DiffAM: Diffusion-based Adversarial Makeup Transfer for Facial Privacy Protection

Yuhao Sun, Lingyun Yu*, Hongtao Xie, Jiaming Li, Yongdong Zhang
University of Science and Technology of China
{syh3327,ljmd}@mail.ustc.edu.cn {yuly,htxie,zhyd73}@ustc.edu.cn

Abstract

*With the rapid development of face recognition (FR) systems, the privacy of face images on social media is facing severe challenges due to the abuse of unauthorized FR systems. Some studies utilize adversarial attack techniques to defend against malicious FR systems by generating adversarial examples. However, the generated adversarial examples, i.e., the protected face images, tend to suffer from subpar visual quality and low transferability. In this paper, we propose a novel face protection approach, dubbed DiffAM, which leverages the powerful generative ability of diffusion models to generate high-quality protected face images with adversarial makeup transferred from reference images. To be specific, we first introduce a makeup removal module to generate non-makeup images utilizing a fine-tuned diffusion model with guidance of textual prompts in CLIP space. As the inverse process of makeup transfer, makeup removal can make it easier to establish the deterministic relationship between makeup domain and non-makeup domain regardless of elaborate text prompts. Then, with this relationship, a CLIP-based makeup loss along with an ensemble attack strategy is introduced to jointly guide the direction of adversarial makeup domain, achieving the generation of protected face images with natural-looking makeup and high black-box transferability. Extensive experiments demonstrate that DiffAM achieves higher visual quality and attack success rates with a gain of **12.98%** under black-box setting compared with the state of the arts. The code will be available at https://github.com/HansSunY/DiffAM.*

1. Introduction

Recent years have witnessed major advances in face recognition (FR) systems based on deep neural networks (DNNs), which have been applied to various scenarios. However, the expanding capabilities of FR systems have raised concerns about the threats they pose to facial privacy.

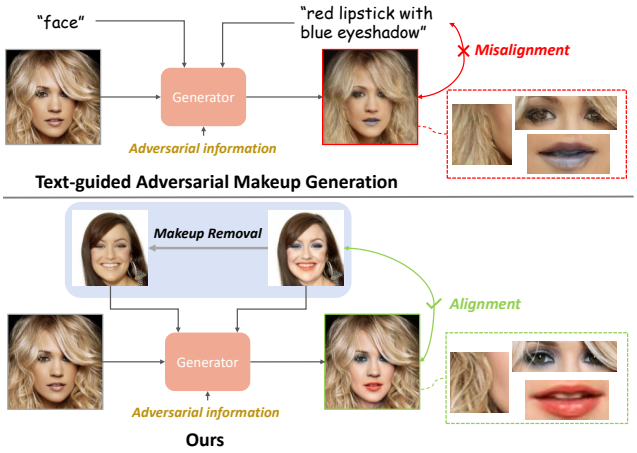

Figure 1. Core idea comparison. Text-guided method generates adversarial makeup simply with a pair of textual prompts. The coarse-grained guidance of text results in unexpected makeup generation (as shown in red boxes). Our method introduces a makeup removal module to transition this task from text-based guidance to image-based guidance and controls the direction and distance of refined adversarial makeup generation (as shown in green boxes).

Particularly, FR systems have the potential for unauthorized surveillance and monitoring, which can analyze social media profiles without consent with the widespread availability of face images on social media[2, 41]. Therefore, it is crucial to find an effective approach to protect facial privacy against unauthorized FR systems.

Many studies leverage adversarial attacks[16] for facial privacy protection, which generate noise-based[6, 39, 49] or patch-based[27, 47] perturbations on face images. Nevertheless, to achieve ideal attack effects, most of the adversarial perturbations generated by these methods are noticeable and cluttered. Consequently, the protected face images tend to suffer from poor visual quality. To gain more natural-looking adversarial examples, makeup-based methods[21, 38, 50] are attracting considerable attention. These methods organize perturbations as makeup, which can generate protected face images with adversarial makeup. However, such

*Corresponding author.

makeup-based methods have the following problems: (1) **Subpar visual quality.** Most of the existing works generate makeup with generative adversarial networks[15] (GANs). The protected face images generated by these GAN-based approaches often have unexpected makeup artifacts and struggle to preserve attributes unrelated to makeup, such as background and hair, leading to poor visual quality. (2) **Weakness in fine-grained makeup generation.** The fine-grained information of generated makeup, like the position, color shade, range and luminosity, may not align consistently with expected makeup, especially for text-guided makeup generation method as shown in Fig. 1. (3) **Low black-box transferability.** Attack effects highly rely on robust makeup generation[32]. Due to the limitation of makeup generation quality proposed above, these methods suffer from low attack success rates under the black-box setting including commercial APIs. In summary, it is still challenging to simultaneously achieve satisfying makeup generation and good attack effects in black-box scenarios.

Diffusion models[19, 43, 44] have shown better performance than GANs in image generation tasks thanks to more stable training process and better coverage of image distribution[10]. Recent works explore to guide diffusion models in CLIP[36] space with textual prompts[1, 25, 28, 31], demonstrating promising results. Thus, it is encouraging to utilize diffusion models to generate protected face images with both high visual quality and transferability. However, for more refined tasks like makeup transfer, textual prompts are too coarse-grained for guidance as illustrated in Fig. 1. So it is worth considering a more fine-grained way of direction guidance in CLIP space for diffusion models.

To address the above problems, we observe that although it is hard to control the refined generation of reference makeup directly with textual prompts, the makeup of reference face image can be easily removed by a fine-tuned diffusion model with guidance of textual prompts in CLIP space[25]. Through this inverse process of makeup generation, domains of makeup and non-makeup can be definitely connected. Following this line of thought, we propose DiffAM, a novel diffusion-based adversarial makeup transfer framework to protect facial privacy. The overall pipeline of DiffAM is shown in Fig. 2. DiffAM aims to generate protected face images with adversarial makeup style transferred from a given reference image. It is designed as two modules, a **text-guided makeup removal module** and an **image-guided makeup transfer module**. In the text-guided makeup removal module, we aim to remove the makeup of reference images, gaining the corresponding non-makeup reference images. This deterministic process simplifies the exploration of the makeup and non-makeup domains' relationship. Notably, the difference between the latent codes of makeup and non-makeup images of reference image in CLIP space indicates the accurate direction from non-makeup domain to makeup domain, providing alignment information for fine-grained makeup transfer. In the image-guided adversarial makeup transfer module, a CLIP-based makeup loss is proposed, combined with an ensemble attack strategy to control the precise generation direction and distance to adversarial makeup domain. In this way, high-quality makeup with strong transferability can be generated with fine-grained cross-domain guidance in CLIP space with diffusion models.

Extensive experiments on the CelebA-HQ[24] and LADN[17] datasets demonstrate the effectiveness of our method in protecting facial privacy against black-box FR models with a gain of **12.98%**, while achieving outstanding visual quality. In summary, our main contributions are:

- A novel diffusion-based adversarial makeup transfer method, called DiffAM, is proposed for facial privacy protection, intending to craft adversarial faces with high visual quality and black-box transferability.
- A text-guided makeup removal module is designed to establish the deterministic relationship between non-makeup and reference makeup domains, offering precise cross-domain alignment guidance for makeup transfer.
- A CLIP-based makeup loss is proposed for refined makeup generation. It consists of a makeup direction loss and a pixel-level makeup loss, which jointly control the direction and distance of makeup generation.

2. Related Works

2.1. Adversarial Attacks on Face Recognition

Due to the vulnerability of DNNs to adversarial examples[16, 45], many methods have been proposed to attack DNN-based face recognition (FR) systems. According to the knowledge about the target FR model, the attacks can be categorized into two main types, white-box attacks[16, 33, 48] and black-box attacks[11, 47, 49, 53].

In white-box attacks, the attacker requires complete information about the target models. However, it is hard to get full access to unauthorized FR systems in real-world scenarios. So black-box attacks, without the limitation of knowledge about the target models, are more suitable in such scenarios. Noise-based methods[11, 12, 49], a common form of black-box attack, can generate transferable adversarial perturbations on face images. But due to the ℓ_∞ constraint of noise, the attack strength cannot be guaranteed. For better attack effect, patch-based methods[27, 40, 47] add abrupt adversarial patches to the limited region of face images. Although these methods attain a measure of privacy protection, the visual quality of the resulting protected face images is often compromised and suffers from weak transferability. Recent works attempt to protect face images with adversarial makeup[21, 38, 50], which is an ideal solution for balancing visual quality and transferability. These

methods hide the adversarial information in the generated makeup style, which can fool FR systems in an imperceptive way. However, existing makeup-based methods tend to suffer from poor visual quality and low transferability. And the attributes unrelated to makeup are hard to be completely preserved. Therefore, in this work, we propose a novel face protection approach DiffAM to improve the quality and black-box transferability of adversarial makeup through fine-grained guidance in CLIP space.

2.2. Makeup Transfer

Makeup transfer[3, 4, 8, 17, 23, 30] aims to transfer makeup styles from the reference faces to the source faces while preserving the original face identity. As a typical image-to-image translation task, many approaches employ generative adversarial networks (GANs) for makeup transfer. BeautyGAN[30] first introduces a dual input/output GAN to achieve makeup transfer and removal simultaneously. Moreover, it proposes a pixel-wise histogram matching loss as guidance for makeup transfer in different face regions which has been subsequently adopted by many methods. LADN[17] adopts multiple overlapping local discriminators and asymmetric losses for heavy facial makeup transfer. Besides the above GAN-based methods, BeautyGlow[4] uses the Glow framework for makeup transfer by decomposing the latent vector into non-makeup and makeup parts. Taking advantage of the good visual properties of makeup transfer, we apply the concept of makeup style to facial privacy protection. The proposed DiffAM organizes the distribution of adversarial information semantically into adversarial makeup, which can minimize the impact on the visual quality of the protected face images while ensuring the effectiveness of attacks on the FR system.

2.3. Diffusion model and Style Transfer

Diffusion models[10, 19, 35, 42] are a class of probabilistic generative models, which have impressive performance in generating high-quality images. They have been applied to various tasks, such as image generation[10, 43], image editing[1, 7, 25], image super-resolution[14, 29] and style transfer[25, 28, 52]. Style transfer is an image-to-image translation[22, 54] task that combines the content of source image and the style of reference image. Existing diffusion-based methods leverage the alignment between text and images in CLIP space for text-driven style transfer[25, 28]. As a subtask of style transfer, makeup transfer can also be guided with text. However, the control of text is too rough for more refined makeup transfer in comparison to global style transfer. Given a reference makeup image, simply using text is insufficient to generate precise makeup, such as the intensity, shape, and position. Considering the limitation of text, we point that makeup removal, the inverse process of makeup transfer, can provide deterministic guidance from non-makeup domain to makeup domain. Moreover, a CLIP-based makeup loss is introduced for image-driven makeup transfer.

3. Method

3.1. Problem Formulation

Black-box attacks on face recognition (FR) systems can be further divided into targeted attacks ($i.e.$, impersonation attacks) and non-targeted attacks ($i.e.$, dodging attacks). For more efficient protection of face images, *we focus on targeted attack which aims to mislead FR systems to recognize the protected faces as the specified target identity.* The targeted attack can be defined as an optimization problem:

$$\min_{x'} L_{adv} = \mathcal{D}(m(x'), m(x^*)), \tag{1}$$

where x' is the protected face image, x^* is the target face image, m represents the feature extractor of FR models, $\mathcal{D}(.)$ represents a distance metric.

Particularly, as for adversarial makeup transfer, the protected face image x' is obtained by transferring makeup style from the reference image y to the clean face image x, which can be formulated as:

$$x' = \mathcal{G}(x, y), \tag{2}$$

where $\mathcal{G}$ is an adversarial makeup transfer network.

3.2. DiffAM

To generate natural-looking and transferable adversarial makeup against FR models, DiffAM aims to explore precise and fine-grained guidance of generation from non-makeup domain to adversarial makeup domain, which is the overlap domain between reference makeup domain and adversarial domain, as shown in Fig. 3(a). To achieve this, DiffAM consists of two stages: text-guided makeup removal and image-guided adversarial makeup transfer, as illustrated in Fig. 2. The details of each component are described as follows.

3.2.1 Text-guided Makeup Removal

Adopting makeup transfer against FR models, an intuitive idea is to directly use text pairs to guide the generation of makeup. However, the coarse-grained text is hard to build the precise relationship between non-makeup domain and reference makeup domain. Concretely, the details of reference makeup, such as color depth, shape, etc., are difficult to control in text, resulting in undesired makeup generation. To eliminate ambiguity caused by textual guidance, we innovatively design the text-guided makeup removal module to remove the makeup style of reference image y and obtain the corresponding non-makeup image $\hat{y}$ with the guidance of text pair in CLIP space. The pair of reference images with and without makeup can connect makeup domain

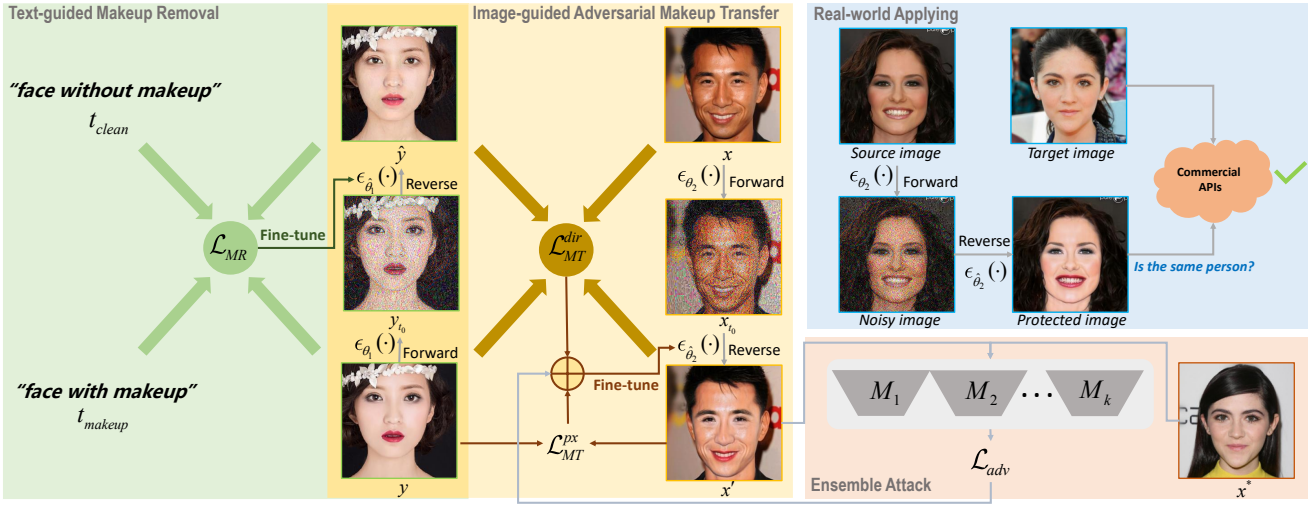

Figure 2. Overview of DiffAM. DiffAM is a two-stage framework that generates protected face image x' by transferring the makeup style of y to x. Specifically, in text-guided makeup removal module, we input a reference image y and obtain the non-makeup $\hat{y}$ through text guidance, determining the precise makeup direction. Then, in image-guided adversarial makeup transfer module, we input a face image x and obtain the adversarial-makeup image x' through image guidance of y and $\hat{y}$, along with an ensemble attack strategy.

and non-makeup domain in CLIP space, as illustrated in Fig. 3(b), providing the deterministic cross-domain guidance for subsequent stage of adversarial makeup transfer.

Given the reference image y, we first convert it to the latent y_{t_0} by forward diffusion process with a pre-trained diffusion model ϵ_{θ_1}. In the reverse diffusion process, the diffusion model ϵ_{θ_1} is fine-tuned for makeup removal to obtain the non-makeup image $\hat{y}$, which is guided by directional CLIP loss $\mathcal{L}_{MR}$ [25]:

$$\mathcal{L}_{MR} = 1 - \frac{\Delta I_y \cdot \Delta T}{\|\Delta I_y\| \|\Delta T\|}, \tag{3}$$

where $\Delta I_y = E_I(\hat{y}(\hat{\theta}_1)) - E_I(y)$ and $\Delta T = E_T(t_{clean}) - E_T(t_{makeup})$. Here, E_I and E_T are the image and text encoders of CLIP model[36], $\hat{y}(\hat{\theta}_1)$ is the sampled image from y_{t_0} with the optimized parameter $\hat{\theta}_1$, t_{clean} and t_{makeup} are the text descriptions for non-makeup and makeup domains, which can be simply set as *"face without makeup"* and *"face with makeup"*.

To preserve identity information and image quality, we introduce the face identity loss $\mathcal{L}_{id}(\hat{y}, y)$ [9] and perceptual loss $\mathcal{L}_{LPIPS}(\hat{y}, y)$ [51]. As for total loss $\mathcal{L}_{total}$, we have

$$\mathcal{L}_{total} = \lambda_{MR}\mathcal{L}_{MR} + \lambda_{id}\mathcal{L}_{id} + \lambda_{LPIPS}\mathcal{L}_{LPIPS}, \tag{4}$$

where λ_{MR}, λ_{id} and λ_{LPIPS} are weight parameters.

It is worth noting that we use deterministic DDIM sampling and DDIM inversion[43] as the reverse diffusion process and forward diffusion process. The reconstruction capability of deterministic DDIM inversion and sampling ensures the effects of makeup removal.

3.2.2 Image-guided Adversarial Makeup Transfer

To protect source image x against FR models, the image-guided adversarial makeup transfer module aims to generate protected face image x' with adversarial makeup transferred from reference image y. The protected face image x' misleads FR models into recognizing it as the target identity x^*, as shown in Fig. 2. After getting y and $\hat{y}$, a CLIP-based makeup loss $\mathcal{L}_{MT}$ coordinating with an ensemble attack strategy is introduced to align the direction between x and x' with the direction between non-makeup domain and adversarial makeup domain. The fine-grained cross-domain alignment ensures the quality and black-box transferability of adversarial makeup style.

Given the source image x, we first get the latent x_{t_0} through deterministic DDIM inversion with another pre-trained diffusion model ϵ_{θ_2}. Then, the diffusion model ϵ_{θ_2} is fine-tuned to generate protected face image x' with guidance of CLIP-based makeup loss $\mathcal{L}_{MT}$ and ensemble attack loss L_{adv}. We also incorporate a makeup-irrelevant information preservation operation for better visual quality during fine-tuning. The details of the fine-tuning process are presented as follows.

CLIP-based Makeup Loss. During the stage of makeup removal, the learned direction from reference makeup domain to non-makeup domain in CLIP space is expressed as $\Delta I_y = E_I(\hat{y}) - E_I(y)$. As shown in Fig. 3(c), We can just reverse the direction to get the guidance of direction from non-makeup domain to reference makeup domain:

$$\Delta I_{ref} = -\Delta I_y = E_I(y) - E_I(\hat{y}), \tag{5}$$

where E_I is the image encoder of CLIP model. In addition to maintaining consistency with the style removal stage

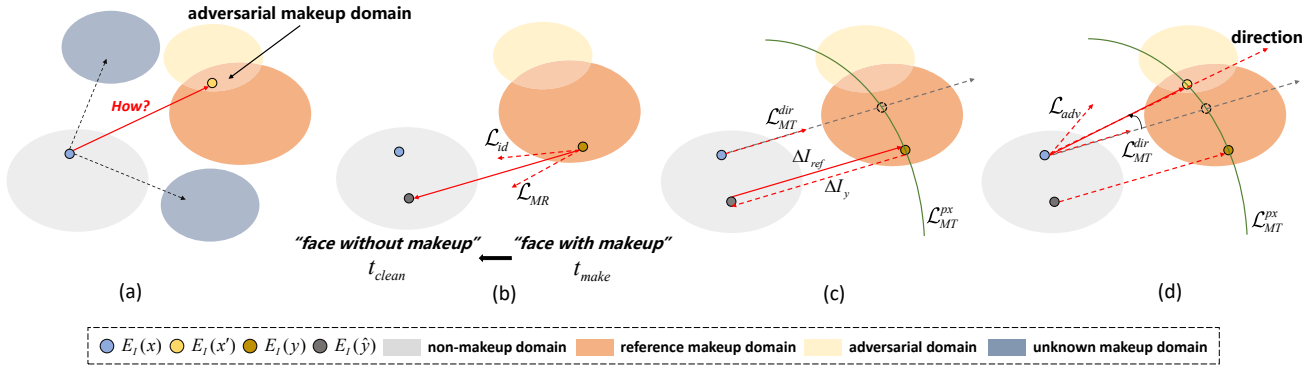

Figure 3. The process of adversarial makeup transfer in CLIP space. (a) It is challenging to directly find a precise path from the non-makeup domain to the adversarial makeup domain. (b) The process of text-guided makeup removal can help establish the relationship between domains. (c) The inverse direction of makeup removal indicates the direction to makeup domain for makeup transfer and the pixel-level makeup loss guides the distance to makeup domain. (d) The direction of ensemble attack and makeup transfer jointly guide the final direction to the adversarial makeup domain.

in CLIP space, E_I has powerful image understanding capabilities, which can facilitate better extraction of semantic information of makeup styles, such as shape and the relative position on the face[13, 25]. In this way, ΔI_{ref} can achieve more semantic and holistic supervision than simple pixel-level guidance or text guidance. To align ΔI_x, the direction between x and x', with ΔI_{ref} in CLIP space, a makeup direction loss is proposed:

$$\mathcal{L}_{MT}^{dir} = 1 - \frac{\Delta I_x \cdot \Delta I_{ref}}{\|\Delta I_x\| \|\Delta I_{ref}\|}, \tag{6}$$

where $\Delta I_x = E_I(x'(\hat{\theta}_2)) - E_I(x)$ and $x'(\hat{\theta}_2)$ is the protected face image generated by fine-tuned diffusion model $\epsilon_{\hat{\theta}_2}$. By aligning image pairs in CLIP space, makeup direction loss controls precise direction for makeup transfer.

Besides the guidance of makeup transfer direction, we also need to consider the makeup transfer distance between makeup domain and non-makeup domain, as shown in Fig. 3(c), which determines the intensity and accurate color of makeup. Therefore, a pixel-level makeup loss L_{MT}^{px}[30] is employed to constrain makeup transfer distance in pixel space. We conduct histogram matching between generated image x' and reference image y on three facial regions as guidance for the intensity of makeup. The pixel-level makeup loss is defined as:

$$\mathcal{L}_{MT}^{px} = \|x' - HM(x', y)\|, \tag{7}$$

where $HM(.)$ represents the histogram matching.

Combining the makeup direction loss and pixel-level makeup loss, the CLIP-based makeup loss is expressed as:

$$\mathcal{L}_{MT} = \lambda_{dir}\mathcal{L}_{MT}^{dir} + \lambda_{px}\mathcal{L}_{MT}^{px}. \tag{8}$$

With the joint guidance of $\mathcal{L}_{MT}^{dir}$ and $\mathcal{L}_{MT}^{px}$ for makeup transfer direction and distance, the generated makeup image x' can precisely fall within the reference makeup domain, achieving excellent makeup transfer effects.

Ensemble Attack. In addition to guidance in makeup transfer direction, there is also a need for guidance in the adversarial direction to find the final adversarial makeup domain, as shown in Fig. 3(d). To solve the optimization problem in Eq. (1), an ensemble attack strategy[21] is introduced. We choose K pre-trained FR models with high recognition accuracy as surrogate models for fine-tuning, aiming to find the direction towards a universal adversarial makeup domain. The ensemble attack loss is formulated as:

$$\mathcal{L}_{adv} = \frac{1}{K}\sum_{k=1}^{K}[1 - \cos(m_k(x'), m_k(x^*))], \tag{9}$$

where m_k represents the k-th pre-trained FR model and we use cosine similarity as the distance metric.

The ensemble attack loss $\mathcal{L}_{adv}$ adjusts the generation direction from the makeup domain to the adversarial makeup domain, improving the transferability of adversarial makeup under black-box settings.

Preservation of Makeup-Irrelevant Information. To ensure the visual quality of protected face images, it is crucial to minimize the impact on makeup-irrelevant information, such as identity and background, during makeup transfer. However, due to fine-tuning the diffusion model ϵ_{θ_2} in the sampling process, the cumulative error between the prediction noise of ϵ_{θ_2} and $\epsilon_{\hat{\theta}_2}$ will increase with denoising steps, resulting some unexpected distortion besides makeup style. To address this problem, we leverage the progressive generation property of diffusion models[19, 34], where coarse-grained information ($e.g.$, layout, shape) is focused at early denoising steps while semantic details at later steps. Makeup style is typically generated in the final steps of denoising process as a kind of fine-grained information of face. Thus, we propose to reduce the time step

T in DDIM inversion and sampling for retention of most makeup-irrelevant information. This simple but effective operation can greatly improve the visual quality of protected face and accelerate the whole process of makeup transfer.

Moreover, the perceptual loss $\mathcal{L}_{LPIPS}(x', x)$ and ℓ_1 loss are further introduced to explicitly control generation quality and pixel similarity:

$$\mathcal{L}_{vis} = \mathcal{L}_{LPIPS}(x', x) + \lambda_{\ell_1} \|x' - x\|. \tag{10}$$

Total Loss Function. By combining all the above loss functions, we have total loss function as follows:

$$\mathcal{L} = \lambda_{MT}\mathcal{L}_{MT} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{vis}\mathcal{L}_{vis}, \tag{11}$$

where λ_{MT}, λ_{adv} and λ_{vis} are weight parameters.

4. Experiments

4.1. Experimental Settings

Datasets. For makeup removal, we randomly sample 200 makeup images from MT dataset[30], which consists of 2719 makeup images and 1115 non-makeup images, for fine-tuning. For adversarial makeup transfer, we randomly sample 200 images from CelebA-HQ dataset[24] for fine-tuning. To evaluate the effectiveness of DiffAM, we choose CelebA-HQ and LADN[17] as our test sets. For CelebA-HQ, we select a subset of 1000 images and divide them into four groups, each of which has a target identity[21]. Similarly, for LADN, we divide the 332 images into four groups for attack on different target identities.

Benchmark. We do comparisons with multiple benchmark schemes of adversarial attacks, including PGD[33], MI-FGSM[11], TI-DIM[12], TIP-IM[49], Adv-Makeup[50], AMT-GAN[21] and CLIP2Protect[38]. PGD, MI-FGSM, TI-DIM and TIP-IM are typical noise-based methods, while Adv-Makeup, AMT-GAN, CLIP2Protect and DiffAM are makeup-based methods that also exploit makeup transfer to generate protected face images.

Target Models. We choose four popular public FR models as the attacked models, including IR152[9], IRSE50[20], FaceNet[37] and MobileFace[5]. Three of them are chosen for ensemble attack during training and the remaining one serves as the black-box model for testing. Meanwhile, we evaluate the performance of DiffAM on commercial FR APIs including Face++[1] and Aliyun[2].

Implemention Details. For text-guided makeup removal and image-guided makeup transfer, we use ADM[10] pre-trained on Makeup Transfer (MT) dataset[30] and CelebA-HQ dataset[24] respectively as the generative model. To fine-tune diffusion models, we use an Adam optimizer[26] with an initial learning rate of 4e-6. It is increased linearly by 1.2 per 50 iterations. As mentioned in Sec. 3.2.2, we set total time step $T = 60$ and (S_{inv}, S_{sam}) = (20, 6), where S_{inv} and S_{sam} represent the discretization steps of DDIM inversion and sampling. The diffusion models are fine-tuned with 6 epochs. All our experiments are conducted on one NVIDIA RTX3090 GPU.

Evaluation Metrics. Following [21, 38], we use attack success rate (ASR) to evaluate the effectiveness of privacy protection of different methods. When calculating the ASR, we set False Acceptance Rate (FAR) at 0.01 for each FR model. In addition, we use FID[18], PSNR(dB) and SSIM[46] to evaluate the image quality of protected face images.

4.2. Comparison Study

Comparison on black-box attacks. Tab. 1 reports quantitative results of black-box attacks against four popular FR models on CelebA-HQ and LADN datasets. We test the performance of targeted attack against four target identities[21], with the results of DiffAM averaged over 5 reference makeup images from MT-dataset, following [38]. To simulate real-world protection scenarios, the target face images used during testing are different images of the same individual compared to the one used during training. The average black-box ASRs of DiffAM are significantly about 28% and 13% higher than SOTA noise-based method TIP-IM and makeup-based method CLIP2Protect. DiffAM also maintains a good attack effectiveness on Facenet, which is difficult to attack using other methods. The results show that DiffAM has strong black-box transferability, which demonstrates the role of DiffAM in accurate guidance to the adversarial makeup domain as we expected.

Comparison on image quality. Tab. 2 reports the evaluations of image quality. We choose Adv-makeup, AMT-GAN and CLIP2Protect, three latest makeup-based methods, as benchmarks for comparison. Adv-makeup has the best performance in all quantitative assessments. This is because Adv-makeup only generates eyeshadow compared to the full-face makeup generation of the others. Although Adv-makeup has minimal image modification, the trade-off is a significantly lower attack success rate as shown in Tab. 1. Compared to AMT-GAN and CLIP2Protect, DiffAM achieves lower FID scores and higher PSNR and SSIM scores, which indicates that the adversarial makeups generated by DiffAM are more natural-looking and have less impact on images at the pixel level.

We also show the qualitative comparison of visual quality in Fig. 4. Note that for text-guided method CLIP2Protect, we use textual prompts, such as *"purple lipstick with purple eyeshadow"*, derived from the reference images to generate makeup. Compared to the noise-based method TIP-IM, DiffAM generates more natural-looking protected face images without noticeable noise patterns.

[1] https://www.faceplusplus.com/face-comparing/

[2] https://vision.aliyun.com/experience/detail?&tagName=facebody&children=CompareFace

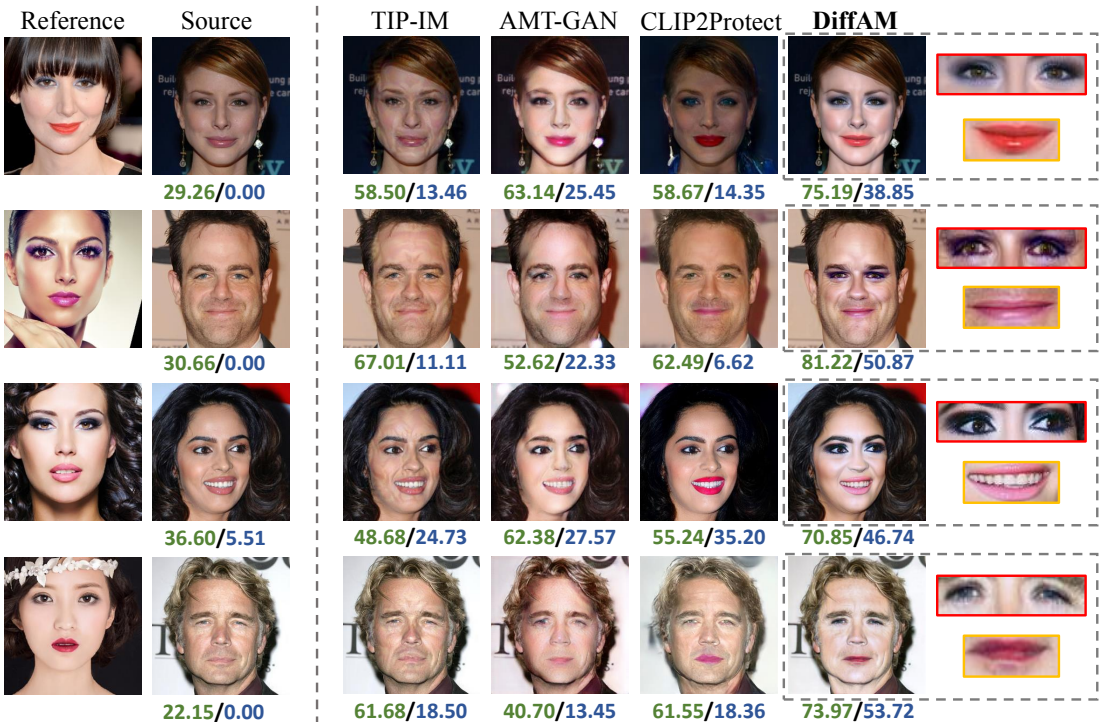

Figure 4. Visualizations of the protected face images generated by different facial privacy protection methods on CelebA-HQ. The green and blue numbers below each image are confidence scores returned by Face++ and Aliyun.

	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
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
	Ours	92.00	63.13	64.67	83.35	95.66	66.75	65.44	92.04	77.88

Table 1. Evaluations of *attack success rate* (ASR) for black-box attacks. For each column, we choose the other three FR models as surrogates to generate protected face images. DiffAM achieves a **12.98%** improvement on average ASR compared to the state of the art.

	FID(↓)	PSNR(↑)	SSIM(↑)
Adv-makeup	4.2282	34.5152	0.9850
AMT-GAN	34.4405	19.5045	0.7873
CLIP2Protect	37.1172	19.3537	0.6025
DiffAM (w/o L_{MT}^{dir})	33.6896	19.3099	0.8651
DiffAM ($T = 200$)	47.3186	18.6768	0.8367
DiffAM ($T = 100$)	32.4767	19.8816	0.8742
Our DiffAM	26.1015	20.5260	0.8861

Table 2. Quantitative evaluations of image quality. Our DiffAM represents the results with L_{MT}^{dir} and $T = 60$.

As for makeup-based methods, AMT-GAN fails to transfer makeup precisely and the generated face images have obvious makeup artifacts. CLIP2Protect struggles to generate accurate makeup corresponding to the given textual prompt and loses most of image details. In contrast, DiffAM stands out for accurate and high-quality makeup transfer, such as lipstick and eyeshadow, thanks to fine-grained supervision of generation direction and distance. Our proposed operation for preserving makeup-irrelevant information also ensures that face image details are well-preserved. Notably, DiffAM is effective in generating makeup for male images, which is a challenge for other makeup-based methods.

4.3. Attack Performance on Commercial APIs

Fig. 5 shows the quantitative results of attacks on commercial APIs Face++ and Aliyun. We randomly selected 100 images each from CelebA-HQ and LADN datasets to protect and report confidence scores returned from APIs. The confidence scores are between 0 to 100, where the

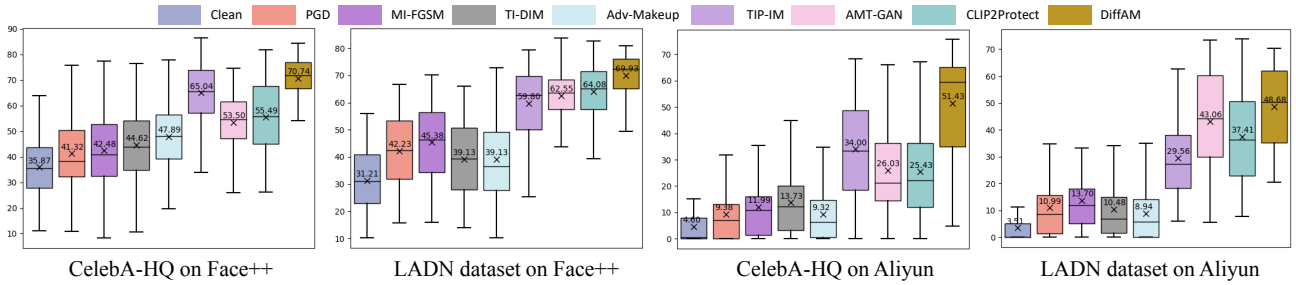

Figure 5. The confidence scores (higher is better) returned from commercial APIs, Face++ and Aliyun. DiffAM has higher and more stable confidence scores than state-of-the-art noise-based and makeup-based facial privacy protection methods.

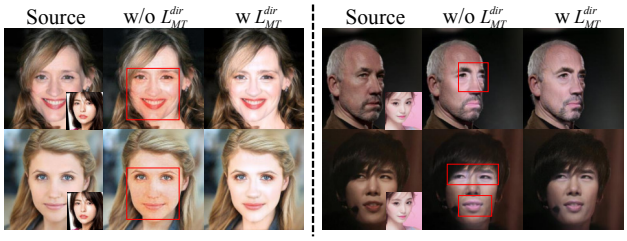

Figure 6. Ablation study for the makeup direction loss. The generated protected face images without makeup direction loss have obvious makeup artifacts (red boxes).

higher score indicates higher similarity between the protected face image and the target image. DiffAM achieves the highest average confidence scores about 70 and 50 on each API and the attack effect is relatively stable across different datasets, which indicates the strong black-box attack capability in real-world scenarios.

4.4. Ablation Studies

Control of Makeup Direction. We verify the importance of makeup direction loss L_{MT}^{dir} for makeup quality in Fig. 6. In the absence of L_{MT}^{dir}, the generated makeup has obvious makeup artifacts (red boxes in Fig. 6), leading to a decrease in image quality. Tab. 2 also illustrates that the generated images with L_{MT}^{dir} have better quantitative results than the ones without L_{MT}^{dir}. This is because, without L_{MT}^{dir}, the generated makeup is only guided by pixel-level makeup loss L_{MT}^{px}. L_{MT}^{px} just supervises makeup generation in different facial segmentation regions individually without global semantic supervision, resulting in inaccurate makeup generation. By applying makeup direction loss L_{MT}^{dir}, precise guidance can be provided for the global generation direction of the makeup, ensuring high-quality and accurate makeup.

Preservation of Makeup-Irrelevant Information. Fig. 7 shows the generated face images under a set of increasing inversion steps. With the increase of DDIM inversion steps, the generated face image has unexpected changes in facial attribute information. Tab. 2 shows the quantitative results at different steps, indicating that it is a simple but effective operation to preserve makeup-irrelevant information by controlling DDIM inversion steps.

Figure 7. Impact of inversion steps on preservation of makeup-irrelevant information. As the number of inversion step T increases, the facial features of the protected face image will change.

Reference	XMY-060	vHX570	vFG137	vRX189	XYH-045	Std.
ASR	76.01	78.56	77.91	79.06	78.23	1.04

Table 3. Impact of different reference makeup styles on ASR. Five reference makeup styles are selected from MT-dataset. Std. denotes standard deviation.

Robustness on different makeup styles. Being able to generate protected face images with any given reference makeup holds more practical value. Thus, we randomly select five reference images from MT-dataset to evaluate the impact of different reference makeup styles on attack effects of DiffAM. As shown in Tab. 3, the change of makeup styles has limited influence on ASR, which indicates the robustness of DiffAM to the change of makeup styles.

5. Conclusion

In this paper, we introduce DiffAM, a novel diffusion-based adversarial makeup transfer method for facial privacy protection. Building upon the generative capabilities of diffusion models, We innovatively introduce a makeup removal module to address uncertainty in text-guided generation. The deterministic cross-domain relationship can be obtained during makeup removal process, enabling fine-grained alignment guidance for adversarial makeup generation with the proposed CLIP-based makeup loss and ensemble attack strategy. Experiments have verified that DiffAM ensures strong black-box attack capabilities against many FR models and commercial APIs, while maintaining high-quality and precise makeup generation.

Acknowledgments. This work is supported by the National Nature Science Foundation of China (U23B2028, 62121002, 62232006, 62032006, 62102127).

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2, 3

[2] Andrew Besmer and Heather Richter Lipford. Moving beyond untagging: photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1563–1572, 2010. 1

[3] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 40–48, 2018. 3

[4] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. Beautyglow: On-demand makeup transfer framework with reversible generative network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10042–10050, 2019. 3

[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023. 3

[8] Han Deng, Chu Han, Hongmin Cai, Guoqiang Han, and Shengfeng He. Spatially-invariant style-codes controlled makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6549–6557, 2021. 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. 4, 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3, 6

[11] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 2, 6

[12] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 2, 6

[13] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 5

[14] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023. 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 2

[17] 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. 2, 3, 6

[18] 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

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 5

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 6

[21] 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, 5, 6

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 3

[23] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5202, 2020. 3

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 6

[25] 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. 2, 3, 4, 5

[26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[27] 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, 2021. 1, 2

[28] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. 2023. 2, 3

[29] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 3

[30] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 645–653, 2018. 3, 5, 6

[31] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023. 2

[32] Yueming Lyu, Yue Jiang, Ziwen He, Bo Peng, Yunfan Liu, and Jing Dong. 3d-aware adversarial makeup generation for facial privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:13438–13453, 2023. 2

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 6

[34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 5

[35] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[36] 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. 2, 4

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 6

[38] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. 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*, pages 20595–20605, 2023. 1, 2, 6

[39] 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. 1

[40] 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. 2

[41] Matthew Smith, Christian Szongott, Benjamin Henne, and Gabriele Von Voigt. Big data privacy issues in public social media. In *2012 6th IEEE international conference on digital ecosystems and technologies (DEST)*, pages 1–6. IEEE, 2012. 1

[42] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2, 3, 4

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[45] 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

[46] 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

[47] 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. 1, 2

[48] Lu Yang, Qing Song, and Yingqi Wu. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. *Multimedia tools and applications*, 80: 855–875, 2021. 2

[49] 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 IEEE/CVF International Conference on Computer Vision*, pages 3897–3907, 2021. 1, 2, 6

[50] 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 *International Joint Conference on Artificial Intelligence*, pages 1252–1258, 2021. 1, 2, 6

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[52] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 3

[53] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020. 2

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3