DEEP VARIATIONAL PRIVACY FUNNEL: GENERAL MODELING WITH APPLICATIONS IN FACE RECOGNITION

Behrooz Razeghi[⋆] *Parsa Rahimi*[⋆†] *Sébastien Marcel*[⋆‡]

[⋆] Idiap Research Institute
[†] École Polytechnique Fédérale de Lausanne (EPFL)
[‡] Université de Lausanne (UNIL)

ABSTRACT

In this study, we harness the information-theoretic Privacy Funnel (PF) model to develop a method for privacy-preserving representation learning using an end-to-end training framework. We rigorously address the trade-off between obfuscation and utility. Both are quantified through the *logarithmic loss*, a measure also recognized as self-information loss. This exploration deepens the interplay between information-theoretic privacy and representation learning, offering substantive insights into data protection mechanisms for both discriminative and generative models. Importantly, we apply our model to state-of-the-art face recognition systems. The model demonstrates adaptability across diverse inputs, from raw facial images to both derived or refined embeddings, and is competent in tasks such as classification, reconstruction, and generation.

***Index Terms*—** Privacy funnel, information leakage, statistical inference, obfuscation, face recognition.

1. INTRODUCTION

In the fields of information theory and computer science, privacy preservation has been a perennial concern, evolving with technology and emerging privacy threats. The advent of big data intensified both the opportunities, such as innovative business models and personalized services, and challenges, including new privacy threats. Current privacy research pivots around a delicate balance between the provable privacy level and maintaining data utility, which may vary significantly depending on the specific application and data properties.

There exist two main types of privacy-preserving mechanisms: '*prior-independent*' and '*prior-dependent*'. Prior-independent mechanisms make minimal assumptions about the data distribution and adversary information, while prior-dependent mechanisms exploit knowledge about the data distribution and the adversary to protect privacy. Anonymization techniques like k-anonymity [1], ℓ-diversity [2], t-closeness [3], differential privacy (DP) [4], and pufferfish [5] aim to preserve data privacy by perturbing data. DP, in particular, is a widely used prior-independent metric that ensures statistical queries' results remain approximately the same regardless of the inclusion of an individual record in the dataset. Conversely, IT privacy [6–14] works on designing mechanisms and metrics that preserve privacy when the statistical properties of the data are partially known or estimated. IT privacy approaches use metrics like f-divergences and Renyi divergence to model the trade-off between privacy (obfuscation) and utility, helping to understand the fundamental privacy limits.

Data-driven privacy mechanisms, like Generative Adversarial Networks (GANs) [15] inspired ones, model the obfuscation-utility trade-off as a game between a defender (privatizer) and an adversary [16–18]. With the continuous improvement in machine learning capabilities, the importance of data-driven privacy mechanisms will increase. Privacy breaches can have serious consequences, hence the need to develop robust privacy-preserving techniques to protect sensitive information.

The primary contributions of our work are as follows:

- To the best of our knowledge, ours is among the first comprehensive studies on Privacy Funnel (PF) modeling within the domain of deep learning. We establish a connection between the information-theoretic foundations of privacy and privacy-preserving representation learning, with a particular emphasis on cutting-edge face recognition systems.
- We introduce a tight variational bound for information leakage which sheds light on the complexities inherent in privacy preservation during deep variational PF (DVPF) learning.
- Our insights into the upper bound of information leakage play as a crucial role in guiding the optimization of privacy-preserving synthetic data generation techniques.
- Our model is proficient in processing both raw image samples and facial image-derived embeddings. Its versatility spans classification, reconstruction, and generation tasks, and its inherent robustness distinguishes it. In alignment with our commitment to furthering research, a comprehensive package will be released, with its particulars detailed in the extended version of our paper.

This research is supported by the Swiss Center for Biometrics Research and Testing.

For the source code visit: `https://gitlab.idiap.ch/biometric/icassp2024.dvpf`.

2. PRIVACY FUNNEL MODEL

Consider two correlated random variables $\mathbf{S}$ and $\mathbf{X}$ with a joint distribution $P_{\mathbf{S},\mathbf{X}}$. The objective of the Privacy Funnel (PF) method [19] is to derive a representation $\mathbf{Z}$ of $\mathbf{X}$ through a stochastic mapping $P_{\mathbf{Z}|\mathbf{X}}$, satisfying the following conditions: (i) $\mathbf{S}-\circ-\mathbf{X}-\circ-\mathbf{Z}$, (ii) representation $\mathbf{Z}$ maximizes the mutual information about $\mathbf{X}$ (i.e., $\mathrm{I}(\mathbf{X};\mathbf{Z})$), and (iii) representation $\mathbf{Z}$ minimizes the mutual information about $\mathbf{S}$ (i.e., $\mathrm{I}(\mathbf{S};\mathbf{Z})$). In essence, the PF method meticulously navigates the balance between the potential information leakage, $\mathrm{I}(\mathbf{S};\mathbf{Z})$, and the utility of the revealed information, $\mathrm{I}(\mathbf{X};\mathbf{Z})$. The functional representation of the Privacy Funnel can be expressed as:

$$\mathsf{PF}\left(R^{\mathrm{s}}, P_{\mathbf{S},\mathbf{X}}\right) \coloneqq \sup_{\substack{P_{\mathbf{Z}|\mathbf{X}}:\\ \mathbf{S}-\circ-\mathbf{X}-\circ-\mathbf{Z}}} \mathrm{I}(\mathbf{X};\mathbf{Z}) \;\; \text{s.t.} \;\; \mathrm{I}(\mathbf{S};\mathbf{Z}) \leq R^{\mathrm{s}}. \quad (1)$$

The PF curve is defined by the values $\mathsf{PF}\left(R^{\mathrm{s}}, P_{\mathbf{S},\mathbf{X}}\right)$ for different R^{s}. We can use a Lagrange multiplier $\alpha \geq 0$ to represent the PF problem by the associated Lagrangian functional: $\mathcal{L}_{\mathrm{PF}}\left(P_{\mathbf{Z}|\mathbf{X}}, \alpha\right) \coloneqq \mathrm{I}(\mathbf{X};\mathbf{Z}) - \alpha\, \mathrm{I}(\mathbf{S};\mathbf{Z})$. Note that the PF model emerges as a specific instance of the CLUB model [20] when the utility information corresponds directly to data $\mathbf{X}$ and the information complexity of the CLUB model exceeds Shannon entropy $\mathrm{H}(P_{\mathbf{X}})$.

Our threat model includes the following assumptions:

- We consider an adversary who is interested in a specific attribute $\mathbf{S}$ related to the data $\mathbf{X}$. This attribute $\mathbf{S}$ could be any function of $\mathbf{X}$, possibly randomized. We restrict $\mathbf{S}$ to represent a discrete attribute, covering prevalent scenarios of interest, such as facial features or identity attributes.
- The adversary has access to the released representation $\mathbf{Z}$ and respects the Markov chain relationship $\mathbf{S}-\circ-\mathbf{X}-\circ-\mathbf{Z}$.
- The mapping $P_{\mathbf{Z}|\mathbf{X}}$, designed by the defender (privatizer), is assumed to be public knowledge. This implies that the adversary is aware of the strategy employed by the defender.

3. DEEP VARIATIONAL PRIVACY FUNNEL

In this section, we introduce our core methodology, the Deep Variational Privacy Funnel (DVPF). Building on the PF principle, this framework utilizes deep neural networks to optimize the information obfuscation-utility trade-offs.

3.1. Parameterized Variational Approximation of $\mathrm{I}(\mathbf{S};\mathbf{Z})$

We provide parameterized variational approximations for information leakage, which include both an explicit tight variational bound and an upper bound. To better understand the nature of information leakage, we can express $\mathrm{I}(\mathbf{S};\mathbf{Z})$ as $\mathrm{I}(\mathbf{X};\mathbf{Z}) - \mathrm{I}(\mathbf{X};\mathbf{Z} \mid \mathbf{S}) = \mathrm{I}(\mathbf{X};\mathbf{Z}) - \mathrm{H}(\mathbf{X} \mid \mathbf{S}) + \mathrm{H}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})$. The conditional entropy $\mathrm{H}(\mathbf{X} \mid \mathbf{S})$ is originated from the nature of data, since it is out of our control. Now, we derive the variational decomposition of $\mathrm{I}(\mathbf{X};\mathbf{Z})$ and $\mathrm{H}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})$. The mutual information $\mathrm{I}(\mathbf{X};\mathbf{Z})$ can be decomposed as:

$$\mathrm{I}(\mathbf{X};\mathbf{Z}) = \mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Z}|\mathbf{X}} \| Q_{\mathbf{Z}} \mid P_{\mathbf{X}}\right) - \mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Z}} \| Q_{\mathbf{Z}}\right), \quad (2)$$

where $Q_{\mathbf{Z}} : \mathcal{Z} \to \mathcal{P}(\mathcal{Z})$ is variational approximation of the latent space distribution $P_{\mathbf{Z}}$. The conditional entropy $\mathrm{H}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})$ can be decomposed as:

$$\begin{aligned} &\mathrm{H}(\mathbf{X} \mid \mathbf{S},\mathbf{Z}) && (3a)\\ &= -\mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{Z}|\mathbf{X}}}\left[\log Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\right]\right] - \mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{X}|\mathbf{S},\mathbf{Z}} \| Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\right) \\ &\leq -\mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{Z}|\mathbf{X}}}\left[\log Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}\right]\right] =: \mathrm{H}^{\mathrm{U}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z}), && (3b)\end{aligned}$$

where $Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}} : \mathcal{S} \times \mathcal{Z} \to \mathcal{P}(\mathcal{X})$ is variational approximation of the optimal uncertainty decoder distribution $P_{\mathbf{X}|\mathbf{S},\mathbf{Z}}$, and the inequality in (3b) follows by noticing that $\mathrm{D}_{\mathrm{KL}}(P_{\mathbf{X}|\mathbf{S},\mathbf{Z}} \| Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}) \geq 0$. Using (2) and (3), the *variational upper bound* of information leakage is given as:

$$\mathrm{I}(\mathbf{S};\mathbf{Z}) \leq \mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Z}|\mathbf{X}} \| Q_{\mathbf{Z}} \mid P_{\mathbf{X}}\right) - \mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Z}} \| Q_{\mathbf{Z}}\right) + \mathrm{H}^{\mathrm{U}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z}). \quad (4)$$

We now employ neural networks to approximate the parameterized variational upper bound of information leakage. Let $P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X})$ represent the family of encoding probability distributions $P_{\mathbf{Z}|\mathbf{X}}$ over $\mathcal{Z}$ for each element of space $\mathcal{X}$, parameterized by the output of a deep neural network $f_{\boldsymbol{\phi}}$ with parameters $\boldsymbol{\phi}$. Analogously, let $P_{\boldsymbol{\varphi}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})$ denote the corresponding family of decoding probability distributions $Q_{\mathbf{X}|\mathbf{S},\mathbf{Z}}$, driven by $g_{\boldsymbol{\varphi}}$. Lastly, $Q_{\boldsymbol{\psi}}(\mathbf{Z})$ denotes the parameterized prior distribution, either explicit or implicit, that is associated with $Q_{\mathbf{Z}}$. Using (2), the parameterized variational approximation of $\mathrm{I}(\mathbf{X};\mathbf{Z})$ can be defined as:

$$\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z}) \coloneqq \mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z} \mid \mathbf{X}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z}) \mid P_{\mathrm{D}}(\mathbf{X})) - \mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\phi}}(\mathbf{Z}) \,\|\, Q_{\boldsymbol{\psi}}(\mathbf{Z})). \quad (5)$$

The parameterized variational approximation of conditional entropy $\mathrm{H}^{\mathrm{U}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})$ in (3b) can be defined as:

$$\mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z}) \coloneqq -\mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}[\log P_{\boldsymbol{\varphi}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})]\right]. \quad (6)$$

Let $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z})$ denote the parameterized variational approximation of information leakage $\mathrm{I}(\mathbf{S};\mathbf{Z})$. Using (4), an upper bound of $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z})$ can be given as:

$$\begin{aligned} \mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z}) &\leq \underbrace{\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})}_{\text{Information Complexity}} + \underbrace{\mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})}_{\text{Information Uncertainty}} + \mathrm{c} && (7a)\\ &=: \mathrm{I}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\varphi}}(\mathbf{S};\mathbf{Z}) + \mathrm{c}, && (7b)\end{aligned}$$

where c is a constant term, independent of the neural networks parameters. This upper bound encourages the model to reduce both the information complexity, represented by $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})$, and the information uncertainty, denoted by $\mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z})$. Consequently, this leads the model to forget or de-emphasize the sensitive attribute $\mathbf{S}$, which subsequently reduces the uncertainty about the useful data $\mathbf{X}$. In essence, this nudges the model towards an accurate reconstruction of the data $\mathbf{X}$.

Now, let us derive another parameterized variational bound of information leakage $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z})$ [20]. We can decompose $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z})$ as follows:

$$\begin{aligned} &\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z}) && (8)\\ &= \underbrace{-\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S} \mid \mathbf{Z}) + \mathrm{H}(P_{\mathbf{S}} \,\|\, P_{\boldsymbol{\xi}}(\mathbf{S}))}_{\text{Prediction Fidelity}} \underbrace{- \mathrm{D}_{\mathrm{KL}}(P_{\mathbf{S}} \,\|\, P_{\boldsymbol{\xi}}(\mathbf{S}))}_{\text{Distribution Discrepancy}}, \end{aligned}$$

where $P_{\boldsymbol{\xi}}(\mathbf{S} \mid \mathbf{Z})$ denotes the corresponding family of decoding

probability distribution $Q_{\mathbf{S}|\mathbf{Z}}$, where $Q_{\mathbf{S}|\mathbf{Z}} : \mathcal{Z} \to \mathcal{P}(\mathcal{S})$ is a variational approximation of optimal decoder distribution $P_{\mathbf{S}|\mathbf{Z}}$. This information decomposition encourages the model to (i) increase uncertainty regarding the sensitive attribute $\mathbf{S}$ upon knowing the released representation $\mathbf{Z}$. Specifically, the goal is to attain maximum entropy for a discrete sensitive attribute $\mathbf{S}$ when all conditional distributions are uniform. This means the adversary, lacking any additional information, can do no better than '*random guessing*'. This scenario equates to a potential lower boundary for $-\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S}\,|\,\mathbf{Z})$ at $-\log_2 N$ and upper boundary for $\mathrm{H}(P_{\mathbf{S}} \,\|\, P_{\boldsymbol{\xi}}(\mathbf{S}))$ at $\log_2 N$, where N represents the possible states (or values, or classes) of $\mathbf{S}$. (ii) Ensure the model's inferred distribution, $P_{\boldsymbol{\xi}}(\mathbf{S})$, aligns tightly with the actual distribution $P_{\mathbf{S}}$. Ideally, the divergence measure, $\mathrm{D}_{\mathrm{KL}}(P_{\mathbf{S}} \| P_{\boldsymbol{\xi}}(\mathbf{S}))$, is minimized to zero when $P_{\boldsymbol{\xi}}(\mathbf{S})$ aligns perfectly with $P_{\mathbf{S}}$. It's essential to recognize that, although the parameterized approximation in (8) doesn't explicitly rely on the information complexity $\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z})$, it is intrinsically linked through the encoder $f_{\boldsymbol{\phi}}$.

3.2. Parameterized Variational Approximation of $\mathrm{I}(\mathbf{X};\mathbf{Z})$

We now quantify information utility by decomposing the mutual information $\mathrm{I}(\mathbf{X};\mathbf{Z})$ and deriving its parameterized variational approximation. The end-to-end parameterized variational approximation associated to the information utility $\mathrm{I}(\mathbf{X};\mathbf{Z})$ can be defined as:

$$\begin{aligned}
\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X};\mathbf{Z}) &:= \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})\right]\right] && \text{(9a)}\\
&\quad - \mathrm{D}_{\mathrm{KL}}\left(P_{\mathrm{D}}(\mathbf{X}) \| P_{\boldsymbol{\theta}}(\mathbf{X})\right) + \mathrm{H}\left(P_{\mathrm{D}}(\mathbf{X}) \| P_{\boldsymbol{\theta}}(\mathbf{X})\right)\\
&\geq \underbrace{-\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})}_{\text{Reconstruction Fidelity}} \underbrace{-\mathrm{D}_{\mathrm{KL}}(P_{\mathrm{D}}(\mathbf{X}) \| P_{\boldsymbol{\theta}}(\mathbf{X}))}_{\text{Distribution Discrepancy}} && \text{(9b)}\\
&=: \mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X};\mathbf{Z}), && \text{(9c)}
\end{aligned}$$

where $\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z}) := \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})\right]\right]$.

3.3. DVPF Objectives

Given (1) and the parameterized approximations detailed earlier, the DVPF Lagrangian functional can be derived. Specifically, considering (8) and (9), we propose this objective:

$$\begin{aligned}
&(\mathsf{P1}):\ \mathcal{L}_{\mathrm{DVPF}}(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\xi},\alpha) := && \text{(10)}\\
&\overbrace{-\mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z}) - \mathrm{D}_{\mathrm{KL}}\left(P_{\mathrm{D}}(\mathbf{X}) \| P_{\boldsymbol{\theta}}(\mathbf{X})\right)}^{\text{Information Utility: } \mathrm{I}^{\mathrm{L}}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X};\mathbf{Z})}\\
&\underbrace{- \alpha \left(- \mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S}\,|\,\mathbf{Z}) + \mathrm{H}\left(P_{\mathbf{S}} \,\|\, P_{\boldsymbol{\xi}}(\mathbf{S})\right) - \mathrm{D}_{\mathrm{KL}}(P_{\mathbf{S}} \| P_{\boldsymbol{\xi}}(\mathbf{S})) \right)}_{\text{Information Leakage: } \mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\xi}}(\mathbf{S};\mathbf{Z})}.
\end{aligned}$$

Considering the upper bound (7), the corresponding objective is:

$$\begin{aligned}
&(\mathsf{P2}):\ \mathcal{L}_{\mathrm{DVPF}}(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\varphi},\alpha) :=\\
&\quad - \mathrm{H}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z}) - \mathrm{D}_{\mathrm{KL}}\left(P_{\mathrm{D}}(\mathbf{X}) \,\|\, P_{\boldsymbol{\theta}}(\mathbf{X})\right)\\
&\quad \underbrace{- \alpha \left(\mathrm{I}_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{X};\mathbf{Z}) + \mathrm{H}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\varphi}}(\mathbf{X} \mid \mathbf{S},\mathbf{Z}) \right)}_{\text{Information Leakage: } \mathrm{I}^{\mathrm{U}}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\varphi}}(\mathbf{S};\mathbf{Z})}. \qquad \text{(11)}
\end{aligned}$$

Figure 1 illustrates the training architecture for (P1). Due to space constraints, we present only the results for (P1).

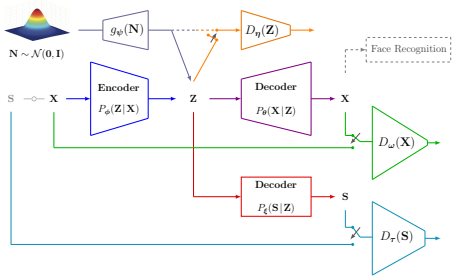

Fig. 1: Training architecture associated with DVPF (P1).

Learning Procedure: The DVPF model (P1) is trained using alternating block coordinate descent across six steps:

(1) **Train Encoder, Utility and Uncertainty Decoders**.

$$\max_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\xi}} \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\theta}}(\mathbf{X}\,|\,\mathbf{Z})\right]\right]$$
$$- \alpha\, \mathbb{E}_{P_{\mathbf{S},\mathbf{X}}}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[\log P_{\boldsymbol{\xi}}(\mathbf{S}\,|\,\mathbf{Z})\right]\right] - \alpha\, \mathbb{E}_{P_{\mathbf{S}}}\left[\log P_{\boldsymbol{\xi}}(\mathbf{S})\right].$$

(2) **Train Latent Space Discriminator**.

$$\min_{\boldsymbol{\eta}} \quad \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log D_{\boldsymbol{\eta}}(\mathbf{Z})\right]\right] + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log(1 - D_{\boldsymbol{\eta}}(\mathbf{Z}))\right].$$

(3) **Train Encoder and Prior Distribution Generator Adversarially**.

$$\max_{\boldsymbol{\phi},\boldsymbol{\psi}} \quad \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}\left[\mathbb{E}_{P_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})}\left[-\log D_{\boldsymbol{\eta}}(\mathbf{Z})\right]\right] + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log(1 - D_{\boldsymbol{\eta}}(\mathbf{Z}))\right].$$

(4) **Train Utility Output Space Discriminator**.

$$\min_{\boldsymbol{\omega}} \mathbb{E}_{P_{\mathrm{D}}(\mathbf{X})}[-\log D_{\boldsymbol{\omega}}(\mathbf{X})] + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}[-\log(1 - D_{\boldsymbol{\omega}}(g_{\boldsymbol{\theta}}(\mathbf{Z})))].$$

(5) **Train Sensitive Attribute Class Discriminator**.

$$\min_{\boldsymbol{\tau}} \mathbb{E}_{P_{\mathbf{S}}}[-\log D_{\boldsymbol{\tau}}(\mathbf{S})] + \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}[-\log(1 - D_{\boldsymbol{\tau}}(g_{\boldsymbol{\xi}}(\mathbf{Z})))].$$

(6) **Train Prior Distribution Generator and Utility Decoder Adversarially**.

$$\max_{\boldsymbol{\psi},\boldsymbol{\theta}} \quad \mathbb{E}_{Q_{\boldsymbol{\psi}}(\mathbf{Z})}\left[-\log(1 - D_{\boldsymbol{\omega}}(g_{\boldsymbol{\theta}}(\mathbf{Z})))\right].$$

4. EXPERIMENTS

In this condensed study, we delve into the methodology of *Embedding-Based Data Learning* for facial image analysis. We've excluded detailed results and discussions, such as the bounds of information leakage, various plots, and methodologies like *Raw Data Transfer Learning with Fine-Tuning* and *End-to-End Raw Data Scratch Learning*. Notably, the generative variational privacy funnel, vital for private synthetic data generation, will be extensively covered in our upcoming extended research version.

We consider the state-of-the-art Face Recognition (FR) backbones with three variants of IResNet [21, 22] architecture (IResNet100, IResNet50, and IResNet18). These architectures have been trained using either the MS1MV3 [23] or WebFace4M/12M [24] datasets. For loss functions, ArcFace [22] and AdaFace [25] methods were employed. Table 1 depicts the Shannon entropy, estimated mutual information between the extracted embeddings $\mathbf{X} \in \mathbb{R}^{512}$ and sensitive attributes $\mathbf{S}$,

Table 1: Evaluation of facial recognition models using various backbones and loss functions. Metrics include entropy, mutual information between embeddings and labels (gender and race), and recognition accuracy on the 'Morph' and 'FairFace' datasets.

				S: Gender						**S**: Race					
				H(**S**)		I(**X**; **S**)		Acc		H(**S**)		I(**X**; **S**)		Acc	
Backbone Dataset	Backbone	Loss Function	Applied Dataset	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
WebFace**4M**	iresnet**18**	AdaFace	Morph			0.610	0.620	0.999	0.996			0.878	0.924	0.998	0.993
WebFace**4M**	iresnet**50**	AdaFace	Morph			0.610	0.620	0.999	0.996			0.873	0.930	0.998	0.992
WebFace**12M**	iresnet**101**	AdaFace	Morph	0.619	0.621	0.605	0.622	0.999	0.996	0.924	0.933	0.873	0.911	0.998	0.992
MS1M-RetinaFace	iresnet**50**	ArcFace	Morph			0.600	0.620	0.999	0.996			0.865	0.910	0.997	0.993
MS1M-RetinaFace	iresnet**100**	ArcFace	Morph			0.597	0.618	0.999	0.997			0.868	0.905	0.997	0.993
WebFace**4M**	iresnet**18**	AdaFace	FairFace			0.930	0.968	0.953	0.923			2.099	2.405	0.882	0.763
WebFace**4M**	iresnet**50**	AdaFace	FairFace			0.932	0.968	0.954	0.931			2.113	2.409	0.883	0.769
WebFace**12M**	iresnet**101**	AdaFace	FairFace	0.999	0.999	0.934	0.969	0.957	0.930	2.517	2.515	2.151	2.417	0.892	0.765
MS1M-RetinaFace	iresnet**50**	ArcFace	FairFace			0.892	0.962	0.950	0.927			1.952	2.355	0.872	0.753
MS1M-RetinaFace	iresnet**100**	ArcFace	FairFace			0.889	0.954	0.951	0.927			1.949	2.348	0.875	0.765

Table 2: Analysis of obfuscation-utility trade-off in facial recognition models using the iresnet-50 architecture. Performance is evaluated across varying information leakage weights α, with significant differences between $\alpha = 0.1$ and $\alpha = 10$. Sensitive attributes considered are 'Gender' and 'Race' with a latent dimensionality of $d_{\mathbf{z}} = 256$. Notations: "WF4M" represents "WebFace4M", and "MS1M-RF" denotes "MS1M-RetinaFace".

	S: Gender						**S**: Race					
	$\alpha = 0.1$			$\alpha = 10$			$\alpha = 0.1$			$\alpha = 10$		
Face Recognition Model	TMR@FMR=10e-1	I(**Z**; **S**)	Acc on **S**	TMR@FMR=10e-1	I(**Z**; **S**)	Acc on **S**	TMR@FMR=10e-1	I(**Z**; **S**)	Acc on **S**	TMR@FMR=10e-1	I(**Z**; **S**)	Acc on **S**
WF4M-i50-Ada-Morph	93.60	0.464	0.992	30.76	0.388	0.843	92.37	0.628	0.997	30.03	0.550	0.857
MS1M-RF-i50-Arc-Morph	94.05	0.485	0.992	58.67	0.335	0.846	94.01	0.635	0.997	58.34	0.558	0.868
WF4M-i50-Ada-FairFace	94.83	0.638	0.925	42.95	0.367	0.576	94.62	0.866	0.946	42.13	0.595	0.756
MS1M-RF-i50-Arc-FairFace	88.28	0.636	0.915	59.91	0.388	0.598	95.57	0.899	0.947	60.33	0.608	0.766

and accuracy of recognition of $\mathbf{S}$, for test and train sets, before applying our DVPF model. For the training phase, we utilized pre-trained models sourced from the aforementioned studies. All input images underwent a standardized pre-processing routine, encompassing alignment, scaling, and normalization. This was in accordance with the specifications of the pre-trained models. We then trained our networks using the Morph dataset [26] and FairFace [27], focusing on different demographic group combinations such as race and gender. A close proximity between $\mathrm{I}(\mathbf{X};\mathbf{S})$ and entropy $\mathrm{H}(\mathbf{S})$ indicates that the embeddings considerably mitigate label uncertainty. Given $\mathrm{I}(\mathbf{X};\mathbf{S}) = \mathrm{H}(\mathbf{X}) + \mathrm{H}(\mathbf{S}) - \mathrm{H}(\mathbf{X},\mathbf{S})$, mutual information serves as a measure of the reduced joint uncertainty about $\mathbf{X}$ and $\mathbf{S}$. It's pivotal to note that $\mathrm{I}(\mathbf{X};\mathbf{S}) \leq \min(\mathrm{H}(\mathbf{X}), \mathrm{H}(\mathbf{S}))$. For the Morph/FairFace datasets, the entropy of sensitive attributes (gender or race) remains consistent across both train/test sets and differing FR model embeddings, emphasizing the same dataset usage throughout experiments. Both Morph and FairFace datasets, featuring 'male' and 'female' gender labels, attain a maximum entropy of $\log_2(2) = 1$. The Morph dataset, with four distinct race labels, reaches a maximum entropy of $\log_2(4) = 2$, while the FairFace dataset, with six race labels, tops at $\log_2(6) = 2.585$. Within Morph, the mutual information for gender mirrors its entropy, suggesting notable preservation of sensitive information in the embeddings. However, for race, values of approximately 0.92-0.93 underscore an imbalanced label distribution, as they don't reach the theoretical $\log_2(4) = 2$. In contrast, the FairFace dataset displays near-maximal entropies for race (~ 2.517 relative to a potential 2.585) and gender (~ 0.999 compared to an ideal 1), illustrating well-balanced racial and gender label distributions.

We applied our DVPF model to the embeddings obtained from the FR models referenced in Table 1. For the accuracy evaluation of our DVPF model within the facial recognition domain, we utilized the challenging IJB-C test dataset [28] as our benchmark. The assessment was initiated with the pre-trained backbones, followed by our DVPF model, which was developed using embeddings from these pre-trained structures. Given space constraints and the consistent performance observed across various IResNet architectures, we present results specific to the IResNet50.

In Table 2, we precisely quantify the disclosed information leakage, represented as $\mathrm{I}(\mathbf{S};\mathbf{Z})$. Additionally, we provide a detailed account of the accuracy achieved in recognizing sensitive attributes from the disclosed representation $\mathbf{Z} \in \mathbb{R}^{256}$, utilizing the support vector classifier optimization. These evaluations are based on test sets derived from either the Morph or FairFace datasets. Moreover, we detail the True Match Rate (TMR) for our models. It's imperative to note that all these evaluations are systematically benchmarked against a predetermined False Match Rate (FMR) of 10^{-1}. When subjecting the 'WF4M-i50-Ada' model to evaluation against the IJB-C dataset—prior to the DVPF model's integration—a TMR of 99.40% at $\mathrm{FMR} = 10\mathrm{e}-1$ was observed. Similarly, for the 'MS1M-RF-i50-Arc' configuration, a TMR of 99.58% was observed on the IJB-C dataset before the integration of the DVPF model, with measurements anchored to the same FMR. Consistent with our expectations, as α increases towards infinity ($\alpha \to \infty$), the information leakage $\mathrm{I}(\mathbf{S};\mathbf{Z})$ decreases to zero. At the same time, the recognition accuracy for the sensitive attribute $\mathbf{S}$ approaches 0.5, indicative of random guessing.

5. CONCLUSION

In this study, we integrate the privacy funnel model for privacy-preserving deep learning, bridging information-theoretic privacy and representation learning. Applied to the state-of-the-art face recognition models, our approach underscores the balance between information obfuscation and utility. The model enhances data protection in discriminative and generative contexts, with an accompanying reproducible software package facilitating further research exploration and adoption.

6. REFERENCES

[1] Latanya Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, 2002.

[2] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *ICDE*, 2006.

[3] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *IEEE ICDE*, 2007.

[4] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*. Springer, 2006, pp. 265–284.

[5] Daniel Kifer and Ashwin Machanavajjhala, "A rigorous and customizable framework for privacy," in *ACM symposium on Principles of Database Systems*, 2012.

[6] Flavio P. Calmon, Ali Makhdoumi, and Muriel Médard, "Fundamental limits of perfect privacy," in *IEEE ISIT*, 2015.

[7] Kousha Kalantari, Lalitha Sankar, and Oliver Kosut, "On information-theoretic privacy with general distortion cost functions," in *IEEE ISIT*, 2017.

[8] Seyed Ali Osia, Ali Taheri, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Hamid R Rabiee, "Deep private-feature extraction," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[9] Ardhendu Tripathy, Ye Wang, and Prakash Ishwar, "Privacy-preserving adversarial networks," in *57th Annual Allerton Conference*, 2019.

[10] Sreejith Sreekumar and Deniz Gündüz, "Optimal privacy-utility trade-off under a rate constraint," in *IEEE International Symposium on Information Theory (ISIT)*, 2019.

[11] Mario Diaz, Hao Wang, Flavio P. Calmon, and Lalitha Sankar, "On the robustness of information-theoretic privacy measures and mechanisms," *IEEE Transactions on Information Theory*, vol. 66, 2019.

[12] Behrooz Razeghi, Flavio P. Calmon, Deniz Gündüz, and Slava Voloshynovskiy, "On perfect obfuscation: Local information geometry analysis," in *IEEE WIFS*, 2020.

[13] Borzoo Rassouli and Deniz Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE TIFS*, vol. 15, 2019.

[14] Amir Ahooye Atashin, Behrooz Razeghi, Deniz Gündüz, and Slava Voloshynovskiy, "Variational leakage: The role of information complexity in privacy leakage," in *ACM WiseML*, 2021.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014.

[16] Harrison Edwards and Amos Storkey, "Censoring representations with an adversary," in *ICLR*, 2016.

[17] Jihun Hamm, "Enhancing utility and privacy with noisy minimax filters," in *IEEE ICASSP*, 2017.

[18] Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal, "Context-aware generative adversarial privacy," *Entropy*, vol. 19, pp. 656, 2017.

[19] Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard, "From the information bottleneck to the privacy funnel," in *IEEE ITW*, 2014.

[20] Behrooz Razeghi, Flavio P Calmon, Deniz Gunduz, and Slava Voloshynovskiy, "Bottlenecks CLUB: Unifying information-theoretic trade-offs among complexity, leakage, and utility," *IEEE TIFS*, vol. 18, 2023.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[22] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE/CVF CVPR*, 2019.

[23] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi, "Lightweight face recognition challenge," in *IEEE/CVF ICCV Workshops*, 2019.

[24] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al., "Webface260m: A benchmark unveiling the power of million-scale deep face recognition," in *IEEE/CVF CVPR*, 2021.

[25] Minchul Kim, Anil K Jain, and Xiaoming Liu, "Adaface: Quality adaptive margin for face recognition," in *IEEE/CVF CVPR*, 2022.

[26] K. Ricanek and T. Tesafaye, "Morph: a longitudinal image database of normal adult age-progression," in *Int. Conf. on Automatic Face and Gesture Recognition*, 2006.

[27] Kimmo Karkkainen and Jungseock Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[28] Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother, "Iarpa janus benchmark - c: Face dataset and protocol," in *ICB*, 2018, pp. 158–165.