FaceShield: Defending Facial Image against Deepfake Threats

Jaehwan Jeong[1] Sumin In[1] Sieun Kim[1] Hannie Shin[1] Jongheon Jeong[1]
Sang Ho Yoon[2] Jaewook Chung[3] Sangpil Kim[1*]

[1]Korea University [2]KAIST [3]Samsung Research

Abstract

The rising use of deepfakes in criminal activities presents a significant issue, inciting widespread controversy. While numerous studies have tackled this problem, most primarily focus on deepfake detection. These reactive solutions are insufficient as a fundamental approach for crimes where authenticity is disregarded. Existing proactive defenses also have limitations, as they are effective only for deepfake models based on specific Generative Adversarial Networks (GANs), making them less applicable in light of recent advancements in diffusion-based models. In this paper, we propose a proactive defense method named ***FaceShield****, which introduces novel defense strategies targeting deepfakes generated by Diffusion Models (DMs) and facilitates defenses on various existing GAN-based deepfake models through facial feature extractor manipulations. Our approach consists of three main components: (i) manipulating the attention mechanism of DMs to exclude protected facial features during the denoising process, (ii) targeting prominent facial feature extraction models to enhance the robustness of our adversarial perturbation, and (iii) employing Gaussian blur and low-pass filtering techniques to improve imperceptibility while enhancing robustness against JPEG compression. Experimental results on the CelebA-HQ and VGGFace2-HQ datasets demonstrate that our method achieves state-of-the-art performance against the latest deepfake models based on DMs, while also exhibiting transferability to GANs and showcasing greater imperceptibility of noise along with enhanced robustness.*

1. Introduction

The advancement of deepfake technology and improved accessibility [3, 22, 25, 37, 51] has led to significant transformations in modern society. Due to the ease of face swapping, it has been applied across various fields, providing both entertainment and convenience. However, its powerful capability to generate realistic content has also enabled malicious users to exploit it for criminal purposes, leading to the creation of fake news and various societal problems.

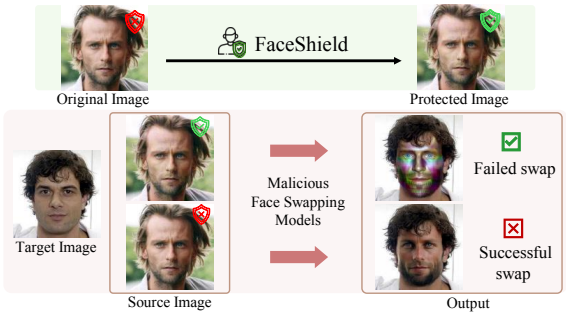

Figure 1. **Protecting Face during Deepfake using *FaceShield*.** Pure images are vulnerable to face swapping, allowing the target image's face to be easily reflected. In contrast, images protected by *FaceShield* conceal facial feature from deepfake.

To address the growing concerns surrounding deepfake technology, various countermeasures have been explored, which can be broadly divided into two categories. The first is deepfake detection techniques [5, 12, 13, 31, 36], which act as passive defenses by classifying whether content is synthetic or authentic. While effective for authenticity verification, these offer only binary results and fail to address advanced threats, such as crimes using realistic fakes. In contrast, proactive defense strategies offer a more comprehensive solution. These approaches involve embedding imperceptible adversarial perturbation into face images to prevent the protected face from being effectively processed by deepfakes. However, most previous research [8, 14, 28, 35, 44] has concentrated on GAN-based models, often targeting individual models, which limits effectiveness against emerging DM-based deepfakes [20, 34, 43, 49]. Although significant research exists on image protection within DMs [6, 23, 24, 29, 30, 42] for image editing, the focus has primarily been on attacking the noising and denoising processes when an image is used as a query (Fig.2a). This leads to targeting the encoder or predicted noise post-UNet processing. However, we observe that such strategies are ineffective for DM-based deepfake models, where the source image influences the output in the form of key-value pairs through attention mechanisms (Fig.2b).

*Corresponding author

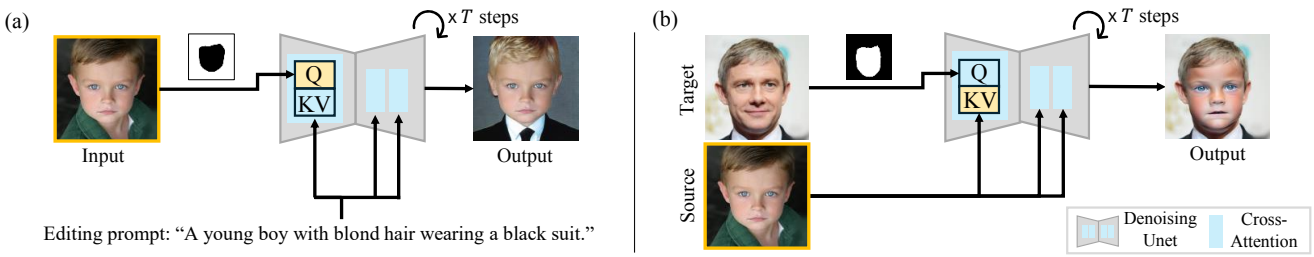

Figure 2. **Image editing and Deepfake processes in DMs**. (a) In DM-based image editing, a single image is input as a query Q and edited based on a prompt condition. (b) In DM-based deepfake, two images are used, with the target image serving as the query Q while the source image acts as the condition for swapping. This condition operates as key K and value V in the cross-attention layer.

In this paper, we focus on attacking state-of-the-art DM-based deepfakes while ensuring applicability to GAN-based models as well (Fig.1). Given the uncertainty surrounding deepfakes that malicious users might employ, we explore approaches to improve the extensibility across different architectures (e.g., GANs, DMs) and model transferability across different pre-trained backbones. Simultaneously, we propose a novel noise update method that enhances imperceptibility while being robust to JPEG compression.

For DM-based deepfake attacks, we leverage the structural properties in which the conditioning image is embedded and integrated into the denoising UNet through attention mechanisms. By utilizing the IP-Adapter [43], commonly employed for inpainting, we extract effective adversarial noise from the embedding of the conditioning image. This perturbation effectively disrupts the propagation of the conditioning process, ensuring that the final output does not replicate the features of the protected image.

To enhance the generalizability of our approach, we target two commonly used facial feature extractors. First, we attack the MTCNN model [48], which uses a cascade pyramid architecture to achieve superior performance and robust detection capabilities. Due to this, it is widely adopted not only in deepfake generation but also across various applications. We leverage the fact that the model scales images to different sizes during face detection. Our perturbation is designed to ensure robustness across various scaling factors and interpolation modes (e.g., `BILINEAR`, `AREA`), leading to superior performance compared to existing methods [19, 46]. Additionally, we target ArcFace [7], a widely adopted pre-trained model for facial feature extraction in deepfake applications. By incorporating both methods into our work, we ensure that our approach disrupts a range of deepfakes commonly used for facial landmark detection and feature extraction, thereby improving the overall robustness of our method against various deepfake systems.

In the noise updating process, we refine the perturbation using two techniques: *Noise Blur*, which measures differences between adjacent pixels for imperceptible refinement, and *Low-pass filtering*, retaining low-frequency components, enhancing robustness against JPEG compression.

To summarize, our main contributions are as follows:

- We introduce a novel attack on deepfakes based on diffusion models. To the best of our knowledge, our proposed method is the first attempt to protect images used as conditions while demonstrating robust performance across various deepfake models by targeting common facial feature extractors.
- We propose a novel noise update mechanism that integrates Gaussian blur technique with the projected gradient descent method, significantly enhancing imperceptibility. Additionally, we implement low-pass filtering to reduce perturbation loss rates during JPEG compression compared to existing methods.
- We demonstrate that our deepfake attack method is robust across various deepfake models, outperforming previous diffusion attacks by achieving higher distortion with significantly less noise.

2. Related Work

Deepfake techniques. With advancements in generative models, deepfake technology has evolved into a specialized field focused on facial synthesis. Previous deepfake models, primarily based on GANs, generally follow a three-stage process: face detection and localization, feature extraction, and face swapping. Among these, studies such as [11, 40, 41, 45, 51] employ MTCNN [48] for face detection and landmark extraction, while the majority of deepfake models, including [3, 21, 22, 40, 51, 52], leverage ArcFace [7] for identity feature extraction. These steps are similarly employed in DM-based deepfakes that have emerged with the progress of diffusion models. Notable examples, including [20, 34, 49], integrate [7] to maintain identity consistency. However, recent work has focused on leveraging the capabilities of diffusion models to develop face-swapping methods [43] that achieve high performance without explicitly following previous approaches.

Enhancing model transferability. In the research on adversarial attacks, various attempts have been made to improve transferability. [9, 38] proposed the model ensemble technique, generating adversarial examples using multiple

models to enhance their effectiveness on unseen models. [39] introduced a method that selectively utilizes specific layers within a model to improve transferability. Similarly, [15, 16, 50] investigated techniques that manipulate intermediate layer feature distributions or amplify activation values to prevent adversarial noise from overfitting to a particular model. Furthermore, [2, 47] explored the use of multiple pre-trained backbones within similar model architectures to enhance transferability across different backbone networks.

3. Method

We propose a novel pipeline, *FaceShield*, to safeguard facial images from being exploited by diverse Deepfake methods through conditional attacks on DMs and facial feature extractor attacks. In this section, we first introduce the foundational adversarial attack framework utilized across our approach (Sec.3.1). We then detail our method for disrupting information flow when a facial image is employed as a conditioning input in DMs (Sec.3.2). Subsequently, we present our approach for preventing accurate facial feature extraction (Sec.3.3). Finally, we introduce our adversarial noise update mechanism, designed to enhance imperceptibility and mitigate degradation from JPEG compression (Sec.3.4).

3.1. Preliminaries

Cross-attention mechanism. To condition generative DMs, the cross-attention mechanism is used, as shown in Fig.2. Similar to self-attention, it involves computations using the query Q, key K, and value V. However, unlike self-attention, where Q, K and V are derived from the same source, cross-attention conditions the process by obtaining Q from the noised image z_t through a learned linear projection ℓ_q, while K and V are derived from the textual or image embedding $C_{\texttt{emb}}$ using learned linear projection ℓ_k and ℓ_v, respectively:

$$Q = \ell_q(z_t), \quad K = \ell_k(C_{\texttt{emb}}), \quad V = \ell_v(C_{\texttt{emb}}), \quad \text{and} \tag{1}$$

$$\texttt{Attention}(Q, K, V) = \texttt{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where d_k are the dimensions of the key vectors.

Projected gradient descent (PGD). PGD is a widely used method for crafting adversarial examples when the user has access to the model parameters. This technique iteratively updates an adversarial perturbation by computing the gradient of a certain loss $\mathcal{L}_{\texttt{adv}}$ with respect to the input. At each step, noise is added in the gradient direction while keeping the perturbation within a predefined bound, ensuring the noise is small but effective:

$$\delta \leftarrow \texttt{Proj}_{\|\delta\|_\infty \leq \eta}\left(\delta - \alpha \cdot \texttt{sign}(\nabla_\delta \mathcal{L}_{\texttt{adv}})\right), \tag{3}$$

where α is the step size and $\texttt{Proj}_{\|\delta\|_\infty \leq \eta}(\cdot)$ projects δ onto the ℓ_∞ ball of radius η. By projecting the adversarial example back onto the valid perturbation space, PGD maintains imperceptibility while disrupting the model's predictions.

3.2. Conditioned Face Attack

We now describe our approach to protecting images, specifically by disrupting the effective transfer of information when they are used as conditioning inputs in DMs. The core of our approach is to effectively interfere with key information using minimal noise, while also ensuring that the model does not overfit by accessing only a minimal number of layers to obtain gradients. To achieve this, we propose two methods that target both the initial projection phase and the final attention mechanism during the image conditioning process within latent diffusion models [27].

Face projector attack. When images are used as conditioning inputs, they are firstly transformed into an embedding vector through a pre-trained model [26]. In this method, we access only the topmost layer $\mathcal{P}$ of the model to disrupt the projection process, causing the image to be projected with incorrect information at the initial stage. For the attack loss function, we consider that converging to a single target value might not ensure consistent convergence speeds or balanced performance. Given that one of our main goals is to design noise applicable to various images, we design our approach to induce random divergence based on the input image, using the $\mathcal{L}_1$ loss function in this process:

$$\mathcal{L}_{\texttt{proj}}(\delta; x) = \|\mathcal{P}(x+\delta) - \mathcal{P}(x)\|_1, \tag{4}$$

where δ is the adversarial noise.

Attention disruption attack. We also focus on identifying the core vectors within the denoising UNet that are most sensitive to conditional inputs. Initially, we analyze the influence of cross-attention across each UNet layer. Based on prior research [33], which shows that different cross-attention layers respond variably to conditioning information, we investigate the impact on perturbation performance for each region. Our findings lead to the conclusion that targeting attacks near mid-layers produces more significant disruption in qualitative metrics compared to using only the up-down layers or the entire layers, as supported by our experimental results in Fig.6. Based on these insights, we propose a novel approach that specifically targets mid-layers during the attack on the diffusion process.

To induce a mismatch in conditioning, we use the mid-layer cross-attention mechanism, as described in Eq. (1). Based on the idea that the condition is conveyed to the query through attention, we calculate the attention score to obtain the strength of attention. This is done by performing operations on the query $Q \in \mathbb{R}^{h \times \texttt{res} \times d}$ and key $K \in \mathbb{R}^{h \times \texttt{seq} \times d}$, where h (number of heads), `res` (resolution), `seq` (sequence length), and d (head dimension). This is followed by a `Softmax` operation along the `seq` dimension to derive the

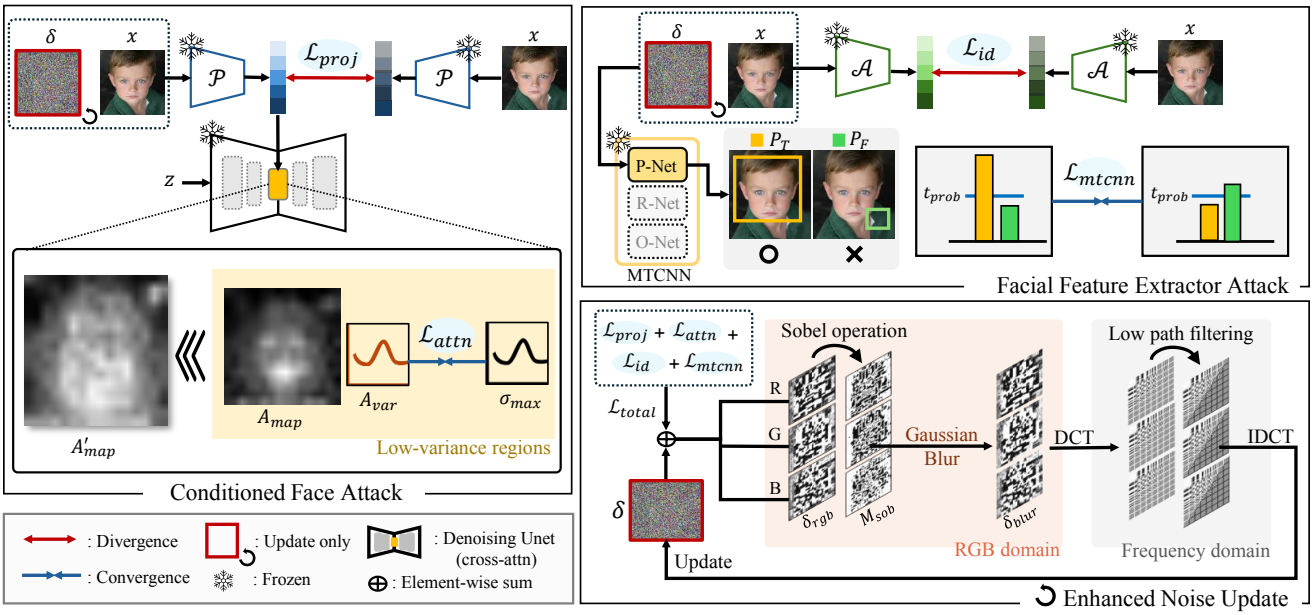

Figure 3. **Overview**. Our method has three main parts: (i) Conditioned face attack, which disrupts feature transfer by targeting the embedding process and the attention map variance in the cross-attention layer; (ii) Facial feature extractor attack, which decreases the probability value of face detection and causes extraction disruptions, and (iii) Enhanced noise update, which improves imperceptibility by applying Gaussian blur to regions with significant intensity changes between adjacent pixels, and increases robustness against JPEG compression distortion by encoding the noise in the low-frequency domain.

attention map $A_{\texttt{map}} \in \mathbb{R}^{h \times \texttt{res} \times \texttt{seq}}$. Exploiting this mechanism, we obtain the variance $A_{\texttt{var}}$, allowing us to evaluate attention strength by the following equation:

$$A_{\texttt{var}} = \frac{1}{\texttt{seq}} \sum_{i=1}^{\texttt{seq}} \left(A_{\texttt{map}}[:, i, :] - \bar{A}_{\texttt{map}} \right)^2 \in \mathbb{R}^{h \times \texttt{res}}, \quad (5)$$

where $\bar{A}_{\texttt{map}} = \frac{1}{\texttt{seq}} \sum_{i'=1}^{\texttt{seq}} A_{\texttt{map}}[:, i', :]$ is the mean of attention map across the $\texttt{seq}$ dimension.

Based on them, we propose an adversarial attack strategy that maximizes $A_{\texttt{var}}$, thereby preventing the proper reflection of conditional information K on Q. In this process, we encode the original image x to use as the query Q and project the same x to obtain the key K, which is then used to calculate $A_{\texttt{var}}$. Thereafter, we find a quantile $P_{t_{\texttt{var}}}$ corresponding to a predefined threshold $t_{\texttt{var}}$ between 0 and 1 to identify the regions exhibiting weak attention. Using this $P_{t_{\texttt{var}}}$, we create a mask $M_{\texttt{var}}$ such that values less than or equal to $P_{t_{\texttt{var}}}$ are set to 1, and values greater than $P_{t_{\texttt{var}}}$ are set to 0. This can be mathematically expressed as follows:

$$M_{\texttt{var}} = \mathbb{1}[A_{\texttt{var}} \leq P_{t_{\texttt{var}}}], \quad (6)$$

where $\mathbb{1}$ is the indicator random variable. Subsequently, we derive $A'_{\texttt{var}}$ from the same process, using the perturbed image $x + \delta$, and perform attention unequalization on the regions defined by $M_{\texttt{var}}$. This method generates missing values by assigning random attention to previously unattended regions between the original images, with the loss function $\mathcal{L}_{\texttt{attn}}$ defined as follows:

$$\mathcal{L}_{\texttt{attn}}(\delta;\, x,\, \sigma_{\texttt{max}}) = \|(\sigma_{\texttt{max}} - A'_{\texttt{var}}) \odot M_{\texttt{var}}\|_2, \quad (7)$$

where $\odot$ is the Hadamard product, and $\sigma_{\texttt{max}}$ denotes the maximum variance that can be obtained from the $\texttt{Softmax}$ output based on the $\texttt{seq}$, following the equation:

$$\sigma_{\texttt{max}} = \frac{1}{\texttt{seq}} \left(\left(1 - \frac{1}{\texttt{seq}} \right)^2 + (\texttt{seq} - 1) \cdot \frac{1}{\texttt{seq}^2} \right). \quad (8)$$

In the supplementary material, we provide the algorithm that outlines the method for calculating $\mathcal{L}_{\texttt{attn}}$.

3.3. Facial Feature Extractor Attack

We design additional perturbation targeting two types of facial extraction, enhancing the applicability of our method not only to DM-based models but also to various other deepfake architecture-based models.

MTCNN attack. We break down the MTCNN attack we propose into three principal stages: (*i*) selecting the resizing scale, (*ii*) enhancing robustness against interpolation, and (*iii*) formulating a loss function to expedite convergence. Through this process, we achieve not only various resize modes but also model transferability.

Firstly, we select a set of appropriate resizing scales $s_i \in S$. This is to ensure that our attack technique effectively targets only the bounding boxes reaching the final layers of MTCNN. The suitable scale values are selected

among the multi-scale factors that the MTCNN model internally uses, and the detailed workings are provided in the supplementary material.

Using the scale factor s_i selected in the previous step, we next scale the input image size $D_{\texttt{adv}} = (h, w)$, where h and w are the image's height and width, to yield $D_{\texttt{scl}} = (s_i \cdot h, s_i \cdot w)$. This results in an intermediate size $D_{\texttt{int}} = D_{\texttt{adv}} \odot D_{\texttt{scl}}$ obtained through element-wise multiplication. The input image $x \in \mathbb{R}^{c \times h \times w}$ is then upscaled to $D_{\texttt{int}}$ using `NEAREST` interpolation. Afterward, we downsample the image to $D_{\texttt{scl}}$ via average pooling, assigning equal weights to each region referenced during interpolation. This approach runs parallel with a direct `BILINEAR` scaling of $D_{\texttt{adv}}$ to $D_{\texttt{scl}}$, thereby ensuring robust noise generation that functions effectively across various interpolation modes.

In the final stage, we perform a targeted attack on the initial P-Net $\mathcal{T}$ to effectively disrupt the cascade pyramid structure. We pass the downsampled adversarial noise-added image $\tilde{x}_{\texttt{adv}} \in \mathbb{R}^{c \times s_i \cdot h \times s_i \cdot w}$ through $\mathcal{T}$, which outputs probabilities $P_{\texttt{T,F}}$ for bounding boxes. To expedite the convergence of the MTCNN loss function $\mathcal{L}_{\texttt{mtcnn}}$, we propose a masking technique that leverages both the existence probabilities $P_{\texttt{T}}$ and the non-existence probabilities $P_{\texttt{F}}$. The mask $M_{\texttt{prob}}$ is constructed to retain indices in $P_{\texttt{T}}$ that exceed the detection threshold $t_{\texttt{prob}}$:

$$M_{\texttt{prob}} = \mathbb{1}[P_{\texttt{T}}(i,j) > t_{\texttt{prob}}], \tag{9}$$

where $\mathbb{1}$ is the indicator random variable. Then, the $\mathcal{L}_{\texttt{mtcnn}}$ converges with the mean squared error loss using $M_{\texttt{prob}}$:

$$\mathcal{L}_{\texttt{mtcnn}}(\delta; x, p_{\texttt{gt}}) = \|(\mathcal{T}(x+\delta) - p_{\texttt{gt}}) \odot M_{\texttt{prob}}\|_2, \tag{10}$$

where $\mathcal{T}(\cdot) = [P_{\texttt{F}}, P_{\texttt{T}}]^T$, $p_{\texttt{gt}} = [t_{\texttt{prob}} + \beta t_{\texttt{prob}} - \beta]^T$, and β is a value between 0 and 1. Additional details are provided along with the algorithm in the supplementary material.

Identity attack. To effectively disrupt the accurate reflection of source face information, we target the ArcFace $\mathcal{A}$ models [7], which are face identity embedding models widely used in deepfake systems. To improve transferability, we ensemble the most commonly used pre-trained backbones within these models. Since $\mathcal{A}$ represents feature vectors extracted from the same person's face as vectors pointing in similar directions, we designed our approach to induce divergence from the original image x by employing cosine similarity loss, thereby effectively obscuring the relevant identity information:

$$\mathcal{L}_{\texttt{id}}(\delta; x) = \frac{\mathcal{A}(x+\delta) \cdot \mathcal{A}(x)}{\|\mathcal{A}(x+\delta)\|_2 \|\mathcal{A}(x)\|_2} - 1. \tag{11}$$

Overall loss operation. Accordingly, the total loss function $\mathcal{L}_{\texttt{total}}$ is defined and used as follows:

$$\begin{aligned}\mathcal{L}_{\texttt{total}} = \lambda_{\texttt{proj}}\mathcal{L}_{\texttt{proj}} + \lambda_{\texttt{attn}}\mathcal{L}_{\texttt{attn}} \\ + \lambda_{\texttt{mtcnn}}\mathcal{L}_{\texttt{mtcnn}} + \lambda_{\texttt{id}}\mathcal{L}_{\texttt{id}},\end{aligned} \tag{12}$$

Algorithm 1: FaceShield

Input: image x, steps N, noise clamp ϵ, step size α, MTCNN detection threshold $t_{\texttt{prob}}$, threshold weight β, CLIP Image Projector $\mathcal{P}$, Mid-layer cross-attention variance in Stable Diffusion $A'_{\texttt{var}}$, MTCNN P-Network $\mathcal{T}$, ArcFace $\mathcal{A}$

Result: protected image $x_{\texttt{adv}}$

Initialize adversarial perturbation $\delta \leftarrow 0$, and protected image $x_{\texttt{adv}} \leftarrow x$

for *n = 1, ..., N* **do**

  $\mathcal{L}_{\texttt{proj}} \leftarrow \|\mathcal{P}(x+\delta) - \mathcal{P}(x)\|_1$

  $\mathcal{L}_{\texttt{attn}} \leftarrow \|(\sigma_{\texttt{max}} - A'_{\texttt{var}}) \odot M_{\texttt{var}}\|_2$, where $\sigma_{\texttt{max}}$ derived from Eq. (8), and $M_{\texttt{var}}$ from Eq. (6)

  $\mathcal{L}_{\texttt{mtcnn}} \leftarrow \|(\mathcal{T}(x+\delta) - p_{\texttt{gt}}) \odot M_{\texttt{prob}}\|_2$, where $p_{\texttt{gt}} = [t_{\texttt{prob}} + \beta t_{\texttt{prob}} - \beta]^T$, and $M_{\texttt{prob}}$ from Eq. (9)

  $\mathcal{L}_{\texttt{id}} \leftarrow \frac{\mathcal{A}(x+\delta)\cdot\mathcal{A}(x)}{\|\mathcal{A}(x+\delta)\|_2\|\mathcal{A}(x)\|_2} - 1$

  Compute the total attack loss: $\mathcal{L}_{\texttt{total}} = \lambda_{\texttt{proj}}\mathcal{L}_{\texttt{proj}} + \lambda_{\texttt{attn}}\mathcal{L}_{\texttt{attn}} + \lambda_{\texttt{mtcnn}}\mathcal{L}_{\texttt{mtcnn}} + \lambda_{\texttt{id}}\mathcal{L}_{\texttt{id}}$

  Update adversarial perturbation: $\delta \leftarrow \alpha \cdot \text{sign}(\nabla_{x_{\texttt{adv}}}\mathcal{L}_{\texttt{total}})$

  $\delta_{\texttt{blur}} \leftarrow$ **GaussianBlur**(δ)

  $\delta'_{\texttt{rgb}} \leftarrow$ **LowPassFilter**$(\delta_{\texttt{blur}})$

  $x_{\texttt{adv}} \leftarrow x_{\texttt{adv}} - \delta'_{\texttt{rgb}}$

  $x_{\texttt{adv}} \leftarrow x + \text{clip}(x_{\texttt{adv}} - x, -\epsilon, \epsilon)$

end

Clip the image range: $x_{\texttt{adv}} \leftarrow \text{clip}(x_{\texttt{adv}}, 0, 255)$

where each λ is a hyperparameter derived from grid searches to control the strength of the respective loss term. Additionally, the sign of λ determines the convergence or divergence of the loss function (i.e., $\lambda_{\texttt{proj}}$ and $\lambda_{\texttt{id}}$ are negative, while $\lambda_{\texttt{attn}}$ and $\lambda_{\texttt{mtcnn}}$ are positive).

3.4. Enhanced Noise Update

We integrate two additional techniques into the standard PGD to enhance robustness by enabling more imperceptible noise updates and preventing the loss of information due to purification techniques.

Gaussian blur. To enhance noise imperceptibility, we introduce a technique that constrains variations between adjacent regions, addressing the limitations of PGD methods (see Eq. (3)) that only regulate overall noise magnitude. This stems from the observation that differences between neighboring pixels can be as perceptible as the total noise itself. To achieve this, we utilize the Sobel operator [17] to emphasize areas of rapid intensity change, generating a mask $M_{\texttt{sob}}$ that highlights image boundaries. Gaussian blur $\mathcal{G}(\cdot)$ is then applied selectively to these regions during noise up-

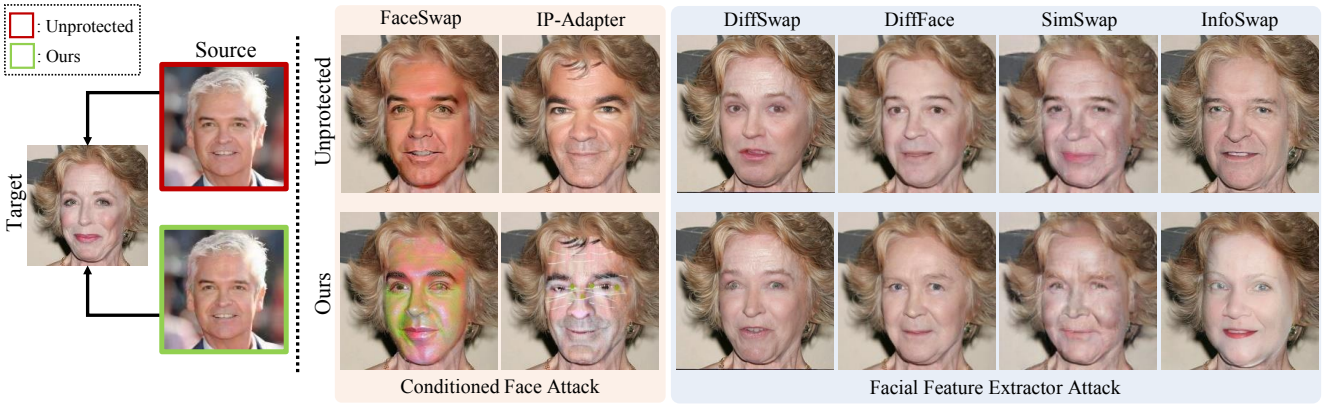

Figure 4. **Qualitative Results**. Protection performance across various deepfake models when our adversarial noise is applied. Models [34, 43] highlighted in the orange box typically exhibit facial distortions due to the influence described in Sec. 3.2, while those [3, 11, 20, 49] in the blue box display newly generated faces that diverge from the source image, attributed to the impact detailed in Sec. 3.3.

dates, ensuring smoother transitions between adjacent pixels and maintaining a consistent visual appearance:

$$\delta_{\texttt{blur}} = \mathcal{G}(\delta) \odot M_{\texttt{sob}} + \delta \odot (1 - M_{\texttt{sob}}). \tag{13}$$

Low pass filtering. To minimize information loss when saving images in `JPEG` format and ensure robustness to bit reduction during the compression process [10], low-frequency components are utilized. At each iteration, the newly updated adversarial noise $\delta_{\texttt{rgb}} \in \mathbb{R}^{c \times h \times w}$, where c (channel), h (height), and w (width), undergoes a padding operation and patchification according to a predefined patch size p. Then a DCT transform [1] is applied to each patch and channel, resulting in $\delta_{\texttt{dct}} \in \mathbb{R}^{c \times h' \times w' \times p \times p}$, where $h' = h/p$ and $w' = w/p$, in the frequency domain. Using a low-pass filtering mask $M_{\texttt{lp}}$, only the low-frequency components of the noise are retained. The noise $\delta'_{\texttt{rgb}} \in \mathbb{R}^{c \times h \times w}$ is then reconstructed back into the RGB domain through an inverse transformation. The effectiveness of this approach is demonstrated through the experimental results presented in the supplementary material, and the overall operation of *FaceShield* is described in Algorithm 1.

4. Experiments

4.1. Setups

Evaluation details. For a fair performance comparison, we use open-source baseline [23, 24, 29, 42] and apply noise to the same dataset under identical hyperparameter settings. The corresponding results are presented in Table 2, while Table 1 provides a performance comparison on diffusion-based deepfakes [20, 34, 43, 49]. The extensibility experiments on GAN-based models [3, 11] are shown in Table 4, where, in the absence of existing attack methods for these models, we validate *FaceShield*'s effectiveness through comparisons with the original images. In cases where the feature extractor fails to detect a face, we adjust the generation process to exclude facial features during reconstruction. Detailed descriptions and an analysis of the resources are provided in the supplementary material.

Datasets. We evaluate our method using two datasets: CelebA-HQ [18] and VGGFace2-HQ [4], both of which have been used in previous studies [3, 11, 34]. The former is the high-resolution version of CelebA, containing 30,000 celebrity face images, while the latter is the high-resolution version of VGGFace2, consisting of 3.3 million face images from 9,131 unique identities. For our experiments, we randomly select 200 identities from each dataset, using 100 images for the source and 100 images for the target.

4.2. Qualitative Results

Performance results across deepfakes. As shown in Fig.4, *FaceShield* demonstrates robustness across various deepfake models. The perturbations result in either (i) pronounced artifacts reflecting non-relevant facial information instead of key features [34, 43], or (ii) a complete misinterpretation of the source face, generating a new, unrelated identity [3, 11, 20, 49].

Comparison with state-of-the-art methods. We compare our method with baselines on DM-based deepfake model [43]. Although the methods [23, 24, 29, 42] that achieved high performance in diffusion adversarial attacks fail to induce visible changes on the deepfake model, ours demonstrates strong protective performance (Fig.5).

4.3. Quantitative Results

Automatic metrics. As shown in Table 1, we compare *FaceShield* to baseline methods across deepfake models [3, 11, 20, 34, 43, 49] using L_2, Identity Score Matching [32] (ISM), and PSNR. The L_2 and PSNR metrics evaluate image quality by comparing deepfake results from clean and protected images, with higher L_2 and lower PSNR indicating more distortion. ISM measures the similarity between

Model	DiffFace [20]				DiffSwap [49]				FaceSwap [34]				IP-Adapter [43]			
Dataset	CelebA-HQ [18]															
Method	$L_2 \uparrow$	ISM $\downarrow$	PSNR $\downarrow$	HE $\uparrow$	$L_2 \uparrow$	ISM $\downarrow$	PSNR $\downarrow$	HE $\uparrow$	$L_2 \uparrow$	ISM $\downarrow$	PSNR $\downarrow$	HE $\uparrow$	$L_2 \uparrow$	ISM $\downarrow$	PSNR $\downarrow$	HE $\uparrow$
AdvDM [24]	0.021	0.471	39.368	4.22	0.068	0.199	28.362	4.68	0.303	0.245	21.615	4.52	0.207	0.235	25.332	2.76
Mist [23]	0.021	0.468	39.443	3.94	0.067	0.201	28.384	4.18	0.287	0.230	22.263	4.78	0.152	0.265	28.213	4.26
PhotoGuard [29]	0.022	0.469	39.194	3.82	0.068	0.201	28.292	4.58	0.282	0.238	22.316	4.44	0.153	0.268	28.101	4.44
SDST [42]	0.021	0.470	39.512	4.08	0.067	0.207	28.383	5.04	0.274	0.261	22.582	4.68	0.147	0.273	28.440	4.32
Ours	**0.044**	**0.243**	**32.052**	**5.76**	**0.072**	**0.163**	**27.833**	**6.20**	**0.336**	**0.194**	**20.759**	**6.16**	**0.350**	**0.072**	**20.266**	**6.60**
Ours (Q=75)	0.043	0.259	32.259	-	0.070	0.169	28.034	-	0.317	0.209	21.286	-	0.326	0.112	20.867	-
Dataset	VGGFace2-HQ [4]															
Method	$L_2 \uparrow$	ISM $\downarrow$	PSNR $\downarrow$	HE $\uparrow$	$L_2 \uparrow$	ISM $\downarrow$	PSNR $\downarrow$	HE $\uparrow$	$L_2 \uparrow$	ISM $\downarrow$	PSNR $\downarrow$	HE $\uparrow$	$L_2 \uparrow$	ISM $\downarrow$	PSNR $\downarrow$	HE $\uparrow$
AdvDM [24]	0.042	0.479	33.064	3.68	0.105	0.215	24.769	4.78	0.419	0.361	18.596	4.38	0.251	0.271	23.250	2.36
Mist [23]	0.041	0.478	33.215	4.26	0.102	0.227	24.964	3.94	0.379	0.259	19.626	4.50	0.181	0.291	26.070	4.10
PhotoGuard [29]	0.043	0.479	32.938	3.96	0.110	0.215	24.272	4.18	0.373	0.266	19.655	4.14	0.180	0.294	26.157	3.82
SDST [42]	0.041	0.483	33.242	5.30	0.107	0.225	24.506	4.58	0.359	0.258	19.996	4.14	0.166	0.292	26.784	4.06
Ours	**0.062**	**0.278**	**29.204**	**6.10**	**0.113**	**0.177**	**24.054**	**6.12**	**0.453**	**0.237**	**17.919**	**6.16**	**0.382**	**0.112**	**19.478**	**6.42**
Ours (Q=75)	0.060	0.308	29.435	-	0.112	0.185	24.201	-	0.421	0.237	18.573	-	0.377	0.167	19.618	-

Table 1. Comparison of perturbation effectiveness among baseline methods on four deepfake models using the CelebA-HQ [18] and VGGFace2-HQ [4] datasets. Our method exhibits the largest distortion in image quality (L2, PSNR) and source similarity (ISM), as well as in human evaluation (HE). Results on JPEG-compressed images (Quality factor 75) further confirm robust protection under compression.

Figure 5. We generate deepfake [43] results from protected images of methods [23, 24, 29, 42]. While these fail to disrupt deepfake generation, our method causes deepfakes to malfunction.

the source face and the deepfake output, with lower values indicating less similarity. We conduct experiments on 100 source-target pairs from CelebA-HQ [18] and VGGFace2-HQ [4], showing that *FaceShield* outperforms baselines across all metrics. We also analyze the noise levels in protected images using LPIPS, PSNR, and SSIM, as shown in Table 2. These image quality metrics, compared between protected and original images, show that our method consistently produces less noise than baseline methods. Additionally, we measure the Frequency Rate (FR), which indicates that most of *FaceShield*'s noise is concentrated in low frequencies. This property helps maintain its effectiveness under JPEG compression. To verify, we compressed the protected images to JPEG Quality 75 and tested across four deepfake models. The results show that while performance slightly decreases, *FaceShield* still outperforms baseline methods, as shown in Table 1, **Ours** (Q=75).

Human evaluation. We conduct a human evaluation (HE) on the same methods and models, using 20 source images and 100 participants recruited via Amazon Mechanical Turk. Participants assess two factors: protection noise visibility (Table 2) and the similarity between the source image and its deepfake output (Table 1). We employ a Likert scale from 1 to 7. For noise visibility, a score of 7 indicates the least visible noise, while for deepfake similarity, a score of 7 reflects a significant deviation from the source identity.

Dataset	CelebA-HQ [18]				
Method	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	FR $\uparrow$	HE $\uparrow$
AdvDM [24]	0.4214	30.4476	0.8438	2.1077	3.86
Mist [23]	0.5492	29.9935	0.8684	1.6583	4.70
PhotoGuard [29]	0.5515	29.9127	0.8669	1.6538	4.82
SDST [42]	0.5409	31.4762	0.9033	1.6767	5.12
Ours	**0.2017**	**32.6289**	**0.9394**	**18.4651**	**5.64**
Dataset	VGGFace2-HQ [4]				
Method	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	FR $\uparrow$	HE $\uparrow$
AdvDM [24]	0.4108	30.2523	0.8436	2.0667	3.66
Mist [23]	0.5208	29.9068	0.8721	1.6872	4.34
PhotoGuard [29]	0.5221	29.8204	0.8712	1.6824	4.62
SDST [42]	0.5060	31.3545	0.9092	1.6892	4.48
Ours	**0.1941**	**31.5799**	**0.9341**	**18.0400**	**5.28**

Table 2. With the same step size, iterations, and noise clamping values applied, our method shows the least distortion across three image quality metrics (LPIPS, SSIM, PSNR, HE) and demonstrates a higher low-frequency content (FR).

4.4. Ablation Study

Effect of gaussian blur on noise. To evaluate the effect of Gaussian blur, one of the enhanced noise update methods, we present qualitative results in the Supplementary Material comparing the blur effect's presence and absence. From this comparison, it is clear that the noise update becomes significantly more invisible at the boundaries of abrupt changes in the noise, as detected through Sobel filtering.

Impact of each loss function. To demonstrate the generalizability of each loss function, we conducted an ablation study across six models using the ISM metric, as shown in Table 3. The results illustrate how *FaceShield* protects faces, as seen in Table 1 and Table 4, with red shading indicating performance degradation when a loss function is removed. This experiment shows that each loss function successfully impacts multiple models, and by combining them into $\mathcal{L}_{\texttt{total}}$, we cover a broader range of deepfakes.

ISM↓	DiffFace	DiffSwap	FaceSwap	IP-Adapter	SimSwap	InfoSwap
w/o $\mathcal{L}_{\texttt{proj}}$	0.241	0.167	0.270	0.135	0.544	0.256
w/o $\mathcal{L}_{\texttt{attn}}$	0.254	0.170	0.223	0.076	0.168	0.252
w/o $\mathcal{L}_{\texttt{mtcnn}}$	0.231	0.174	0.166	0.047	0.183	0.354
w/o $\mathcal{L}_{\texttt{id}}$	0.446	0.175	0.217	0.040	0.512	0.430
$\mathcal{L}_{\texttt{total}}$	0.243	0.163	0.194	0.072	0.184	0.237

Table 3. Each model's performance is measured using the ISM score, confirming that each loss function ensures transferability across various deepfakes. As a result, the integrated $\mathcal{L}_{\texttt{total}}$ is capable of covering a wider range.

Attack effectiveness of mid-layers. We qualitatively show that focusing the diffusion attack on the mid-layers of the denoising UNet [27] is more effective than applying it to the entire layers or the up/down layers, as shown in Fig.6. The experiment is conducted by applying noise $\delta = 4/255$ to create protected images, which are then passed through the deepfake model [43] to compare the resulting outputs.

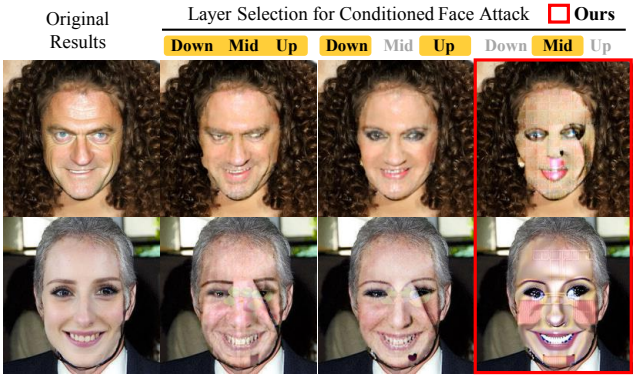

Figure 6. A comparison of conditioned face attack results across UNet [27] layers shows the best protection when targeting mid-layer cross-attention.

MTCNN resize evaluation. To demonstrate the superiority of our method across various resizing modes and model transferability, we conduct an ablation study comparing it to the conventional `BILINEAR` scaling method [46]. We evaluate performance using different scaling methods from the `OpenCV` and `PIL` libraries, with 3,000 images from both the CelebA-HQ and VGGFace2-HQ datasets. Transferability is assessed through experiments conducted on both PyTorch and TensorFlow versions. The evaluation is based on the final bounding box detection probabilities from MTCNN [48], and the results in the supplementary material confirm that our method outperforms existing approaches.

4.5. Applications

Extensibility on GAN-based deepfake models. We also conduct additional experiments on the GAN-based diffusion model [3, 11]. The experimental conditions are the same as those in Table 2, and the results, as shown in Table 4, indicate a degradation in model performance. Qualitative assessments are provided in Fig. 4.

Model	SimSwap [3]			InfoSwap [11]		
Dataset	CelebA-HQ [18]					
Method	$L_2 \uparrow$	ISM ↓	PSNR ↓	$L_2 \uparrow$	ISM ↓	PSNR ↓
Original	0.000	0.544	80.000	0.000	0.431	80.000
Ours	0.070	0.184	26.921	0.129	0.237	30.220
Dataset	VGGFace2-HQ [4]					
Method	$L_2 \uparrow$	ISM ↓	PSNR ↓	$L_2 \uparrow$	ISM ↓	PSNR ↓
Original	0.000	0.681	80.000	0.000	0.565	80.000
Ours	0.067	0.314	27.305	0.142	0.356	29.044

Table 4. Applicability of *FaceShield* to other deepfake frameworks. Our method, when applied to GAN-based models, not only reduces image quality (L_2, PSNR) but also significantly lowers source similarity (ISM).

Transferability with different weights. To demonstrate that *FaceShield* ensures robust transferability not only across structurally different models but also to models with similar architectures but different pre-trained weights, we evaluated its performance on various versions of IP-Adapter [43]. The results, which can be found in the supplementary material, confirm the superior transferability performance of our method.

5. Conclusion

In this study, we propose *FaceShield*, an invisible facial protection technique designed to attack various deepfakes. Through comparisons with multiple baseline methods, we demonstrate that *FaceShield* offers superior protection with significantly lower resource costs, particularly for deepfake models utilizing the latest diffusion techniques. Furthermore, its design integrates diverse transferability enhancement strategies, ensuring consistent performance not only across various pretrained versions but also across diffusion-based models with different architectures. This robustness extends to entirely different architectures, including GAN-based models. Additionally, by incorporating an improved noise update mechanism that ensures invisibility while minimizing information loss, *FaceShield* proves to be a practical and effective solution for preventing the misuse of facial images across a wide range of deepfake systems.

Limitations and Future Work. Although we introduce a method to enhance robustness against JPEG compression and resizing, other purification techniques still exist, which may lead to the potential loss of our protective noise information. Therefore, we plan to further strengthen the noise to effectively counter a broader range of purification methods.

6. Acknowledgment

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025, 47%; Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469, 25%), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00521602, 25%), Institute of Information & communications Technology Planning & Evaluation (IITP) & ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT) (No.RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University), 1%; IITP-2025-RS-2024-00436857, 1%; IITP-2025-RS-2025-02304828, Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents, 1%), and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

References

[1] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. 6

[2] Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[3] 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. 1, 2, 6, 8

[4] Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. Simswap++: Towards faster and high-quality identity swapping. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(1):576–592, 2024. 6, 7, 8

[5] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1133–1143, 2024. 1

[6] June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. *arXiv preprint arXiv:2410.05694*, 2024. 1

[7] 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. 2, 5

[8] Junhao Dong and Xiaohua Xie. Visually maintained image disturbance against deepfake face swapping. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1

[9] 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 conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2

[10] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 6

[11] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3404–3413, 2021. 2, 6, 8

[12] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 1

[13] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023. 1

[14] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 989–997, 2022. 1

[15] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2019. 3

[16] Nathan Inkawhich, Kevin Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. In *International Conference on Learning Representations*, 2020. 3

[17] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2): 358–367, 1988. 5

[18] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6, 7, 8

[19] Edgar Kaziakhmedov, Klim Kireev, Grigorii Melnikov, Mikhail Pautov, and Aleksandr Petiushko. Real-world attack on mtcnn face detection system. In *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, pages 0422–0427. IEEE, 2019. 2

[20] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seung Wook Kim, and Kwanghee Lee.

Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recognit.*, 163:111451, 2022. 1, 2, 6, 7

[21] Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, and Ran He. Faceinpainter: High fidelity face adaptation to heterogeneous domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5089–5098, 2021. 2

[22] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 1, 2

[23] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023. 1, 6, 7

[24] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023. 1, 6, 7

[25] Kunlin Liu, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Wenbo Zhou, and Weiming Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognit.*, 141:109628, 2020. 1

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

[27] 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 (CVPR)*, pages 10684–10695, 2022. 3, 8

[28] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 236–251. Springer, 2020. 1

[29] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *International Conference on Machine Learning*, 2023. 1, 6, 7

[30] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023. 1

[31] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 1

[32] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023. 6

[33] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 3

[34] Feifei Wang. Face swap via diffusion model. *arXiv preprint arXiv:2403.01108*, 2024. 1, 2, 6, 7

[35] Run Wang, Ziheng Huang, Zhikai Chen, Li Liu, Jing Chen, and Lina Wang. Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. *arXiv preprint arXiv:2206.00477*, 2022. 1

[36] Tianyi Wang and Kam Pui Chow. Noise based deepfake detection via multi-head relative-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14548–14556, 2023. 1

[37] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021. 1

[38] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7639–7648, 2021. 2

[39] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 3

[40] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7632–7641, 2022. 2

[41] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7642–7651, 2022. 2

[42] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 6, 7

[43] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 6, 7, 8

[44] Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 53–62, 2020. 1

[45] Ge Yuan, Maomao Li, Yong Zhang, and Huicheng Zheng. Reliableswap: Boosting general face swapping via reliable supervision. *arXiv preprint arXiv:2306.05356*, 2023. 2

[46] Chongyang Zhang, Yu Qi, and Hiroyuki Kameda. Multi-scale perturbation fusion adversarial attack on mtcnn face

detection system. In *2022 4th International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 142–146. IEEE, 2022. 2, 8

[47] Jianping Zhang, Zhuoer Xu, shiwen cui, Changhua Meng, Weibin Wu, and Michael Lyu. On the robustness of latent diffusion models, 2024. 3

[48] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 2, 8

[49] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023. 1, 2, 6, 7

[50] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018. 3

[51] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4834–4844, 2021. 1, 2

[52] Yixuan Zhu, Wenliang Zhao, Yansong Tang, Yongming Rao, Jie Zhou, and Jiwen Lu. Stableswap: Stable face swapping in a shared and controllable latent space. *IEEE Transactions on Multimedia*, 2024. 2