DeepProtect: Proactive Face-Swapping Defense using Identity Blending and Attribute Distortion
Supplementary Material

Eungi Lee† Seung-hyeok Back† Hyung-Il Kim* Seok Bong Yoo*
Chonnam National University, Gwangju, Korea
{st0421,aiback856336,hyungil.kim,sbyoo}@jnu.ac.kr

1. Analysis of Hyperparameter Impact

Our method involves several hyperparameters. While these hyperparameters may seem extensive, most of them primarily affect computational efficiency rather than the core performance of the model. For example, m does not significantly impact results as long as a sufficient number of samples are used to represent the attribute groups. Similarly, the low-rank adaptation (LoRA) [5] rank r shows negligible effect beyond a certain point, making it more of a trade-off between efficiency and capacity. The regularization weight λ_R also only needs to be large enough to maintain semantic order; increasing it further does not degrade performance. In this section, we analyze the impact of each hyperparameter and justify our selected values based on empirical observations.

1.1. On the Necessity of CLIP-based Retrieval

While the goal is to dilute identity, it is not immediately obvious why the candidate set must be constrained by CLIP feature similarity. A simple workaround could be to bypass similarity constraints entirely by selecting a latent vector from an unrelated face or by random sampling. However, as shown in Fig. 1, such strategies introduce limitations. When the selected latent code lacks visual compatibility with the input, the generator often fails to reconstruct coherent images, leading to unnatural outputs. In contrast, latent codes retrieved based on high CLIP similarity (above a threshold τ) retain sufficient appearance-level consistency, which enables the generator to preserve the input's visual characteristics after fine-tuning while still allowing for identity-level changes. These results indicate that visual similarity, as captured in the CLIP space, is essential not only for maintaining perceptual realism but also for enabling effective adaptation of the generator. Thus, the CLIP-based candidate selection plays a pivotal role in achieving identity obfuscation without compromising visual fidelity.

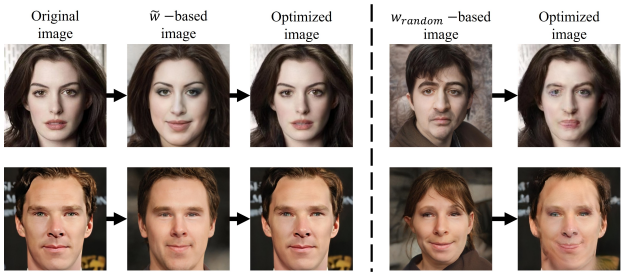

Figure 1. Illustration of reconstruction outcomes after generator tuning using a candidate-guided latent code and a randomly initialized one.

1.2. CLIP similarity threshold τ

These observations motivate the need for a quantitative evaluation of how well the generator can reconstruct the input image from latent codes that do not correspond to it. To this end, we define the reconstruction success rate (RSR) as a metric that measures how well the generator can reconstruct an input image when optimized starting from a different latent code. In other words, given a latent code not associated with the input image, we optimize the generator to reproduce the input image as accurately as possible. Reconstruction is considered successful if the perceptual similarity between the optimized output and the input image meets a predefined criterion. Specifically, we adopt the early stopping condition from PTI [13], where reconstruction is deemed successful if the LPIPS loss falls below 0.06.

For evaluation, we randomly sample 500 identities from the VGGFace2-HQ [3] dataset as query inputs. For each query, candidate latent codes are ranked by their CLIP similarity. Then, starting from the highest-ranked candidate, latent codes are sampled incrementally at intervals of 0.05 in

† Equal contribution.
* Corresponding authors.
Eungi Lee is currently with the Electronics and Telecommunications Research Institute (ETRI), Korea.

similarity score, up to a maximum of 20 samples per query, for reconstruction attempts. This setup allows us to quantitatively assess the feasibility of reconstructing the input image from latent codes with varying degrees of visual similarity, thus demonstrating the impact of the CLIP similarity threshold τ on reconstruction performance.

Metric	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.65$	$\tau = 0.7$	$\tau = 0.75$	$\tau = 0.8$	$\tau = 0.85$
RSR	81.3	90.8	93.4	95.7	**96.1**	**96.1**	**96.1**

Table 1. Reconstruction success rate (RSR) for different τ values.

Table 1 reports the reconstruction success rate (RSR) for varying levels of the CLIP similarity threshold τ. As τ increases, the RSR improves consistently, indicating that latent codes with higher CLIP similarity to the input are more likely to result in successful reconstructions. However, increasing τ comes at the cost of reduced candidate diversity, as higher thresholds limit the pool of available latent codes. To balance reconstruction quality and candidate diversity, we select $\tau = 0.75$ as our default setting.

1.3. Optimal m Value for Attribute Vector Extraction

To extract meaningful attribute vectors in the identity space, we first need to determine an appropriate group size m for linear discriminant analysis (LDA). Since LDA relies on grouped data to compute a direction that maximizes between-group variance while minimizing within-group variance, the number of samples in each group significantly affects the quality of the resulting attribute direction. We therefore conduct experiments with varying m to analyze its impact. As shown in Fig. 2, the effect of estimated attribute directions stabilizes once m reaches around 30. However, increasing m also leads to higher computational cost. For a balance between semantic clarity and efficiency, we set $m = 30$ for all experiments.

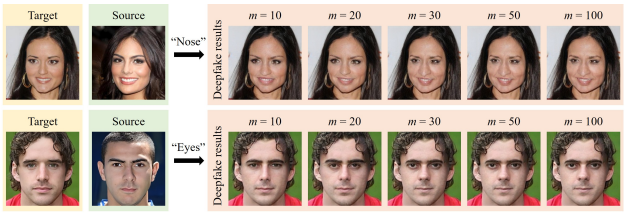

Figure 2. Comparison of deepfake attribute distortion according to the number of features m used for attribute direction extraction.

1.4. Effect of ID-lock Weight

Table 2 shows the preservation-disruption score (PDS) for images generated with varying $\lambda_{id-lock}$ values. The source metric-SSIM decreases from 0.911 to 0.855 as the $\lambda_{id-lock}$ value increases, indicating that the visual similarity between the original and protected images decreases as the $\lambda_{id-lock}$ value becomes larger. Similarly, the deepfake metric-SSIM also decreases continuously from 0.745 to 0.688 as the $\lambda_{id-lock}$ value increases. PDS is calculated as the difference between the source metric-SSIM and the deepfake metric-SSIM, reaching its maximum value at $\lambda_{id-lock} = 0.1$. This indicates the optimal balance point between source preservation and deepfake distortion, where the protected images maintain visual consistency with the original images while effectively disrupting deepfake generation.

$\lambda_{id-lock}$	0	0.01	0.1	0.5	1
Source metric-SSIM† ↑	0.911	0.908	0.902	0.873	0.855
Deepfake metric-SSIM‡ ↓	0.745	0.727	0.710	0.701	0.688
PDS ↑	0.166	0.181	**0.192**	0.172	0.161

Table 2. Experimental results evaluating the effect of different values of $\lambda_{id-lock}$ on preservation-disruption score (PDS) on the CelebA-HQ dataset.

1.5. Adaptive Regularization with λ_R

This section uses λ_R to balance the separation between the top and bottom groups and the preservation of semantic order. λ_R can be flexibly adjusted depending on the characteristics of the attribute. For gradually changing attributes like age, a larger λ_R helps preserve semantic continuity, while a smaller value is more suitable for binary traits like gender. This adaptability contributes to the overall efficiency of the proposed approach. In our experiments, we set λ_R to 1, as the attribute manipulations in our method are strong enough to necessitate preserving semantic order.

Figure 3 presents the visual results of this analysis. When $\lambda_R = 0$, the deepfake output follows the original LDA direction vector, and the distortion occurs somewhat randomly without reflecting the given prompt. In contrast, when $\lambda_R = 1$, the deepfake output is distorted according to the attributes specified by the prompt.

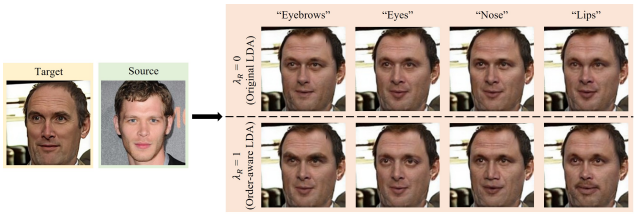

Figure 3. Comparison of attribute distortion in deepfake outputs according to prompt and λ value in order-aware LDA.

1.6. Fine-Tuning Results Across LoRA Layers

To prevent identity extraction from the original images, we blend the identity and ensure that the protected images

are generated to be visually similar to the original images. Through our experiments, we directly confirmed that using a pretrained StyleGAN2 [7] for reconstruction often fails to accurately restore the eyes of the original images due to its inductive bias.

Therefore, we fine-tune the latent code, which consists of 18 style vectors with 512 dimensions each, using LoRA. Of the 18 layers, layers zero through two optimize fine details critical for generating realistic facial structures and head poses, layers three through seven optimize important identity-related features, and layers eight through seventeen maintain real color distributions and background information [18]. Figure 4 visually compares the reconstruction performance under four different conditions. Among these, we apply LoRA optimization only to the middle block, considering both reconstruction quality and computational cost.

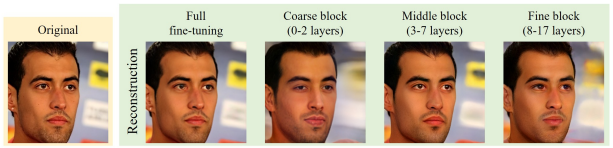

Figure 4. Visualization of the effect of LoRA optimization on the coarse, middle, and fine blocks. Reconstruction using only the middle block demonstrates superior effectiveness.

1.7. Complexity Comparison Based on LoRA Rank r Values

To reduce computational cost, we apply LoRA to the affine transform module of the middle layers in StyleGAN2. The original affine transform weight matrix $\theta_0 \in \mathbb{R}^{512\times b}$, where b is the output dimension of the affine transform, is decomposed into $\theta_1 \in \mathbb{R}^{512\times r}$ and $\theta_2 \in \mathbb{R}^{r\times b}$, where the LoRA rank $r \ll \min(512, b)$, and only θ_1 and θ_2 are updated during training. Table 3 shows the results of experiments with different r values. We conduct our experiments on the VGGFace2-HQ dataset. As r increases, the computational complexity rises, but the quality of the reconstructed images improves. However, when the r value is greater than 8, the training time and the number of trainable parameters increase while SSIM shows little improvement. Based on these results, we set the LoRA rank r to 8.

Complexity	$r = 2$	$r = 4$	$r = 8$	$r = 16$
Training time (s) ↓	11	12	13	15
Trainable params (K) ↓	12.288	24.576	49.152	98.304
Source metric-SSIM †↑	0.896	0.904	0.915	0.917

Table 3. Complexity comparison for different LoRA rank r values.

2. Observations

2.1. Observational Study in Deepfake Outputs

To study proactive defenses against face-swapping deepfakes, we first applied various existing manipulation methods to original images and then examined the resulting deepfake outputs. We conducted several experiments, including inserting visible watermarks into the original images, manipulating expressions using StyleCLIP [11], and altering styles with CLIP2Protect [16]. However, these approaches had little to no impact on the deepfake outputs. To better understand how facial attribute manipulations affect face-swapping deepfakes, we further investigated identity feature representations. In particular, we focused on ArcFace [4], a widely-used model that extracts high-dimensional identity embeddings. By intentionally modifying facial attributes that influence these embeddings, we were able to observe notable degradations in the resulting deepfake quality. Figure 5 illustrates how such manipulations—unlike visible watermarks or expression/style changes—can more effectively disrupt the identity consistency of the generated outputs.

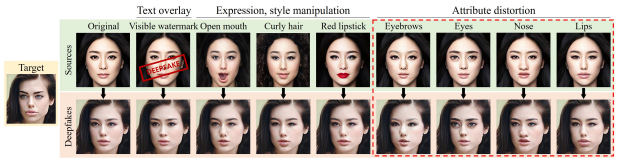

Figure 5. Observational analysis of deepfake output distortion. While existing manipulations such as text overlay and facial expression modifications have little impact on the resulting deepfakes, intentional attribute distortions lead to noticeable alterations in the generated outputs.

2.2. Analysis of the Impact of ϵ Value

Figure 6. Comparison of the impact on protected images for different ϵ values.

This section aims to ensure robust protection while main-

taining visual fidelity between the protected images and the original ones. However, existing noise-based proactive defenses are evaluated at a noise level of above $\epsilon = 0.05$, where the protected images often exhibit noticeable artifacts. Figure 6 visualizes and compares images protected with ϵ values of 0.02 and 0.05. The second-row images, with $\epsilon = 0.02$, exhibit imperceptible noise. In contrast, the images in the third row, with $\epsilon = 0.05$, show clearly visible noise—even without direct comparison to the originals. This indicates that, regardless of protection performance, such visible noise can negatively impact the user experience. Therefore, to ensure visually acceptable results, we constrain the perturbation magnitude to $\epsilon = 0.02$ in our experiments.

3. Additional Results

3.1. Visual Results of Protected Images

Figure 7 presents additional experiments for Fig. 6 in the main paper. It compares the results of various deepfake models on images protected by existing methods and by our method using the VGGFace2-HQ dataset. For comparison, we used state-of-the-art face-swapping models, including SimSwap [1], FaceDancer [14], BlendFace [17], FaceSwapper [10] and DiffFace [8]. Compared to existing methods, our protected images remain visually similar to the original ones, while still providing sufficient protection against deepfake distortion.

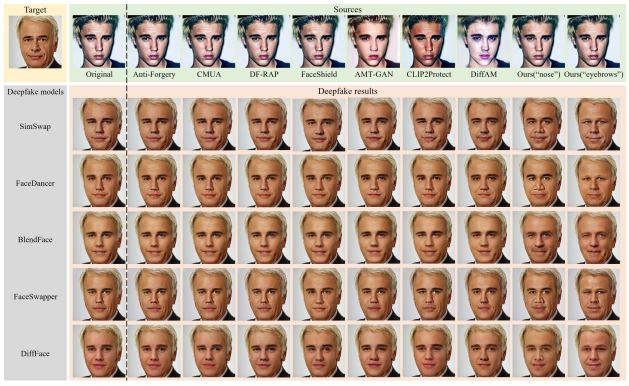

Figure 7. Visualization of protected images and deepfake results for each prompt in the proposed method.

3.2. Ablation Study for Identity Blending and Attribute Distortion

As shown in Fig. 8, the proposed method is visually analyzed through an ablation study, highlighting the effects of the individual modules. All components of our proposed method—the identity blending module, the attribute distortion module, and their integrated version—produce protected images that are visually similar to the originals. First, images protected using only the identity blending module cause global distortions in the resulting deepfakes. While this module alone is sufficient to protect the original identity, the generated outputs may still appear relatively natural. Next, images protected using the attribute distortion module show localized distortions in the deepfake outputs, specifically targeting the regions corresponding to the prompt attributes. The final image shows the result of integrating the two previously described modules. The protected image remains visually similar to the original, while the deepfake output is significantly distorted, achieving robust identity protection.

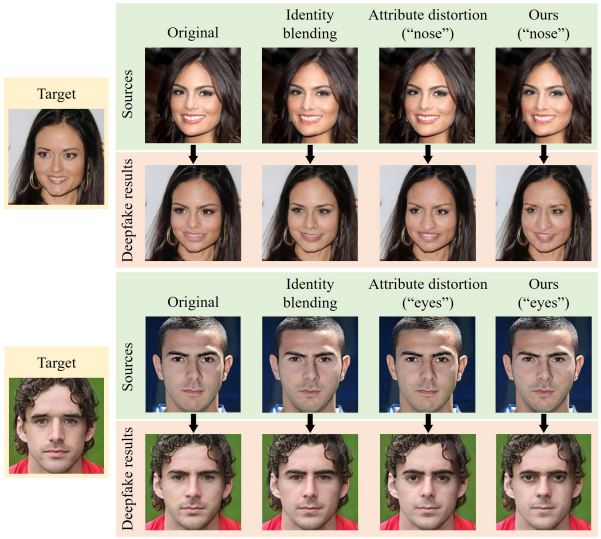

Figure 8. Experiments demonstrating the effectiveness of identity blending and attribute distortion in protecting against deepfake.

3.3. Evaluation under Image Transformations

We also examine robustness under dynamic real-world scenarios where protected images are shared on social media, leading to image transformations (e.g., compression or resizing). Following DF-RAP [12], we upload our protected images to Facebook, Instagram, and X, then download and apply face-swapping. Owing to identity blending, our method remains robust under such post-processing. Figure 9 demonstrates that the protection is preserved even after social media compression.

In addition to such social media compressions, we further assess robustness under common image degradations. Specifically, we apply Gaussian noise (σ=0.08), Gaussian blur (σ=1, kernel size 7×7), and JPEG compression (quality factor = 25). As shown in Fig. 10, DeepProtect maintains its protection effectiveness under all conditions. This robustness is primarily attributed to identity blend-

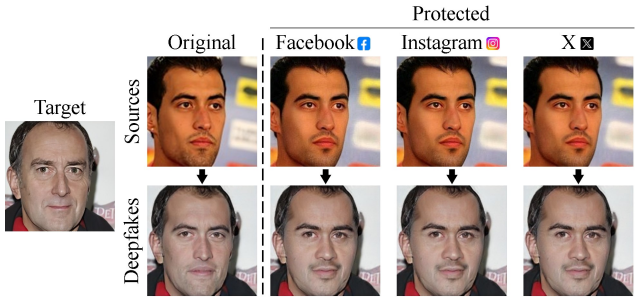

Figure 9. Visualization of protection performance in real-world scenario when protected images are uploaded to various social media platforms.

ing, which fundamentally alters the original facial identity in a high-dimensional latent space. As a result, the core identity information remains protected even under low-level degradations like noise and compression. We also evaluate

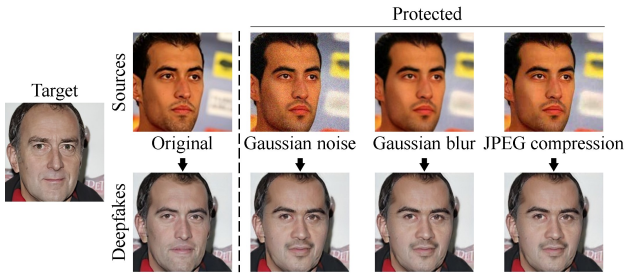

Figure 10. Robustness to image degradations applied before deepfake generation.

robustness under varying JPEG compression by applying quality factors from 0 to 70 (Fig. 11). Despite severe degradation, our method consistently preserves protection across all settings.

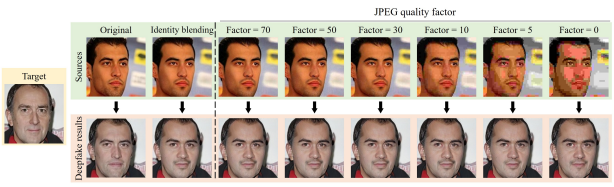

Figure 11. Visualization of protection performance against compression under various JPEG quality factors.

3.4. Effectiveness of the Proposed Method on Various Datasets

We provide additional visual results in Figs. 13, 14, 15, and 16. Figure 13 shows visual results for male subjects from the CelebA-HQ dataset, protected using the 'lips' prompt. Figure 14 presents results for female subjects from the CelebA-HQ dataset, protected using the 'eyebrows' prompt. Figures 15 and 16 show results for male and female subjects from the VGGFace2-HQ dataset, protected using the 'nose' and 'eyes' prompts, respectively.

4. The Effects of Identity Blending

4.1. Visual Analysis of Identity Retrieval Results

To verify the effectiveness of the proposed identity blending module, including both blending and LoRA optimization, we analyze the retrieval results based on reconstructed images. Figure 12 shows the retrieval results for images similar to the query images from the VGGFace2-HQ dataset. The numbers above each image indicate the similarity scores with the respective query images. Green numbers represent images that match or have high identity similarity to the query, while red numbers indicate images that mismatch with the query and have low identity similarity. The top row shows the retrieval results for unprotected original images, where images with the same or similar identities are retrieved, resulting in high similarity scores. The middle row presents the retrieval results using reconstructed images from latent codes with identity blending but without LoRA optimization. Because the optimization was not applied, the query images are not visually similar to the originals. The retrieval results also consist of mismatched identities. Finally, the bottom row uses the results of identity blending with LoRA optimization. While the query images remain visually similar to the original query images, the retrieval results still consist of mismatched images, similar to the blended results. These results demonstrate that the proposed identity blending module effectively prevents identity extraction, while the protected images remain visually similar to the original images.

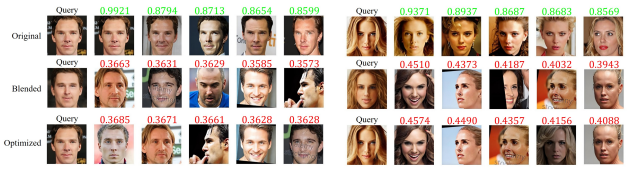

Figure 12. Visualization of top-5 identity retrieval results using the VGGFace2-HQ. Green: match, red: mismatch.

4.2. Dodging Attack Performance on LFW across Face Recognition Models

To quantitatively evaluate the identity protection effectiveness of the identity blending module, we conducted experiments on the LFW dataset using various face recognition models. Expanding on the results provided in the CLIP2Protect [16], we included our method to compare its effectiveness in the same experimental settings. The

	Face recognition models				
Method	IRSE50	IR152	FaceNet	MobileFace	**Average**
MI-FGSM	70.2	58.4	59.2	68.0	63.9
TI-DIM	79.0	67.4	74.4	79.2	75.0
TIP-IM	81.4	71.8	76.0	82.2	77.8
CLIP2Protect	86.6	73.4	83.8	85.0	82.2
DeepProtect (w/o Adversarial watermarking)	91.3	77.1	86.4	88.6	85.6

Table 4. Rank-1-U dodging attack results on LFW across face recognition models.

primary evaluation metric used was the untargeted identity success rate (Rank-1-U), which measures the proportion of cases where the top-1 retrieved candidate does not match the original identity, indicating successful identity protection. Table 4 compares the performance of DeepProtect against existing approaches across four face recognition models: IRSE50 [6], IR152 [4], FaceNet [15], and MobileFace [2]. Our achieves an average success rate of 85.6%, demonstrating superior robustness and generalization in preventing identity extraction compared to prior methods. This suggests that DeepProtect can effectively hinder face-swapping deepfake systems that rely on identity extraction for image generation.

5. Algorithm: Full Procedure of DeepProtect

The pseudo-code of our proposed method is provided to illustrate the full procedure. Algorithm 1 illustrates the process of identity blending for diluting identity, along with the optimization procedure using LoRA. Through the identity blending module, images with diluted identity are fully protected against deepfake generation by embedding an adversarial watermark that induces facial attribute distortions. Details of the adversarial watermark can be found in Algorithm 2.

6. Discussion

6.1. Applicability

The DeepProtect method is designed as a proactive, server-side defense mechanism that protects images before they are shared, similar to widely used server-side services for image editing. Recent trends indicate that users prioritize output quality over inference speed when using such services. In the context of deepfake defense, the critical factors are robust protection and high visual fidelity, rather than real-time performance. In DeepProtect, retrievals, identity blending, and adversarial watermarking processes take less than 1s, and generator optimization takes less than 13s. Although the optimization step adds inference time, it reflects an essential trade-off to achieve strong protection and high image realism across face-swapping deepfakes. Therefore, real-time inference is not critical during the preupload protection phase; instead, ensuring robust defense and maintaining visual quality are prioritized to meet user expectations.

6.2. Vulnerabilities of CLIP Feature-Based Retrieval

In the proposed method, CLIP features are extracted using a FaRL model pre-trained on the LAION-Face20M dataset when performing candidate searches based on CLIP similarity. LAION-Face20M is a large-scale dataset containing over 20 million diverse facial images, encompassing data collected from various races, ages, genders, and environments. A model pre-trained on such a large-scale dataset can exhibit robust generalization performance across diverse scenarios when extracting features, playing a critical role in distinguishing subtle differences between similar faces.

Additionally, the pre-built feature bank used for retrieval is built on the VGGFace2-HQ dataset. The VGGFace2-HQ dataset is a high-quality, large-scale facial image dataset

Algorithm 1 Pseudo-code for identity blending and generator optimization

Require: I (source image), G_{inv} (GAN inversion model), $\mathcal{C}$ (candidate set), E_{id} (identity encoder), $G_{initial}$ (initial generator), $\lambda_{id-lock}$ (weight for identity-lock loss), α (learning rate)
Ensure: Fine-tuned generator G and reconstructed image with identity blended I_{recon}

$w \leftarrow G_{inv}(I)$
$\tilde{w} \leftarrow w$
for $l = 3$ to 7 **do**
  $\tilde{w}[l] \leftarrow \arg\max_{w_j \in \mathcal{C}} \cos\left(w[l],\ w_j[l]\right)$ ▷ w_j denotes the j-th latent code in the $\mathcal{C}$
end for
$I_{initial} \leftarrow G_{initial}(\tilde{w})$
$G \leftarrow G_{init}$
Apply LoRA to the affine transform module of the middle layers in G
while $\mathcal{L}_{LPIPS} > 0.06$ **do**
  $I_{generated} \leftarrow G(\tilde{w})$
  Calculate the $\mathcal{L}_2$ and $\mathcal{L}_{LPIPS}$ with I and $I_{generated}$
  $\mathcal{L}_{id-lock} \leftarrow 1 - \cos\left(E_{id}(I_{initial}), E_{id}(I_{generated})\right)$
  $\mathcal{L}_{total} \leftarrow \mathcal{L}_2 + \mathcal{L}_{LPIPS} + \lambda_{id-lock}\mathcal{L}_{id-lock}$
  $\theta_{new} \leftarrow \theta_{old} - \alpha \cdot \nabla_\theta \mathcal{L}_{total}$ ▷ θ denotes the LoRA weights of the generator G
end while
$I_{recon} \leftarrow G(\tilde{w})$

Algorithm 2 Pseudo-code for adversarial watermarking

Require: I (input image), v_{attr} (attribute direction), E_{id} (identity encoder), T (number of update steps), α (step size), ϵ (ℓ_∞-norm bound)
Ensure: Adversarial watermark W_{attr}

$W_{attr} \leftarrow 0$
$z^{id} \leftarrow E_{id}(I)$
$p \leftarrow z^{\text{id}} \cdot v_{attr}$
$v_{target} \leftarrow -\text{sign}(p) \cdot v_{attr}$
for $t = 1$ to T **do**
    $I_t \leftarrow I + W_{\text{attr}}$
    $z_t \leftarrow E_{\text{id}}(I_t)$
    $\mathcal{L}_t \leftarrow z_t \cdot v_{\text{target}}$
    $g_t \leftarrow \nabla_{W_{\text{attr}}} \mathcal{L}_t$
    $W_{\text{attr}} \leftarrow W_{\text{attr}} + \alpha \cdot \text{sign}(g_t)$
end for
$W_{\text{attr}} \leftarrow \max(\min(W_{\text{attr}}, \epsilon), -\epsilon)$

that includes a diverse set of identities in race, age, gender, lighting conditions, and occluded images. Using a diverse dataset like this also provides a solid foundation for the generalization of retrieval.

Specifically, we conducted our experiments using a pre-constructed feature bank containing 9,630 images, which suggests that the feature bank has sufficient generalizability. Additionally, the feature bank can be further expanded in the future.

6.3. Risk for Identity Leakage

This section is using a pretrained GAN inversion model to obtain the latent code w for the input image. Using the obtained latent code w, we then search for visually similar feature candidates within the feature bank, based on CLIP similarity in the CLIP space. During this process, multiple images of the same individual may be retrieved. This outcome is a natural consequence of CLIP similarity-based searches, which identify visually similar individuals. In this context, concerns may arise regarding the risk of identity leakage. However, when performing searches within the feature bank and replacing latent vectors at the style channel-wise using the selected candidates, we ensure that individuals with the same identity are excluded from the replacement process by leveraging the identity information present in the feature bank. This means that the style vectors used in the replacement are always drawn from distinct identities, effectively eliminating the risk of identity leakage.

6.4. Dependency on Style Generator

Our approach strongly relies on the Style Generator. This is because we utilize the high-dimensional latent space of StyleGAN, which can simultaneously embed both identity and appearance, allowing for seamless identity blending without structural distortion of the original source image. Such a latent space is specifically designed to separate and independently control the individual features of identity and appearance. As a result, it enables us to go beyond merely generating images, enabling the independent replacement of various features. This mechanism can also be applied to generative models other than StyleGAN, as long as a latent space that can simultaneously embed both identity and appearance is available. If a new feature space is proposed in the future, our proposed method could be utilized even more broadly.

6.5. Limitations

This section focuses on defending against face-swapping deepfakes, which often involve the unauthorized misuse of widely shared images from online sources, potentially leading to digital crimes and privacy violations. This approach can also be effectively extended to various deepfake methods, such as face reenactment, which rely heavily on identity as a critical component. However, the robust protection may not be achieved for face attribute manipulation deepfakes that do not rely on identity as a primary element. This is because the generation mechanisms of these deepfakes differ from those targeted by our approach, meaning that certain aspects of the proposed method may not be applicable for protection. Developing a universal defense capable of protecting against a wide range of facial manipulation mechanisms remains a significant challenge, highlighting a important area for future research and development.

6.6. Future Work

We aim to enhance our proposed method by leveraging diffusion models, which have recently gained significant attention as powerful generative models. In the forward process of diffusion models, we plan to introduce adversarial noise that specifically distorts attributes directly related to deepfake transformations. During the denoising phase, a mechanism similar to id-lock loss will be employed, ensuring that the reconstructed images remain visually similar to the original inputs while returning entirely unrelated identity features. This approach will effectively integrate our mechanism into the diffusion framework, extending our method beyond the current reliance on StyleGAN[9].

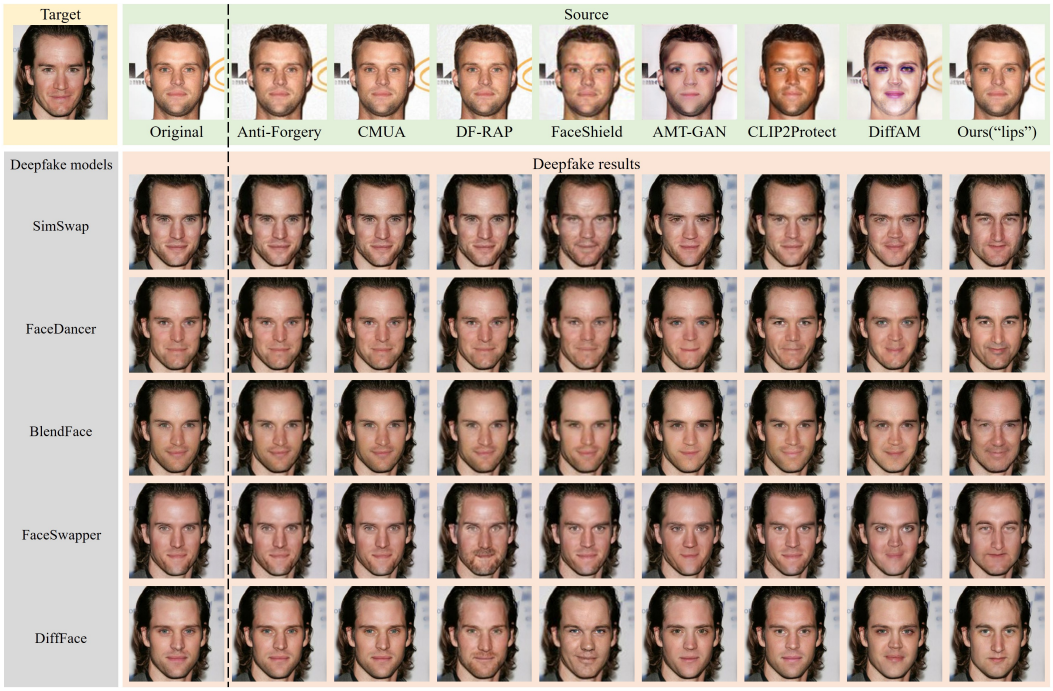

Figure 13. Visualization of existing methods and our proposed method on male images from the CelebA-HQ dataset. protected using the 'lips' prompt.

Figure 14. Visualization of existing methods and our proposed method on female images from the CelebA-HQ dataset. protected using the 'eyebrows' prompt.

Figure 15. Visualization of existing methods and our proposed method on male images from the VGGFace2-HQ dataset. protected using the 'nose' prompt.

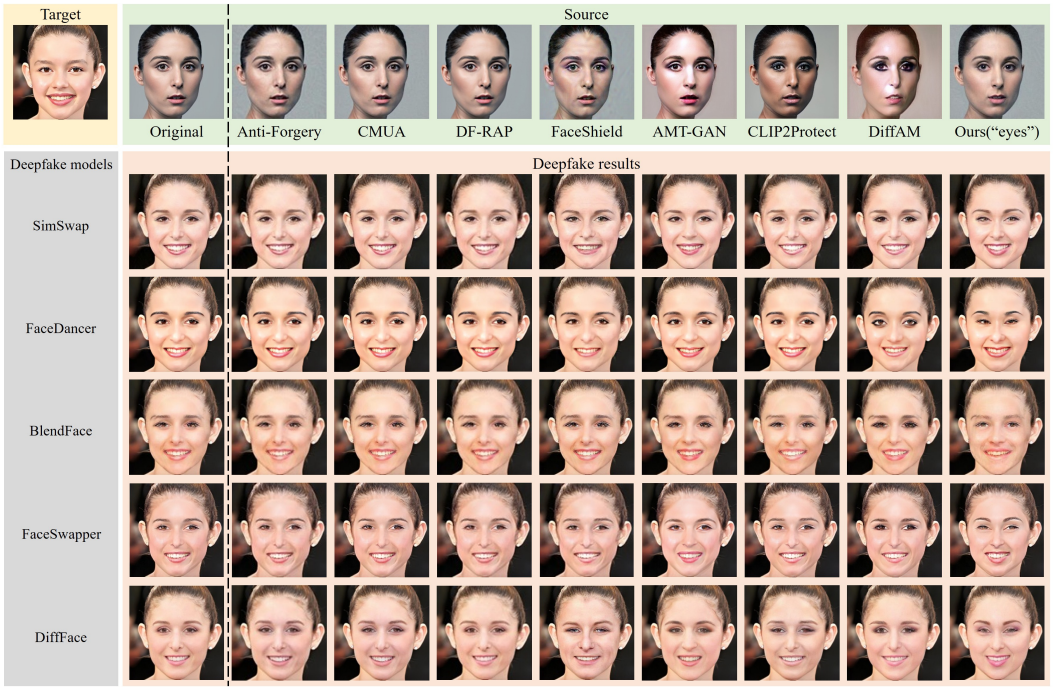

Figure 16. Visualization of existing methods and our proposed method on female images from the VGGFace2-HQ dataset. protected using the 'eyes' prompt.

References

[1] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 4

[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. 6

[3] Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. Simswap++: Towards faster and high-quality identity swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):576–592, 2023. 1

[4] 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. 3, 6

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[6] 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

[7] 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. 3

[8] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163:111451, 2025. 4

[9] Woo Kyung Kim, Youngseok Lee, Jooyoung Kim, and Honguk Woo. Llm-based skill diffusion for zero-shot policy adaptation. *Advances in Neural Information Processing Systems*, 37:6749–6775, 2024. 7

[10] Qi Li, Weining Wang, Chengzhong Xu, Zhenan Sun, and Ming-Hsuan Yang. Learning disentangled representation for one-shot progressive face swapping. *IEEE transactions on pattern analysis and machine intelligence*, 2024. 4

[11] 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

[12] Zuomin Qu, Zuping Xi, Wei Lu, Xiangyang Luo, Qian Wang, and Bin Li. Df-rap: A robust adversarial perturbation for defending against deepfakes in real-world social network scenarios. *IEEE Transactions on Information Forensics and Security*, 2024. 4

[13] 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. 1

[14] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023. 4

[15] 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

[16] 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, 5

[17] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7634–7644, 2023. 4

[18] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 3