Disguise without Disruption: Utility-Preserving Face De-identification

Zikui Cai[1 *], **Zhongpai Gao**[2], **Benjamin Planche**[2], **Meng Zheng**[3], **Terrence Chen**[2], **M. Salman Asif**[1], **Ziyan Wu**[2]

[1]University of California, Riverside, CA
[2]United Imaging Intelligence, Burlington, MA
[3]Rensselaer Polytechnic Institute, Troy, NY
{zcai032, sasif}@ucr.edu, {first.last}@uii-ai.com, zhengm5@rpi.edu

Abstract

With the rise of cameras and smart sensors, humanity generates an exponential amount of data. This valuable information, including underrepresented cases like AI in medical settings, can fuel new deep-learning tools. However, data scientists must prioritize ensuring privacy for individuals in these untapped datasets, especially for images or videos with faces, which are prime targets for identification methods. Proposed solutions to de-identify such images often compromise non-identifying facial attributes relevant to downstream tasks. In this paper, we introduce *Disguise*, a novel algorithm that seamlessly de-identifies facial images while ensuring the usability of the modified data. Unlike previous approaches, our solution is firmly grounded in the domains of differential privacy and ensemble-learning research. Our method involves extracting and substituting depicted identities with synthetic ones, generated using variational mechanisms to maximize obfuscation and non-invertibility. Additionally, we leverage supervision from a mixture-of-experts to disentangle and preserve other utility attributes. We extensively evaluate our method using multiple datasets, demonstrating a higher de-identification rate and superior consistency compared to prior approaches in various downstream tasks.

Introduction

Global privacy laws safeguard personal data, including regulations like GDPR (Voigt and Von dem Bussche 2017) in Europe, HIPPA (HIP 2003) and CCPA (CCP 2018) in the US, and PIPL (PIP 2021) in China. Particularly stringent for medical information and data from medical settings, these rules tightly control storage and distribution of patient images to ensure confidentiality. Yet, this data holds valuable potential, *e.g.*, automating medical procedures and new AI-driven diagnoses. To tap into it, scientists explore techniques for using sensitive images without compromising identity. Most methods focus on face obfuscation (Newton, Sweeney, and Malin 2005), blurring (Frome et al. 2009), pixelation (Zhou and Pun 2020), warping (Korshunov and Ebrahimi 2013)), affecting image saliency. Face-swapping (Hukkelås, Mester, and Lindseth 2019; Maximov, Elezi, and Leal-Taixé 2020; Cao et al. 2021; Proença 2021; Agarwal, Chattopadhyay, and Wang 2021) is emerging as a promising solution.

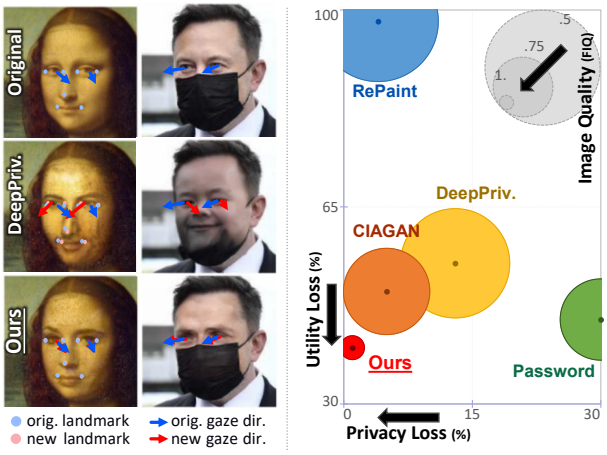


Figure 1: *Disguise* anonymizes face images while preserving their utility (*i.e.*, attributes relevant to downstream tasks). For instance, facial landmarks and gaze direction are better preserved compared to prior art: the red dots for landmarks and red arrows for gazes in new images above are better aligned with the blue ones in the original images. We outperform prior art by a large margin along various axes, including privacy, utility, and image quality. For the latter, small radius indicates higher FIQ score and better image quality.

Popularized through the notion of *deepfakes* (Westerlund 2019), these deep-learning models are trained to replace any face in an image or video by another one (user-provided or AI-generated), while trying to preserve the overall saliency or specific facial attributes, such as perceived gender, expression, or hair color. While recent solutions can generate convincing results, they are not suitable for the targeted use cases as they lack formal *privacy* and *utility* guarantees for the resulting images. Face-swapping methods evade the confidentiality of the ID provider since the swapped face leaks the source ID. In addition, they lack proper mechanisms to maximize de-identification and minimize identity leakage of the target ID. Furthermore, they do not emphasize on maintaining the *utility* of resulting images, *i.e.*, they do not guarantee that the altered images can have the same function as the original ones for various downstream tasks. For ex-

*This work was primarily carried out during the internship of Zikui Cai at United Imaging Intelligence, Burlington, MA 01803.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ample, a dataset would become *useless* for analysis if relevant non-biometric features are corrupted (*e.g.*, facial expressions have changed for sentiment analysis tasks) or for training recognition models if the altered images no longer match their annotations (*e.g.*, facial landmarks, gaze directions, head-pose orientations, *etc.*).

In this work, we aim to address the challenge of anonymizing images of individuals while ensuring privacy and maintaining high data utility. To this end, we propose *Disguise* (Deep Identity Swapper Guaranteeing Utility with Implicit Supervision from Experts), a de-identification method built upon face-swapping technology that offers formal guarantees regarding identity obfuscation and utility retention. Our main contributions are as follows:

- We propose a simple yet effective framework for face de-identification which generates natural faces with distinct identities from the original ones, while maintaining non-biometric attributes unchanged.
- Unlike prior art that pre-discards original face IDs, we condition synthetic faces on the original ID vectors and maximize the distance to the original identities while ensuring differential privacy (Dwork, Roth et al. 2014), with randomization to prevent re-identification.
- We demonstrate superior results than state-of-the-art methods through extensive evaluation regarding the de-identification rate, utility preservation, and image quality of the resulting data over a large number of metrics.

Related Work

Face Swapping. The topic of face swapping has received significant attention in research and is highly relevant, as evidenced by the large body of works dedicated to it (Nirkin, Keller, and Hassner 2019; Li et al. 2020; Perov et al. 2020; Chen et al. 2020; Zhu et al. 2021; Xu et al. 2022). However, it presents inherent and important differences compared to face anonymization/de-identification. Face swapping aims to change the original identity to a specified target identity, whereas face anonymization shall not rely on actual identities, as it would otherwise compromise both target and source individuals. Moreover, the two domains consider different performance indicators and evaluation metrics. Anonymization aims at providing privacy-preserving guarantees, including face anonymization rate and non-re-identifiability (Gross et al. 2005; Liu et al. 2021; Croft, Sack, and Shi 2021; Tölle et al. 2022), which implies additional mechanisms compared to the face-swapping methods that prioritize preserving facial attributes while reckoning the visual quality of the injected identity (Nirkin, Keller, and Hassner 2019; Xu et al. 2022).

Face Anonymization. Although traditional methods such as blacking out, pixelation, and Gaussian blur (Boyle, Edwards, and Greenberg 2000; Gross et al. 2006, 2009; Neustaedter, Greenberg, and Boyle 2006; Newton, Sweeney, and Malin 2005) are effective in removing privacy-sensitive information, they drastically alter the original data distribution, resulting in a significant loss in *utility*. In other words, these methods generate anonymized images that are not suitable for downstream tasks such as gaze estimation (Kellnhofer et al. 2019; Zhang et al. 2020), head-pose prediction (Zhou and Gregson 2020; Hempel, Abdelrahman, and Al-Hamadi 2022), facial-landmarks regression (Deng et al. 2020; King 2009), and expression estimation (Wen et al. 2021; Savchenko 2022) due to the lack of necessary visual information.

A significant amount of research on face anonymization approaches the problem as an image inpainting task, where the face region is first erased and then replaced with another. Early methods (Gross et al. 2005; Padilla-López, Chaaraoui, and Flórez-Revuelta 2015) use a database of real faces to aggregate the new identity, while more recent methods (Hukkelås, Mester, and Lindseth 2019; Maximov, Elezi, and Leal-Taixé 2020; Liu et al. 2021) use generative models to synthesize fake identities based on the learned distribution. DeepPrivacy (Hukkelås, Mester, and Lindseth 2019) is one of the pioneering works in this field, which reconstructs the missing face by taking the masked face and facial landmarks as inputs. However, the reconstructed face distribution suffers from bias as it is solely conditioned on its training data, leading to a tendency to generate smiling, young-looking faces. CIAGAN (Maximov, Elezi, and Leal-Taixé 2020) is another work that uses facial masks and landmarks to generate new faces. However, it tends to generate faces with duplicated identities due to the length limitation of the one-hot vector. RePaint (Lugmayr et al. 2022), a recent method based on diffusion models, generates photo-realistic faces with large facial variances, but it fails to maintain the utility of the faces and is sensitive to input distributions.

Some methods (Gu et al. 2020; Cao et al. 2021; Proença 2021) have focused on making the anonymization process reversible, such as *Password* (Gu et al. 2020) and *RiDDLE* (Li et al. 2023), which generate anonymized faces conditioned on a password that can be used to de-anonymize them. While such a feature can be desirable in some scenarios, it violates privacy regulations like GDPR (Voigt and Von dem Bussche 2017) that protects *pseudonymous* data (data that has been de-identified from the data's subject but can be re-identified as needed by the use of additional information). In this work, we propose to anonymize faces in an irreversible manner. Other solutions (Li and Lin 2019; Chen et al. 2021; Li and Clifton 2021; Liu et al. 2021) incorporate notions of differential privacy (Duchi, Jordan, and Wainwright 2013; Dwork, Roth et al. 2014; Abadi et al. 2016) by adding adequately-calibrated random noise either at training or inference time, ensuring privacy levels linked to their parameter ϵ. Or directly optimize in the latent space of StyleGAN (Barattin et al. 2023). However, they often neglect utility preservation (*e.g.*, they edit image background and utility attributes) and require complex post-processing, making them not readily applicable to anonymization tasks.

Methodology

In this section, we formalize our objectives, theoretically ground our work, and finally describe our proposed solution.

Problem Formulation

Privacy Utility Dual Optimization. Let $\mathcal{X} \subset \mathbb{R}^{3\times H\times W}$ be the image space, with $x \in \mathcal{X}$ an image depicting an indi-

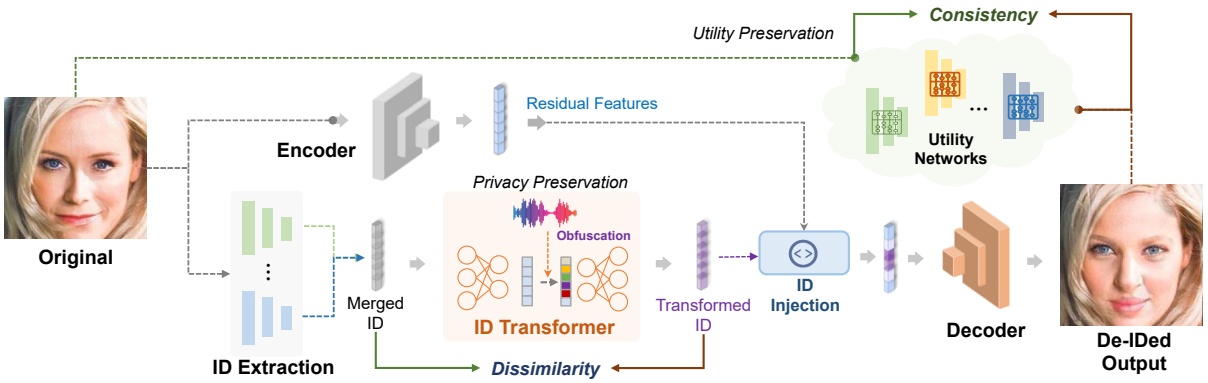


Figure 2: Illustration of the training process for the proposed *Disguise* framework. More discussions in Methodology Section.

vidual. Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space, with $\mathcal{Z} \subset \mathbb{R}^{n_{\mathcal{Z}}}$ space of identity-distilled facial features (*i.e.*, facial features that uniquely identify an individual) and $d_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a distance function attached to space $\mathcal{Z}$. Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be another metric space, with $\mathcal{Y} \subset \mathbb{R}^{n_{\mathcal{Y}}}$ containing utility-distilled facial features (*i.e.*, features that are useful to downstream tasks) and $d_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a distance function relating to $\mathcal{Y}$. We note $f_{\mathcal{Z}} : \mathcal{X} \to \mathcal{Z}$ and $f_{\mathcal{Y}} : \mathcal{X} \to \mathcal{Y}$ the objective labeling functions respective to each domain.

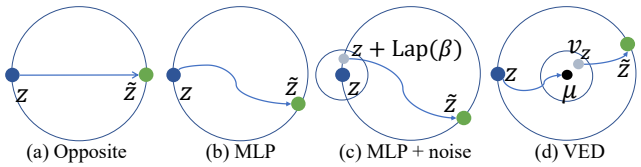
We define a conditional generative function $G : \mathcal{X} \to \mathcal{X}$ parameterized by θ, that takes $x \in \mathcal{X}$ as input and returns an edited version $G(x) = \widetilde{x}$. Our goal is to learn a G such that *utility* is maximized (*i.e.*, $f_{\mathcal{Y}}(x) = f_{\mathcal{Y}}(\widetilde{x})$) and *privacy* is maximized (*i.e.*, $f_{\mathcal{Z}}(x)$ is distant from $f_{\mathcal{Z}}(\widetilde{x})$). In other terms, the output of G should contain the same utility attributes as the input and contain identity attributes different from the input beyond recognition. Formally, we want G to achieve Pareto optimality (Sener and Koltun 2018; Momma, Dong, and Liu 2022) w.r.t. the aforementioned multiple objectives (*i.e.*, identity obfuscation and utility preservation), accounting for their possible competition (depending on downstream tasks, utility and identity attributes may overlap), thus minimizing the following objective:

$$\min_{\theta} \Big(-\mathbb{E}_{x \in \mathcal{X}} \left[d_{\mathcal{Z}} \left(f_{\mathcal{Z}}(x), f_{\mathcal{Z}} \circ G_{\theta}(x) \right) \right], \\ \mathbb{E}_{x \in \mathcal{X}} \left[d_{\mathcal{Y}} \left(f_{\mathcal{Y}}(x), f_{\mathcal{Y}} \circ G_{\theta}(x) \right) \right] \Big)^{\intercal} \quad (1)$$

Before tackling the challenges of multi-objective optimization that such a task brings, one has to consider how to model the unknown objective distance and labeling functions $d_{\mathcal{Z}}, f_{\mathcal{Z}}$ and $d_{\mathcal{Y}}, f_{\mathcal{Y}}$ for the identity and utility space respectively. We argue that identity and utility are conceptually subjective, *i.e.*, different authoritative entities have different definitions and target features assigned to each concept. *e.g.*, given a picture of a person, each human or algorithmic agent will rely on different features (facial landmarks, eye color, *etc.*) and their own subjective judgment to certify the person's identity, as there is no absolute objective function to perform the ill-posed mapping of a facial picture to an identity. Similarly, the concept of *utility* is conditioned by a set of target tasks or the agents in charge of said tasks. *e.g.*, an image with the person's face completely blurred could still be *used* by a person-detection algorithm, but would be *useless* for facial landmark regression.

Therefore, we propose to rely on predefined agents (*experts*) to provide the identity and utility definitions to guide the optimization of our model (Gross et al. 2005). We thus consider some parameterized models $h_{\mathcal{Z}}$ and $h_{\mathcal{Y}}$ pre-optimized to approximate their respective objective labeling functions $f_{\mathcal{Z}}$ and $f_{\mathcal{Y}}$. Note that we make no assumption on the architecture or training of each of these models (we demonstrate with various state-of-the-art identity extraction and recognition models). Without loss of generality and to account for individual bias, we define $H_{\mathcal{Z}} = \left\{ h^i_{\mathcal{Z}} \right\}_{i=1}^{k_{\mathcal{Z}}}$ and $H_{\mathcal{Y}} = \left\{ h^i_{\mathcal{Y}} \right\}_{i=1}^{k_{\mathcal{Y}}}$ as sets of $k_{\mathcal{Z}}$ and $k_{\mathcal{Y}}$ unique models which differ in terms of architecture and/or training regime, *c.f.* mixture-of-experts theory (Miller and Uyar 1996; Masoudnia and Ebrahimpour 2014; Dai et al. 2021). We demonstrate in this paper how these identification/utilization experts can be leveraged in an adversarial/collaborative framework to train g towards a satisfying optimum.

Identity Obfuscation Guarantees. To provide formal de-identification guarantees, we ground our work in the extensive theory on ϵ-differential privacy (ϵ-DP) and ϵ-local-differential privacy (ϵ-LDP, relevant when obfuscation should be performed without global knowledge) applied to identity-swapping functions (Duchi, Jordan, and Wainwright 2013; Dwork, Roth et al. 2014; Abadi et al. 2016; Yu et al. 2020; Liu et al. 2021; Croft, Sack, and Shi 2021; Tölle et al. 2022; Qiu et al. 2022). Let $\psi : \mathcal{Z} \to \mathcal{Z}$ be a function that performs ID obfuscation, *i.e.*, taking an identity vector z and returning a new one $\widetilde{z}$ that maximizes $d_{\mathcal{Z}}(z, \widetilde{z})$. We consider that an approximate but randomized function $\psi^{\epsilon} : \mathcal{Z} \to \mathcal{Z}$ satisfies ϵ-LDP if, for any two adjacent inputs $z, z' \in \mathcal{Z}$ and for any subset of outputs $Z_s \subseteq \text{range}(\psi^{\epsilon})$, it holds that $\mathrm{P}\left(\psi^{\epsilon}(z)\right) \leq e^{\epsilon} \mathrm{P}\left(\psi^{\epsilon}(z')\right)$. Given $\Delta\psi = \sup_{z,z' \in \mathcal{Z}} \|\psi(z) - \psi(z')\|_1$ the sensitivity of ψ, Laplace noise is commonly leveraged to define an ϵ-DP version of the function: $\psi^{\epsilon}(z) \triangleq \psi(z) + \left(\text{Lap}\left(\Delta\psi/\epsilon\right)\right)^{n_{\mathcal{Z}}}$ (Duchi, Jordan, and Wainwright 2013; Dwork, Roth et al. 2014; Abadi et al. 2016; Liu et al. 2021). We demonstrate that to ensure ϵ-LDP, the $d_{\mathcal{Z}}$-maximization property of the identity-obfuscation function has to be relaxed. The manifold of identity vectors generated by an identification function $h_{\mathcal{Z}}$ is bounded by the range of said function. In such



(a) Opposite (b) MLP (c) MLP + noise (d) VED

Figure 3: Identity transformation. The identity vector is normalized to the surface of a unit n-sphere.

a space and for any Euclidean distance $d_{\mathcal{Z}}$, a non-relaxed version of ψ would be the bijective (and thus non-private) function ψ_{opp} mapping an ID vector to its opposite. No other function (*e.g.*, ψ^{ϵ}) could ensure $d_{\mathcal{Z}}$-maximization. Therefore, in this work, we consider the inherent trade-off between maximizing swapping-based identity obfuscation and ensuring differential privacy, and we propose a variety of solutions ψ^{ϵ} tailored to different needs (as illustrated in Figure 3, and more details in Proposed Solution Section).

Non Re-identifiability. Another important aspect to consider in privacy-preserving applications is *non-invertibility*. If the de-identified data can be re-identified with additional information, then the operation is not truly anonymization but *pseudonymization*. For example, with the correct password for Password (Gu et al. 2020) and RiDDLE (Li et al. 2023), or using the opposite ID for ψ_{opp}, the original ID is compromized. We empirically demonstrate that the proposed obfuscation solutions achieve varying degrees of robustness to such re-identification efforts.

In the remaining of the section, we explain how we define and train g to ensure privacy-preserving non-invertible identity swapping in images and utility preservation.

Proposed Solution

The proposed architecture can be defined as the composition of a face-swapping model $g : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$, an identity extractor $h_{\mathcal{Z}} : \mathcal{X} \to \mathcal{Z}$, and an identity obfuscation function $\psi^{\epsilon} : \mathcal{Z} \to \mathcal{Z}$, such that $G(x) = g\left(x, \psi^{\epsilon} \circ h_{\mathcal{Z}}(x)\right)$. Given a facial image x, $h_{\mathcal{Z}}$ extracts the vector z encoding the identity of the depicted person. This vector z is passed to the privacy-enabling function ψ^{ϵ}, which returns a synthetic identity $\widetilde{z}$ that maximizes obfuscation. Finally, the face-swapper model g edits the original image x to inject the fake identity $\widetilde{z}$, resulting in an image $\widetilde{x}$ where the original visual identifying attributes are replaced by those encoded in $\widetilde{z}$. Additionally, during its training, g relies on the feedback of tasks-specific models $h^i_{\mathcal{Y}} : \mathcal{X} \to \mathcal{Y}$ to ensure that the utility of $\widetilde{x}$ is maintained compared to x. We expand on each block in the following paragraphs.

Identity Extraction. As mentioned in Section , we propose to extract the identity information from facial images via model ensembling (Miller and Uyar 1996; Masoudnia and Ebrahimpour 2014; Dai et al. 2021), to ensure generalizability as well as to limit the impact of models' bias (as we assume no control over the architecture or training regimen of selected identity-expert models). Therefore, given a set $H_{\mathcal{Z}} = \left\{h^i_{\mathcal{Z}}\right\}_{i=1}^{k_{\mathcal{Z}}}$ of ID extractors, we define $h_{\mathcal{Z}}$ as the ensemble method $h_{\mathcal{Z}}(x) = \text{MLP}_{\theta_z}\left(\left\|_{i=1}^{k_{\mathcal{Z}}} h^i_{\mathcal{Z}}(x)\right.\right)$, *i.e.*, concatenating (symbol $\|$) the $k_{\mathcal{Z}}$ predicted vectors together and merging them into $z \in \mathcal{Z}$ via a multilayer perceptron (MLP) with parameters θ_z and $\tanh$ as final activation.

Identity Transformation. A variety of techniques can be considered to perform the identity transformation ψ, as shown in Figure 3. If we were to maximize the distance between the original and obfuscating IDs, the optimal function would be $\psi_{\text{opp}}(z) = -z$ since $\arg\max_{z'} d_{\mathcal{Z}}(z, z') = -z$ in our normalized Euclidean identity space. However, such a function is reversible, making it easy to re-identify the original individual by taking the opposite of the pseudonymized ID. A more secure solution would be a parametric function, *e.g.*, $\psi_{\text{mlp}}(z) = \text{MLP}_{\theta_\psi}(z)$, trained to optimally fool $h_{\mathcal{Z}}$. As a non-explicit function, ψ_{mlp} is more challenging to invert, though not impossible with the access to the model or its parameters θ_ψ (*c.f.* gradient-based attacks (Fredrikson, Jha, and Ristenpart 2015; Wang, Si, and Wu 2015)). To increase robustness and ensure ϵ-LDP, we can add dimension-wise noise to the inner operation, *i.e.*, $\psi^{\epsilon}_{\text{mlp}}(z) = \text{MLP}_{\theta_\psi}\left(z + (\text{Lap}\,(\beta))^{n_{\mathcal{Z}}}\right)$, with $\beta = \frac{\Delta\psi_{\text{mlp}}}{\epsilon}$. The larger β is set (*i.e.*, the smaller ϵ is), the more noise is applied to the original ID vector before further MLP-based obfuscation. Therefore, larger β provides stricter privacy guarantee and robustness but adversarial affects the ability of $\psi^{\epsilon}_{\text{mlp}}$ to learn how to fool identification experts $H_{\mathcal{Z}}$.

To better navigate this trade-off and guarantee a more continuous space for the noise application, we leverage the properties inherent to variational autoencoders (VAEs) (Kingma and Welling 2013). We introduce a variational encoder-decoder (VED) to transform the identity vector, *i.e.*, $\psi^{\epsilon}_{\text{ved}}(z) = \text{VED}_{\theta_{\psi'}}(z)$. This model's encoder predicts the parameters (μ, σ) of the latent data distribution (assumed to be Gaussian). A latent vector v_z is picked as $\mu + \sigma\eta$ with $\eta \sim (\mathcal{N}(0,1))^{n_v}$ (*c.f.* reparameterization (Kingma, Salimans, and Welling 2015)) then passed to the decoder. While a VAE decoder would reconstruct the input identity from v_z, our VED decoder should generate a new, distant identity. During inference, we sample v_z as $\mu + \sigma\,(\text{Lap}\,(\alpha))^{n_v}$ to meet ϵ-LDP, with n_v dimension of latent space and $\alpha = \frac{\Delta\psi_{\text{ved}}}{\epsilon}$. To train either of these models, we enforce cosine dissimilarity between the original and generated ID vectors:

$$\mathcal{L}_{\text{deid}} = 1 + \frac{z \cdot \widetilde{z}}{\|z\|_2 \|\widetilde{z}\|_2}. \tag{2}$$

For the VED model, we add to this criterion the usual Kullback–Leibler divergence (KLD) loss $\mathcal{L}_{\text{kld}}$ (Kingma, Salimans, and Welling 2015; Kingma and Welling 2013).

Face Swapping. Once the fake identity vector $\widetilde{z}$ is generated, it is passed to the face-swapping model g, along with the original image x. Similar to existing solutions (Chen et al. 2020; Perov et al. 2020; Liu et al. 2021), g is composed of three modules: (1) an image encoder that extracts identity-unrelated features ν; (2) an ID injector that aggregates ν and $\widetilde{z}$ into a vector encoding the content of the obfuscated image $\widetilde{x}$; (3) a decoder conditioned on this vector that generates $\widetilde{x}$. These existing works also share similar losses that we borrow and adapt:

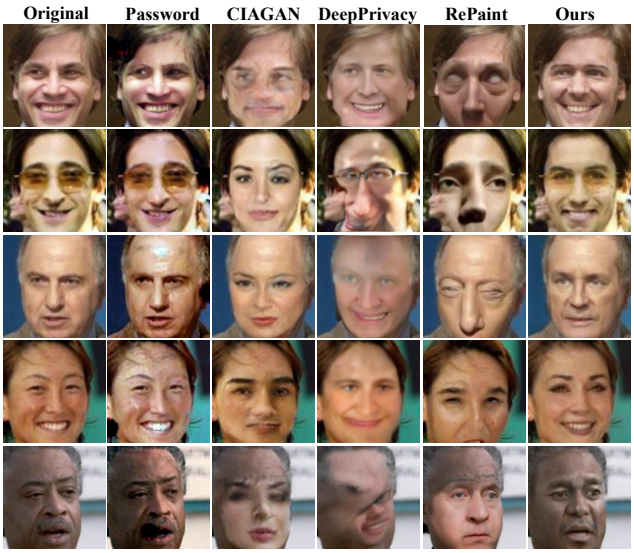


Figure 4: Qualitative results of different methods. Ours preserves utility while anonymizing identities.

$$\mathcal{L}_{\text{mix}} = \|g(x,\widetilde{z}) - g(x,z)\|_1; \tag{3}$$

$$\mathcal{L}_{\text{gen}} = \sum_{i=1}^{k_d} \log\left(1 - D_i(x,\widetilde{x})\right); \tag{4}$$

$$\mathcal{L}_{\text{id}} = \sum_{\hat{z}\in\{z,\widetilde{z}\}} \Big(1 - \frac{\hat{z}\cdot\hat{z}_h}{\|\hat{z}\|_2\|\hat{z}_h\|_2}\Big); \tag{5}$$

with $\hat{z}_h = h_{\mathcal{Z}}\left(g(x,\hat{z})\right)$. Here, $\mathcal{L}_{\text{mix}}$ is a mixing loss to ensure implicit disentanglement of ID features (encoded in z or $\widetilde{z}$) and residual features (*i.e.*, ν). $\mathcal{L}_{\text{gen}}$ pits the generator against k_d discriminators D to ensure realistic results preserving image saliency, *c.f.* recent GAN solutions (Wang et al. 2018; Hukkelås, Mester, and Lindseth 2019; Chen et al. 2020) (we also use their weak-feature matching loss, further ensuring the high-level semantic alignment between the image pairs). Finally, $\mathcal{L}_{\text{id}}$ enforces cosine similarity between the injected identity $\hat{z}$ and the one observed by the identification model $h_{\mathcal{Z}}$ in the resulting image. Combined together, along with $\mathcal{L}_{\text{deid}}$ and $\mathcal{L}_{\text{kld}}$ (using weighting hyperparameters), these losses form the overall objective for our privacy-enforcing face-swapping solution G.

Utility Preservation. Existing face-swapping methods (Hukkelås, Mester, and Lindseth 2019; Chen et al. 2020; Perov et al. 2020; Liu et al. 2021) claim that their adversarial and feature-matching losses ensure the preservation of non-identifying content. However, such supervision is too weak to guarantee that the images will maintain their utility w.r.t. downstream tasks, especially for tasks relying on small attention regions (*e.g.*, gaze estimation). We thus complement the aforementioned objective with a criterion that leverages the implicit expertise of tasks-relevant models $H_{\mathcal{Y}}$, as $\mathcal{L}_{\text{uti}} = \sum_{i=1}^{k_{\mathcal{Y}}} \lambda_{\text{uti},i} \|h^{i,l}_{\mathcal{Y}}(x) - h^{i,l}_{\mathcal{Y}}(\widetilde{x})\|_1$, with $h^{i,l}_{\mathcal{Y}}(\cdot)$ the features returned by the last differential non-softmax layer l of model $h^i_{\mathcal{Y}}$, and $\lambda_{\text{uti}} \in \mathbb{R}^{k_{\mathcal{Y}}}$ hyperparameters weighting the task/expert contributions. Hence, $\mathcal{L}_{\text{uti}}$ imposes that altered images contains the same utility attributes as original images, as expected by tasks-relevant models.

Note that the entire solution $G(x) = g\left(x, \psi^{\epsilon} \circ h_{\mathcal{Z}}(x)\right)$ is end-to-end differentiable, thus single-pass trainable. In practice, we leverage its modularity and train each component separately before jointly fine-tuning. Scalar hyperparameters weigh the contribution of each loss to the total objective (we fix $\{\lambda_{\text{id}}, \lambda_{\text{deid}}, \lambda_{\text{mix}}, \lambda_{\text{uti,eye}}, \lambda_{\text{uti,emo}}, \lambda_{\text{kld}}\} = \{30, 30, 10, 2, 2, 0.2\}$).

Experiments

We now describe our experimental setup and compare with other methods in terms of privacy robustness and data usability. More details in supplementary material.

Experimental Protocol

Datasets. We use multiple datasets for training and evaluation. We train our models on VGGFace2 dataset (Cao et al. 2018), which totals 3.31 million images with 9,131 identities. Evaluation datasets include LFW (Huang et al. 2008) (13,233 face images and 5,749 identities) for utility and de-identification performance, CelebA-HQ (Karras et al. 2017) (30,000 face images) for utility evaluation, and WFLW (Wu et al. 2018) (10,000 face images) for the training usability w.r.t. the downstream task of landmark detection.

Identity and Utility Models. To demonstrate the genericity of our method, we consider a variety of pretrained face-identification networksand of utility networks over different recognition tasks. As identity experts, we use ArcFace (Deng et al. 2019), AdaFace (Kim, Jain, and Liu 2022), FaceNet (Schroff, Kalenichenko, and Philbin 2015), and SphereFace (Liu et al. 2017). Either ArcFace ($h^{\text{arc}}_{\mathcal{Z}}$), AdaFace ($h^{\text{ada}}_{\mathcal{Z}}$), or both ($h^{\text{mix}}_{\mathcal{Z}}$) are used to guide g during training (*c.f.* Equation **??**); FaceNet and SphereFace are used only for evaluation. For the downstream tasks, we use ETH-XGaze (Zhang et al. 2020) (noted $h^{\text{eye}}_{\mathcal{Y}}$) for gaze estimation, DAN (Wen et al. 2021) ($h^{\text{emo}}_{\mathcal{Y}}$) for facial expression recognition, or both ($h^{\text{mix}}_{\mathcal{Y}}$) to provide utility feedback during training. During evaluation, we use L2CS-Net (Abdelrahman et al. 2022) for gaze estimation, DeepFace (DF) (Serengil and Ozpinar 2021) for emotion recognition, and RetinaFace (Deng et al. 2020) and Dlib (King 2009) for landmark detection.

Metrics. We employ the commonly-used validation rate and verification accuracy as metrics for evaluating privacy preservability (Schroff, Kalenichenko, and Philbin 2015; Liu et al. 2017; Deng et al. 2019; Kim, Jain, and Liu 2022). The validation rate is defined as the true positive rate (TPR) at certain false positive rate (FPR), *e.g.*, TPR @ FPR=1e-3. Verification accuracy is the percentage of image pairs correctly classified as the same/different person using the best ℓ_2 distance threshold. The verification accuracy of random guessing is thus 50%, which is what anonymization aims at. To measure utility preservation, we use ℓ_2 pixel distance and normalized mean error (NME) for facial landmark detection, mean absolute error (MAE) for gaze estimation, and accuracy for emotion recognition. For image quality, we use SER-FIQ (Terhorst et al. 2020).

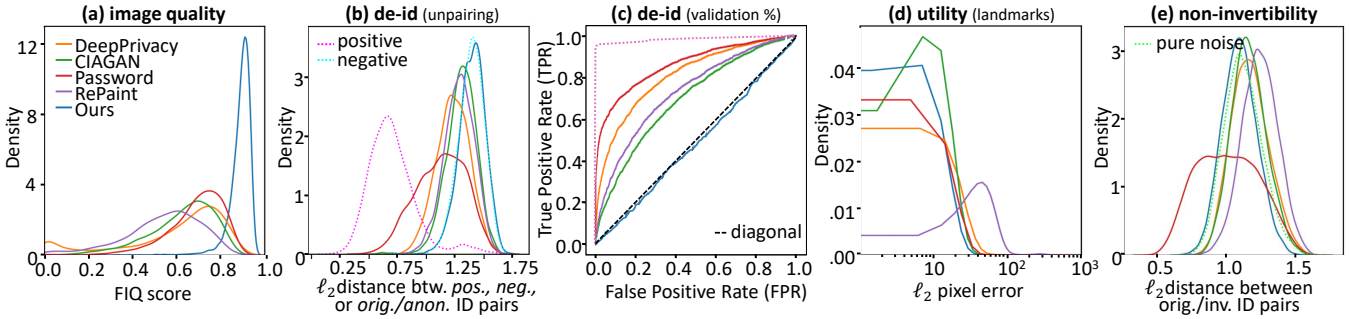


Figure 5: *Disguise* outperforms existing methods in various aspects, including image quality, de-id rate, and utility. For non-invertibility, our solution is close to other methods that completely erase the original IDs (*i.e.*, recovering pure Gaussian noise).

Methods	TPR (%) @ FPR=10^{-3} / Accuracy (%) ↓				FIQ ↑
	FaceNet	Sph.Face	AdaFace	avg	SER
Original	93.8 / 97.1	87.9 / 96.2	95.4 / 97.7	92.4 / 97.0	.77
DeepPriv	7.3 / 73.8	2.9 / 70.9	4.6 / 68.6	4.9 / 71.1	.67
DeepPriv.2	1.7 / 62.5	1.0 / 61.5	2.2 / 62.2	1.6 / 62.1	.68
CIAGAN	1.8 / 64.5	1.0 / 59.0	5.6 / 71.0	2.8 / 64.8	.58
Password	31.7 / 79.1	17.1 / 73.5	51.0 / 84.0	33.3 / 78.9	.69
RePaint	2.8 / 67.7	1.1 / 63.5	3.6 / 68.5	2.5 / 66.6	.54
Ours	**0.03 / 50.0**	**0.03 / 50.0**	**0.00 / 50.0**	**0.02 / 50.0**	**.90**

Table 1: Identification / validation rate and image quality evaluation over edited LFW data.

Comparison. We consider various de-identification methods, including DeepPrivacy (Hukkelås, Mester, and Lindseth 2019), DeepPrivacy2 (Hukkelås and Lindseth 2023), CIAGAN (Maximov, Elezi, and Leal-Taixé 2020), Password (Gu et al. 2020), and RePaint (Lugmayr et al. 2022). For readibility of the tables, we denote different versions as "Ours (a, b, c)" where a fixes the identity model(s) $h^a_{\mathcal{Z}}$ used, b the transformation function ψ^{ϵ}_b, and c the utility model(s) $h^c_{\mathcal{Y}}$. For simplicity, we use "Ours (arc, ved, eye)" as our default method unless otherwise mentioned. We demonstrate the impact of different transformation models and identity/utility experts in ablation studies.

Privacy: Obfuscation Evaluation

De-identification performance. As shown in Table 1, we achieve near perfect de-id rate, *i.e.*, with a validation rate close to 0 and verification accuracy close to 50%, outperforming other methods by a significant margin, and is even more secure than randomly picking replacement images from the dataset. Figure 5(b) presents the ℓ_2 distance histogram for original positive pairs, original negative pairs, and original-anonymized positive pairs on LFW (Huang et al. 2008), and Figure 5(c) shows the ROC curves of validation rate. We observe that *Disguise* creates image pairs that are close to the negative distribution, hence perfect obfuscation. We also achieve the highest facial image quality, see Figures 1 and 4 for visual reference. Among other comparing methods, it is worth noticing that Password fails to de-identify images, hence the highest validation rate. CIAGAN and RePaint are better than Password in de-identification, however they suffer from low facial image quality due to high artifacts and distortions.

Original and anonymized ID de-correlation. We consider scenarios where malicious attackers attempt to link anonymized IDs with their original IDs, allowing them to perform inversion inference on the anonymized IDs and recover the original ones. We use encoder-decoder networks to learn the correlation on existing original-anonymized image pairs. Figure 5(e) shows the results of using MLPs to decode obfuscated IDs from CelebA-HQ (Karras et al. 2017) while trained on LFW (Huang et al. 2008) using original IDs as supervision. While methods like DeepPrivacy, CIAGAN, and RePaint are inherently robust to inversion attacks since the original face region is entirely erased, and their networks are solely tasked with inpainting the blank region, our method still offers de-correlation on par with these methods, suggesting that our method is also resilient to inversion attacks.

Utility: Usability Evaluation

Utility corruption in anonymized images. Our method demonstrates superior utility preservation compared to others across datasets (Table 2). We highlight our approach's excellence through qualitative comparison (Figure 1 and 4). DeepPrivacy lacks facial attribute preservation, exhibiting bias towards smiles and youth. CIAGAN bears heavy artifacts; Password yields blurry and easily re-identifiable outcomes. RePaint excels with in-distribution faces (RePaint is trained on CelebA-HQ thus has improved performance on the same dataset in Table 2), but it fails elsewhere and doesn't retain original attributes. For challenging scenarios, like heavy occlusion (*e.g.*, masks), CIAGAN and DeepPrivacy falter, unlike our effective face-swapping model.

Usability of anonymized images as training data. We have demonstrated utility attribute non-corruption by comparing performance of pretrained task-specific models on obfuscated versus original data. Now, we advance toward the initial motivation of data anonymization for new solutions, evaluating how utility networks trained from scratch on anonymized data perform on real, unseen samples. Ideally, these privacy-preprocessed models should match performance of those trained on original, non-obfuscated data. Taking facial landmark detection as an example on the WFLW dataset (Wu et al. 2018) (98 landmarks per image), we split data into training/testing sets (7,500/2,500) and generate obfuscated training data using mentioned methods (test data remains unaltered). We use an HRNetv2-W18

Dataset	Methods	Facial landmarks (L2 pixel distance ↓)								Gaze estimation (MAE ↓)				Emotion	
		RetinaFace (5 points)				Dlib (68 points)				L2CS-Net		ETH-XGaze		(Acc. % ↑)	
		All	Eyes	Nose	Mouth	All	Eyes	Nose	Mouth	Pitch	Yaw	Pitch	Yaw	DAN	DF
LFW	DeepPriv	23.9	13.1	9.9	16.5	263.0	32.7	25.1	89.0	7.7	13.6	8.0	16.3	27.1	34.3
	DeepPriv2	31.2	18.4	14.4	19.6	385.7	59.9	49.9	120.6	9.2	12.2	7.8	15.1	22.4	30.2
	CIAGAN	14.6	9.3	5.5	9.2	348.2	59.0	31.2	97.7	8.8	14.6	7.8	16.9	32.5	36.9
	Password	17.4	10.4	7.7	11.1	204.8	**26.5**	**19.3**	**55.4**	10.5	24.7	7.7	11.5	45.9	43.4
	RePaint	66.1	30.8	32.2	47.3	1103.1	133.5	152.1	432.0	11.3	18.1	9.2	18.8	17.3	19.4
	Ours	**12.9**	**7.7**	**5.6**	**8.2**	**203.3**	28.8	19.8	60.0	**6.8**	**8.4**	**5.6**	**10.0**	**46.2**	**47.0**
CelebAHQ	DeepPriv	13.1	4.8	4.3	10.8	293.9	30.4	24.2	113.0	7.0	8.7	6.7	10.1	41.0	45.5
	CIAGAN	14.9	10.3	4.6	8.6	365.6	79.2	35.6	91.5	8.7	13.7	7.5	13.4	38.0	44.4
	RePaint	9.9	**3.0**	4.6	7.5	249.1	**22.7**	29.7	95.9	6.2	8.0	5.5	8.0	50.4	55.8
	Ours	**6.7**	3.4	**3.3**	**4.3**	**196.0**	25.1	**20.0**	**59.9**	**5.6**	**6.2**	**4.6**	**5.9**	**61.9**	**59.6**

Table 2: Utility performance of anonymization methods over diverse downstream tasks on LFW and CelebA-HQ datasets.

Methods	Normalized Mean Error (NME) ↓						
	all	pose	illu	occ	blur	mu	exp
Original	.039	.068	.039	.047	.045	.038	.043
DeepPrivacy	.058	.100	.057	.072	.066	.060	.066
CIAGAN	.055	.087	.054	.064	.061	.053	.060
Ours	**.047**	**.079**	**.047**	**.056**	**.054**	**.046**	**.050**

Table 3: Usability of de-identified datasets for the training of task-specific models (facial landmark detection on WFLW).

model (Wang et al. 2021) for the task, trained for 60 epochs with Adam optimizer (Kingma and Ba 2014) ($\beta_1 = 0$, $\beta_2 = 0.999$), learning rate 10^{-4}, and batch size 64. Table 3 shows models on obfuscated data perform worse (higher NME of facial landmarks) than the one on original data. Our anonymized data model demonstrates the smallest accuracy drop, confirming higher utility preservation for downstream tasks while maintaining privacy.

Ablation Study

Here we demonstrate the impact of different transformation models, identity and utility experts.

Impact of transformation models on re-identifiability. As justified in Section and experimentally measured in Table 4, ψ_{opp} would suffer high re-identification, *i.e*. we can recover the original ID using the opposite of transformed ID. MLP-based transformations outperforms opposite transformation but VED-based transformations yield the best results in terms of de-identification and non-invertibility, confirming the superiority of our proposed solution.

The introduction of stochastic operations in alignment with ϵ-LDP further strengthens the method. As shown in Table 5, the higher the amount of β or α noise introduced (*i.e*., the lower ϵ), the more robust to attacks the method becomes, but the lower the original de-id rate (the noisier the data, the harder it is to synthesize an ID that maximizes obfuscation). This negative impact is however better mitigated by the proposed VED. We provide further insights in annex.

Effects of using multiple ID extractors. As shown in Table 4, MLP transforms relying on multiple identity extractors, *i.e*., "(mix, mlp, eye)", outperform versions with only one ID expert. We attribute the increased robustness to the combined knowledge of the two algorithms which capture more varied ID-related features that are then obfuscated.

Proposed Solutions *(id, trans, util)*	TPR (%) @ FPR=1e-3 ↓ (LFW data)			
	Swapped		Inverted	
	FaceNet	Sph.Face	FaceNet	Sph.Face
(arc, opp, Ø)	0.63	0.03	67.03	53.07
(arc, ved, emo)	0.23	**0.00**	**12.03**	7.10
(arc, ved, eye)	0.03	0.03	13.03	**6.43**
(arc, mlp, eye)	**0.00**	**0.00**	52.90	45.97
(arc, mlp, emo)	0.03	**0.00**	49.77	45.97
(arc, mlp, mix)	**0.00**	**0.00**	50.23	44.67
(mix, mlp, eye)	0.07	**0.00**	36.70	34.70

Table 4: Re-identifiability of our ID transformation methods.

Methods		TPR (%) @ FPR=1e-3 ↓ (LFW data)			
		Swapped		Inverted	
Net	Noise	FaceNet	Sph.Face	FaceNet	Sph.Face
MLP	$\beta = 0.0$	**0.00**	**0.00**	52.90	45.97
	$\beta = 0.5$	0.40	**0.00**	23.20	20.80
	$\beta = 0.9$	5.90	2.50	**3.07**	**1.30**
VED	$\alpha = 1.0$	**0.03**	0.03	13.03	6.43
	$\alpha = 2.0$	0.37	**0.00**	7.43	3.20
	$\alpha = 3.0$	0.37	0.10	**5.37**	**2.23**

Table 5: Effect of $\psi^{\epsilon}_{\cdot}$ noise w.r.t. (re-)identifiability.

Conclusion and Discussion

We introduced *Disguise*, a privacy-enhancing face de-identification model that ensures both depicted people's privacy and image usability. Our experiments demonstrate its effectiveness in pre-processing sensitive data for inference or training. Rooted in privacy and mixture-of-experts theory, it outperforms prior methods in re-identification robustness and utility preservation.

Limitations. Our model, primarily for face obfuscation, doesn't address other unique identifiers like glasses or haircuts. ID-extracting methods like $H_{\mathcal{Z}}$ (Bhanu, Kumar et al. 2017) or multi-objective learning (Désidéri 2012; Momma, Dong, and Liu 2022) might improve recognition of overlapping identity and utility features.

Acknowledgments

Z. Cai and M. Asif were supported in part by AFOSR award FA9550-21-1-0330 and ONR award N00014-19-1-2264.

References

2003. Health Insurance Portability and Accountability Act. U.S. Department of Health and Human Services. 45 CFR Parts 160, 162, and 164.

2018. California Consumer Privacy Act. California Legislative Information. Cal. Civ. Code §1798.100 et seq.

2021. Personal Information Protection Law. National People's Congress of the People's Republic of China.

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *ACM CCS*.

Abdelrahman, A. A.; Hempel, T.; Khalifa, A.; and Al-Hamadi, A. 2022. L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments. *arXiv:2203.03339*.

Agarwal, A.; Chattopadhyay, P.; and Wang, L. 2021. Privacy preservation through facial de-identification with simultaneous emotion preservation. *SIVP*, 15(5).

Barattin, S.; Tzelepis, C.; Patras, I.; and Sebe, N. 2023. Attribute-preserving Face Dataset Anonymization via Latent Code Optimization. In *CVPR*.

Bhanu, B.; Kumar, A.; et al. 2017. *Deep learning for biometrics*, volume 7. Springer.

Boyle, M.; Edwards, C.; and Greenberg, S. 2000. The effects of filtered video on awareness and privacy. In *CSCW*.

Cao, J.; Liu, B.; Wen, Y.; Xie, R.; and Song, L. 2021. Personalized and Invertible Face De-identification by Disentangled Identity Information Manipulation. In *ICCV*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

Chen, J.-W.; Chen, L.-J.; Yu, C.-M.; and Lu, C.-S. 2021. Perceptual Indistinguishability-Net (PI-Net): Facial image obfuscation with manipulable semantics. In *CVPR*.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. Simswap: An efficient framework for high fidelity face swapping. In *ACM International Conference on Multimedia*.

Croft, W. L.; Sack, J.-R.; and Shi, W. 2021. Obfuscation of images via differential privacy: From facial images to general images. *P2PNA*, 14(3).

Dai, Y.; Li, X.; Liu, J.; Tong, Z.; and Duan, L.-Y. 2021. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*.

Désidéri, J.-A. 2012. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6).

Duchi, J. C.; Jordan, M. I.; and Wainwright, M. J. 2013. Local privacy and statistical minimax rates. In *FOCS*. IEEE.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4).

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM CCS*.

Frome, A.; Cheung, G.; Abdulkader, A.; Zennaro, M.; Wu, B.; Bissacco, A.; Adam, H.; Neven, H.; and Vincent, L. 2009. Large-scale privacy protection in google street view. In *ICCV*.

Gross, R.; Airoldi, E.; Malin, B.; and Sweeney, L. 2005. Integrating utility into face de-identification. In *International Workshop on Privacy Enhancing Technologies*. Springer.

Gross, R.; Sweeney, L.; Cohn, J.; Torre, F. d. l.; and Baker, S. 2009. Face de-identification. In *Protecting privacy in video surveillance*. Springer.

Gross, R.; Sweeney, L.; De la Torre, F.; and Baker, S. 2006. Model-based face de-identification. In *CVPR workshop*.

Gu, X.; Luo, W.; Ryoo, M. S.; and Lee, Y. J. 2020. Password-conditioned anonymization and deanonymization with face identity transformers. In *ECCV*. Springer.

Hempel, T.; Abdelrahman, A. A.; and Al-Hamadi, A. 2022. 6D Rotation Representation For Unconstrained Head Pose Estimation. *arXiv:2202.12555*.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images*.

Hukkelås, H.; and Lindseth, F. 2023. Deepprivacy2: Towards realistic full-body anonymization. In *WACV*.

Hukkelås, H.; Mester, R.; and Lindseth, F. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *ISVC*. Springer.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv:1710.10196*.

Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*.

Kim, M.; Jain, A. K.; and Liu, X. 2022. AdaFace: Quality Adaptive Margin for Face Recognition. In *CVPR*.

King, D. E. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Kingma, D. P.; Salimans, T.; and Welling, M. 2015. Variational dropout and the local reparameterization trick. *NeurIPS*, 28.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv:1312.6114*.

Korshunov, P.; and Ebrahimi, T. 2013. Using warping for privacy protection in video surveillance. In *DSP*, 1–6. IEEE.

Li, D.; Wang, W.; Zhao, K.; Dong, J.; and Tan, T. 2023. RiDDLE: Reversible and Diversified De-identification with Latent Encryptor. In *CVPR*.

Li, L.; Bao, J.; Yang, H.; Chen, D.; and Wen, F. 2020. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *CVPR*.

Li, T.; and Clifton, C. 2021. Differentially private imaging via latent space manipulation. In *SP*.

Li, T.; and Lin, L. 2019. Anonymousnet: Natural face de-identification with measurable privacy. In *CVPR workshop*.

Liu, B.; Ding, M.; Xue, H.; Zhu, T.; Ye, D.; Song, L.; and Zhou, W. 2021. Dp-image: differential privacy for image data in feature space. *arXiv:2103.07073*.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *CVPR*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*.

Masoudnia, S.; and Ebrahimpour, R. 2014. Mixture of experts: a literature survey. *ARTR*, 42(2).

Maximov, M.; Elezi, I.; and Leal-Taixé, L. 2020. Ciagan: Conditional identity anonymization generative adversarial networks. In *CVPR*.

Miller, D. J.; and Uyar, H. 1996. A mixture of experts classifier with learning based on both labelled and unlabelled data. *NeurIPS*, 9.

Momma, M.; Dong, C.; and Liu, J. 2022. A multi-objective/multi-task learning framework induced by Pareto stationarity. In *ICML*. PMLR.

Neustaedter, C.; Greenberg, S.; and Boyle, M. 2006. Blur filtration fails to preserve privacy for home-based video conferencing. *TOCHI*, 13(1).

Newton, E. M.; Sweeney, L.; and Malin, B. 2005. Preserving privacy by de-identifying face images. *TKDE*, 17(2).

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*.

Padilla-López, J. R.; Chaaraoui, A. A.; and Flórez-Revuelta, F. 2015. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9).

Perov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Umé, C.; Dpfks, M.; Facenheim, C. S.; RP, L.; Jiang, J.; et al. 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv:2005.05535*.

Proença, H. 2021. The uu-net: Reversible face de-identification for visual surveillance video footage. *TCSVT*, 32(2).

Qiu, Y.; Niu, Z.; Song, B.; Ma, T.; Al-Dhelaan, A.; and Al-Dhelaan, M. 2022. A Novel Generative Model for Face Privacy Protection in Video Surveillance with Utility Maintenance. *Applied Sciences*, 12(14): 6962.

Savchenko, A. V. 2022. Video-Based Frame-Level Facial Analysis of Affective Behavior on Mobile Devices Using EfficientNets. In *CVPR*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *NeurIPS*, 31.

Serengil, S. I.; and Ozpinar, A. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *ICEET*.

Terhorst, P.; Kolf, J. N.; Damer, N.; Kirchbuchner, F.; and Kuijper, A. 2020. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *CVPR*.

Tölle, M.; Köthe, U.; André, F.; Meder, B.; and Engelhardt, S. 2022. Content-Aware Differential Privacy with Conditional Invertible Neural Networks. In *DeCaF*. Springer.

Voigt, P.; and Von dem Bussche, A. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed.*

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; Liu, W.; and Xiao, B. 2021. Deep High-Resolution Representation Learning for Visual Recognition. *TPAMI*, 43(10).

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*.

Wang, Y.; Si, C.; and Wu, X. 2015. Regression model fitting under differential privacy and model inversion attack. In *IJCAI*.

Wen, Z.; Lin, W.; Wang, T.; and Xu, G. 2021. Distract your attention: multi-head cross attention network for facial expression recognition. *arXiv:2109.07270*.

Westerlund, M. 2019. The emergence of deepfake technology: A review. *TIM Review*, 9(11).

Wu, W.; Qian, C.; Yang, S.; Wang, Q.; Cai, Y.; and Zhou, Q. 2018. Look at Boundary: A Boundary-Aware Face Alignment Algorithm. In *CVPR*.

Xu, Y.; Deng, B.; Wang, J.; Jing, Y.; Pan, J.; and He, S. 2022. High-resolution face swapping via latent semantics disentanglement. In *CVPR*.

Yu, J.; Xue, H.; Liu, B.; Wang, Y.; Zhu, S.; and Ding, M. 2020. Gan-based differential private image privacy protection framework for the internet of multimedia things. *Sensors*, 21(1).

Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; and Hilliges, O. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In *ECCV*.

Zhou, J.; and Pun, C.-M. 2020. Personal privacy protection via irrelevant faces tracking and pixelation in video live streaming. *TIFS*, 16.

Zhou, Y.; and Gregson, J. 2020. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv:2005.10353*.

Zhu, Y.; Li, Q.; Wang, J.; Xu, C.-Z.; and Sun, Z. 2021. One shot face swapping on megapixels. In *CVPR*.