Machine Pareidolia: Protecting Facial Image with Emotional Editing

Binh M. Le, Simon S. Woo*

Sungkyunkwan University, Suwon, South Korea
bmle@g.skku.edu, swoo@g.skku.edu

Abstract

The proliferation of facial recognition (FR) systems has raised privacy concerns in the digital realm, as malicious uses of FR models pose a significant threat. Traditional countermeasures, such as makeup style transfer, have suffered from low transferability in black-box settings and limited applicability across various demographic groups, including males and individuals with darker skin tones. To address these challenges, we introduce a novel facial privacy protection method, dubbed **MAP**, a pioneering approach that employs human emotion modifications to disguise original identities as target identities in facial images. Our method uniquely fine-tunes a score network to learn dual objectives, target identity and human expression, which are jointly optimized through gradient projection to ensure convergence at a shared local optimum. Additionally, we enhance the perceptual quality of protected images by applying local smoothness regularization and optimizing the score matching loss within our network. Empirical experiments demonstrate that our innovative approach surpasses previous baselines, including noise-based, makeup-based, and freeform attribute methods, in both qualitative fidelity and quantitative metrics. Furthermore, **MAP** proves its effectiveness against an online FR API and shows advanced adaptability in uncommon photographic scenarios.

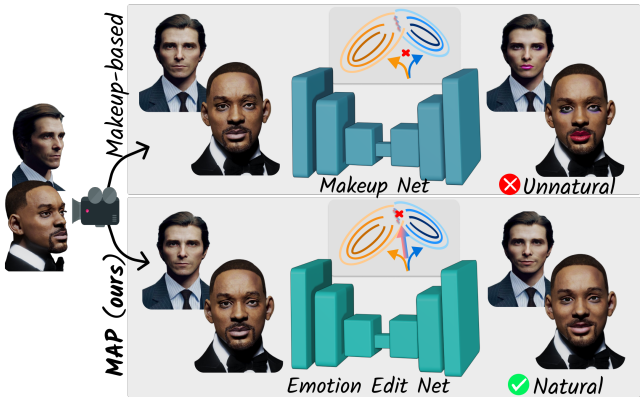

Figure 1: Comparison of makeup-based baselines with our **MAP**. Baseline methods (top) use makeup styles obscure original identities, optimizing objectives independently, which reduces efficiency and limits applicability across demographics (*e.g.*, males). In contrast, **MAP** (bottom) leverages human emotion modifications and a unified optimization strategy to disguise original identities as target identities, achieving universal robustness.

Introduction

Recent advances in deep learning-based face recognition (FR) systems have enabled their widespread adoption in applications like biometrics (Meden et al. 2021), security (Wang et al. 2017), and criminal investigation (Phillips et al. 2018). However, these advancements also pose significant privacy risks in the digital realm. Malicious uses of FR, such as unauthorized surveillance (Wenger et al. 2023; Hill 2024), tracking relationships, and monitoring activities on social platforms (Hill 2022; Shoshitaishvili, Kruegel, and Vigna 2015), highlight the urgent need for effective methods to protect facial images from unauthorized FR systems.

An effective facial privacy protection method should achieve an optimal balance between maintaining *natural appearance* and ensuring *robust privacy*. Early approaches (Zhou et al. 2024, 2023; Zhong and Deng 2022; Yang et al. 2021) overlaid bounded adversarial perturbations on original images, but these often rendered images unnatural, negatively impacting user experience. Subsequent methods (Na, Ji, and Kim 2022; Kakizaki and Yoshida 2019; Hu et al. 2022; Shamshad, Naseer, and Nandakumar 2023; Sun et al. 2024) employed unrestricted adversarial examples to thwart FR systems. Among these, adversarial makeup style transfer (Yin et al. 2021; Hu et al. 2022; Shamshad, Naseer, and Nandakumar 2023; Sun et al. 2024) has gained substantial attention for its natural edits. However, existing makeup-based techniques face key limitations (Fig. 1 - Top): (i) Unnatural edits: makeup style transfer yields unnatural results, particularly for demographics such as males and darker-skinned individuals; (ii) Suboptimal dual-task learning: simultaneously optimizing unrelated objectives (adversarial identity and makeup style) can trigger a tug-of-war between conflicting gradients, causing negative transfer and leading to unnatural and non-robust outcomes.

Recently, diffusion models (Ho, Jain, and Abbeel 2020) have gained attention in image editing due to their train-

*Simon S. Woo is the corresponding author.

ing stability and ability to capture the full data distribution through their simple denoising process (Dhariwal and Nichol 2021). Guided diffusion models, enhanced with pretrained CLIP text embeddings (Radford et al. 2021), have also shown promising results across various applications (Kim, Kwon, and Ye 2022; Liu et al. 2023; Sun et al. 2024). Nevertheless, effectively injecting adversarial noise into attribute edits for facial privacy while minimally altering the original image to maintain utility remains an open challenge.

To this end, we propose MAchine Pareidolia (**MAP**) - a novel solution that exploits the human psychological marvel of *pareidolia*, where we perceive familiar faces in ambiguous patterns (Liu et al. 2014). **MAP** transforms facial action units to dupe FR systems into recognizing target identities, leveraging their inherent tendency to misread subtle emotional tweaks. Unlike prior approaches, our medium-to-high-frequency emotion modifications are seamless across all demographics, avoiding the unnatural, exaggerated appearance of makeup-based techniques and global mismatches like skin tone shifts (Fig. 1). Inspired by research on multitasking learning (Hsieh et al. 2024; Yu et al. 2020), we introduce a synergistic gradient adjustment strategy in the following way to resolve the conflict between adversarial identity and expression gradients: whenever layer-wise minibatch gradients oppose empirical gradients of the other loss, we decompose the minibatch gradients into parallel and orthogonal components relative to the empirical gradients, retaining only the orthogonal component to forge a cohesive path to a shared optimal region. To preserve the natural allure of protected images, we use Laplacian smoothness regularization, delicately maintaining the relative positions of facial landmarks to prevent catastrophic distortions.

Extensive experiments on the CelebA-HQ and LADN datasets showcase **MAP**'s superior privacy protection - boosting black-box success rates by up to 11% - while delivering exceptional perceptual quality and universal applicability across different demographics and photographic scenarios. Moreover, **MAP** exhibits a more favorable balance between perceptual quality and identity obfuscation when compared to recent free-form approaches. Our contributions are summarized as follows:

- We propose **MAP**, a psychology-inspired approach that subtly changes facial expressions to disguise original identities as target ones, thwarting malicious FR systems.
- We introduce a novel synergistic gradient adjustment strategy to harmonize identity and emotion edits, guiding the model to a shared optimal region. To prevent catastrophic distortions from expression edits, we propose Laplacian smoothness regularization to preserve the relative positions of facial landmarks.
- Extensive experiments on diverse datasets demonstrate our **MAP** outperforms prior work, including freeform protection baselines, showcasing **MAP**'s applicability across demographics and photographic styles.

Related Works

Face Protection Method. Due to privacy concerns regarding user identities on online social platforms, various obfuscation methods have been proposed to conceal identities (Meden et al. 2021). Lately, the advent of deep neural networks has facilitated more advanced approaches to shield users from unauthorized facial recognition (FR) systems. Early strategies often employed noise-based adversarial samples (Zhou et al. 2024, 2023; Zhong and Deng 2022; Yang et al. 2021; Oh, Fritz, and Schiele 2017), which involved adding carefully designed adversarial perturbations to the original face images to mislead hostile FR models. Oh, Fritz, and Schiele (2017) developed a game-theoretical framework to derive guarantees on privacy levels in white-box settings. Additionally, TIP-IM (Yang et al. 2021) created adversarial identity masks that can be imposed on facial images to obscure the original identity against black-box FR models. However, such perturbations are usually noticeable to observers and can compromise the user experience.

Recently, strategies that utilize unbounded adversarial samples, which do not constrain the perturbation norm in pixel space, have emerged, resulting in improved image quality (Yin et al. 2021; Hu et al. 2022; Shamshad, Naseer, and Nandakumar 2023; Sun et al. 2024; Liu, Lau, and Chellappa 2023; Li et al. 2024). Among these, makeup-based methods conceal perturbations under the guide of natural makeup features. However, this approach may not be suitable for certain demographic groups, such as males. Another line of research employs generative models to traverse various image attributes, thereby creating adversarial images (Joshi et al. 2019; Khedr, Xiong, and He 2023; Liu, Lau, and Chellappa 2023; He et al. 2024; Du et al. 2024). Specifically, Li et al. (2024) propose a two-step optimization in the low-dimensional latent space of a GAN model (Karras 2019) to construct a protected image. Similarly, DiffProtect (Liu, Lau, and Chellappa 2023) optimizes the latent vector of the source image adversarially to mask the original identity. However, since latent vectors encode a wide range of entangled semantic features, these methods can inadvertently alter global features such as image lighting or saturation, misaligning edited faces with the original scene.

Diffusion Models and Multimodal Guidance. The rise of probabilistic generative models, particularly score-based diffusion models (Ho, Jain, and Abbeel 2020), has been fueled by their training stability and scalability, sparking their adoption across diverse vision tasks. These include image generation (Rombach et al. 2022; Dhariwal and Nichol 2021), image editing (Meng et al. 2021; Wang, Zhao, and Xing 2023), image restoration (Xia et al. 2023), *ect.*. To control generated content and style, multimodal approaches leverage pretrained text or vision encoders to guide the score network (Zhang, Rao, and Agrawala 2023; Kim, Kwon, and Ye 2022). Notably, CLIP, a dual network pretrained on text-image pairs, is widely used for style transfer by navigating its shared latent space. However, effectively injecting adversarial noise into CLIP-guided style edits to enhance privacy remains an open challenge.

Methods

Problem Statement

Let $\boldsymbol{x} \in \mathcal{D}$ be an original face image. A pretrained face recognition model maps input image $\boldsymbol{x}$ to a hypersphere space as $z = f(\boldsymbol{x}) : \mathcal{D} \to \mathcal{Z} \subset \mathbb{R}^d$. The similarity between two facial images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, measured by f, is denoted by $\Phi(z_i, z_j | f)$ and typically adopts cosine similarity between z_i and z_j. Throughout the paper, we use $\bar{v}$ to denote the Euclidean normalization (or unit vector) of a general vector v. Let $\mathcal{L}$ and $\mathcal{L}^{\mathcal{B}_t}$ represent the expected objective value over the entire dataset $\mathcal{D}$ and a subset $\mathcal{B}_t \subset \mathcal{D}$, respectively.

Black-box attacks on face recognition systems are generally categorized into targeted attacks (impersonation attacks) and non-targeted attacks (dodging attacks). Following the settings in (Sun et al. 2024), we specifically focus on targeted attacks for more efficient protection of identity images. In this scenario, an operator $\mathcal{T}$ parameterized by $\boldsymbol{w}$ transforms an original face image $\boldsymbol{x}_o$ into a perturbed image $\boldsymbol{x}_p$, where $\boldsymbol{x}_p = \mathcal{T}_{\boldsymbol{w}}(\boldsymbol{x}_o)$, such that $\boldsymbol{x}_p$ successfully impersonates a target face $\boldsymbol{x}_t$, without knowledge of the target face recognition model. At the same time, we ensure that the transformed image $\boldsymbol{x}_p$ does not significantly deviate from the natural image manifold, maintaining its usability. Formally, the optimization problem that we aim to solve is:

$$\max_{\boldsymbol{w}} \Phi(\underbrace{f(\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{x}_o))}_{z_p}, \underbrace{f(\boldsymbol{x}_t)}_{z_t}) \; s.t. \; \mathcal{H}(\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{x}_o), \boldsymbol{x}_o) \leq \varepsilon \quad (1)$$

where ε bounds the extent of image modifications. For a noise-based approach, the ℓ_p norm of $(\boldsymbol{x}_p - \boldsymbol{x}_o)$ is typically used for $\mathcal{H}$, though this can introduce noticeable artifacts that compromise user experience. In the above objective equation, $\Phi(z_p, z_t | f)$ is generally unknown since f is a black-box model, and we substitute it with a set of M surrogate models $\{f_i\}_{i=1}^M$ for optimization.

MAP: Emotion-based Approach

In response to the limitations of prior works as identified in previous sections, our proposed method leverages emotion-based editing that exposes **three solid advantages** compared with prior studies: This technique involves adjusting subtle facial action units that correspond to different emotional states without altering core demographic attributes, thus demographic-agnostic. Secondly, unlike makeup transfer, which primarily introduces low-frequency artifacts, action units involve medium to high frequencies, making them suitable for infusing subtle adversarial noises, as illustrated in Fig. 2. Lastly, as a consequence of the second advantage, our method has unnoticeable interference in non-face areas, maintaining the faithfulness of the original image. Hence, our objective is to inject adversarial noises into the emotional change of a person and dually optimize them, while maintaining the original looks.

Given a pretrained score network $\boldsymbol{s}_{\boldsymbol{w}}$, let $\boldsymbol{x}_o$ be a face image drawn from a training dataset $\mathcal{D}$. We first transform $\boldsymbol{x}_o$ into a stochastic latent variable of $\boldsymbol{s}_{\boldsymbol{w}}$ at timestep τ using a Gaussian transition (Ho, Jain, and Abbeel 2020):

$$\boldsymbol{x}_o^\tau = \sqrt{\alpha_\tau}\boldsymbol{x}_o + (1 - \alpha_\tau)\boldsymbol{r}, \quad \boldsymbol{r} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad (2)$$

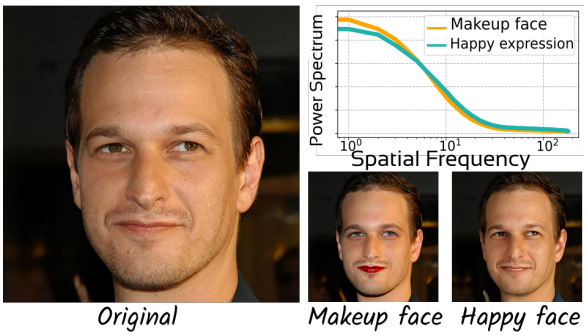

Figure 2: Comparison between makeup transfer-based and emotion-based approaches in terms of frequency changes (with azimuthal integral). Makeup transfer primarily edits the image in the low-frequency range, while our emotion-based approach targets medium to high frequencies, resulting in a more natural appearance and being better suited for obfuscating identity adversarial noises.

where $\alpha_\tau = \Pi_{i=1}^\tau (1 - \beta_i)$ with β_i's are variance schedule. To generate an edited image, we adopt fast detemisnistic reverse DDIM process (Song, Meng, and Ermon 2021):

$$\boldsymbol{x}_o^{\tau-1} = \sqrt{\alpha_{\tau-1}}\tilde{x_o} + \sqrt{1 - \alpha_{\tau-1}}\boldsymbol{s}_{\boldsymbol{w}}(\boldsymbol{x}_o^\tau, \tau), \quad (3)$$

where $\tilde{x_o} = (\boldsymbol{x}_o^\tau - \sqrt{1 - \alpha_{\tau-1}}\boldsymbol{s}_{\boldsymbol{w}}(\boldsymbol{x}_o^\tau, \tau))/\sqrt{\alpha_t}$. After τ backward steps, we obtain the edited image, denoted as the protected image $\boldsymbol{x}_p = \boldsymbol{x}_o^0$.

Dual Objectives. During the training of $\boldsymbol{s}_{\boldsymbol{w}}$ on $\mathcal{D}$, we employ a set of M surrogate face recognition models $\{f_i\}_{i=1}^M$ to bring $\boldsymbol{x}_p$ and $\boldsymbol{x}_t$ closer together. To achieve this, we minimize the cosine discrepancy between $\boldsymbol{x}_p$ and $\boldsymbol{x}_t$ in the latent space of each f_i using an angular divergence loss:

$$\mathcal{L}_{\text{A}}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}_o \sim \mathcal{D}} \left(\frac{1}{M} \sum_{i=1}^{M} 1 - \langle \overline{f_i}(\boldsymbol{x}_p), \overline{f_i}(\boldsymbol{x}_t) \rangle \right), \quad (4)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\overline{f_i}(\boldsymbol{x})$ is the ℓ_2 normalization of $f_i(\boldsymbol{x})$. Meanwhile, to conceal visual adversarial artifacts stemming from Eq. 4 in $\boldsymbol{x}_p$, we apply a demographic-agnostic transformation through emotional alteration of $\boldsymbol{x}_p$. Unlike previous methods (Shamshad, Naseer, and Nandakumar 2023; Sun et al. 2024), this transformation subtly adjusts action units (Ekman 1982; Tian, Kanade, and Cohn 2001), targeting their medium-to-high-frequency components, to effectively mask adversarial artifacts while preserving the utility of $\boldsymbol{x}_p$. Inspired by (Gal et al. 2022), we leverage the semantic information encapsulated in the pretrained CLIP (Radford et al. 2021) model. This induces alignment between the embeddings of the reference image $\boldsymbol{x}_o$ and the generated image $\boldsymbol{x}_p$, and those of the reference text `"source"` and target text `"target"` within the CLIP space. Formally, the emotion objective is defined as follows:

$$\begin{aligned} \Delta_{txt} &= \mathcal{E}_{txt}(\texttt{"target"}) - \mathcal{E}_{txt}(\texttt{"source"}), \\ \Delta_{vis} &= \mathcal{E}_{vis}(\boldsymbol{x}_p) - \mathcal{E}_{vis}(\boldsymbol{x}_o), \\ \mathcal{L}_{\text{E}}(\boldsymbol{w}) &= \mathbb{E}_{\boldsymbol{x}_o \sim \mathcal{D}} \left(1 - \langle \overline{\Delta}_{txt}, \overline{\Delta}_{vis} \rangle\right), \end{aligned} \quad (5)$$

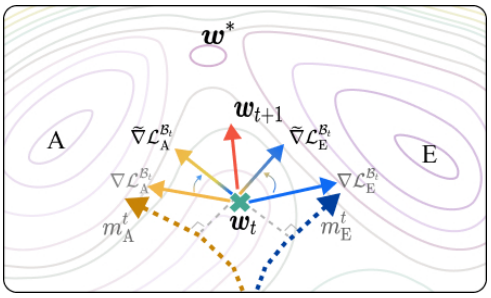

Figure 3: **Illustration of our dual objectives optimization strategies.** Naively optimizing unrelated tasks (identity ← and emotion →) may lead to negative transfer, canceling out each other's gradients. Our approach can render a new update (↑) that helps guide the model towards optimal values.

where $\mathcal{E}_{\text{txt}}$ and $\mathcal{E}_{\text{vis}}$ are textual and visual encoders, respectively. Here, we define `"target"` as *"a photo of [emotion] face"* and `"source"` as *"a photo of face"*.

Efficient Dual Objective Optimization. Thus far, we have optimized two objectives, $\mathcal{L}_{\text{A}}$ and $\mathcal{L}_{\text{E}}$, simultaneously within the diffusion model $\boldsymbol{s_w}$. However, since these objectives in Eqs. 4 and 5 address distinct transformations: *identity* and *emotion*, respectively, their concurrent optimization may lead to a conflict between opposing gradients. This tug-of-war may yield suboptimal results; for instance, while $\boldsymbol{x}_p$ may effectively mask the original identity, the resulting appearance could appear unnatural, or vice versa. Inspired by recent studies (Hsieh et al. 2024; Yu et al. 2020), we propose managing the gradient updates from both losses at every layer of the network. Let $\nabla\mathcal{L}_{\text{A}}(\boldsymbol{w}_l)$ and $\nabla\mathcal{L}_{\text{E}}(\boldsymbol{w}_l)$ represent the gradients at the l^{th} layer of $\boldsymbol{s_w}$, derived from the empirical losses in Eqs. 4 and 5, respectively. At training iteration t, with a minibatch $\mathcal{B}_t$ of images, we compute gradients for each loss as $\nabla\mathcal{L}_{\text{A}}^{\mathcal{B}_t}(\boldsymbol{w}_l)$ and $\nabla\mathcal{L}_{\text{E}}^{\mathcal{B}_t}(\boldsymbol{w}_l)$. We then adjust any conflicting gradients, $\nabla\mathcal{L}_{\text{A/E}}^{\mathcal{B}_t}(\boldsymbol{w}_l)$, which exacerbate the discrepancies between $\nabla\mathcal{L}_{\text{A}}(\boldsymbol{w}_l)$ and $\nabla\mathcal{L}_{\text{E}}(\boldsymbol{w}_l)$. Formally:

$$\widetilde{\nabla}\mathcal{L}_{\text{A}}^{\mathcal{B}_t} = \nabla\mathcal{L}_{\text{A}}^{\mathcal{B}_t} - \min\left\{0, \langle\nabla\mathcal{L}_{\text{A}}^{\mathcal{B}_t}, \overline{\nabla\mathcal{L}}_{\text{E}}\rangle\right\}\cdot\overline{\nabla\mathcal{L}}_{\text{E}}, \quad (6)$$

$$\widetilde{\nabla}\mathcal{L}_{\text{E}}^{\mathcal{B}_t} = \nabla\mathcal{L}_{\text{E}}^{\mathcal{B}_t} - \min\left\{0, \langle\nabla\mathcal{L}_{\text{E}}^{\mathcal{B}_t}, \overline{\nabla\mathcal{L}}_{\text{A}}\rangle\right\}\cdot\overline{\nabla\mathcal{L}}_{\text{A}}. \quad (7)$$

Intuitively, for $\nabla\mathcal{L}_{\text{A}}^{\mathcal{B}_t}(\boldsymbol{w}_l)$, whenever it forms an angle with $\nabla\mathcal{L}_{\text{E}}(\boldsymbol{w}_l)$ greater than 90° (thus increasing update differences), we decompose $\nabla\mathcal{L}_{\text{A}}^{\mathcal{B}_t}$ into two independent components: one parallel and one orthogonal with $\nabla\mathcal{L}_{\text{E}}(\boldsymbol{w}_l)$. We retain only the orthogonal component as described by Eq. 6. For $\nabla\mathcal{L}_{\text{A/E}}(\boldsymbol{w}_l)$, as computing the full gradient on the entire dataset is computationally prohibitive, we estimate these gradients using an exponential moving average (EMA) that accumulates historical minibatch gradients:

$$\boldsymbol{m}_{\text{A/E}}^t(\boldsymbol{w}_l) = \lambda\boldsymbol{m}_{\text{A/E}}^{t-1}(\boldsymbol{w}_l) + (1-\lambda)\nabla\mathcal{L}_{\text{A/E}}^{\mathcal{B}_t}(\boldsymbol{w}_l), \quad (8)$$

where λ is a hyperparameter. To ensure Eq. 8 closely approximates $\nabla\mathcal{L}_{\text{A/E}}$, we present the following theorem.

Momentum Bound Theorem. *Suppose that the loss functions in Eqs. 4 and 5 satisfy the following assumptions: (1) their gradients $\nabla\mathcal{L}_{A/E}(\boldsymbol{w})$ are bounded, i.e., $||\nabla\mathcal{L}_{A/E}(\boldsymbol{w}^t)|| \leq G$; (2) the stochastic gradients are L-Lipschitz, i.e, $||\nabla\mathcal{L}_{A/E}(\boldsymbol{w}^t) - \nabla\mathcal{L}_{A/E}(\boldsymbol{v}^t)|| \leq L||\boldsymbol{w}^t - \boldsymbol{v}^t||$, $\forall\boldsymbol{w}^t, \boldsymbol{v}^t$; (3) the gradient's variance is bounded, i.e., there exists a constant $M > 0$ for any data batch $\mathcal{B}_t$ such that $\mathbb{E}\left[\|\nabla\mathcal{L}_{A/E}^{\mathcal{B}_t}(\boldsymbol{w}) - \nabla\mathcal{L}_{A/E}(\boldsymbol{w})\|_2^2\right] \leq M, \quad \forall\boldsymbol{w} \in \mathbb{R}^d$. Assume that $\boldsymbol{s_w}$ uses SGD as the base optimizer with a learning rate η to update the model parameter with adjusted gradients in Eqs. 6 and 7. Then, by setting $\lambda = 1 - \left(M^{1/2}/LG\right)^{2/3}\eta^{2/3}$, after $T > C'\eta^{-2/3}$ training iterations (C' is a constant), with probability $1-\delta$, we obtain: $\|\boldsymbol{m}_{A/E}^T - \nabla\mathcal{L}_{\text{A/E}}(\boldsymbol{w}^T)\|_2 \leq \mathcal{O}(\eta^{\frac{1}{3}}L^{\frac{1}{3}}G^{\frac{1}{3}}M^{\frac{1}{3}}\log^{\frac{1}{2}}(1/\delta))$.*

This theorem establishes that the error bound between the full gradient $\nabla\mathcal{L}(\boldsymbol{w}_{\text{A/E}}^t)$ and its EMA estimate $\boldsymbol{m}_{\text{A/E}}^t$, at $\mathcal{O}(\gamma^{1/3})$, reduces to $\mathcal{O}(T^{-1/6})$ with $\gamma = \mathcal{O}(1/\sqrt{T})$ in large-T non-convex settings, making $\boldsymbol{m}_t$ an effective approximation. Our synergistic gradient adjustment strategy ensures that updates to $\boldsymbol{s_w}$ navigate towards an overlapping optimal area of two unrelated objectives. Using Eqs. 6 and 7, adversarial identity updates are effectively redirected into emotional updates and vice versa, ensuring that the output $\boldsymbol{x}_p$ appears natural while incorporating robust target identity's attributes. We provide an illustration in Fig. 3.

Training Convergence Theorem. *Suppose that the loss functions defined in Eqs. 4 and 5 satisfy the following two conditions: (1) their gradients $\nabla\mathcal{L}_{A/E}(\boldsymbol{w})$ are bounded; (2) the stochastic gradients are L-Lipschitz. Suppose $\boldsymbol{s_w}$ is updated using SGD with adjusted gradients in Eqs. 6 and 7. Let the learning rate η_t be $\frac{\eta_0}{\sqrt{t}}$, we have:*

$\frac{1}{T}\Sigma_{t=1}^{T}\mathbb{E}_{\boldsymbol{x}_o\sim\mathcal{D}}\left[||\nabla\mathcal{L}_{\text{A/E}}(\boldsymbol{w}^t)||_2\right] \leq \mathcal{O}(\frac{\log T}{\sqrt{T}})$.

For non-convex stochastic optimization, the above theorem shows that training $\boldsymbol{s_w}$ with projected gradients in Eqs. 6 and 7 have a convergence rate $\mathcal{O}(\log T/\sqrt{T})$, but enjoys better performance compared to naively training.

Visual Perception Objective. To enhance the quality of $\boldsymbol{x}_p$, we adopt a methodology aligned with prior works (Sun et al. 2024; Shamshad, Naseer, and Nandakumar 2023), utilizing LPIPS and ℓ_1 loss ($\mathcal{L}_{\text{LPIPS}}$ and $\mathcal{L}_1$). Distinct from methods that rely on makeup transfer, our method allows for the adjustment of facial attributes through action units, thereby highlighting the shortcomings of pixel-wise perceptual losses. Furthermore, as shown in Fig. 4, relying solely on traditional perceptual losses leads to undesirable alterations in the original facial features, such as the length of the eyebrow, resulting in visually unappealing results. To address this, we introduce a Laplacian smoothness loss to ensure that any facial deformations are natural and maintain the integrity of facial landmarks' relative positions,

$$\mathcal{L}_{\text{LS}}(\boldsymbol{w}) = \mathbb{E}_{\substack{\boldsymbol{x}_o\sim\mathcal{D}\\ i\sim\mathcal{K}}}||\Delta v_o^i - \Delta v_p^i||, \Delta v^i = v^i - \mathbb{E}_{\mathcal{N}_i}v^j, \quad (9)$$

where $v_{o/p}^i$ represent the facial landmarks of the original and protected images, identified by a pretrained landmark regression model. $\mathcal{N}_i$ denotes the neighbors of vertex v^i formed by the Delaunay triangulation algorithm (Delaunay 1934), and $\mathcal{K}$ is a set of selected landmarks for applying the smoothness loss. An illustration of $\mathcal{L}_{\text{LS}}$ is shown in Fig. 4 - Top. In our experiments, we also observe that applying

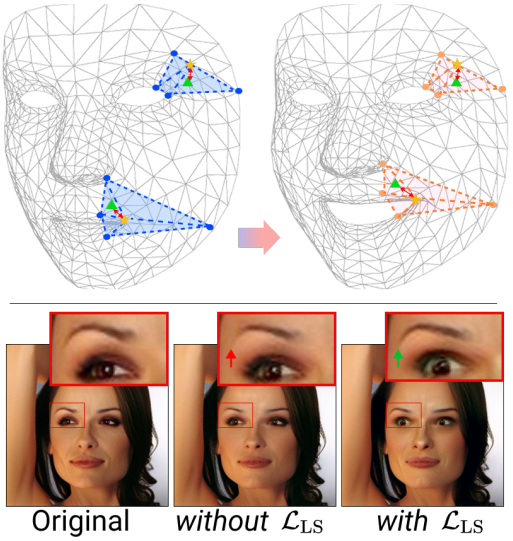

Figure 4: **Top:** Illustration of Laplacian Smoothness Regularization for landmarks (★) 54^{th} and 26^{th}. ▲ represents the average of neighbors (●) of ★. **Bottom:** Effects of Laplacian Smoothness; without it, the eyebrow is eroded compared to the original image.

Laplacian smoothness makes our training more stable, and helps avoid mode collapse compared to not using it.

Score matching. While the score network $\boldsymbol{s_w}$ is fine-tuned by objectives in Eqs. 4, 5, and 9, so far, those objectives are applied to data space (timestep of 0) of $\boldsymbol{s_w}$. Meanwhile, study by Ho, Jain, and Abbeel (2024) has shown that a pretrained score network serves as a prior regularization during generation, which prompts us to fine-tune the score network using the facial training data by minimizing score matching loss in its latent space as follows:

$$\mathcal{L}_D(\boldsymbol{w}) = \sum_{t=1}^{\tau} \mathbb{E}_{\boldsymbol{x}_o \sim \mathcal{D}, \epsilon_t} \left[||\epsilon_t - \boldsymbol{s_w}(\boldsymbol{x}_o^t, t)|| \right], \tag{10}$$

where ϵ_t is the noise added to $\boldsymbol{x}_o$ to produce $\boldsymbol{x}_o^t$. During training, we alternately optimize $\mathcal{L}_D$ with other objectives.

End-to-end training procedure. Our end-to-end training objective is provided as follows:

$$\mathcal{L} = \gamma_a \mathcal{L}_{\text{A}} + \gamma_e \mathcal{L}_{\text{E}} + \gamma_{\text{lpips}} \mathcal{L}_{\text{LPIPS}} + \gamma_1 \mathcal{L}_1 + \gamma_{\text{ls}} \mathcal{L}_{\text{LS}} + \gamma_d \mathcal{L}_{\text{D}}, \tag{11}$$

where γ_i's are hyperparameters that balance the contributions of each objective. During training, the adversarial emotion term is optimized using projected gradients, and the score matching is alternately optimized with other terms.

Experiments

Experimental Settings

Datasets. We evaluate our approach using face verification and identification tests, alongside impersonation attack scenarios (Sun et al. 2024). For face verification, we employ the CelebA-HQ (Karras et al. 2018) and LADN (Gu et al. 2019) datasets for impersonation attacks. Following the standard protocol by Hu et al. (2022), in CelebA-HQ, we use 1,000 images to impersonate 4 distinct target identities. In LADN, we divide the available 332 images into 4 groups, each group targeting different identities for impersonation.

Target model. Consistent with prior studies (Sun et al. 2024), we assess facial privacy protection by attacking four FR models with diverse backbones under black-box settings. These models include IRSE50, IR152, FaceNet, and MobileFace. In each experiment, three models serve as surrogates for training, and one for black-box evaluation.

Evaluation metric. To assess privacy protection effectiveness, we employ the Protection Success Rate (PSR), defined as the fraction of protected faces misclassified by the black-box FR system. We set the threshold at the False Acceptance Rate (FAR) of 0.01 for each model. Additionally, we evaluate the image quality of protected faces using standard metrics, including FID (Heusel et al. 2017), PSNR (dB), and SSIM (Wang et al. 2004).

Implementation Details. We fine-tune our score network on 200 images from each dataset, using AdamW as the optimizer with a learning rate of 4e-6 and 5e-7 for CelebA and LADN, respectively. The training is conducted over 10 epochs with a batch size of 4 on 4 NVIDIA RTX 3090 24GB GPUs. To mask original identities, we employ the "surprised" expression, which is context-dependent and can convey either positive or negative emotions. We set $\{\gamma_{\text{lpips}} = 0.1, \gamma_1 = 0.5, \gamma_{\text{ls}} = 4, \gamma_d = 0.05\}$ for CelebA-HQ and double them for LADN; also $\{\lambda = 0.95, \gamma_a = 0.5, \gamma_e = 0.08\}$ is commonly used for both datasets.

Experimental Results

Face Verification Task. Our quantitative results in terms of Protection Success Rate (PSR) for impersonation attacks under face verification tasks are presented in Table 1. As shown, **MAP** demonstrates superior performance across black-box models on both the CelebA-HQ and LADN datasets. Compared to noise-based and makeup-based approaches, it significantly improves performance by 38% and 11%, respectively, on average.

Face Identification Task. We select 500 subjects from CelebA-HQ, each represented by a pair of images, and additionally integrate four target identities into the gallery set. For face identification, we measure the Rank-N targeted identity success rate, which determines whether the target image $\boldsymbol{x}_t$ appears at least once among the top N candidates in the gallery. We present PSR at both Rank-1 and Rank-5 settings in Table 2. Our approach consistently outperforms recent SoTA baselines in both settings, with average performance improvements of 33% and 23%, respectively.

Quantitative Comparison. We report the performance of various methods in terms of perception quality in Table 3. The results are averaged across both CelebA-HQ and LADN datasets on the verification task. While Adv-Makeup (Yin et al. 2021) update, which only synthesizes eyeshadow to obscure the source identity, reveals the highest scores in all quantitative measures, it has minimal PSR. Our method exhibits the lowest FID scores compared to all other makeup-based baselines and achieves the highest PSR gain.

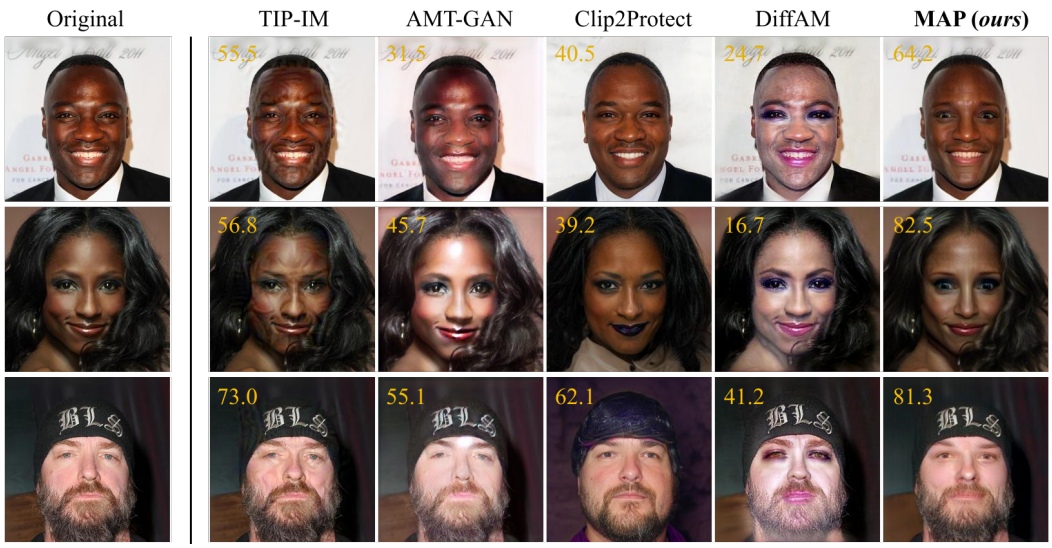

Figure 5: Visualizations of protected face images generated by different facial privacy protection methods on CelebA-HQ. The yellow numbers in each image represent confidence scores returned by Face++. Unlike makeup-based approaches, which may not be suitable for all demographics, our method successfully protects images against malicious FR systems through emotion editing (top to bottom: surprise, surprise, happy), while preserving original details like color grading and background.

Method		CelebA-HQ				LADN-Dataset				Avg.
		IRSE50	IR152	FaceNet	MobileFace	IRSE50	IR152	FaceNet	MobileFace	
	Clean	7.29	3.80	1.08	12.68	2.71	3.61	0.60	5.11	4.61
Noise-based	PGD	36.87	20.68	1.85	43.99	40.09	19.59	3.82	41.09	25.60
	MI-FGSM	45.79	25.03	2.58	45.85	48.90	25.57	6.31	45.01	30.63
	TI-DIM	63.63	36.17	15.30	57.12	56.36	34.18	22.11	48.30	41.64
	TIP-IM	54.40	37.23	40.74	48.72	65.89	43.57	63.50	46.48	50.06
Makeup-based	Adv-Makeup	21.95	9.48	1.37	22.00	29.64	10.03	0.97	22.38	14.72
	AMT-GAN	76.96	35.13	16.62	50.71	89.64	49.12	32.13	72.43	52.84
	CLIP2Protect	81.10	48.42	41.72	75.26	91.57	53.31	47.91	79.94	64.90
	DiffAM	92.00	63.13	64.67	83.35	95.66	66.75	65.44	92.04	77.88
Emotion-based	**MAP (*ours*)**	**93.30**	**78.98**	**72.35**	**92.50**	**96.65**	**91.16**	**86.43**	**96.49**	**88.48**

Table 1: Protect Success Rate (PSR) with FAR@1e-2 for black-box setting on CelebA-HQ and LADN dataset.

Method	IRSE50		IR152		FaceNet		MobileFace	
	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T
TIP-IM	16.2	51.4	21.2	56.0	8.1	35.8	9.6	24.0
CLIP2Protect	24.5	64.7	24.2	65.2	12.5	38.7	11.8	28.2
SD4Privacy	15.6	26.8	23.4	41.2	33.6	53.8	31.8	49.8
GIFT	21.2	57.2	34.6	49.4	33.2	65.6	41.2	67.6
Adv-CPG	24.4	56.4	33.8	51.2	36.6	67.4	43.4	70.4
MAP (*ours*)	**57.9**	**83.2**	**54.0**	**61.8**	**45.0**	**70.7**	**50.3**	**74.5**

Table 2: Protection success rate (PSR) of impersonation attacks under the face identification task on CelebA-HQ.

Comparison with Freeform Approaches. We explore the trade-off between quantitative (PSR under face verification task) and qualitative performance by simply adjusting the values of γ_d. Additionally, we include the performance of five freeform attribute baselines, DiffProtect (Liu, Lau, and Chellappa 2023), SD4Privacy (An et al. 2024), Adv-Diffusion (Liu et al. 2024), GIFT (Li et al. 2024), and Adv-CPG (Wang, Zhang, and Yuan 2025). As shown in Fig. 6,

Method	FID ↓	PSNR ↑	SSIM ↑	PSR Gain ↑
Adv-Makeup	4.22	34.51	0.985	0
AMT-GAN	34.44	19.50	0.787	38.12
CLIP2Protect	26.62	19.31	0.750	50.18
DiffAM	26.10	20.52	0.886	63.13
MAP (*ours*)	24.21	29.07	0.876	73.76

Table 3: Quantitative evaluations of image quality. PSR Gain is absolute gain in PSR relative to Adv-Makeup.

MAP provides a higher PSR at similar FID scores compared to all freeform baselines. Meanwhile, Fig. 6 also shows that with $\gamma_d = 0$, our method has similar quantitative performance to GIFT but cannot achieve its best performance, indicating the crucial role of optimizing the score matching.

Real-world Effectiveness. To validate the effectiveness of **MAP** against unauthorized FR systems, we conducted experiments using a commercial API, namely Face++, with a face verification test. The API returns confidence scores between two images on a scale from 0 to 100, where a

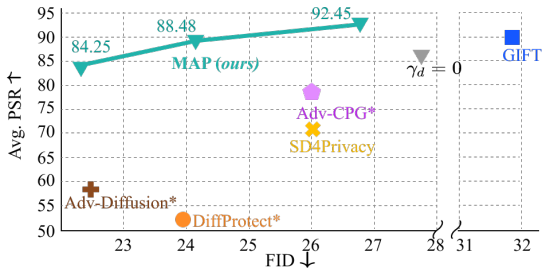

Figure 6: Trade-off between PSR and FID. **MAP** achieves a better trade-off and outperforms five freeform baselines: DiffProtect, SD4Privacy, Adv-Diffusion, GIFT, and Adv-CPG.

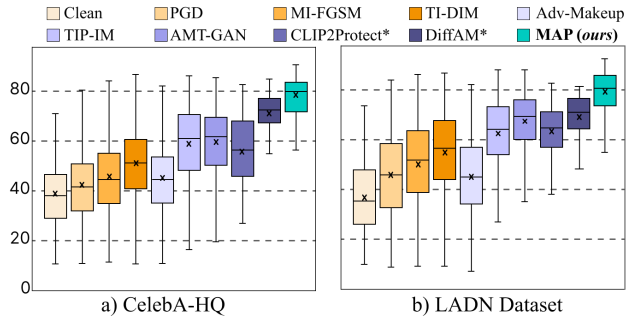

Figure 7: Average confidence score from the real-world face verification API, Face++, for impersonation attacks.

Settings	FID↓	PSR↑ with FAR		
		@1e-1	@1e-2	@1e-3
w/o emotion	24.1	95.6	85.0	67.0
w/o grads proj.	24.5	90.5	71.2	50.9
w/o EMA	22.5	93.6	76.5	56.8
w/o $\mathcal{L}_{\text{LS}}$	23.6	94.8	78.8	56.4
MAP (*ours*)	21.5	95.4	80.3	60.5

Table 4: Ablation studies on different optimization settings.

higher score indicates greater similarity. We run the experiment with 1,000 images from the CelebA dataset and 324 images from the LADN dataset. As shown in Fig. 7, our proposed method achieved the highest confidence scores, with improvements over the second-best method of 8% and 10% on the CelebA-HQ and LADN datasets, respectively.

Impact of Loss Components. We evaluate on the largest black-box model, IR152 (Deng et al. 2019), using the first identity from the CelebA-HQ dataset to validate the effect of each proposed loss component. As shown in Table 4, without our proposed gradient projection, the model exhibits suboptimal performance, illustrating the negative transfer effects during optimization. Meanwhile, obfuscating the target identity without the use of makeup or emotional transformations can obtain higher PSR yet results in poorer qualitative scores due to erosion of facial attributes.

Uncommon Portrait Photography Styles. We compare our **MAP** method against baselines across uncommon photographic styles, namely monochrome, Rembrandt lighting, and backlighting, as illustrated in Fig. 8. As shown, unlike makeup-based methods, our approach exhibits minimal impact on monochrome images while preserving fine details, color grading, and object naturalness in protected images across all tested styles. This demonstrates the robustness and versatility of our method across diverse scenarios.

Figure 8: Illustration of uncommon photography styles: monochrome, Rembrandt lighting, and backlighting.

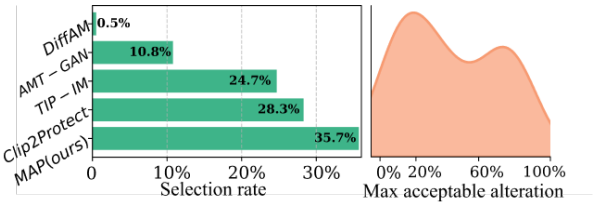

Figure 9: Our human evaluation from 25 participants.

Human Study. To evaluate practical applicability, we conducted a user study to compare users' preferences for our method against previous baselines. We asked 25 participants to answer 25 questions about their preferred editing method for use on social networks while all method names were replaced with dummy text. Our method was favored in 35.7% of the cases (Fig.9-left). Additionally, most users preferred alterations of 20% or less (Fig.9-right), highlighting the suitability of **MAP**'s subtle modifications compared to makeup, hairstyle, or facial attribute eddits.

Conclusion

This paper presents a new approach to disguise the original identity as a target identity in arbitrary portrait images, enhancing privacy protection for facial images on online platforms. Departing from prior methods, we transform expressions in the source image to redirect recognition from the original identity to a target one, thwarting adversarial FR systems. To reconcile conflicts between emotion and identity learning, we project their mini-batch gradients onto mutual empirical gradients, ensuring cohesive optimization. We also introduce a perceptual objective, Laplacian smoothness, combined with score matching loss to maintain image naturalness. Experiments show our method surpasses noise-based, makeup-based, and freeform attribute approaches in quantitative metrics and qualitative fidelity, underscoring its potential as a robust privacy safeguard.

Acknowledgments

This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (RS-2022-II221199, RS-2022-II220688, RS-2019-II190421, RS-2023-00230337, RS-2024-00437849, RS-2021-II212068, and RS-2025-02263841). Also, this work was supported by the Cyber Investigation Support Technology Development Program (No.RS-2025-02304983) of the Korea Institute of Police Technology (KIPoT), funded by the Korean National Police Agency. Lastly, this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2024-00356293).

References

An, J.; Zhang, W.; Wu, D.; Lin, Z.; Gu, J.; and Wang, W. 2024. Sd4privacy: exploiting stable diffusion for protecting facial privacy. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Delaunay, B. 1934. Sur la sphère vide. *Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk*, 793–800.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Du, R.; Li, Y.; Li, M.; Peng, J.; Zhu, Y.; and Ma, C. 2024. Multi-attribute Semantic Adversarial Attack Based on Cross-layer Interpolation for Face Recognition. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE.

Ekman, P. 1982. Methods for measuring facial action. *Handbook of methods in nonverbal behavior research*, 45–90.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Gu, Q.; Wang, G.; Chiu, M. T.; Tai, Y.-W.; and Tang, C.-K. 2019. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International conference on computer vision*, 10481–10490.

He, X.; Zhu, M.; Chen, D.; Wang, N.; and Gao, X. 2024. Diff-privacy: Diffusion-based face privacy protection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hill, K. 2022. The secretive company that might end privacy as we know it. In *Ethics of Data and Analytics*, 170–177. Auerbach Publications.

Hill, K. 2024. Two Students Created Face Recognition Glasses. It Wasn't Hard. *The New York Times*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Jain, A.; and Abbeel, P. 2024. Gradient Guidance for Diffusion Models: An Optimization Perspective. *Advances in neural information processing systems*.

Hsieh, Y.-G.; Thornton, J.; Ndiaye, E.; Klein, M.; Cuturi, M.; and Ablin, P. 2024. Careful with that Scalpel: Improving Gradient Surgery with an EMA. In *International Conference on Machine Learning*, 19085–19100. PMLR.

Hu, S.; Liu, X.; Zhang, Y.; Li, M.; Zhang, L. Y.; Jin, H.; and Wu, L. 2022. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15014–15023.

Joshi, A.; Mukherjee, A.; Sarkar, S.; and Hegde, C. 2019. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4773–4783.

Kakizaki, K.; and Yoshida, K. 2019. Adversarial image translation: Unrestricted adversarial examples in face recognition systems. *arXiv preprint arXiv:1905.03421*.

Karras, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv preprint arXiv:1812.04948*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of gans for improved quality, stability, and variation. arXiv 2017. *arXiv preprint arXiv:1710.10196*, 1–26.

Khedr, Y. M.; Xiong, Y.; and He, K. 2023. Semantic adversarial attacks on face recognition through significant attributes. *International Journal of Computational Intelligence Systems*, 16(1): 196.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2426–2435.

Li, M.; Wang, J.; Zhang, H.; Zhou, Z.; Hu, S.; and Pei, X. 2024. Transferable adversarial facial images for privacy protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10649–10658.

Liu, D.; Wang, X.; Peng, C.; Wang, N.; Hu, R.; and Gao, X. 2024. Adv-diffusion: imperceptible adversarial face identity attack via latent diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 3585–3593.

Liu, J.; Lau, C. P.; and Chellappa, R. 2023. Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection. *arXiv preprint arXiv:2305.13625*.

Liu, J.; Li, J.; Feng, L.; Li, L.; Tian, J.; and Lee, K. 2014. Seeing Jesus in toast: neural and behavioral correlates of face pareidolia. *Cortex*, 53: 60–77.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023. More control for free! image synthesis with semantic diffusion

guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 289–299.

Meden, B.; Rot, P.; Terhörst, P.; Damer, N.; Kuijper, A.; Scheirer, W. J.; Ross, A.; Peer, P.; and Štruc, V. 2021. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16: 4147–4183.

Meng, C.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Na, D.; Ji, S.; and Kim, J. 2022. Unrestricted black-box adversarial attack using gan with limited queries. In *European Conference on Computer Vision*, 467–482. Springer.

Oh, S. J.; Fritz, M.; and Schiele, B. 2017. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 1491–1500. IEEE.

Phillips, P. J.; Yates, A. N.; Hu, Y.; Hahn, C. A.; Noyes, E.; Jackson, K.; Cavazos, J. G.; Jeckeln, G.; Ranjan, R.; Sankaranarayanan, S.; et al. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24): 6171–6176.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shamshad, F.; Naseer, M.; and Nandakumar, K. 2023. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20595–20605.

Shoshitaishvili, Y.; Kruegel, C.; and Vigna, G. 2015. Portrait of a privacy invasion. *Proceedings on Privacy Enhancing Technologies*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Sun, Y.; Yu, L.; Xie, H.; Li, J.; and Zhang, Y. 2024. DiffAM: Diffusion-based Adversarial Makeup Transfer for Facial Privacy Protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24584–24594.

Tian, Y.-I.; Kanade, T.; and Cohn, J. F. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2): 97–115.

Wang, J.; Zhang, H.; and Yuan, Y. 2025. Adv-cpg: A customized portrait generation framework with facial adversarial attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21001–21010.

Wang, Y.; Bao, T.; Ding, C.; and Zhu, M. 2017. Face recognition in real-world surveillance videos with deep learning method. In *2017 2nd international conference on image, vision and computing (icivc)*, 239–243. IEEE.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Zhao, L.; and Xing, W. 2023. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7677–7689.

Wenger, E.; Shan, S.; Zheng, H.; and Zhao, B. Y. 2023. Sok: Anti-facial recognition technology. In *2023 IEEE Symposium on Security and Privacy (SP)*, 864–881. IEEE.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13095–13105.

Yang, X.; Dong, Y.; Pang, T.; Su, H.; Zhu, J.; Chen, Y.; and Xue, H. 2021. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3897–3907.

Yin, B.; Wang, W.; Yao, T.; Guo, J.; Kong, Z.; Ding, S.; Li, J.; and Liu, C. 2021. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *successfullyProceedings of the Thirtieth International Joint Conference on Artificial Intelligence (*, 1252–1258.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 5824–5836.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhong, Y.; and Deng, W. 2022. Opom: Customized invisible cloak towards face privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3590–3603.

Zhou, Z.; Hu, S.; Li, M.; Zhang, H.; Zhang, Y.; and Jin, H. 2023. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6311–6320.

Zhou, Z.; Li, M.; Liu, W.; Hu, S.; Zhang, Y.; Wan, W.; Xue, L.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024. Securely fine-tuning pre-trained encoders against adversarial examples. In *2024 IEEE Symposium on Security and Privacy (SP)*, 3015–3033. IEEE.