CLIP2Protect: Protecting Facial Privacy using Text-Guided Makeup via Adversarial Latent Search

Fahad Shamshad Muzammal Naseer Karthik Nandakumar
Mohamed Bin Zayed University of AI, UAE
{fahad.shamshad, muzammal.naseer, karthik.nandakumar}@mbzuai.ac.ae

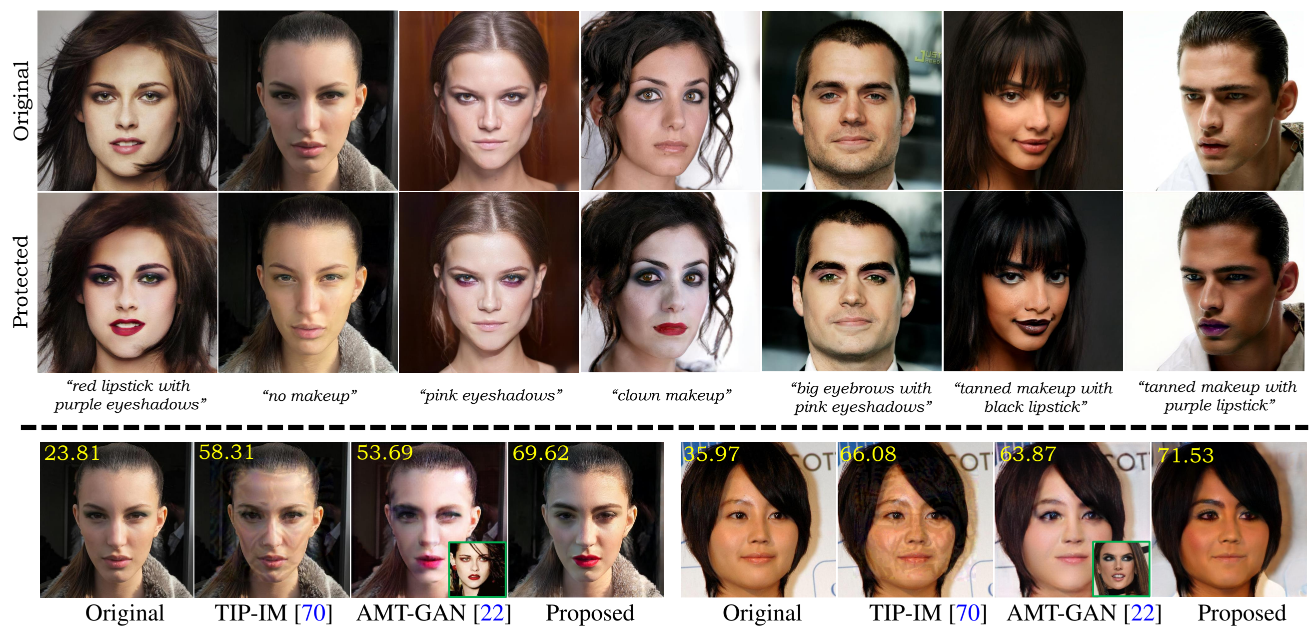

Figure 1. The proposed approach crafts "naturalistic" and transferable text-guided adversarial faces to deceive black-box face recognition systems. First row shows original images that need to be protected and second row shows corresponding protected images along with the user-defined makeup text prompts that guide the adversarial search. Comparison against existing methods is shown in the third row. The yellow text represents the confidence score (higher is better) output by a commercial API (Face++), when matching the protected image against the target identity shown in the bottom right. The reference image used by [22] for makeup transfer is shown at the bottom corner of the corresponding adversarial image.

Abstract

The success of deep learning based face recognition systems has given rise to serious privacy concerns due to their ability to enable unauthorized tracking of users in the digital world. Existing methods for enhancing privacy fail to generate "naturalistic" images that can protect facial privacy without compromising user experience. We propose a novel two-step approach for facial privacy protection that relies on finding adversarial latent codes in the low-dimensional manifold of a pretrained generative model. The first step inverts the given face image into the latent space and finetunes the generative model to achieve an accurate reconstruction of the given image from its latent code. This step produces a good initialization, aiding the generation of high-quality faces that resemble the given identity. Subsequently, user-defined makeup text prompts and identity-preserving regularization are used to guide the search for adversarial codes in the latent space. Extensive experiments demonstrate that faces generated by our approach have stronger black-box transferability with an absolute gain of 12.06% over the state-of-the-art facial privacy protection approach under the face verification task. Finally, we demonstrate the effectiveness of the proposed approach for commercial face recognition systems. Our code is available at https://github.com/fahadshamshad/Clip2Protect.

1. Introduction

Deep learning based face recognition (FR) systems [43, 61] have found widespread usage in multiple applications,

Table 1. Comparison among different facial privacy protection methods w.r.t. the natural outputs, black box setting, experiments under face verification and identification tasks, unrestricted (semantically meaningful), and more flexible text guided adversaries.

	Adv-Makeup [71]	TIP-IM [70]	AMT-GAN [22]	Ours
Natural outputs	Yes	Partially	Partially	Yes
Black box	Yes	Yes	Yes	Yes
Verification	Yes	No	Yes	Yes
Identification	No	Yes	No	Yes
Unrestricted	Yes	No	Yes	Yes
Text guided	No	No	No	Yes

including security [63], biometrics [38], and criminal investigation [45], outperforming humans in many scenarios [12, 48, 61]. Despite positive aspects of this technology, FR systems seriously threaten personal security and privacy in the digital world because of their potential to enable mass surveillance capabilities [1, 67]. For example, government and private entities can use FR systems to track user relationships and activities by scraping face images from social media profiles such as Twitter, Linkedin, and Facebook [18, 20]. These entities generally use proprietary FR systems, whose specifications are unknown to the public (*black box model*). Therefore, there is an urgent need for an effective approach that protects facial privacy against such unknown FR systems.

An ideal facial privacy protection algorithm must strike the right balance between naturalness and privacy protection [70, 77]. In this context, "naturalness" is defined as the *absence of any noise artifacts that can be easily perceived by human observers* and the *preservation of human-perceived identity*. "Privacy protection" refers to the fact that the protected image must be capable of *deceiving a black-box malicious FR system*. In other words, the protected image must closely resemble the given face image and be artifact-free for a human observer, while at the same time fool an unknown automated FR system. Since failure to generate naturalistic faces can significantly affect user experience on social media platforms, it is a necessary precondition for adoption of a privacy-enhancement algorithm.

Recent works exploit adversarial attacks [57] to conceal user identity by overlaying noise-constrained (bounded) adversarial perturbations on the original face image [6,53,74]. Since the adversarial examples are generally optimized in the *image space*, it is often difficult to simultaneously achieve naturalness and privacy [70]. Unlike noise-based methods, unrestricted adversarial examples are not constrained by the magnitude of perturbation in the image space and have demonstrated better perceptual realism for human observers while being adversarially effective [3, 55, 68, 76].

Several efforts have been made to generate unrestricted adversarial examples that mislead FR systems (see Tab. 1) [22, 25, 39, 72]. Among these, adversarial makeup based methods [22, 72] are gaining increasing attention as they can embed adversarial modifications in a more natural way. These approaches use generative adversarial networks [15] (GANs) to adversarially transfer makeup from a given *reference image* to the user's face image while impersonating a *target identity*. However, existing techniques based on adversarial makeup transfer have the following limitations: (i) adversarial toxicity in these methods hamper the performance of the makeup transfer module, thereby resulting in unnatural faces with makeup artifacts (see Fig. 1); (ii) the use of a reference image to define the desired makeup style affects the practicality of this approach; (iii) for every new target identity, these approaches require end-to-end retraining from scratch using large makeup datasets; and (iv) most of these methods primarily aim at impersonation of the target identity, whereas the desired privacy objective is dodging, *i.e.,* multiple images of the user's face scraped from different social media sites must not match with each other.

To mitigate the above problems, we propose a new approach to protect user facial privacy on online platforms (Sec. 3). The proposed approach aims to search for ***adversarial latent codes*** in a low-dimensional manifold learned by a generative model trained to generate face images [2, 27]. Our main contributions are:

- **Facial Privacy-protection Framework Using Adversarial Latent Codes:** Given a face image, we propose a novel two-step method to search for adversarial latent codes, which can be used by a generative model (*e.g.*, StyleGAN) to produce face images with high visual quality that matches human-perceived identity, while deceiving black-box FR systems.
- **Adversarial Makeup Transfer using Textual Prompts:** A critical component of the above framework is a technique for leveraging user-defined textual (makeup) prompts to traverse over the latent manifold of the generative model and find transferable adversarial latent codes. Our approach effectively hides attack information in the desired makeup style, without the need for any large makeup dataset or retraining of models for different target identities.
- **Identity Preserving Regularization:** We propose a regularizer that preserves identity-related attributes within the latent space of the generative model and ensures that the protected face image visually resembles the original face.

Extensive experiments (Sec. 4.1) for both *face verification* and *identification* scenarios demonstrate the effectiveness of our approach against black-box FR models and online commercial facial recognition APIs (Sec. 4.2). Furthermore, we provide detailed ablative analysis to dissect the performance of different components of our approach (Sec. 4.3).

2. Related Work

Obfuscation Methods: Obfuscation is the most widely used technique [38] to protect user's facial privacy. Earlier obfuscation approaches typically degrade the quality of the original face image by applying simple operations such as masking [52, 64], filtering [33, 78], and image transformations [8, 36, 62]. While these relatively simple obfuscation techniques are reasonable for surveillance applications, they are ill-suited for online/social media platforms where user experience is critical [41]. Though deep learning based obfuscation approaches generate more realistic images [4, 7, 56, 58], they often result in a change of identity compared to the original image and occasionally produce undesirable artifacts [30, 31, 34].

Noise-based Adversarial Examples: Adversarial attacks have been used to protect users from unauthorized `FR` models. Some methods [6, 53] rely on data poisoning to deceive targeted `FR` models, but are less practical because access to the training data or the gallery set of the unknown `FR` system is often not available. Other approaches have used game-theory perspective [42] in white-box settings or person-specific privacy masks (one mask per person) to generate protected images at the cost of acquiring multiple images of the same user [77]. In contrast, we aim to fool the *black box* `FR` model using only single image. In TIP-IM [70], targeted optimization was used to generate privacy masks against unknown `FR` models by introducing a naturalness constraint. While this approach provides effective privacy, it generates output images with perceptible noises that can affect the user experience [70].

Unrestricted Adversarial Examples: Unrestricted adversarial attacks (UAAs) are not constrained by the perturbation norm and can induce large but semantically meaningful perturbations. These attacks have been extensively studied in image classification literature [3, 35, 55, 68, 73, 76] and it has been shown that outputs generated via UAAs are less perceptible to human observers as compared to noise-based adversarial attacks. Motivated by this observation, patch-based unrestricted attacks have been proposed to generate wearable adversarial accessories like colorful glasses [54], hat [29] or random patch [69] to fool the `FR` model, but such synthesized patches generally have weak transferability due to the limited editing region and the large visible pattern compromises naturalness and affects user experience. Recently, generative models [24, 50] have been leveraged to craft UAAs against `FR` models. However, these generative approaches are either designed for the white-box settings [46, 79] or show limited performance in query-free black-box settings [25]. Makeup-based UAAs [17, 72] have also been proposed against `FR` systems by embedding the perturbations into a natural makeup effect. These makeup based attacks have also been exploited to protect the user privacy by applying adversarial makeup on the user face image [22]. However, interference between adversarial perturbations and makeup transfer can produce undesirable makeup artifacts in the output images. Moreover, these attacks generally assume access to large makeup datasets for training models and require a reference makeup image. In contrast, our approach finds adversarial faces on the natural image manifold in black-box setting *via guidance from makeup text prompt*, which makes it less susceptible to artifacts (see Fig. 1) and more practical.

Vision-Language Modelling: Cross-modal vision-language modelling has attracted significant attention in recent years [13]. OpenAI introduced CLIP [47] that is trained on 400 million image-text pairs using contrastive objective and maps both image and text in a joint multimodal embedding space. With powerful representation embedding of CLIP, several methods have been proposed to manipulate images with text-guidance. StyleCLIP [44] and DiffusionCLIP [28, 40] leverage the powerful generative capabilities of StyleGAN and diffusion models to manipulate images with text prompts. Other similar works include HairCLIP [66], CLIP-NeRF [60], CLIPstyler [32], and CLIPDraw [14]. While these methods focus on the text-guidance ability of CLIP, our approach aims to find the adversarial latent codes in a generative model's latent space for privacy protection against black-box `FR` models.

3. Proposed Approach for Facial Privacy

Our goal is to protect user facial privacy on online platforms against unknown (black-box) `FR` models without compromising on the user's online experience. The proposed approach finds protected faces by adversarially exploring the low-dimensional latent space of a pretrained generative model that is trained on natural face images. To avoid artifacts in the protected image, we restrict the search for adversarial faces close to the clean image manifold learned by the generative model. Moreover, we propose to optimize only over identity-preserving latent codes in the latent space. This effectively preserves human-perceived identity during attack while offering high privacy against automated systems. Further, we employ natural makeup-like perturbations via guidance from a text prompt, which provides more flexibility to the user compared to reference image-based adversarial makeup transfer [22].

3.1. Preliminaries

Let $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^n$ denote the given original/real face image. Let $f(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathbb{R}^d$ be a `FR` model that extracts a fixed-length normalized feature representation. Let $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) = D(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2))$ be a distance metric that measures the dissimilarity between two face images $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ based on their respective representations $f(\boldsymbol{x}_1)$ and $f(\boldsymbol{x}_2)$. Generally a `FR` system can operate in two modes: *verification* and *identification*. A face verification

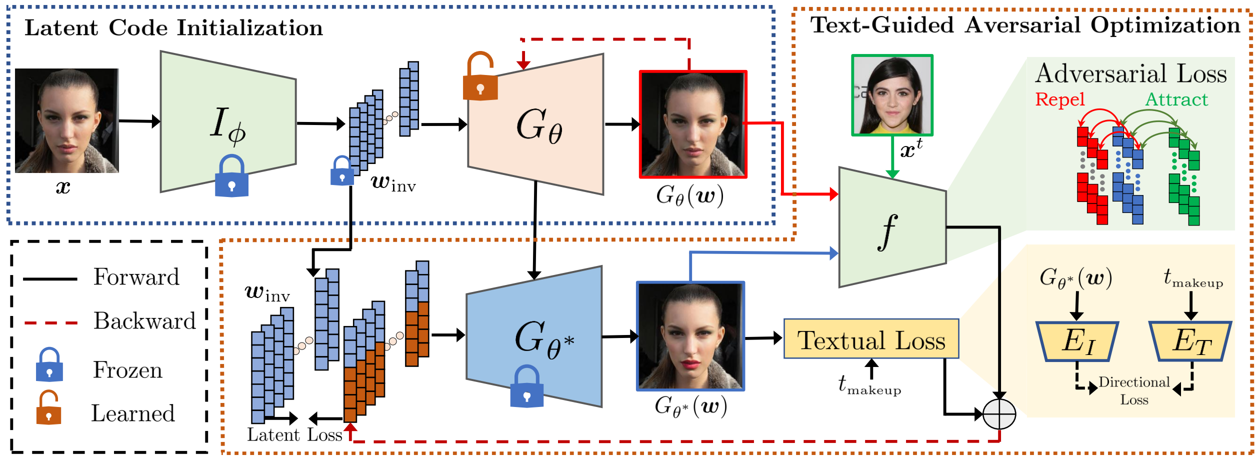

Figure 2. Overall pipeline of the proposed approach to protect users facial privacy. Our proposed approach searches for the adversarial latent codes on the generative manifold to reconstruct an adversarial face that is capable of fooling unknown FR systems for privacy protection. Our approach allows "makeup" editing in an adversarial manner through user defined textual prompts and thereby enhance the user's online experience. Our text-guided objective searches for such latent codes while keeping the original identity preserved.

system predicts that two faces belong to the same identity if $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq \tau$, where τ is the system threshold. On the other hand, a (*closed set*) face identification system compares the input image (probe) against a set of face images (gallery) and outputs the identity whose representation is most similar to that of the probe. Since the attacker can employ verification or identification to determine the user identity using black-box FR models, a protection approach should conceal the user's identity in both scenarios.

User privacy can be protected by misleading the malicious FR model through *impersonation* or *dodging* attacks. In the context of verification, impersonation (false match) implies that the protected face matches with the face of a specific target identity and dodging (false non-match) means that the protected face does not match with some other image of the same person. Similarly, for face identification, impersonation ensures that the protected image gets matched to a specified target identity in the gallery set, while dodging prevents the protected face from matching with images of the same person in the gallery.

Problem Statement: Given the original face image $\boldsymbol{x}$, our goal is to generate a protected face image $\boldsymbol{x}^p$ such that $\mathcal{D}(\boldsymbol{x}^p, \boldsymbol{x})$ is large (for successful dodging attack) and $\mathcal{D}(\boldsymbol{x}^p, \boldsymbol{x}^t)$ is small (for successfully impersonating a target face $\boldsymbol{x}^t$), where $\mathcal{O}(\boldsymbol{x}) \neq \mathcal{O}(\boldsymbol{x}^t)$ and $\mathcal{O}$ is the oracle that gives the true identity labels. At the same time, we want to minimize $\mathcal{H}(\boldsymbol{x}^p, \boldsymbol{x})$, where $\mathcal{H}$ quantifies the degree of unnaturalness introduced in the protected image $\boldsymbol{x}^p$ in relation to the original image $\boldsymbol{x}$. Formally, the optimization problem that we aim to solve is:

$$\min_{\boldsymbol{x}^p} \mathcal{L}(\boldsymbol{x}^p) = \mathcal{D}(\boldsymbol{x}^p, \boldsymbol{x}^t) - \mathcal{D}(\boldsymbol{x}^p, \boldsymbol{x}) \quad (1)$$
$$\text{s.t. } \mathcal{H}(\boldsymbol{x}^p, \boldsymbol{x}) \leq \epsilon$$

(a) Original (b) Encoder Inversion (c) Generator finetuning

Figure 3. Generator finetuning allows near-perfect reconstructions of LFW dataset sample. This is crucial for the online experience of users. Matching scores returned by Face++ API are 62.38 and 98.96 for encoder and generator-finetuned inversions, respectively.

where ϵ is a bound on the adversarial perturbation. For noise-based approach, $\mathcal{H}(\boldsymbol{x}^p, \boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}^p\|_p$, where $\|\cdot\|_p$ denotes the L_p norm. However, direct enforcement of the perturbation constraint leads to visible artifacts, which affects visual quality and user experience. Constraining the solution search space to a natural image manifold using an effective image prior can produce more realistic images. Note that the distance metric $\mathcal{D}$ is unknown since our goal is to deceive a black-box FR system.

3.2. Makeup Text-Guided Adversarial Faces

Our approach restricts the solution space of the protected face $\boldsymbol{x}^p$ to lie close to the clean face manifold $\mathcal{X}$. This manifold can be learned using a generative model trained on real human faces. Specifically, let $G_\theta(\boldsymbol{w}) : \mathcal{W} \rightarrow \mathbb{R}^n$ denote the pretrained generative model with weights θ, where $\mathcal{W}$ is the latent space. Our proposed approach consists of two stages: (i) latent code initialization (Sec. 3.2.1) and (ii) text-guided adversarial optimization (Sec. 3.2.2). The overall pipeline of the proposed approach is shown in Fig. 2.

3.2.1 Latent Code Initialization

The latent code initialization stage is based on GAN inversion, which aims to invert the original image $\boldsymbol{x}$ into the latent space $\mathcal{W}$, i.e., find a latent code $\boldsymbol{w}_{\text{inv}} \in \mathcal{W}$ such that $\boldsymbol{x}_{\text{inv}} = G_\theta(\boldsymbol{w}_{\text{inv}}) \approx \boldsymbol{x}$. To achieve this, we first use an encoder-based inversion called e4e [59] to infer $\boldsymbol{w}_{\text{inv}}$ in $\mathcal{W}$ from $\boldsymbol{x}$ *i.e.,* $\boldsymbol{w}_{\text{inv}} = I_\phi(\boldsymbol{x})$, where $I_\phi : \mathcal{X} \to \mathcal{W}$ is the pretrained encoder with weights ϕ (see Fig. 2).

We use StyleGAN trained on a high-resolution dataset of face images as the pretrained generative model G_θ due to its powerful synthesis ability and the disentangled structure of its latent space. A significant challenge during inversion is preserving the identity of the original image *i.e.,* $\mathcal{O}(\boldsymbol{x}) = \mathcal{O}(\boldsymbol{x}_{\text{inv}})$. Generally, optimization and encoder-based inversion approaches struggle to preserve identity after reconstruction [49] (see Fig. 3b). Moreover, when using these approaches, the inversion error can be large for out-of-domain face images with extreme poses and viewpoints, which are quite common in social media applications. Therefore, these approaches cannot be applied directly to invert $\boldsymbol{x}$. Instead, motivated by the recent observation [49] that slight changes to the pretrained generator weights do not harm its editing abilities while achieving near-perfect reconstructions, we finetune the pretrained generator weights θ instead of the encoder weights ϕ. Specifically, we fix $\boldsymbol{w}_{\text{inv}} = I_\phi(\boldsymbol{x})$ and fine-tune G_θ using the following loss:

$$\theta^* = \arg\min_\theta \mathcal{L}_{\text{LPIPS}}(\boldsymbol{x}, G_\theta(\boldsymbol{w}_{\text{inv}})) + \lambda_2 \mathcal{L}_2(\boldsymbol{x}, G_\theta(\boldsymbol{w}_{\text{inv}})),$$

where $\mathcal{L}_{\text{LPIPS}}$ is the perceptual loss and $\mathcal{L}_2$ denotes the pixel-wise similarity. The final inverted image $\boldsymbol{x}^*_{\text{inv}}$ (see Fig. 3c) can be obtained by performing a forward pass of $\boldsymbol{w}_{\text{inv}}$ through fine-tuned generator *i.e.,* $\boldsymbol{x}^*_{\text{inv}} = G_{\theta^*}(\boldsymbol{w}_{\text{inv}})$.

3.2.2 Text-guided adversarial optimization

Given the inverted latent code $\boldsymbol{w}_{\text{inv}}$ and fine-tuned generator $G_{\theta^*}(.)$, our goal is to adversarially perturb this latent code $\boldsymbol{w}_{\text{inv}}$ in the low-dimensional generative manifold $\mathcal{W}$ to generate a protected face that fools the black-box `FR` model, while imitating the makeup style of the text prompt t_{makeup}.

To achieve these objectives, we investigate the following questions: (i) how to effectively extract makeup style information from t_{makeup} and apply it to the face image $\boldsymbol{x}$ in an *adversarial* manner?, (ii) how to regularize the optimization process so that the output face image is not qualitatively impaired?, (iii) how to craft effective adversarial perturbations that mislead *black-box* `FR` models?, and (iv) how to preserve the human-perceived identity $\mathcal{O}(\boldsymbol{x})$ of the original face image while ensuring high privacy?

The first issue can be addressed by aligning the output adversarial image with the text prompt t_{makeup} in the embedding space of a pretrained vision-language model. The second issue is addressed by enforcing the adversarial latent code to remain close to initialization $\boldsymbol{w}_{\text{inv}}$. The third issue is solved by crafting transferable text-guided adversarial faces on a white-box surrogate model (or an ensemble of models) with the goal of boosting the fooling rate on the black-box `FR` model. Finally, we leverage the disentangled nature of latent space in the generative model and incorporate an identity-preserving regularization to effectively maintain the original visual identity. We now present the details of the loss functions used to incorporate the above ideas.

Textual Loss: A key ingredient of the proposed approach is text-based guidance to inconspicuously hide the adversarial perturbations into the makeup effect. This can be naively achieved by aligning the representation of t_{makeup} and the adversarial face $G_{\theta^*}(\boldsymbol{w})$ in the common embedding space of a pre-trained vision-language model (*e.g.* CLIP [47]). However, this approach will transform the *whole* output image to follow the makeup style of t_{makeup}, which results in low diversity. Therefore, we use a directional CLIP loss that aligns the CLIP-space direction between the text-image pairs of the original and adversarial images. Specifically,

$$\mathcal{L}_{\text{clip}} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|}, \tag{2}$$

where $\Delta T = E_T(t_{\text{makeup}}) - E_T(t_{\text{src}})$ and $\Delta I = E_I(G_{\theta^*}(\boldsymbol{w})) - E_I(\boldsymbol{x})$. Here, E_T and E_I are the text and image encoders of the CLIP model and t_{src} is the semantic text of the input image $\boldsymbol{x}$. Since we are dealing with faces, t_{src} can be simply set as "face". This loss localizes makeup transfer (*e.g.* red lipstick) without affecting privacy.

Adversarial Loss: Our goal is to traverse over the latent space $\mathcal{W}$ to find adversarial latent codes on the generative manifold whose face feature representation lies close to that of target image and far away from the original image itself *i.e.,* $\mathcal{D}(\boldsymbol{x}^p, \boldsymbol{x}) > \mathcal{D}(\boldsymbol{x}^p, \boldsymbol{x}^t)$. Hence, the adversarial loss is:

$$\mathcal{L}_{\text{adv}} = \mathcal{D}(G_{\theta^*}(\boldsymbol{w}), \boldsymbol{x}^t) - \mathcal{D}(G_{\theta^*}(\boldsymbol{w}), \boldsymbol{x}), \tag{3}$$

where $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - \cos[f(\boldsymbol{x}_1), f(\boldsymbol{x}_2))]$ is the cosine distance. Since the malicious `FR` model is unknown in the *black-box* setting, Eq. 3 cannot be solved directly. Instead, following AMT-GAN [22], we perform adversarial optimization on an ensemble of white-box surrogate models to imitate the decision boundary of the unknown `FR` model.

Identity Preservation Loss: The optimization over the generative manifold ensures that the protected image $\boldsymbol{x}^p$ is natural *i.e.,* artifact-free, however, it does not explicitly enforce the protected image to preserve the identity of the original image with respect to the human observer. To mitigate the issue, we take advantage of the semantic control exhibited by StyleGAN in its latent space. The latent code $\boldsymbol{w} \in \mathcal{W}$ impacts image generation by controlling different level of semantics in the output image. Specifically, latent codes corresponding to the initial layers of StyleGAN control high-level aspects such as pose, general hairstyle, and face shape [27]. Adversarially perturbing these latent layers can change these attributes, resulting in a change of identity (see Sec. 4.3). Latent codes corresponding to deeper layers of StyleGAN are associated with fine-level control such as makeup style [2]. Therefore, we perturb only those latent codes associated with deeper layers of StyleGAN, thereby restricting the adversarial faces to the identity preserving manifold. We further constrain the latent code to stay close to its initial value $\boldsymbol{w}_{\text{inv}}$ using the following regularization:

$$\mathcal{L}_{\text{latent}} = \|(\boldsymbol{w} \odot \boldsymbol{m}_{id}) - (\boldsymbol{w}_{\text{inv}} \odot \boldsymbol{m}_{id})\|_2, \quad (4)$$

where $\odot$ denotes element-wise product and $\boldsymbol{m}_{id}$ is an identity preservation mask that is 0 for the initial layers and 1 only for the deeper layers of the latent code. StyleGAN has 18 layers, each having a dimension of 512. The identity preservation mask is set to 1 only from layer 8 to 18. Finally, combining the three loss functions, we have

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} + \lambda_{\text{latent}}\mathcal{L}_{\text{latent}}, \quad (5)$$

where $\lambda_{\text{adv.}}$, λ_{clip}, and λ_{latent} are hyperparameters. Note that $\mathcal{L}_{\text{adv}}$ accounts for the adversarial objective in Eq. 1, while the text-guided makeup transfer ($\mathcal{L}_{\text{clip}}$) and identity-preserving regularization ($\mathcal{L}_{\text{latent}}$) implicitly enforce the naturalness constraint in Eq. 1.

4. Experiments

Implementation details: In all experiments, we use StyleGAN2 pretrained on the FFHQ face dataset as our generative model. For adversarial text guidance, we use a vision transformer-based CLIP model. For generator fine-tuning in the latent code initialization step, we use 450 iterations with value of λ_2 in Eq. 2 set to 0.5. For the makeup text input, we collect 40 text prompts based on the makeup style of diverse nature (details in supplementary material). For adversarial optimization, we use an Adam optimizer with β_1 and β_2 set to 0.9 and 0.999, respectively, and a learning rate of 0.01. We run the optimizer for 50 iterations to craft protected faces. We set the value of λ_{adv}, λ_{clip}, and λ_{latent} to 1, 0.5, and 0.01, respectively. All our experiments are conducted on a A100 GPU with 40 GB memory.

Datasets: We perform experiments for both face verification and identification settings. *Face verification*: We use CelebA-HQ [26] and LADN [16] for the impersonation attack. We select subset of 1,000 images from CelebA-HQ and report average results over 4 target identities provided by [22]. Similarly, for LADN, we divide the 332 images available into 4 groups, where images in each group aim to impersonate the target identities provided by [22]. For dodging attack, we use CelebA-HQ [26] and LFW [23] datasets. Specifically, we select 500 subjects at random and each subject has a pair of faces. *Face identification*: For impersonation and dodging, we use CelebA-HQ [26] and LFW [23] as our evaluation set. For both datasets, we randomly select 500 subjects, each with a pair of faces. We assign one image in the pair to the gallery set and the other to the probe set. Both impersonation and dodging attacks are performed on the probe set. For impersonation, we insert 4 target identities provided by [22] into the gallery set. A more detailed description of all datasets and pre-processing steps is provided in the supplementary material.

Target Models: We aim to protect user facial privacy by attacking four FR model with diverse back bones in the black-box settings. The target models include IRSE50 [21], IR152 [9], FaceNet [51], and MobileFace [5]. Following standard protocol, we align and crop the face images using MTCNN [75] before giving them as input to FR models. Further, we also report privacy protection performance based on commercial FR API including Face++ and Tencent Yunshentu FR platforms.

Evaluation metrics: Following [70], we use protection success rate (PSR) to evaluate the proposed approach. PSR is defined as the fraction of protected faces missclassified by the malicious FR system. To evaluate PSR, we use the thresholding and closed set strategies for face verification and identification, respectively. For face identification, we also use Rank-N targeted identity success rate (*Rank-N-T*) and untargeted identity success rate (*Rank-N-U*), where *Rank-N-T* means that target image $\boldsymbol{x}^t$ will appear at least once in the top N candidates shortlisted from the gallery and *Rank-N-U* implies that the top N candidate list does not have the same identity as that of original image $\boldsymbol{x}$. We also report results of PSNR (dB), SSIM, and FID [19] scores to evaluate the imperceptibility of method. Large PSNR and SSIM [65] indicates better match with the original images, while low FID score indicates more realistic images. For commercial APIs, we directly report the confidence score returned by the respective servers.

Baseline methods: We compare our approach with recent noise-based and makeup based facial privacy protection approaches. Noise based methods include PGD [37], MI-FGSM [10], TI-DIM [11], and TIP-IM [70], whereas makeup-based approaches are Adv-Makeup [71] and AMT-GAN [22]. We want to highlight that TIP-IM and AMT-GAN are considered the state-of-the-art (SOTA) for face privacy protection against *black-box* FR systems in noise-based and unrestricted settings, respectively. TIP-IM also incorporate multi-target objective in its optimization to find the optimal target image among multiple targets. For fair comparison, we use its single target variant.

4.1. Experimental Results

In this section, we present experimental results of our approach in *black-box* settings on four different pretrained

Table 2. Protection success rate (PSR %) of *black-box* impersonation attack under the face verification task. For each column, the other three FR systems are used as surrogates to generate the protected faces.

Method	CelebA-HQ				LADN-Dataset				Average
	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
Inverted	5.57	2.77	0.60	13.32	6.80	4.51	0.25	11.66	5.68
PGD [37]	36.87	20.68	1.85	43.99	40.09	19.59	3.82	41.09	25.60
MI-FGSM [10]	45.79	25.03	2.58	45.85	48.90	25.57	6.31	45.01	30.63
TI-DIM [11]	63.63	36.17	15.30	57.12	56.36	34.18	22.11	48.30	41.64
Adv-Makeup$_{\text{(IJCAI'21)}}$ [71]	21.95	9.48	1.37	22.00	29.64	10.03	0.97	22.38	14.72
TIP-IM$_{\text{(ICCV'21)}}$ [70]	54.40	37.23	40.74	48.72	65.89	43.57	**63.50**	46.48	50.06
AMT-GAN$_{\text{(CVPR'22)}}$ [22]	76.96	35.13	16.62	50.71	89.64	49.12	32.13	72.43	52.84
Ours	**81.10**	**48.42**	**41.72**	**75.26**	**91.57**	**53.31**	47.91	**79.94**	**64.90**

Table 3. Protection success rate (PSR %) of *black-box* dodging (top) and impersonation (bottom) attacks under the face identification task for LFW dataset [23]. For each column, the other three FR systems are used as surrogates to generate the protected faces. R1-U: Rank-1-Untargeted, R5-U: Rank-5-Untargeted, R1-T: Rank-1-Targeted, R5-T: Rank-5-Targeted.

Method	IRSE50		IR152		FaceNet		MobileFace		Average	
	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U
MI-FGSM [10]	70.2	42.6	58.4	41.8	59.2	34.0	68.0	47.2	63.9	41.4
TI-DIM [11]	79.0	51.2	67.4	54.0	74.4	52.0	79.2	61.6	75.0	54.7
TIP-IM$_{\text{(ICCV'21)}}$ [70]	81.4	52.2	71.8	54.6	76.0	49.8	82.2	63.0	77.8	54.9
Ours	**86.6**	**59.4**	**73.4**	**56.6**	**83.8**	**51.2**	**85.0**	**66.8**	**82.2**	**58.5**
	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T
MI-FGSM [10]	4.0	10.2	3.2	14.2	9.0	18.8	8.4	22.4	6.15	16.4
TI-DIM [11]	4.0	13.6	7.8	19.6	18.0	32.8	21.6	39.0	12.85	26.25
TIP-IM$_{\text{(ICCV'21)}}$ [70]	8.0	28.2	11.6	31.2	25.2	**56.8**	34.0	51.4	19.7	41.9
Ours	**11.2**	**37.8**	**16.0**	**51.2**	**27.4**	54.0	**39.0**	**61.2**	**23.4**	**51.05**

Method	FID ↓	PSR Gain ↑
Adv-Makeup [71]	4.23	0
TIP-IM [70]	38.73	35.34
AMT-GAN [22]	34.44	38.12
Ours	26.62	50.18

Table 4. FID comparison. PSR Gain is absolute gain in PSR relative to Adv-Makeup.

FR models under face verification and identification tasks. To generate protected images, we use three FR models as a surrogate to imitate the decision boundary of the fourth FR model. All results are averaged over 5 text based makeup styles that are provided in the supplementary material.

For face verification experiments, we set the system threshold value at 0.01 false match rate for each FR model *i.e.,* IRSE50 (0.241), IR152 (0.167), FaceNet (0.409), and MobileFace (0.302). Quantitative results in terms of PSR for impersonation attack under the face verification task are shown in Tab. 2. Our approach is able to achieve an average absolute gain of about 12% and 14% over SOTA unrestricted [22] and noise-based [70] facial privacy protection methods, respectively. Qualitative results are shown in Fig. 1 which shows that protected faces generated by our approach are more realistic. Results for dodging attacks under face verification are provided in the supplementary material. In Tab. 3, we also provide PSR vales under the face identification task for dodging (untargeted) and impersonation attacks. Our approach consistently outperforms recent methods at both *Rank-1* and *Rank-5* settings. We emphasize that we are the first to show effectiveness of generative models in offering untargeted privacy protection (dodging) in a more practical identification setting. Since AMT-GAN and Adv-Makeup are originally trained to impersonate target identity under the verification task, we have not included them in Tab. 3. Qualitative results for LFW and CelebA are provided in the supplementary material.

We report FID scores (lower is better) of our approach in Tab. 4 for CelebA and LADN datasets to measure naturalness. Adv-Makeup has the lowest FID score as it only applies makeup to the eye region without changing the rest of the face. However, this kind of restriction results in poor PSR. Our method has lower FID compared to TIP-IM and AMT-GAN and achieves the highest PSR. We provide PSNR and SSIM results in the supplementary material.

4.2. Effectiveness in Real-World Applications

We now show the effectiveness of our approach to protect facial images (through targeted impersonation) against commercial API such as Face++ and Tencent Yunshentu FR platform operating in the verification mode. These APIs return confidence scores between 0 to 100 to measure whether two images are similar or not, where a high confidence score indicates high similarity. As the training data and model parameters of these propriety FR models are unknown, it

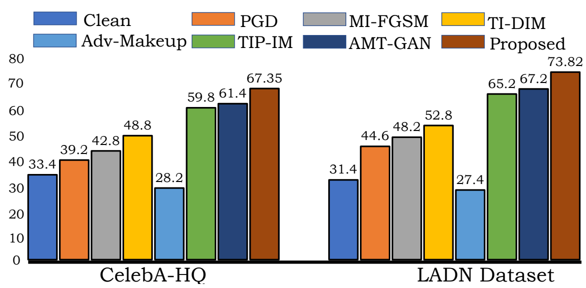

Figure 4. Average confidence score (higher is better) returned by a real-world face verification API, Face++, for impersonation attack. Our approach has a higher confidence score than state-of-the-art makeup and noise-based facial privacy protection methods.

Table 5. Impact of λ_{latent} on FID score and PSR.

λ_{latent}	0.5	0.1	0.05	0.01	0.005	0.0001	0
FID	11.6	21.4	25.2	27.8	30.1	38.4	43.2
PSR (%)	31.2	39.0	57.4	76.2	83.8	90.0	93.6

effectively mimics a real-world scenario. We protect 100 faces randomly selected from CelebA-HQ using the baselines and the proposed method. In Fig. 4, we show the average confidence score returned by Face++ against these images. These results indicate that our method has a high PSR compared to baselines. We defer more details and results for Tencent Yunshentu API to supplementary material.

4.3. Ablation Studies

Next, we report some ablations to evaluate the contributions of our loss components.

Makeup based text guidance: As shown in Fig. 5 (top), in the absence of text guidance, resulting images may contain artifacts due to increased perturbations induced by the adversarial objective. Text-guidance effectively hides the perturbations in the makeup, leading to more natural looking images. It also provides the user more flexibility to select a desired makeup style compared to a reference image.

Identity preserving regularization: Optimizing over the whole latent space provides more degrees of freedom and increases the PSR. However, it does not explicitly enforce adversarial optimization to preserve the user identity as shown in Fig. 5 (bottom). The proposed identity preserving regularization effectively preserves identity, while imitating the desire makeup style.

Impact of latent loss weight: Decreasing the weight assigned to the latent loss λ_{latent} results in an increase in both the FID score and PSR (and vice versa). Allowing the latent to deviate more from the initial inverted latent code of the given face image often results in artifacts caused by the adversarial loss, degrading naturalness but aiding privacy.

Robustness against textual variations. Finally, we evaluate the impact of different textual styles on the PSR. We select five text-based makeup styles to protect 1000 images of CelebA-HQ using our method. Results in Tab. 6 shows that PSR does not change significantly (low standard deviation) for different makeup styles, indicating robustness of our approach *wrt* different text-based makeup styles.

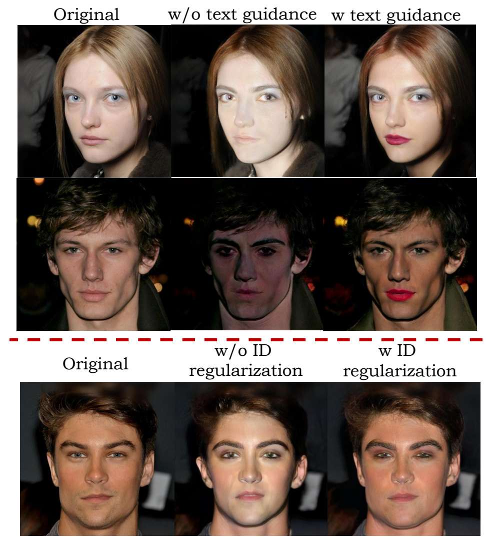

Figure 5. Top: Effect of makeup-based text guidance on the visual quality of the output images. Output images are able to impersonate the target identity for face verification. Text-prompt is *"tanned makeup with red lipstick"*. Bottom: Optimizing over all latent codes changes the identity of the protected image. Our identity-preserving regularization enforces the adversarial optimization to search for latent codes that hide the perturbations in the makeup effect while simultaneously preserving visual identity.

Table 6. Impact of different textual makeup styles on PSR. Makeup styles are *"tanned"*, *"pale"*, *"pink eyeshadows"*, *"red lipstick"*, and *"Matte"*. Std. denotes standard deviation.

	t^1_{makeup}	t^2_{makeup}	t^3_{makeup}	t^4_{makeup}	t^5_{makeup}	Std.
PSR	74.1	77.3	78.4	78.7	79.2	1.24

5. Conclusion

We have proposed a framework to protect privacy of face images on online platforms by carefully searching for adversarial codes in the low-dimensional latent manifold of a pre-trained generative model. We have shown that incorporating a makeup text-guided loss and an identity preserving regularization effectively hides the adversarial perturbations in the makeup style, provides images with high quality, and preserves human-perceived identity. While this approach is robust to the user-defined text-prompt and target identity, it would be beneficial if the text-prompt and target identity can be automatically selected based on the given face image. Limitations of our method include high computational cost at the time of protected face generation.

References

[1] Shane Ahern, Dean Eckles, Nathaniel S Good, Simon King, Mor Naaman, and Rahul Nair. Over-exposed? privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 357–366, 2007. 2

[2] Amit H Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Oren Patashnik, and Daniel Cohen-Or. State-of-the-art in the architecture, methods and applications of stylegan. In *Computer Graphics Forum*, volume 41, pages 591–611. Wiley Online Library, 2022. 2, 6

[3] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. *arXiv preprint arXiv:1904.06347*, 2019. 2, 3

[4] Jia-Wei Chen, Li-Ju Chen, Chia-Mu Yu, and Chun-Shien Lu. Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6478–6487, 2021. 3

[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 6

[6] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations*, 2020. 2, 3

[7] William L Croft, Jörg-Rüdiger Sack, and Wei Shi. Differentially private facial obfuscation via generative adversarial networks. *Future Generation Computer Systems*, 129:358–379, 2022. 3

[8] Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. Fast geometrically-perturbed adversarial faces. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1979–1988. IEEE, 2019. 3

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pages 9185–9193, 2018. 6, 7

[11] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*, pages 4312–4321, 2019. 6, 7

[12] Hang Du, Hailin Shi, Dan Zeng, Xiao-Ping Zhang, and Tao Mei. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM Computing Surveys (CSUR)*, 54(10s):1–42, 2022. 2

[13] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022. 3

[14] Kevin Frans, Lisa B Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021. 3

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014. 2

[16] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10481–10490, 2019. 6

[17] Nitzan Guetta, Asaf Shabtai, Inderjeet Singh, Satoru Momiyama, and Yuval Elovici. Dodging attack using carefully crafted natural makeup. *arXiv preprint arXiv:2109.06467*, 2021. 3

[18] Rebecca Heilweil. The world's scariest facial recognition company explained. *Vox, May*, 8, 2020. 2

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[20] Kashmir Hill. The secretive company that might end privacy as we know it. *The New York Times*, 18:2020, 2020. 2

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6

[22] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. 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*, pages 15014–15023, 2022. 1, 2, 3, 5, 6, 7

[23] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 6, 7

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[25] Kazuya Kakizaki and Kosuke Yoshida. Adversarial image translation: Unrestricted adversarial examples in face recognition systems. *arXiv preprint arXiv:1905.03421*, 2019. 2, 3

[26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 6

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 8110–8119, 2020. 2, 6

[28] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3

[29] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021. 3

[30] Zhenzhong Kuang, Zhiqiang Guo, Jinglong Fang, Jun Yu, Noboru Babaguchi, and Jianping Fan. Unnoticeable synthetic face replacement for image privacy protection. *Neurocomputing*, 457:322–333, 2021. 3

[31] Zhenzhong Kuang, Huigui Liu, Jun Yu, Aikui Tian, Lei Wang, Jianping Fan, and Noboru Babaguchi. Effective de-identification generative adversarial network for face anonymization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3182–3191, 2021. 3

[32] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 3

[33] Tao Li and Min Soo Choi. Deepblur: A simple and effective method for natural image obfuscation. *arXiv preprint arXiv:2104.02655*, 1, 2021. 3

[34] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3

[35] Fangcheng Liu, Chao Zhang, and Hongyang Zhang. Towards transferable unrestricted adversarial examples with minimum changes. *arXiv preprint arXiv:2201.01102*, 2022. 3

[36] Suolan Liu, Lizhi Kong, and Hongyuan Wang. Face detection and encryption for privacy preserving in surveillance video. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 162–172. Springer, 2018. 3

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*, 2018. 6, 7

[38] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 2021. 2, 3

[39] Dongbin Na, Sangwoo Ji, and Jong Kim. Unrestricted black-box adversarial attack using gan with limited queries. *arXiv preprint arXiv:2208.11613*, 2022. 2

[40] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[41] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016. 3

[42] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE, 2017. 3

[43] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 1

[44] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[45] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018. 2

[46] Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge Belongie, and Ser-Nam Lim. Robustness and generalization via generative adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15711–15720, 2021. 3

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 5

[48] Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun-Cheng Chen, Vishal M Patel, Carlos D Castillo, and Rama Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, 2018. 2

[49] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 5

[50] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[51] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6

[52] Sachith Seneviratne, Nuran Kasthuriarachchi, Sanka Rasnayaka, Danula Hettiachchi, and Ridwan Shariffdeen. Does a face mask protect my privacy?: Deep learning to predict protected attributes from masked face images. In *Australasian Joint Conference on Artificial Intelligence*, pages 91–102. Springer, 2022. 3

[53] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pages 1589–1604, 2020. 2, 3

[54] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3):1–30, 2019. 3

[55] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 3

[56] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018. 3

[57] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 2

[58] Huan Tian, Tianqing Zhu, and Wanlei Zhou. Fairness and privacy preservation for facial images: Gan-based methods. *Computers & Security*, 122:102902, 2022. 3

[59] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 5

[60] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 3

[61] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 1, 2

[62] Shunxin Wang, Una M Kelly, and Raymond NJ Veldhuis. Gender obfuscation through face morphing. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2021. 3

[63] Ya Wang, Tianlong Bao, Chunhui Ding, and Ming Zhu. Face recognition in real-world surveillance videos with deep learning method. In *2017 2nd international conference on image, vision and computing (icivc)*, pages 239–243. IEEE, 2017. 2

[64] Yinggui Wang, Jian Liu, Man Luo, Le Yang, and Li Wang. Privacy-preserving face recognition in the frequency domain. 2022. 3

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[66] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18072–18081, 2022. 3

[67] Emily Wenger, Shawn Shan, Haitao Zheng, and Ben Y Zhao. Sok: Anti-facial recognition technology. *arXiv preprint arXiv:2112.04558*, 2021. 2

[68] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018. 2, 3

[69] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11845–11854, 2021. 3

[70] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV'21)*, pages 3897–3907, 2021. 1, 2, 3, 6, 7

[71] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, pages 1252–1258, 2021. 2, 6, 7

[72] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021. 2, 3

[73] Shengming Yuan, Qilong Zhang, Lianli Gao, Yaya Cheng, and Jingkuan Song. Natural color fool: Towards boosting black-box unrestricted attacks. *arXiv preprint arXiv:2210.02041*, 2022. 3

[74] Jiaming Zhang, Jitao Sang, Xian Zhao, Xiaowen Huang, Yanfeng Sun, and Yongli Hu. Adversarial privacy-preserving filter. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1423–1431, 2020. 2

[75] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 6

[76] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1039–1048, 2020. 2, 3

[77] Yaoyao Zhong and Weihong Deng. Opom: Customized invisible cloak towards face privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 3

[78] Jizhe Zhou and Chi-Man Pun. Personal privacy protection via irrelevant faces tracking and pixelation in video live streaming. *IEEE Transactions on Information Forensics and Security*, 16:1088–1103, 2020. 3

[79] Zheng-An Zhu, Yun-Zhong Lu, and Chen-Kuo Chiang. Generating adversarial examples by makeup attacks on face recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2516–2520. IEEE, 2019. 3