems. Indeed, as attacks diversify, observing the choice of a *specific* morphing algorithm can become a more expressive indicator of the author of the morph.

### On morphing attack detection

In Chapter 4, we compared the use of handcrafted features and supervised learning approaches in the context of GAN-based morph detection. We observed that approaches based on Local Binary Patterns, which aim to extract texture information from images and have been previously used for landmark-based morph detection, do not achieve strong detection performance on GAN-based morphs. In contrast, supervised learning approaches perform much better, with the direct transfer of a general GAN-image detection model being the best performing approach. This suggests that transfer learning from preexisting models can be an effective way to approach morphing attack detection. However, as these observations were made in the digital domain, evaluation in the print-scan domain is necessary to gain a better understanding of the real-world performance of the detector.

We also discussed the interpretability of the considered methods. Handcrafted features have the advantage of being explainable by design and enable the visualization of saliency maps that highlight the most important regions of the image for classification decisions. Nevertheless, while these maps look visually compelling, they are not necessarily very informative overall, particularly as they do not help in understanding failure cases. The use of Grad-CAM also enables the generation of saliency maps for detectors based on supervised features. However, the resulting maps remain difficult to interpret. Moreover, we noted that if the salient signal in deep morphs is localized in the frequency domain (the so-called GAN-signature), this signal will not be localized in the image domain, implying that localized saliency maps might not be conceptually valid.

In Chapter 5, we demonstrated the effectiveness of attack-agnostic features extracted from vision foundation models for morphing attack detection. In a supervised setting, we highlighted that MAD systems developed by training simple probes on top of these features generally exhibit better generalization capabilities than supervised approaches relying on end-to-end CNN training. This was observed in the context of generalization to unseen attacks, novel source datasets, and from the digital domain to the print-scan domain.

In an anomaly detection setting, foundational features also enable effective detection. However, it is important to note that this specific analysis was restricted to the digital domain, and further investigation in one-class settings is required to strengthen this claim.

This study provides additional insights into the use of foundational features for synthetic image detection. It confirms previous observations made on CLIP models and strengthens them by showing their preserved relevance when introducing cheapfakes into the problem. Moreover, the evaluation of additional vision foundation models such as DINOv2 and AIM shows that CLIP is not the only capable model for synthetic image detection. However, we observed that while CLIP features tend to perform consistently across various evaluation scenarios, other representations show more variable effectiveness, sometimes outperforming CLIP by a wide margin and other times performing poorly, depending on the scenario.

The effectiveness of DNADet in one-class modeling is also noteworthy, as this model has several advantages over other attack-agnostic extractors. Notably, its model size is orders of magnitude smaller, and it requires much less training data. Its strengths likely stem from a very effective pretraining task. This makes DNADet an interesting model for further MAD experimentation, as it is feasible to retrain and adapt within the resources of an academic research lab, unlike other models that may require extensive data and computation typically available only in industry settings.

Another important takeaway from this study is the observation that generalization to novel source datasets appears to be at least as challenging, if not more so, than generalization to unseen attacks. This suggests that in generalization studies, when training then testing on two completely distinct datasets using both different sources and different attacks (as is commonly the case in video deepfake detection literature), one cannot necessarily assume that performance loss is caused by novel attacks rather than by out-of-distribution source samples. Ideally, an independent analysis of both generalization challenges would be useful to fully understand the exact shortcomings of the proposed detectors.

In the context of MAD, we have shown that the challenge of generalization to novel source datasets can be mitigated through the use of a more diverse source dataset, even when using relatively carelessly generated and not highly realistic morphs. Using this type of data is necessary to develop MAD solutions capable of targeting a plurality of passport standards. However, we also note that real-world solutions targeting a single passport standard could potentially be sufficient, and hence it might not be necessary to create MAD systems that generalize to novel source datasets.

### On the source attribution of morphing attacks

In Chapter 6, we developed a deeper understanding of the structure of attack-agnostic feature spaces by evaluating their benefits for training downstream morphing attribution models. We gained insights into how these representations cluster (or do not cluster) different attacks through T-SNE visualizations for qualitative analysis and quantitatively by assessing how well natural clusters correspond to different groups of attacks.

We then evaluated the effectiveness of the considered feature spaces in a closed-set attribution setting. This highlighted that even for attack-agnostic extractors that do not naturally cluster samples per attack, there is typically linear separability of the different attacks, enabling effective closed-set attribution systems.

Regarding the best-performing representations, we observed that the DNADet pretraining process generally led to representations that cluster attacks well. However, a significant drawback was that print-scan versions of attacks were completely separated from their digital counterparts. These observations were consistent both for the pretrained DNADet model itself and when using the DNADet objective for fine-tuning other attack-agnostic extractors on a dataset of bona fide faces. Fine-tuning typically improved the representations in terms of attack clustering ability and attribution performance, except notably in the scenario of digital to print-scan domain generalization. Importantly, after fine-tuning, all considered representations were still not equivalent, suggesting that the initial representations before fine-tuning matter significantly.

In practice, we observed that DINOv2 representations, both before and after fine-tuning, generally led to the best closed-set attribution performance. Moreover, before fine-tuning, this extractor also enabled the best digital to print-scan generalization of the attribution system. This aligns with similar observations made in Chapter 5 in the context of detection.

The significance of this work is partial, as extending it to open-set scenarios would be crucial for implementing attribution systems in real-world settings. Currently, the state of research does not enable to offer operational systems to the justic system. Additionally, for forensic applications, developing likelihood-ratio-based attribution systems will be necessary. We highlight future research directions on these aspects later in this section. However, establishing these initial closed-set morphing attack attribution systems is a crucial step towards addressing more realistic scenarios. It demonstrates the feasibility of discriminating between different algorithms, which is not guaranteed *a priori*. Furthermore, achieving this using attack-agnostic representations is a promising outcome, as these representations are more likely to remain valid and useful for representing samples produced by unseen attacks in open-set scenarios, compared to supervised representations.

### 7.1.2 Data representations in morphing attack detection

It remains to highlight elements of answers to our central research question :

> *What are the most effective representations of image data, satisfying requirements of discriminative power, generality, robustness, and interpretability, in the context of face morphing attack detection and attribution?*

### Discriminative power

In terms of discriminative power, we observed in Chapter 4 significant limitations of handcrafted features, specifically LBPs. Despite their effectiveness in landmark-based morph detection, they did not perform satisfactorily on GAN-based morphs. While exploring alternative feature designs might improve performance, the rapid evolution of deep generative AI algorithms makes it unsustainable to rely heavily on handcrafted features, which require slow experimentation process, doomed to fall behind generative advancements.

Chapters 4 and 5 confirmed that supervised learning generally leads to highly effective detectors in simple evaluation scenarios, though this methodology has limitations in generality and robustness, which will be discussed later. We also demonstrated that foundational features, despite not using synthetic data during representation learning, are surprisingly effective for both morphing attack detection and attribution, in particular being applicable to represent both cheapfakes and deepfakes.

The discriminative power of attack-agnostic features shown in chapters 5 and 6 is however a relatively novel finding, previously only observed in parallel works focusing on CLIP features for deepfake detection. Our experiments show that this phenomenon extends beyond CLIP to various pretrained representations, some of which even offer advantages over CLIP in specific scenarios.

In conclusion, handcrafted features lack discriminative power, while supervised and foundational features are effective in simple evaluation scenarios.

### Generality

In terms of generality, Chapters 5 and 6 demonstrated the advantages of foundational features over supervised ones. While both types of representations perform similarly in simple scenarios, foundational features are advantageous in generalization scenarios.

For generalization to unseen attacks, features from AIM and CLIP show comparable performance to supervised features on morphs from constrained source datasets (FRGC), but outperform them on morphs from the more diverse FFHQ dataset. For generalization to novel source datasets, attack-agnostic features are much more practical: CLIP and DINOv2 features, and to a lesser extent AIM features, significantly outperform supervised ones. Moreover, AIM and DNADet features benefit greatly from training on the more diverse FFHQ morphs, enabling derived models to generalize well to FRLL. However, source dataset generalization in the context of attribution remains challenging, with no method proving itself particularly effective.

In one-class scenarios, AIM and DNADet features are particularly appropriate, while CLIP remains competitive with reference supervised methods. Additionally, DNADet features provide a well-structured space that separates attacks into distinct clusters. Thus, approaches based on DNADet seem promising for the most challenging scenarios of one-class detection and open-set attribution, although they suffer from an important robustness limitation, which will be discussed later.

Why are supervised representations less appropriate than foundational ones for generalization? We propose the following intuition: first, as discussed in Chapter 1, deep synthetic images contain very salient signatures, which, while difficult to extract manually from single images, become visible when averaging spectra of multiple images. Second, supervised DNNs excel at picking up patterns in the training data. Third, while datasets used for training supervised detectors are large, they are "depth-large" and not "breadth-large", meaning they contain many samples per class but an overall low number of classes. Therefore, many training samples come from the exact same generation algorithm. They all contain similar GAN signatures, which that the detector can easily pick up on during training but are not general. To some extent, those can be considered spurious correlations between the samples, because while useful for this particular set of attacks, they are neither general nor robust. This is analogous to training an object recognition network where all representatives from each class have been captured in the same location using the same camera, making it easy for the network to exploit location and camera correlations for effective detection, but those correlations are not helpful for generalization.

This phenomenon can probably be mitigated through augmentation, as done in works such as [69], but it seems to not be enough. In contrast and by design, foundational features *cannot* exploit spurious correlations. Therefore, provided they enable good base discriminative power, they will also provide better performance in generalization scenarios.

In conclusion, foundational features showcase better generalization properties than supervised ones, although the best choice of attack-agnostic representation depends on the generalization challenge. Nevertheless, CLIP features are performing consistently across all scenarios, which makes them promising for the developement of generalized MAD solutions.

### Robustness

In terms of robustness, we focused specifically on print-scan post-processing, which is crucial in the context of morphing attack forensic analysis. For detection, we observed that most attack-agnostic extractors lead to higher robustness to print-scan degradation compared to the reference supervised CNN. DINOv2 features, in particular, demonstrated the best robustness, followed by CLIP. A notable exception is DNADet-based models. The explanation for this phenomenon becomes clear when examining the T-SNE visualizations from Chapter 6: as a byproduct of the pretraining task, print-scan versions of attacks are grouped in completely separate clusters from their digital counterparts. In other words, while the pretraining task effectively separates different attacks, it goes "too far", resulting in the isolation of print-scan and digital versions of the same image from each other.

In the context of attribution, we also observed that DINOv2 features are the most suitable for generalization from digital to print-scan. However, this robustness capability significantly degrades when fine-tuning the extractor using the DNADet pretraining task, aligning with our previous observations.

We hypothesize that the robustness of DINOv2 and CLIP to print-scan post-processing is due to their focus on semantic traces, which are not drastically altered by the print-scan process, rather than on statistical ones. Initial works on DINOv2 have shown that the attention maps within the model are effective image segmenters, indicating that the model learns to identify objects and high-level concepts in images. The same applies to CLIP, which necessarily focuses on semantic aspects, given its goal of finding a common representation for paired images and captions.

In contrast, an extractor like AIM, which learns the statistical distribution of images, is naturally more likely to express a lot of statistical information that can be altered by print-scanning. Similarly, DNADet, given its pretraining task, specifically aims to ignore content information and focus on degradation-related information, which mostly affects statistical image properties.

This reflection also aligns with our assumption in the generality analysis that supervised models tend to focus on spurious signature-like correlations, which are typically degraded by print-scanning, thus leading to a low robustness to print-scan degradation for supervised models.

In conclusion, we observe an overall advantage of foundational features over supervised ones in terms of robustness to print-scan post-processing. Additionally, we highlight that such robustness can be improved to some extent by focusing on representations that rather focus on semantic traces.

### Interpretability

Our analysis of interpretability was relatively limited and focused on handcrafted and supervised features in Chapter 4. We can highlight two insights from this study. First, saliency map visualization methods such as Grad-CAM might be impractical for synthetic image detection, assuming that the salient traces are statistical rather than semantic. Indeed, saliency map visualization relies on the assumption of locality of the salient information, which is incompatible with statistical traces that are localized in the frequency domain hence not in the image domain. Second, while handcrafted features are typically presented as more interpretable, the extent of their explainability can be debated. As illustrated in our work, while LBP features enable straightforward explanations by identifying the most important textural patterns for detection and localizing the corresponding regions, such visualizations were not particularly helpful in understanding the detector's failure cases.

A study and discussion of the interpretability of foundational features would be necessary for a complete analysis. Intuitively, we expect saliency map visualization to suffer from the same issues as in the context of supervised features (lack of locality of the information). However, it is possible that for representations focusing more on semantic content, such as DINOv2 or CLIP, additional insights could be obtained by computing Grad-CAM visualizations on the morph images or by visualizing the internal attention maps.

We also reflect that a significant part of the need for interpretability arises from concerns that the model might accidentally exploit irrelevant cues from the input data in its final decision. In the context of foundational features, this concern is somewhat mitigated by the demonstrated generality of these representations, as they have been successfully adapted to a wide variety of downstream tasks. However, another crucial need for interpretability is to provide human-level explanations when a model is used in an evaluative forensic setting. Currently, it appears that the field of deep learning is not yet mature enough for such settings and should be limited to investigative purposes. Progress in explainable AI is essential to enable the use of deep learning models in evaluative contexts.

## 7.2 Future directions

### Open-set attribution

A natural extension of this thesis would be towards open-set source attribution of morphing attacks. As observed in Chapter 6, the considered attack-agnostic extractors tend to separate distinct attacks relatively well. This should enable the development of **detectors of novel attacks** by modeling the feature distribution of samples from known attacks (for example, using a GMM as for the one-class detector from Chapter 5) and identifying novel attacks as out-of-distribution samples.

More interestingly, the T-SNE visualizations (particularly Figure 6.3) suggest that some of the considered representations reflect the ground truth hierarchical relationships between different attacks. This is especially true for DNADet representations and all representations after fine-tuning using the DNADet pretraining task. For instance, attacks based on StyleGAN interpolation (SG2-W and SG2-W+) are represented close to each other and are also relatively close to MIPGAN samples, which are also StyleGAN-based. Similarly, landmark-based attacks group together to form a higher-level “landmark morphs” cluster. This phenomenon is also present in the print-scan domain, where landmark-based methods form one higher-level cluster, and deep morph methods (MIPGAN and MorDIFF) form another.

While T-SNE visualizations should not be overly interpreted and only provide qualitative insights, this phenomenon hints that it should be feasible to find an appropriate distance metric in the DNADet feature space. This metric could express the level of similarity between two samples and potentially be used for **source linking**. If the intra-class distance between samples generated using the same method is typically much lower than the inter-class distance between samples generated using distinct methods, a decision threshold could be determined to separate pairs of samples from similar sources and those from dissimilar sources.

We also want to emphasize that the proposed closed-set attribution can already have real-world applications. One such application, mentioned in Chapter 1, is verifying whether a public generator's restrictive license has been infringed. If synthetic images that resemble the generator outputs appeared to be used in an unauthorized context, an investigation into potential license infringement would need to determine whether those images were indeed produced by the licensed generator, or by another generator as would likely be claimed by the creator of the images for their defense.

In the specific context of morphing attacks, closed-set attribution can be used to track trends in real-world occurrences of morphing attacks. In practice, the number of different algorithm families remains quite limited (landmark-based, GAN-based, diffusion-based, and possibly inversion-based). A high-level attribution system that differentiates between these four families could help maintain an overview of the dominant types of algorithms in use. Moreover, closed-set attribution could potentially identify series of morphs that appear to come from different sources, possibly indicating distinct organizations.

Indeed, even though novel algorithms may always be encountered, it is reasonable to assume that most attacks will match one of the broad families of publicly described algorithms. Thus, partial information can already be gained through high-level closed-set attribution systems, especially for forensic intelligence purposes, and even though a proper open-set attribution system would be preferable.

Extending to open-set attribution might notably require to establish more concrete application domains, as these will dictate the necessary level of attribution. The creation and publication of large-scale deepfake datasets specifically designed for attribution tasks could also be valuable. Existing datasets are usually designed for detection, contain a limited number of attribution classes, and do not always describe the generation process in enough detail for low-level attribution settings.

### Exploiting semantic traces for increased robustness

In Chapter 5, we observed that attack-agnostic extractors such as CLIP and DINOv2 lead to stronger inherent robustness of the downstream detector to the print-scan process. We hypothesized that this is due to these representations naturally emphasizing semantic content as a byproduct of their pretraining tasks, and this semantic content is better preserved during print-scanning compared to statistical traces.

This inherent robustness is an advantage over supervised features, whose robustness can only be partially enforced through train-time data augmentation, a methodology that is not very practical in the context of print-scan robustness.

Further work should first confirm this hypothesis of reliance on semantic traces by CLIP and DINO-based detectors through the visualization of saliency maps, such as the GradCAM visualization from Chapter 4, and by investigating internal attention maps. The original article on DINO [86] showcases that these attention maps effectively behave as image segmentation systems, highlighting specific regions of interest in the input image.

It would then be interesting to consider a simple downstream detector based on the fusion of a statistical-oriented representation (e.g., AIM or DNADet) with a semantic-oriented representation (e.g., DINO or CLIP). The former has typically been more effective in most evaluation scenarios but suffers from poor robustness to print-scanning. A fusion might enable a model that benefits from the strengths of both types of features, i.e., good discriminative power and increased robustness to print-scanning.

### Extending one-class detectors to the print-scan domain

In Chapter 5, we demonstrated that a one-class detector based on DNADet features, targeting a single passport format (i.e., single source dataset), performed effectively in the digital domain. However, we expect that this detector will classify print-scan bona fide samples as attacks. This issue arises because DNADet representations separate not only different attacks but also different domains (digital and print-scan), as illustrated in the T-SNE visualization from Chapter 6.

Given the well-structured representation space, it seems feasible to extend this one-class detector to the print-scan domain by including print-scanned bona fide samples in the training data. This inclusion should enable the detector to perform well in both domains.

The open questions for this direction are twofold. First, how many print-scan bona fide samples are necessary for proper modeling of their feature distribution? The print-scan process is relatively time-consuming, which could limit the quantity of training data. However, since no print-scan attack samples are required for training, the overall amount of necessary print-scan data might remain manageable.

Second, will the resulting model generalize well to samples printed and scanned using different hardware than that seen during training? If generalization is poor, it suggests that a variety of printers and scanners should be used to generate the training bona fide print-scan data, which could pose a challenge.

We also highlight a recent work on simulating print-scan textures [145], which could greatly facilitate the generation of sufficient amounts of bona fide training samples. This approach could benefit not only our one-class model but also the data augmentation of supervised models. Therefore, print-scan robustness might be globally improved through the increasing availability of simulated print-scan training data.

## 7.3 Effective countermeasures

The work in this thesis demonstrates some promising approaches for improving technical countermeasures to the phenomenon of deepfakes, especially in terms of generality.

However, the problem is far from being solved due to remaining issues in robustness, limited interpretability, overall performance that does not reach acceptable thresholds for real-world implementations, and significant open questions about how future deepfake generators will continue to challenge current methods. It is unlikely that a single "universal" detector or a comprehensive open-set source attribution system will ever be developed, given the constantly evolving landscape of generation techniques. Moreover, even if such a universal detector were achievable, it would become an easy target for generators to optimize against, by adding the goal of fooling this universal detector as an supplementary objective during generator training.

Therefore, it is important to emphasize that deepfake-related issues can only be successfully addressed through a combination of approaches that go beyond technical solutions alone. Only a multifaceted approach can provide sufficient resilience against deepfake technology. We highlight three key areas of focus: pragmatic technical solutions, education, and standardization.

#### Resilience through pragmatic technical solutions

While a universal deepfake detector is overly idealistic, a more pragmatic approach focuses on systems that can handle the majority of scenarios encountered on online platforms.

Instead of solely maximizing the generalization capability of detectors, a more practical solution is to monitor all major families of deepfake generators and develop a large, diverse dataset representative of the deepfakes likely to appear online. This dataset can then be used to train supervised detectors that, although they may not generalize beyond the tracked families, can effectively identify most deepfake content in real-world settings. Currently, a centralization phenomenon exists where most online AI-generated content is produced by a few popular generators; typically either commercial ones (DALL-E, Flux, ChatGPT) or widely used open-source models (Stable Diffusion, Llama). Detectors designed to accurately detect the output from these well-known models would be highly beneficial, for example, in the automated flagging of deepfake media on social networks. Importantly, such detectors should be updated regularly to include emerging generators through retraining or continuous training.

Additionally, maintaining multiple detectors based on orthogonal methodologies could enhance the resilience against potential adversarial attacks. Relying on a single detector poses a security risk, as it allows adversaries to optimize a novel generator with the goal of specifically bypassing that detector. In contrast, using a diverse collection of detectors makes adversarial training considerably more complex and less worthwhile. Thus, by leveraging a wide range of semantic and statistical traces (including generator signatures, inconsistent eye blinking, audiovisual discrepancies, metadata anomalies, content-metadata inconsistencies, various foundational representations, . . . ), the challenge for adversaries becomes prohibitively complex.

### Resilience through education and training

Communication channels should be established to ensure that individuals involved with deepfake media, such as personnel from media organizations, legal professionals, police and border control officers, or governmental office workers, stay up to date with technological advancements.

Specifically, they should be informed about current developments in generative AI to better understand the capabilities and limitations of existing generators. Similarly, they should gain a good understanding of the strengths and weaknesses of existing deepfake detectors to avoid overestimating the value of a detector's prediction.

They could also be trained to recognize specific types of common manipulations when applicable. For morphing attacks, paying attention to potential ghosting artifacts in landmark-based morphs or overly plastic textures in deep morphs can help identify manipulations. The perceptual study performed in [37] shows that evaluators familiar with morphing attack technology are more accurate at detecting morphs.

For deepfakes, common semantic flaws described in Chapter 1 could be listed, potentially leading to the creation of a basic checklist for media evaluators to use as a first pass.

If specific detectors targeting common manipulations exist, they could be used in a second pass for automated testing of the analyzed media. In such cases, it would be valuable to develop public software accessible to non-technical users to automatically run these tests.

The flow of information should ideally be multidirectional: academic institutions and industry actors could provide insights on technological progress in generation and detection. Agencies facing morphing attacks or deepfake manipulations in the field could also offer feedback on their concrete observations, sharing information among themselves and back to academia to better inform technological mitigation research.

This sharing of information can be challenging due to data privacy concerns, particularly with passport information. Recent research [146] suggests using federated learning for morphing attack detection, which enables training a singular detection model in a decentralized manner by multiple agencies and research labs through the sharing of model weight updates rather than sensitive biometric data. While still in its infancy, this approach could facilitate information sharing in a privacy-compatible way.

This entire process would need regular updates to account for ongoing advancements in both deepfake generation and detection fields.

### Resilience through standardization

Standards and protocols must be established to prevent unnoticed tampering with sensitive media. Specifically for morphing attacks, this involves capturing passport photographs in controlled governmental settings. In Switzerland, most cantons have already implemented this process, suggesting that a nationwide extension could be achieved within a reasonable timeframe. However, for maximum effectiveness, these standards should be developed at an international level, or at least at the European level as a first step, where many countries are currently lagging.

More generally, in the context of deepfakes, initiatives such as the C2PA project or the JPEG Trust standard (presented in Chapter 1) offer robust approaches. Ensuring complete and secure traceability of digital media would significantly simplify the detection of suspicious online content.

Establishing and implementing these standards will be a long-term process. But as current trends in generative AI indicate that the variety and prevalence of deepfake media will continue to rise, proper standardization will drastically increase resilience even against future, unforeseen developments in deepfake technology. Standardization efforts should thus be made a priority, while in the meantime other mentioned solutions, both technical and non-technical, should be used as temporary mitigations against harmful uses of deepfakes.

# A Appendix

## A.1 Published work

1. L. Colbois and S. Marcel, “Evaluating the Effectiveness of Attack-Agnostic Features for Morphing Attack Detection”, in *2024 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2024, pp. 1–9. DOI: 10.1109/IJCB62174.2024.10744532. [Online]. Available: https://ieeexplore.ieee.org/document/10744532 (visited on 02/05/2025)

2. L. Colbois, H. O. Shahreza, and S. Marcel, “Approximating Optimal Morphing Attacks using Template Inversion”, in *2023 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2023, pp. 1–9. DOI: 10.1109/IJCB57857.2023.10448752. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10448752 (visited on 04/08/2024)

3. E. Sarkar, P. Korshunov, L. Colbois, *et al.*, “Are GAN-based morphs threatening face recognition?”, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 2959–2963. DOI: 10.1109/ICASSP43922.2022.9746477. [Online]. Available: https://ieeexplore.ieee.org/document/9746477 (visited on 08/27/2024)

4. L. Colbois and S. Marcel, “On the detection of morphing attacks generated by GANs”, in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2022, pp. 1–5. DOI: 10.1109/BIOSIG55365.2022.9897046. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9897046 (visited on 02/03/2024)

5. L. Colbois, T. de Freitas Pereira, and S. Marcel, “On the use of automatically generated synthetic image datasets for benchmarking face recognition”, in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Aug. 2021, pp. 1–8. DOI: 10.1109/IJCB52358.2021.9484363. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9484363 (visited on 04/08/2024)

### Under review

1. L. Colbois and S. Marcel, “Generator attribution of morphing attacks using attack-agnostic feature representations”, Under review at IEEE Access, Apr. 2024

2. H. Otroshi Shahreza, L. Colbois, and S. Marcel, “On the Generation of Face Morphs by Inversion of Optimal Morph Embeddings”, Under review at IEEE Transactions in Information Forensics and Security (TIFS), Sep. 2024

## A.2 Given talks

1. L. Colbois, “Morphing de visage par hypertrucage”, Outreach talk, presented at the IA Workshop from Entente Forensique Francophone, Apr. 29, 2024

2. L. Colbois, “Generator Attribution of Morphing Attacks using Attack-Agnostic Feature Representations”, Internal institute-wide talk (Idiap Research Institute), Apr. 16, 2024

3. L. Colbois, “Attack-agnostic feature representations in morphing attack attribution”, Visiting talk (GRIP Lab, University Federico II of Naples), Feb. 22, 2024

4. L. Colbois, “Deepfake detection”, Outreach talk, presented at the Meeting between Argovian Police and the School of Criminal Sciences (University of Lausanne (UNIL)), Dec. 7, 2023

5. L. Colbois and H. Otroshi Shahreza, “Approximating optimal morphing attacks using template inversion”, Poster, presented at the 2023 IEEE International Joint Conference on Biometrics (IJCB), Sep. 2023

6. L. Colbois, “Morphing attacks through optimal embedding inversion : Vulnerability analysis”, Poster, presented at the UNIL ESC Summer School 2023, Aug. 2023

7. L. Colbois, “Deepfake attribution”, presented at the European Association for Biometrics (EAB) & Center for Identification Technology Research (CITeR) Biometrics Workshop, May 30, 2023

8. L. Colbois, “Deepfakes : Technologie et enjeux”, Outreach talk, presented at the Cycle de Conférences Grand Public "L’IA Changera-t-Elle Nos Vies?" - Deepfakes : Un Monde de Faux-Semblants (UNIL/CHUV Musée de la main), Oct. 11, 2022

9. L. Colbois, “Introduction to the forensic framework - with a face recognition example”, Internal institute-wide talk (Idiap Research Institute), Sep. 21, 2022

10. L. Colbois, “On the detection of morphing attacks generated by GANs”, presented at the 2022 International Conference of the Biometrics Special Interest Group (BIOSIG), Aug. 31, 2022

11. L. Colbois, “Towards reliable and interpretable detection of face deepfakes”, presented at the UNIL ESC Summer School 2022, Aug. 22, 2022

12. L. Colbois, “On the use of automatically generated synthetic image datasets for benchmarking face recognition”, Video Capsule, presented at the UNIL ESC Summer School 2021, Aug. 2021

13. L. Colbois, “On the use of automatically generated synthetic image datasets for benchmarking face recognition”, presented at the 2021 IEEE International Joint Conference on Biometrics (IJCB), Aug. 6, 2021

## A.3 Published software

Publications mentioned in Appendix A.1 are all accompanied by a software package for reproducibility purposes, which is linked in the article. We highlight the following packages, focused on morphing attack generation, vulnerability evaluation, and detection:

1. Generation of morphing attacks :  
   https://gitlab.idiap.ch/biometric/morphgen
2. Morphing attack vulnerability analysis :  
   https://gitlab.idiap.ch/bob/bob.paper.ijcb2023_inversion_morphing
3. Morphing attack detection using attack-agnostic representations :  
   https://https://gitlab.idiap.ch/bob/bob.paper.ijcb2024_agnostic_features_mad

For additional links and guides to reproducibility of the work presented in this thesis:  
https://gitlab.idiap.ch/bob/bob.thesis.lcolbois

# Bibliography

[1] C. Roux, R. Bucht, F. Crispino, *et al.*, "The Sydney declaration – Revisiting the essence of forensic science through its fundamental principles", *Forensic Science International*, vol. 332, p. 111 182, Mar. 1, 2022, ISSN: 0379-0738. DOI: 10.1016/j.forsciint.2022.111182. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379073822000123 (visited on 12/08/2024).

[2] D.-O. Jaquet-Chiffelle and E. Casey, "A formalized model of the Trace", *Forensic Science International*, vol. 327, p. 110 941, Oct. 1, 2021, ISSN: 0379-0738. DOI: 10.1016/j.forsciint.2021.110941. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379073821002619 (visited on 04/11/2025).

[3] O. Ribaux, "Police scientifique : le renseignement par la trace", in *Police scientifique le renseignement par la trace*, ser. Collection sciences forensiques, Lausanne: PPUR, Presses polytechniques et universitaires romandes, 2014, ISBN: 978-2-88915-061-8.

[4] L. Colbois, H. O. Shahreza, and S. Marcel, "Approximating Optimal Morphing Attacks using Template Inversion", in *2023 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2023, pp. 1–9. DOI: 10.1109/IJCB57857.2023.10448752. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10448752 (visited on 04/08/2024).

[5] L. Colbois and S. Marcel, "On the detection of morphing attacks generated by GANs", in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2022, pp. 1–5. DOI: 10.1109/BIOSIG55365.2022.9897046. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9897046 (visited on 02/03/2024).

[6] L. Colbois and S. Marcel, "Evaluating the Effectiveness of Attack-Agnostic Features for Morphing Attack Detection", in *2024 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2024, pp. 1–9. DOI: 10.1109/IJCB62174.2024.10744532. [Online]. Available: https://ieeexplore.ieee.org/document/10744532 (visited on 02/05/2025).

[7] C. Cortes and V. Vapnik, "Support-vector networks", *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1, 1995, ISSN: 1573-0565. DOI: 10.1007/BF00994018. [Online]. Available: https://doi.org/10.1007/BF00994018 (visited on 11/19/2024).

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", in *Proceedings of the 25th International Conference on*

*Neural Information Processing Systems - Volume 1*, ser. NIPS'12, Red Hook, NY, USA: Curran Associates Inc., Dec. 3, 2012, pp. 1097–1105.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations", in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Nov. 21, 2020, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html (visited on 11/19/2024).

[10] M. Oquab, T. Darcet, T. Moutakanni, *et al.* "DINOv2: Learning Robust Visual Features without Supervision". arXiv: `2304.07193 [cs]`. (Apr. 14, 2023), [Online]. Available: http://arxiv.org/abs/2304.07193 (visited on 01/15/2024), pre-published.

[11] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning Transferable Visual Models From Natural Language Supervision", in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 1, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html (visited on 01/15/2024).

[12] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection", *Information Fusion*, vol. 64, pp. 131–148, Dec. 1, 2020, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2020.06.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253520303110 (visited on 04/27/2021).

[13] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey", *ACM Computing Surveys*, vol. 54, no. 1, 7:1–7:41, Jan. 2, 2021, ISSN: 0360-0300. DOI: 10.1145/3425780. [Online]. Available: https://doi.org/10.1145/3425780 (visited on 05/06/2021).

[14] D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes". arXiv: `1312.6114 [stat]`. (Dec. 10, 2022), [Online]. Available: http://arxiv.org/abs/1312.6114 (visited on 02/02/2025), pre-published.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative Adversarial Nets", in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (visited on 07/06/2020).

[16] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics", in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Jun. 1, 2015, pp. 2256–2265. [Online]. Available: https://proceedings.mlr.press/v37/sohl-dickstein15.html (visited on 11/13/2024).

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". arXiv: `1710.10196 [cs]`. (Feb. 26, 2018), [Online]. Available: http://arxiv.org/abs/1710.10196 (visited on 02/02/2025), pre-published.

[18] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks”, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 4396–4405. DOI: 10.1109/CVPR.2019.00453. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8953766.

[19] A. Razavi, A. van den Oord, and O. Vinyals, “Generating diverse high-fidelity images with VQ-VAE-2”, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 1331, Red Hook, NY, USA: Curran Associates Inc., Dec. 8, 2019, pp. 14 866–14 876.

[20] A. Vahdat and J. Kautz, “NVAE: A deep hierarchical variational autoencoder”, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20, Red Hook, NY, USA: Curran Associates Inc., Dec. 6, 2020, pp. 19 667–19 679, ISBN: 978-1-7138-2954-6.

[21] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models”, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20, Red Hook, NY, USA: Curran Associates Inc., Dec. 6, 2020, pp. 6840–6851, ISBN: 978-1-7138-2954-6.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models”, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10 674–10 685. DOI: 10.1109/CVPR52688.2022.01042. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9878449 (visited on 02/06/2025).

[23] E. Hoogeboom, J. Heek, and T. Salimans, “Simple diffusion: End-to-end diffusion for high resolution images”, in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 3, 2023, pp. 13 213–13 232. [Online]. Available: https://proceedings.mlr.press/v202/hoogeboom23a.html (visited on 11/13/2024).

[24] M. Mirza and S. Osindero. “Conditional Generative Adversarial Nets”. arXiv: 1411.1784. (Nov. 6, 2014), [Online]. Available: http://arxiv.org/abs/1411.1784 (visited on 11/14/2024), pre-published.

[25] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation”, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8789–8797. DOI: 10.1109/CVPR.2018.00916. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8579014 (visited on 02/06/2025).

[26] L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models”, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 3813–3824. DOI: 10.1109/ICCV51070.2023.00355. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10377881 (visited on 11/13/2024).

[27] A. Ramesh, M. Pavlov, G. Goh, *et al.*, “Zero-Shot Text-to-Image Generation”, in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 1, 2021, pp. 8821–8831. [Online]. Available: https://proceedings.mlr.press/v139/ramesh21a.html (visited on 11/13/2024).

[28] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. “Hierarchical Text-Conditional Image Generation with CLIP Latents”. arXiv: 2204.06125. (Apr. 13, 2022), [Online]. Available: http://arxiv.org/abs/2204.06125 (visited on 11/14/2024), pre-published.

[29] F. Abbas and A. Taeihagh, “Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence”, *Expert Systems with Applications*, vol. 252, p. 124 260, Oct. 15, 2024, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2024.124260. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424011266 (visited on 02/07/2025).

[30] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, “FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping”, Sep. 15, 2020. arXiv: 1912.13457 [cs]. [Online]. Available: http://arxiv.org/abs/1912.13457 (visited on 02/24/2022).

[31] R. Chen, X. Chen, B. Ni, and Y. Ge, “SimSwap: An Efficient Framework For High Fidelity Face Swapping”, *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2003–2011, Oct. 12, 2020. DOI: 10.1145/3394171.3413630. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413630 (visited on 01/09/2025).

[32] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures”, *ACM Transactions on Graphics*, vol. 38, no. 4, 66:1–66:12, Jul. 12, 2019, ISSN: 0730-0301. DOI: 10.1145/3306346.3323035. [Online]. Available: https://doi.org/10.1145/3306346.3323035 (visited on 02/24/2022).

[33] Y. Nirkin, Y. Keller, and T. Hassner, “FSGAN: Subject Agnostic Face Swapping and Reenactment”, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 7183–7192. DOI: 10.1109/ICCV.2019.00728. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9010341 (visited on 02/06/2025).

[34] Y. Shen, C. Yang, X. Tang, and B. Zhou, “InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2004–2018, Apr. 2022, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2020.3034267. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9241434 (visited on 02/02/2025).

[35] L. Colbois, T. de Freitas Pereira, and S. Marcel, “On the use of automatically generated synthetic image datasets for benchmarking face recognition”, in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Aug. 2021, pp. 1–8. DOI: 10.1109/IJCB52358.2021.9484363. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9484363 (visited on 04/08/2024).

[36] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, “Are GAN-based morphs threatening face recognition?”, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 2959–2963. DOI: 10.1109/ICASSP43922.2022.9746477. [Online]. Available: https://ieeexplore.ieee.org/document/9746477 (visited on 08/27/2024).

[37] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, “MIP-GAN—Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN”, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 365–383, Jul. 2021, ISSN: 2637-6407. DOI: 10.1109/TBIOM.2021.3072349.

[38] P. Korshunov, H. Chen, P. N. Garner, and S. Marcel, “Vulnerability of Automatic Identity Recognition to Audio-Visual Deepfakes”, in *2023 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2023, pp. 1–10. DOI: 10.1109/IJCB57857.2023.10449189. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10449189 (visited on 02/02/2025).

[39] A. Romano, “Why Reddit’s face-swapping celebrity porn craze is a harbinger of dystopia”, *Vox*, Jan. 31, 2018. [Online]. Available: https://www.vox.com/2018/1/31/16932264/reddit-celebrity-porn-face-swapping-dystopia (visited on 01/10/2022).

[40] S. Samuel, “A guy made a deepfake app to turn photos of women into nudes. It didn’t go well.”, *Vox*, Jun. 27, 2019. [Online]. Available: https://www.vox.com/2019/6/27/18761639/ai-deepfake-deepnude-app-nude-women-porn (visited on 01/10/2022).

[41] L. Llach, “Spanish town shocked by AI nudes of teenage girls: But is it a crime?”, *EuroNews*, Sep. 24, 2023. [Online]. Available: https://www.euronews.com/next/2023/09/24/spanish-teens-received-deepfake-ai-nudes-of-themselves-but-is-it-a-crime (visited on 11/22/2024).

[42] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 22 500–22 510. DOI: 10.1109/CVPR52729.2023.02155. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10204880 (visited on 11/22/2024).

[43] J. Vincent, “Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news”, *The Verge*, Apr. 17, 2018. [Online]. Available: https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed (visited on 01/10/2022).

[44] O. Schwartz, “You thought fake news was bad? Deep fakes are where truth goes to die”, *The Guardian*, Nov. 12, 2018, ISSN: 0261-3077. [Online]. Available: https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth (visited on 01/10/2022).

[45] J. Wakefield, “Deepfake presidents used in Russia-Ukraine war”, *BBC News*, Mar. 18, 2022. [Online]. Available: https://www.bbc.com/news/technology-60780142 (visited on 11/22/2024).

[46] F. Carmichael, “How a fake network pushes pro-China propaganda”, *BBC News*, Aug. 5, 2021. [Online]. Available: https://www.bbc.com/news/world-asia-china-58062630 (visited on 01/11/2022).

[47] “Facebook uncovers Chinese network behind fake expert”, *BBC News*, Dec. 2, 2021. [Online]. Available: https://www.bbc.com/news/world-asia-china-59456548 (visited on 01/11/2022).

[48] C. Christie. “German Art Activists Get Passport Using Digitally Altered Photo of Two Women Merged Together”, Vice. (Oct. 9, 2018), [Online]. Available: https://www.vice.com/en/article/pa9vyb/peng-collective-artists-hack-german-passport (visited on 02/24/2022).

[49] N. Damer, M. Fang, P. Siebke, J. N. Kolf, M. Huber, and F. Boutros, “MorDIFF: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Diffusion Autoencoders”, in *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, Apr. 2023, pp. 1–6. DOI: 10.1109/IWBF57495.2023.10157869.

[50] C. Vaccari and A. Chadwick, “Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News”, *Social Media + Society*, vol. 6, no. 1, p. 2 056 305 120 903 408, Jan. 1, 2020, ISSN: 2056-3051. DOI: 10.1177/2056305120903408. [Online]. Available: https://doi.org/10.1177/2056305120903408 (visited on 01/10/2022).

[51] “Jugé lundi pour outrage à magistrat, Dieudonné se dit victime d’un « deepfake »”, *Le Figaro*, Mar. 14, 2021. [Online]. Available: https://www.lefigaro.fr/culture/juge-lundi-pour-outrage-a-magistrate-dieudonne-se-dit-victime-d-un-deepfake-20210314 (visited on 01/11/2022).

[52] “Evaluating digital open source imagery: A guide for judges and fact-finders”, TRUE Project, 2024. [Online]. Available: www.trueproject.co.uk/osguide (visited on 02/02/2025).

[53] R. Mahaim, *Motion 23.3563 - Réglementer les "deep fakes"*, May 4, 2023. [Online]. Available: https://www.parlament.ch/en/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20233563 (visited on 01/13/2025).

[54] F. Regazzi, *Motion 24.4464 - Pour une stratégie contre l’usage abusif de nos images*, Dec. 19, 2024. [Online]. Available: https://www.parlament.ch/en/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20244464 (visited on 01/16/2025).

[55] R. Mahaim, *Motion 24.4018 - Limiter la diffusion des applications qui créent des nus dégradants*, Sep. 25, 2024. [Online]. Available: https://www.parlament.ch/en/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20244018 (visited on 01/16/2025).

[56] TA-SWISS, “Eyes and Ears on the Test Bench - Condensed version of the TA-SWISS study, "Deepfakes and Manipulated Realities"”, TA-SWISS, p. 24. [Online]. Available: https://www.ta-swiss.ch/en/deepfakes.

[57] Q. Jacquemin. “Deepfakes pornographiques : Le droit pénal suisse prend-il en compte cette nouvelle menace ?”, LexTech Institute. (May 9, 2023), [Online]. Available: https://www.lextechinstitute.ch/deepfakes-pornographiques-le-droit-penal-suisse-prend-il-en-compte-cette-nouvelle-menace/?lang=en (visited on 01/15/2025).

[58] B. Allyn, “’The New York Times’ takes OpenAI to court. ChatGPT’s future could be on the line”, *NPR*, Jan. 14, 2025. [Online]. Available: https://www.npr.org/2025/01/14/nx-s1-5258952/new-york-times-openai-microsoft (visited on 01/19/2025).

[59] H. R. Hasan and K. Salah, “Combating Deepfake Videos Using Blockchain and Smart Contracts”, *IEEE Access*, vol. 7, pp. 41 596–41 606, 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2905689. [Online]. Available: https://ieeexplore.ieee.org/document/8668407 (visited on 11/18/2024).

[60] *C2PA Technical Specification*. [Online]. Available: https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html (visited on 02/02/2025).

[61] Z. Xiang, J. Horváth, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, “Forensic Analysis of Video Files Using Metadata”, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2021, pp. 1042–1051. DOI: 10.1109/CVPRW53098.2021.00115. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9523116 (visited on 02/02/2025).

[62] A. Bharati, D. Moreira, J. Brogan, *et al.*, “Beyond Pixels: Image Provenance Analysis Leveraging Metadata”, in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2019, pp. 1692–1702. DOI: 10.1109/WACV.2019.00185. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8658404 (visited on 02/02/2025).

[63] R. Padilha, T. Salem, S. Workman, F. A. Andaló, A. Rocha, and N. Jacobs, “Content-Aware Detection of Temporal Metadata Manipulation”, *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1316–1327, 2022, ISSN: 1556-6021. DOI: 10.1109/TIFS.2022.3159154. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9733363 (visited on 02/02/2025).

[64] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking”, in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2018, pp. 1–7. DOI: 10.1109/WIFS.2018.8630787.

[65] H. Qi, Q. Guo, F. Juefei-Xu, *et al.*, “DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms”, in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA: Association for Computing Machinery, Oct. 12, 2020, pp. 4318–4327, ISBN: 978-1-4503-7988-5. [Online]. Available: https://doi.org/10.1145/3394171.3413707 (visited on 01/31/2022).

[66] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales. “DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation”. arXiv: 2010.00400 `[cs]`. (Dec. 14, 2020), [Online]. Available: http://arxiv.org/abs/2010.00400 (visited on 02/06/2025), pre-published.

[67] L. Li, J. Bao, T. Zhang, *et al.*, “Face X-Ray for More General Face Forgery Detection”, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 5000–5009. DOI: 10.1109/CVPR42600.2020.00505. [Online]. Available: https://ieeexplore.ieee.org/document/9157215 (visited on 02/06/2025).

[68] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do GANs Leave Artificial Fingerprints?”, in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Mar. 2019, pp. 506–511. DOI: 10.1109/MIPR.2019.00103.

[69] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-Generated Images Are Surprisingly Easy to Spot… for Now”, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 8692–8701. DOI: 10.1109/CVPR42600.2020.00872. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9156876 (visited on 02/02/2025).

[70] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, “Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art”, in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2021, pp. 1–6. DOI: 10.1109/ICME51207.2021.9428429. [Online]. Available: https://ieeexplore.ieee.org/document/9428429 (visited on 08/27/2024).

[71] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, “Intriguing properties of synthetic images: From generative adversarial networks to diffusion models”, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 973–982, Jun. 2023. DOI: 10.1109/CVPRW59228.2023.00104. [Online]. Available: https://ieeexplore.ieee.org/document/10209003/ (visited on 02/03/2024).

[72] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, “Leveraging Frequency Analysis for Deep Fake Image Recognition”, in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Nov. 21, 2020, pp. 3247–3258. [Online]. Available: https://proceedings.mlr.press/v119/frank20a.html (visited on 02/03/2022).

[73] F. Lugstein, S. Baier, G. Bachinger, and A. Uhl, “PRNU-based Deepfake Detection”, in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&amp;MMSec ’21, New York, NY, USA: Association for Computing Machinery, Jun. 21, 2021, pp. 7–12, ISBN: 978-1-4503-8295-3. DOI: 10.1145/3437880.3460400. [Online]. Available: https://doi.org/10.1145/3437880.3460400 (visited on 11/20/2024).

[74] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper. “Unmasking DeepFakes with simple Features”. arXiv: 1911.00686 `[cs]`. (Mar. 4, 2020), [Online]. Available: http://arxiv.org/abs/1911.00686 (visited on 02/06/2025), pre-published.

[75] O. Giudice, L. Guarnera, and S. Battiato, “Fighting Deepfakes by Detecting GAN DCT Anomalies”, *Journal of Imaging*, vol. 7, no. 8, p. 128, 8 Aug. 2021, ISSN: 2313-433X. DOI: 10.3390/jimaging7080128. [Online]. Available: https://www.mdpi.com/2313-433X/7/8/128 (visited on 02/06/2025).

[76] S. McCloskey and M. Albright. "Detecting GAN-generated Imagery using Color Cues". arXiv: 1812.08247 [cs]. (Dec. 19, 2018), [Online]. Available: http://arxiv.org/abs/1812.08247 (visited on 02/06/2025), pre-published.

[77] A. Aghasanli, D. Kangin, and P. Angelov, "Interpretable-through-prototypes deepfake detection for diffusion models", in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2023, pp. 467–474. DOI: 10.1109/ICCVW60793.2023.00053. [Online]. Available: https://ieeexplore.ieee.org/document/10350382 (visited on 11/20/2024).

[78] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What Makes Fake Images Detectable? Understanding Properties that Generalize", in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 103–120, ISBN: 978-3-030-58574-7. DOI: 10.1007/978-3-030-58574-7_7.

[79] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On The Detection of Synthetic Images Generated by Diffusion Models", in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095167. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10095167 (visited on 11/20/2024).

[80] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints", in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 7555–7565, ISBN: 978-1-7281-4803-8. DOI: 10.1109/ICCV.2019.00765. [Online]. Available: https://ieeexplore.ieee.org/document/9010964/ (visited on 02/03/2022).

[81] T. Yang, Z. Huang, J. Cao, L. Li, and X. Li, "Deepfake Network Architecture Attribution", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, pp. 4662–4670, 4 Jun. 28, 2022, ISSN: 2374-3468. DOI: 10.1609/aaai.v36i4.20391. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20391 (visited on 08/27/2024).

[82] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu, "Hierarchical Fine-Grained Image Forgery Detection and Localization", in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 3155–3165, ISBN: 979-8-3503-0129-8. DOI: 10.1109/CVPR52729.2023.00308. [Online]. Available: https://ieeexplore.ieee.org/document/10205306/ (visited on 08/06/2024).

[83] H. H. Nguyen, J. Yamagishi, and I. Echizen, "How Close Are Other Computer Vision Tasks to Deepfake Detection?", in *2023 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2023, pp. 1–10. DOI: 10.1109/IJCB57857.2023.10448744. [Online]. Available: https://ieeexplore.ieee.org/document/10448744 (visited on 02/02/2025).

[84] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models", in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 24 480–24 489, ISBN: 979-8-3503-0129-8. DOI: 10.1109/CVPR52729.2023.02345. [Online]. Available: https://ieeexplore.ieee.org/document/10204883/ (visited on 01/10/2024).

[85] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the Bar of AI-generated Image Detection with CLIP", in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2024, pp. 4356–4366. DOI: 10.1109/CVPRW63382.2024.00439. [Online]. Available: https://ieeexplore.ieee.org/document/10677873 (visited on 02/06/2025).

[86] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging Properties in Self-Supervised Vision Transformers", in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9630–9640. DOI: 10.1109/ICCV48922.2021.00951. [Online]. Available: https://ieeexplore.ieee.org/document/9709990 (visited on 02/02/2025).

[87] A. El-Nouby, M. Klein, S. Zhai, *et al.* "Scalable Pre-training of Large Autoregressive Image Models". arXiv: 2401.08541 [cs]. (Jan. 16, 2024), [Online]. Available: http://arxiv.org/abs/2401.08541 (visited on 06/27/2024), pre-published.

[88] H. Li, J. Yang, K. Wang, *et al.* "Scalable Autoregressive Image Generation with Mamba". arXiv: 2408.12245 [cs]. (Dec. 11, 2024), [Online]. Available: http://arxiv.org/abs/2408.12245 (visited on 02/06/2025), pre-published.

[89] F. Tassone, L. Maiano, and I. Amerini, "Continuous fake media detection: Adapting deepfake detectors to new generative techniques", *Computer Vision and Image Understanding*, vol. 249, p. 104 143, Dec. 1, 2024, ISSN: 1077-3142. DOI: 10.1016/j.cviu.2024.104143. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314224002248 (visited on 02/06/2025).

[90] D. Tariang, R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Synthetic Image Verification in the Era of Generative Artificial Intelligence: What Works and What Isn't There yet", *IEEE Security & Privacy*, vol. 22, no. 3, pp. 37–49, May 2024, ISSN: 1558-4046. DOI: 10.1109/MSEC.2024.3376637. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10492675 (visited on 11/21/2024).

[91] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[92] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 1, 2020, ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. [Online]. Available: https://doi.org/10.1007/s11263-019-01228-7 (visited on 11/21/2024).

[93] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps", in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, Red Hook, NY, USA: Curran Associates Inc., Dec. 3, 2018, pp. 9525–9536.

[94] M. Jacquet and C. Champod, “Automated face recognition in forensic science: Review and perspectives”, *Forensic Science International*, vol. 307, p. 110 124, Feb. 1, 2020, ISSN: 0379-0738. DOI: 10.1016/j.forsciint.2019.110124. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379073819305365 (visited on 08/26/2021).

[95] B. Khoo, R. C.-W. Phan, and C.-H. Lim, “Deepfake attribution: On the source identification of artificially generated images”, *WIREs Data Mining and Knowledge Discovery*, vol. n/a, no. n/a, e1438, 2022, ISSN: 1942-4795. DOI: 10.1002/widm.1438. [Online]. Available: http://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1438 (visited on 12/10/2021).

[96] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, “KoDF: A Large-scale Korean DeepFake Detection Dataset”, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 10 724–10 733. DOI: 10.1109/ICCV48922.2021.01057. [Online]. Available: https://ieeexplore.ieee.org/document/9710066 (visited on 02/06/2025).

[97] S. Jia, X. Li, and S. Lyu, “Model Attribution of Face-Swap Deepfake Videos”, *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2356–2360, Oct. 16, 2022. DOI: 10.1109/ICIP46576.2022.9897972. [Online]. Available: https://ieeexplore.ieee.org/document/9897972/ (visited on 01/18/2023).

[98] M. Ferrara, A. Franco, and D. Maltoni, “The magic passport”, in *IEEE International Joint Conference on Biometrics*, Sep. 2014, pp. 1–7. DOI: 10.1109/BTAS.2014.6996240.

[99] C. Busch, “Morphing Attack Detection - State of the Art and Challenges”, presented at the 20th IAPR/IEEE International Summer School for Advanced Studies on Biometrics (Alghero, Italy), Jun. 6, 2023. [Online]. Available: https://christoph-busch.de/files/Busch-MAD-230606.pdf (visited on 09/26/2024).

[100] “Cantonal passport offices and application procedures”, FEDPOL, Sep. 13, 2024. [Online]. Available: https://www.fedpol.admin.ch/fedpol/en/home/pass---identitaetskarte/pass/passstellen.html (visited on 09/26/2024).

[101] “Face Analysis Technology Evaluation (FATE) MORPH”, NIST, Oct. 5, 2024. [Online]. Available: https://pages.nist.gov/frvt/html/frvt_morph.html (visited on 01/16/2024).

[102] A. Makrushin, T. Neubert, and J. Dittmann, “Automatic Generation and Detection of Visually Faultless Facial Morphs:” in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Porto, Portugal: SCITEPRESS - Science and Technology Publications, 2017, pp. 39–50, ISBN: 978-989-758-225-7 978-989-758-226-4 978-989-758-227-1. DOI: 10.5220/0006131100390050. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006131100390050 (visited on 09/08/2023).

[103] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann, “Extended StirTrace benchmarking of biometric and forensic qualities of morphed face images”, *IET Biometrics*, vol. 7, no. 4, pp. 325–332, 2018, ISSN: 2047-4946. DOI: 10.1049/iet-bmt.2017.0147. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2017.0147 (visited on 02/06/2025).

[104] N. Damer, A. M. Saladié, A. Braun, and A. Kuijper, “MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network”, in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Oct. 2018, pp. 1–10. DOI: 10.1109/BTAS.2018.8698563.

[105] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, and C. Busch, “Can GAN Generated Morphs Threaten Face Recognition Systems Equally as Landmark Based Morphs? - Vulnerability and Detection”, in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, Apr. 2020, pp. 1–6. DOI: 10.1109/IWBF49977.2020.9107970.

[106] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, “Diffusion Autoencoders: Toward a Meaningful and Decodable Representation”, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10 609–10 619. DOI: 10.1109/CVPR52688.2022.01036. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9878402 (visited on 02/06/2025).

[107] Z. Blasingame and C. Liu. “Leveraging Diffusion For Strong and High Quality Face Morphing Attacks”. arXiv: 2301.04218 [cs]. (Jun. 8, 2023), [Online]. Available: http://arxiv.org/abs/2301.04218 (visited on 09/12/2023), pre-published.

[108] P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing”, *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, Jul. 1, 2003, ISSN: 0730-0301. DOI: 10.1145/882262.882269. [Online]. Available: https://dl.acm.org/doi/10.1145/882262.882269 (visited on 10/21/2024).

[109] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN”, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 8107–8116. DOI: 10.1109/CVPR42600.2020.00813. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9156570 (visited on 02/06/2025).

[110] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8578166 (visited on 10/24/2024).

[111] U. Scherhag, A. Nautsch, C. Rathgeb, *et al.*, “Biometric Systems under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting”, in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2017, pp. 1–7. DOI: 10.23919/BIOSIG.2017.8053499.

[112] S. Venkatesh, K. Raja, R. Ramachandra, and C. Busch, “On the Influence of Ageing on Face Morph Attacks: Vulnerability and Detection”, in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2020, pp. 1–10. DOI: 10.1109/IJCB48548.2020.9304856. [Online]. Available: https://ieeexplore.ieee.org/document/9304856 (visited on 10/03/2024).

[113] M. Ferrara, A. Franco, D. Maltoni, and C. Busch, “Morphing Attack Potential”, in *2022 International Workshop on Biometrics and Forensics (IWBF)*, Apr. 2022, pp. 1–6. DOI: 10.1109/IWBF55382.2022.9794509. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9794509 (visited on 02/06/2025).

[114] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros, “Privacy-friendly Synthetic Data for the Development of Face Morphing Attack Detectors”, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2022, pp. 1605–1616. DOI: 10.1109/CVPRW56347.2022.00167. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9857354 (visited on 01/17/2024).

[115] K. Raja, M. Ferrara, A. Franco, *et al.*, “Morphing Attack Detection-Database, Evaluation Platform, and Benchmarking”, *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4336–4351, 2021, ISSN: 1556-6013, 1556-6021. DOI: 10.1109/TIFS.2020.3035252. [Online]. Available: https://ieeexplore.ieee.org/document/9246583/ (visited on 02/03/2024).

[116] L. DeBruine and B. Jones, *Face Research Lab London Set*, figshare, May 30, 2017. DOI: 10.6084/m9.figshare.5047666.v5. [Online]. Available: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666/5 (visited on 01/20/2022).

[117] P. Phillips, P. Flynn, T. Scruggs, *et al.*, “Overview of the face recognition grand challenge”, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, Jun. 2005, 947–954 vol. 1. DOI: 10.1109/CVPR.2005.268.

[118] U. Scherhag, C. Rathgeb, and C. Busch, “Face Morphing Attack Detection Methods”, in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, Eds., Cham: Springer International Publishing, 2022, pp. 331–349, ISBN: 978-3-030-87664-7. DOI: 10.1007/978-3-030-87664-7_15. [Online]. Available: https://doi.org/10.1007/978-3-030-87664-7_15 (visited on 07/03/2024).

[119] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2002.1017623.

[120] J. Lukas, J. Fridrich, and M. Goljan, “Digital camera identification from sensor pattern noise”, *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, Jun. 2006, ISSN: 1556-6021. DOI: 10.1109/TIFS.2006.873602. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1634362 (visited on 10/24/2024).

[121] J. E. Tapia and C. Busch, “Face Feature Visualisation of Single Morphing Attack Detection”, *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6, Apr. 19, 2023. DOI: 10.1109/IWBF57495.2023.10157534. [Online]. Available: https://ieeexplore.ieee.org/document/10157534/ (visited on 02/03/2024).

[122] M. Fang, F. Boutros, and N. Damer. “Unsupervised Face Morphing Attack Detection via Self-paced Anomaly Detection”. arXiv: 2208.05787 [cs]. (Aug. 11, 2022), [Online]. Available: http://arxiv.org/abs/2208.05787 (visited on 06/19/2024), pre-published.

[123] M. Ivanovska and V. Štruc, “Face Morphing Attack Detection with Denoising Diffusion Probabilistic Models”, *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6, Apr. 19, 2023. DOI: 10.1109/IWBF57495.2023.10156877. [Online]. Available: https://ieeexplore.ieee.org/document/10156877/ (visited on 02/03/2024).

[124] U. M. Kelly, L. Spreeuwers, and R. Veldhuis, “Worst-Case Morphs: A Theoretical and a Practical Approach”, in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2022, pp. 1–5. DOI: 10.1109/BIOSIG55365.2022.9896965.

[125] H. Otroshi 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*, vol. 19, pp. 4585–4600, 2024, ISSN: 1556-6013, 1556-6021. DOI: 10.1109/TIFS.2024.3381820. [Online]. Available: https://ieeexplore.ieee.org/document/10478940/ (visited on 10/07/2024).

[126] H. O. Shahreza and S. Marcel, “Face reconstruction from facial templates by learning latent space of a generator network”, in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS ’23, Red Hook, NY, USA: Curran Associates Inc., May 30, 2024, pp. 12 703–12 720.

[127] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou. “Arc2Face: A Foundation Model for ID-Consistent Human Faces”. arXiv: 2403.11641. (Aug. 22, 2024), [Online]. Available: http://arxiv.org/abs/2403.11641 (visited on 11/24/2024), pre-published.

[128] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization”, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980 (visited on 02/06/2025).

[129] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2021.3087709. [Online]. Available: https://ieeexplore.ieee.org/document/9449988 (visited on 11/26/2024).

[130] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, “ElasticFace: Elastic Margin Loss for Deep Face Recognition”, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2022, pp. 1577–1586. DOI: 10.1109/CVPRW56347.2022.00164. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9857083 (visited on 02/06/2025).

[131] *Best practice technical guidelines for automated border control ABC systems*, 2015. [Online]. Available: https://www.frontex.europa.eu/publications/best-practice-technical-guidelines-for-automated-border-control-abc-systems-3ZjKHL.

[132] 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", in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Red Hook, NY, USA: Curran Associates Inc., Dec. 4, 2017, pp. 6629–6640, ISBN: 978-1-5108-6096-4.

[133] M. Ramon, A. K. Bobak, and D. White, "Super-recognizers: From the lab to the world and back again", *British Journal of Psychology*, vol. 110, no. 3, pp. 461–479, 2019, ISSN: 2044-8295. DOI: 10.1111/bjop.12368. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12368 (visited on 02/01/2025).

[134] M. Ramon, M. Vowels, and M. Groh, "Deepfake Detection in Super-Recognizers and Police Officers", *IEEE Security & Privacy*, vol. 22, no. 3, pp. 68–76, May 2024, ISSN: 1558-4046. DOI: 10.1109/MSEC.2024.3371030. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10480398 (visited on 02/01/2025).

[135] D. J. Robertson, R. S. S. Kramer, and A. M. Burton, "Fraudulent ID using face morphs: Experiments on human and automatic recognition", *PLOS ONE*, vol. 12, no. 3, e0173319, Mar. 22, 2017, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0173319. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0173319 (visited on 02/01/2025).

[136] G. Stein, J. C. Cresswell, R. Hosseinzadeh, *et al.*, "Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models", presented at the Thirty-Seventh Conference on Neural Information Processing Systems, Nov. 2, 2023. [Online]. Available: https://openreview.net/forum?id=08zf7kTOoh (visited on 11/26/2024).

[137] L. Spreeuwers, M. Schils, and R. Veldhuis, "Towards Robust Evaluation of Face Morphing Detection", in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1027–1031. DOI: 10.23919/EUSIPCO.2018.8553018. [Online]. Available: https://ieeexplore.ieee.org/document/8553018 (visited on 11/28/2024).

[138] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing", in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, Sep. 2012, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/6313548 (visited on 11/26/2024).

[139] P. Korshunov and S. Marcel, "Improving Generalization of Deepfake Detection with Data Farming and Few-Shot Learning", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Dec. 2021.

[140] A. G. Howard, M. Zhu, B. Chen, *et al.* "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". arXiv: 1704.04861 [cs]. (Apr. 17, 2017), [Online]. Available: http://arxiv.org/abs/1704.04861 (visited on 02/06/2025), prepublished.

[141] H. Zhang, R. Ramachandra, K. Raja, and C. Busch, “Generalized Single-Image-Based Morphing Attack Detection Using Deep Representations from Vision Transformer”, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2024, pp. 1510–1518. DOI: 10.1109/CVPRW63382.2024.00158. [Online]. Available: https://ieeexplore.ieee.org/document/10678107 (visited on 02/06/2025).

[142] A. Paszke, S. Gross, F. Massa, *et al.*, “PyTorch: An imperative style, high-performance deep learning library”, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 721, Red Hook, NY, USA: Curran Associates Inc., Dec. 8, 2019, pp. 8026–8037.

[143] J. B. Hirschberg and A. Rosenberg, “V-Measure: A conditional entropy-based external cluster evaluation”, presented at the EMNLP, Columbia University, 2007. DOI: 10.7916/D80V8N84. [Online]. Available: https://academiccommons.columbia.edu/doi/10.7916/D80V8N84 (visited on 02/15/2024).

[144] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy (visited on 02/06/2025).

[145] J. E. Tapia, M. Russo, and C. Busch. “Generating Automatically Print/Scan Textures for Morphing Attack Detection Applications”. arXiv: 2408.09558 [cs]. (Aug. 18, 2024), [Online]. Available: http://arxiv.org/abs/2408.09558 (visited on 12/17/2024), prepublished.

[146] M. Robledo-Moreno, G. Borghi, N. Di Domenico, A. Franco, K. Raja, and D. Maltoni, “Towards Federated Learning for Morphing Attack Detection”, in *2024 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2024, pp. 1–10. DOI: 10.1109/IJCB62174.2024.10744518. [Online]. Available: https://ieeexplore.ieee.org/document/10744518 (visited on 02/03/2025).

[147] L. Colbois and S. Marcel, “Generator attribution of morphing attacks using attack-agnostic feature representations”, Under review at IEEE Access, Apr. 2024.

[148] H. Otroshi Shahreza, L. Colbois, and S. Marcel, “On the Generation of Face Morphs by Inversion of Optimal Morph Embeddings”, Under review at IEEE Transactions in Information Forensics and Security (TIFS), Sep. 2024.

[149] L. Colbois, “Morphing de visage par hypertrucage”, Outreach talk, presented at the IA Workshop from Entente Forensique Francophone, Apr. 29, 2024.

[150] L. Colbois, “Generator Attribution of Morphing Attacks using Attack-Agnostic Feature Representations”, Internal institute-wide talk (Idiap Research Institute), Apr. 16, 2024.

[151] L. Colbois, “Attack-agnostic feature representations in morphing attack attribution”, Visiting talk (GRIP Lab, University Federico II of Naples), Feb. 22, 2024.

[152] L. Colbois, “Deepfake detection”, Outreach talk, presented at the Meeting between Argovian Police and the School of Criminal Sciences (University of Lausanne (UNIL)), Dec. 7, 2023.

[153] L. Colbois and H. Otroshi Shahreza, “Approximating optimal morphing attacks using template inversion”, Poster, presented at the 2023 IEEE International Joint Conference on Biometrics (IJCB), Sep. 2023.

[154] L. Colbois, “Morphing attacks through optimal embedding inversion : Vulnerability analysis”, Poster, presented at the UNIL ESC Summer School 2023, Aug. 2023.

[155] L. Colbois, “Deepfake attribution”, presented at the European Association for Biometrics (EAB) & Center for Identification Technology Research (CITeR) Biometrics Workshop, May 30, 2023.

[156] L. Colbois, “Deepfakes : Technologie et enjeux”, Outreach talk, presented at the Cycle de Conférences Grand Public "L’IA Changera-t-Elle Nos Vies?" - Deepfakes : Un Monde de Faux-Semblants (UNIL/CHUV Musée de la main), Oct. 11, 2022.

[157] L. Colbois, “Introduction to the forensic framework - with a face recognition example”, Internal institute-wide talk (Idiap Research Institute), Sep. 21, 2022.

[158] L. Colbois, “On the detection of morphing attacks generated by GANs”, presented at the 2022 International Conference of the Biometrics Special Interest Group (BIOSIG), Aug. 31, 2022.

[159] L. Colbois, “Towards reliable and interpretable detection of face deepfakes”, presented at the UNIL ESC Summer School 2022, Aug. 22, 2022.

[160] L. Colbois, “On the use of automatically generated synthetic image datasets for benchmarking face recognition”, Video Capsule, presented at the UNIL ESC Summer School 2021, Aug. 2021.

[161] L. Colbois, “On the use of automatically generated synthetic image datasets for benchmarking face recognition”, presented at the 2021 IEEE International Joint Conference on Biometrics (IJCB), Aug. 6, 2021.

# Acronyms

**ABC** Automated Border Control. 51, 53, 89, 91

**APCER** Attack Presentation Classification Error Rate. 70

**BPCER** Bona Fide Presentation Classification Error Rate. 70

**CNN** Convolutional Neural Network. 13, 41, 69, 79, 92, 99, 110, 113

**D-EER** Detection Equal Error Rate. 71, 112–115, 117–120

**DET** Detection Error Tradeoff. 71, 114

**DiffAE** Diffusion Autoencoder. 62, 67

**DNN** Deep Neural Network. 11–13, 15, 16, 19, 38, 41, 45, 46, 146

**FFHQ** Flicker-Faces HQ. 67, 68, 115, 119

**FMMPMR** Fully Mated Morph Presentation Match Rate. 64

**FRGC** Face Recognition Grand Challenge. 67, 68, 115, 119

**FRLL** Face Research Lab London. 65–68, 117

**FRS** Face Recognition System. 52–54, 61, 63–65, 71, 77, 78, 141, 142

**GAN** Generative Adversarial Network. 16, 20, 37, 43, 44, 51, 55, 59, 69, 110, 111

**GMM** Gaussian Mixture Model. 110, 113, 148

**GPU** Graphical Processing Unit. 12

**LBP** Local Binary Pattern. 69, 98

**LLM** Large Language Model. 26

**MAD** Morphing Attack Detection. 54, 66, 67, 69, 70, 98–100, 108, 110, 111, 113, 115, 120, 121

**MAP** Morphing Attack Potential. 64, 65

**MMPMR** Mated Morph Presentation Match Rate. 63, 64

**PCA** Principal Component Analysis. 113

**PRNU** Photo Response Non-Uniformity. 38, 69

**SMDD** Synthetic Morphing Attack Developement. 66

**SVM** Support Vector Machine. 13, 113

**VAE** Variational Autoencoder. 20

**ViT** Vision Transformer. 41