Enhancing Facial Privacy Protection via Weakening Diffusion Purification

Supplementary Material

In this supplementary, we first review diffusion models in Sec. 1 as they form the foundation of our proposed framework. Next, in Sec. 2 and Sec. 3, we introduce the four target identities and face recognition (FR) models used in our experiments. Sec. 4 describes the weight factor of adversarial loss. Then, we further assess the effectiveness of our approach via some ablation studies in Sec. 5. Finally, we present additional visualization results for a more comprehensive assessment in Sec. 6.

1. Background: Latent Diffusion Model

Diffusion models [6, 11, 14] consist of two processes: (1) a T-step forward diffusion process that progressively corrupts the input image x with Gaussian noise until it approaches a Gaussian distribution x_T at step T; (2) a reverse denoising process, which seeks to recover x from x_T by gradual reducing noise over T reverse steps. Unlike the denoising diffusion probabilistic model (DDPM) [6], the latent diffusion model (LDM) [11] operates in the latent rather than pixel space. In LDM, an autoencoder first compresses the image into a lower-dimensional latent representation z. The diffusion process then applies noise and denoising within this latent space. Finally, the latent representation is decoded back to the original image space. The forward process in LDM is defined as:

$$q\left(z_t \mid z_{t-1}\right)=\mathcal{N}\left(z_t ; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}\right) \ , \quad (1)$$

where $\beta_t \in (0,1]$ are parameters control the noise level at each diffusion step t. An important property of the forward process is that z_t can be directly sampled at any time t given the original latent variable z_0 using:

$$q\left(z_t \mid z_0\right)=\mathcal{N}\left(z_t ; \sqrt{\bar{\alpha}_t} z_0,\left(1-\bar{\alpha}_t\right) \mathbf{I}\right) \ , \quad (2)$$

where $\bar{\alpha}_t=\prod_{s=1}^{t}\left(1-\beta_s\right)$. Given that the reverse process $q\left(z_{t-1} \mid z_t\right)$ is intractable due to its dependence on the unknown data distribution $q\left(z_0\right)$, it can be approximated using a parameterized Gaussian transition model conditioned on a context $\mathcal{C}$, which is formulated as follows:

$$p_\theta\left(z_{t-1} \mid z_t, \mathcal{C}\right)= \mathcal{N}\left(z_{t-1} ; \mu_\theta\left(z_t, t, \mathcal{C}\right), \Sigma_\theta\left(z_t, t, \mathcal{C}\right)\right) \ , \quad (3)$$

where μ_θ and Σ_θ are mean and covariance matrix. The mean μ_θ can be expressed as:

$$\mu_\theta\left(z_t, t, \mathcal{C}\right)=\frac{1}{\sqrt{\alpha_t}}\left(z_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta\left(z_t, t, \mathcal{C}\right)\right) \ . \quad (4)$$

Target 1 Target 2 Target 3 Target 4

Figure 1. Target identities used for impersonation. The first row contains images used for training, while the second includes images used for testing.

Here $\epsilon_\theta\left(z_t, t, \mathcal{C}\right)$ is the model's prediction of the noise added at time step t, given the conditioning information $\mathcal{C}$. After training the model $\epsilon_\theta\left(z_t, t, \mathcal{C}\right)$, the following sampling method can be employed:

$$z_{t-1}=\mu_\theta\left(z_t, t, \mathcal{C}\right)+\sigma_t z, \quad z \sim \mathcal{N}(0,1) \ . \quad (5)$$

To accelerate image generation, Song *et al*. [14] introduce the denoising diffusion implicit model (DDIM), which employs a non-Markovian reverse process, as shown below:

$$z_{t-1}=\sqrt{\bar{\alpha}_{t-1} / \bar{\alpha}_t} z_t+ \sqrt{\bar{\alpha}_{t-1}}\left(\sqrt{1 / \bar{\alpha}_{t-1}-1}-\sqrt{1 / \bar{\alpha}_t-1}\right) \epsilon_\theta\left(z_t, t, \mathcal{C}\right) \ . \quad (6)$$

Equation 6 is derived from Equation 5 by eliminating the stochastic noise component ($\sigma_t = 0$), following the DDIM's principle, and substituting it with a deterministic process.

2. Target Images

The proposed model is designed to generate protected face images that deceive malicious FR models into misidentifying those protected faces as a specified target identity. Fig. 1 presents the four target identities provided by [8] mentioned in the Experiments Section (Sec. 4 in the main paper). To better mimic real-world scenarios, we ensure that the target images used during training differ from those used during testing.

3. Face Recognition Models

For fair comparisons, we adopt publicly available pretrained FR models following [8]. Three of these mod-

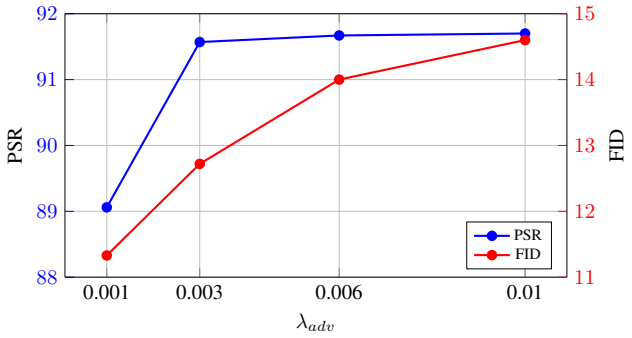

Figure 2. Quantitative study on the parameter settings of the weight factor of adversarial loss.

els are based on ArcFace [3], the state-of-the-art FR algorithm, which processes facial images at a resolution of 112×112 and encodes them into 512-dimensional feature vectors. These models differ in their neural architectures and training datasets: IR152 [5] employs ResNet-152, IRSE50 [7] uses ResNet-50, and MobileFace [2] is built on MobileFaceNet. Facenet [12], on the other hand, leverages InceptionResnet [16] and follows the original training protocols outlined in its paper, using an input resolution of 160×160. To assess the models' effectiveness, we report their FR accuracy on the CelebA-HQ dataset: IR152: 90.70%, IRSE50: 90.80%, MobileFace: 83.00%, and Facenet: 91.20%.

4. Parameter Settings

We evaluated the impact of varying the adversarial loss weight, λ_{adv}, on both privacy protection and image quality in Fig. 2. The results indicate that increasing λ_{adv} slightly improves privacy protection performance, as reflected by higher protection success rate (PSR). However, this comes at the expense of image quality, as evidenced by deteriorating Fréchet inception distance (FID). Conversely, lower λ_{adv} yields better image quality but significantly weaker privacy protection.

5. Ablation Studies

5.1. Optimizing Latent Codes across Timesteps

As mentioned in Sec. 3.4 in the main paper, the denoising model projects the perturbed noise back toward the natural data manifold during the reverse diffusion process. One potential solution to prevent the purification effect could be considering z_i from all timesteps as the latent code, optimizing it throughout the adversarial latent code learning process. Instead of utilizing the learned unconditional embedding proposed in our approach, we conducted an additional experiment by optimizing z_i across multiple timesteps from t to 0. The results show that the PSR improves from 91.57 to 93.37 when optimizing from z_3 to z_1, comparable to and slightly better than our method. However, this improvement comes at the cost of image quality, as indicated by the increase in FID from 12.72 to 15.71, suggesting that the generated images exhibit more structural changes. Additionally, the computational complexity increases significantly, with generation time rising from 15 seconds when optimizing only z_3, to 23 seconds when optimizing from z_3 to z_1, and up to 40 seconds when optimizing from z_5 to z_1 (Experiments were conducted using MobileFace as the target model).

The comparable performance of null-text guidance indicates that it implicitly approximates the impact of optimizing the latent codes at different timesteps while offering substantial benefits in preserving image quality and computational efficiency.

	IRSE50	IR152	Facenet	Mobileface
Ours w/o smoothing	88.87	67.25	59.53	91.57
$\text{Gauss}_{3\times3}$	88.47	67.20	59.23	91.47
$\text{Gauss}_{5\times5}$	87.61	66.73	58.73	90.26
$\text{Gauss}_{7\times7}$	87.06	66.35	57.93	88.56
$\text{Mean}_{5\times5}$	86.66	65.75	57.33	87.86

Table 1. Protection success rate (PSR) of our method against adaptive adversaries.

5.2. Effectiveness Against Adaptive Adversaries

An adaptive privacy adversary with advanced knowledge may deploy additional mechanisms to bypass the protection method. To evaluate the resilience of our approach under such adaptive scenarios, we assess its effectiveness against common image-smoothing techniques. Table 1 presents the results of applying Gaussian filters with kernel sizes of 3×3, 5×5, and 7×7, as well as a mean filter with a 5×5 kernel—widely used methods in the adversarial robustness domain. Despite slight degradation, PSR remains relatively high after smoothing, indicating that our approach maintains robust protection against these countermeasures.

5.3. Protection Performances on Commercial APIs

In Fig. 3, we further evaluate the protection performance of our proposed approach alongside other benchmarks using two commercial FR APIs, i.e., Face++[1] and Tencent[2], to simulate real-world conditions. We randomly select 100 images from the CelebA-HQ [9] and 100 images from LADN [4] datasets for protection, recording the confidence scores returned by each API. These scores range from 0 to 100, with higher values indicating greater similarity between the protected image and the target identity. The results show that our method achieves the highest confidence score com-

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

[2] https://cloud.tencent.com/product/facerecognition

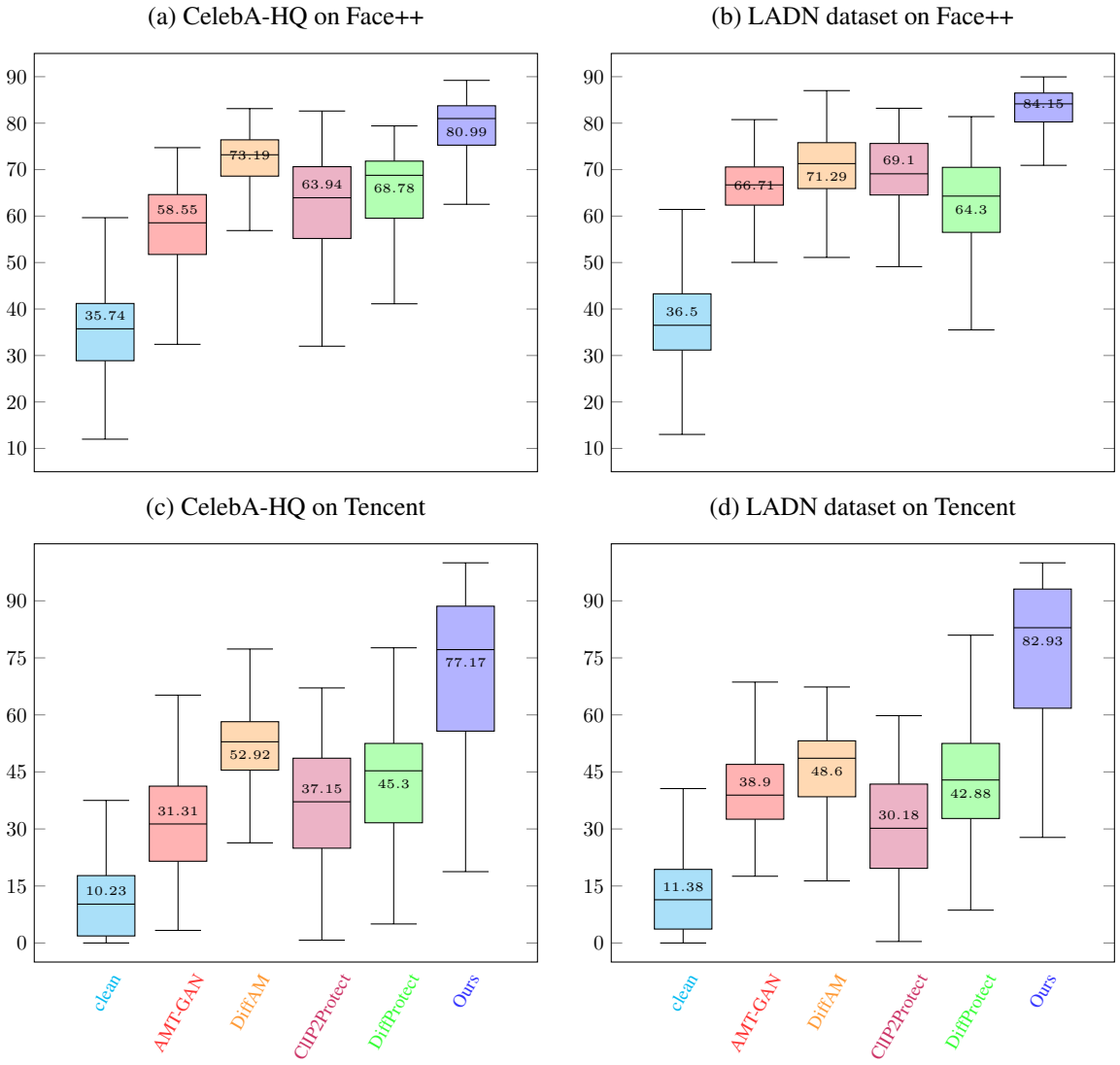

Figure 3. The confidence scores returned from Face++ and Tencent APIs. The higher confidence score indicates better protection performance. Our approach has a higher confidence score compared to four state-of-the-art methods, i.e., AMT-GAN [8], DiffAM[15], CIIP2Protect[13], and DiffProtect[10].

pared to other approaches.

6. More Visualization Results

Impersonation. To show the effectiveness of our proposed method in impersonating different identities, we visually compare the protected face images generated by ours and recent methods in Fig. 4. Compared to makeup-based methods, i.e., AMT-GAN [8], DiffAM [15] and CLIP2Protect [13], which change the makeup styles of the input images and intensify makeup in special parts of the face, our method can better preserve image styles. Compared to DiffProtect [10], which changes the facial expressions of the input images and smooths them out, ours preserves facial and hair details and adds perturbation only to identity-related features.

Obfuscation. A visual comparison between images generated using a combination of impersonation and obfuscation loss functions and those generated solely with the obfuscation loss function is shown in Fig. 5. The results demonstrate that the images generated with both losses simultaneously appear more natural and exhibit fewer distortions. This suggests that incorporating an impersonation objective with obfuscation enhances the visual quality of the generated images, producing faces that maintain more realistic features and preserve coherence in appearance.

Figure 4. Visual assessment of the protected images generated by previous methods and our approach for impersonation. Target images for each group are shown on the left side. Original images are selected from the CelebA-HQ [9] dataset.

(a) Original Images (b) Impersonation (c) Obfuscation (d) Original Images (e) Impersonation (f) Obfuscation

Figure 5. Visual assessment of the protected images generated by both impersonation and obfuscation losses and those generated with only the obfuscation loss. The synthesized target image is shown on the left side. (a) and (d) show original images, which are selected from the CelebA-HQ [9] dataset. (b) and (e) show protected images generated with both impersonation and obfuscation losses. (c) and (f) show protected images generated with only obfuscation loss.

7. Limitations and Future Directions

Given an input and target image, our approach generates the protected image in approximately 15 seconds on average, outperforming DiffProtect [10] (≈19 seconds) and CLIP2Protect [13] (≈30 seconds). All experiments were conducted on a single Nvidia GeForce RTX 4090. Despite its faster performance, the protection time of our approach can be further reduced by leveraging multiple GPUs and parallel computing optimizations. While AMT-GAN [8] and DiffAM [15] generate protected images in under one second, they require re-training the entire model for each new target identity, making them less flexible in practical scenarios.

In future work, we plan to replace the current surrogate model-based training paradigm, which involves iterative image reconstruction during latent code optimization, with a more efficient attack strategy that operates directly within the semantic space of the UNet proposed by An *et al*. [1]. This shift is expected to accelerate the execution time of our method significantly.

References

[1] Jinyang An, Wanqian Zhang, Dayan Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. Sd4privacy: Exploiting stable diffusion for protecting facial privacy. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2024. 5

[2] 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. 2

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4690–4699, 2019. 2

[4] 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

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[7] 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. 2

[8] 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, 3, 4, 5

[9] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 4, 5

[10] Jiang Liu, Chun Pong Lau, and Rama Chellappa. DiffProtect: Generate adversarial examples with diffusion models for facial privacy protection. *arXiv preprint arXiv:2305.13625*, 2023. 3, 4, 5

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[12] 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. 2

[13] 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. 3, 4, 5

[14] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1

[15] Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, and Yongdong Zhang. DiffAM: Diffusion-based adversarial makeup transfer for facial privacy protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24584–24594, 2024. 3, 4, 5

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 2