Tsinghua Science and Technology
Doi: 10.26599/TST.2026.90100020

Research Article

Towards Robust and Highly Stealthy Proactive Deepfake Defense on Consumer Electronic Devices

Chengsheng Yuan, Youqiang Cao, YuChen Wu, Fengjun Xiao(✉), Xinting Li(✉), Zhihua Xia

Abstract The integration of AIGC technologies into consumer electronics has democratized the creation of highly realistic facial deepfakes, posing unprecedented threats to personal identity security and undermining trust in digital ecosystem. Current proactive defense methods primarily operate by injecting perturbations directly into the pixel space of facial images to disrupt deepfake synthesis. However, such pixel-level modifications often introduce noticeable artifacts, making them easily detectable by human observers and vulnerable to common image transformations. To tackle this issue, this paper proposes DiffDefend, a novel three-stage proactive defense framework based on low-frequency perturbations in the diffusion latent space. The framework first employs a stable diffusion model to encode the input image into a latent representation that preserves high-fidelity reconstruction capability. Adversarial perturbations are then introduced specifically into the low-frequency subband of the latent code, which is extracted via discrete wavelet transform (DWT/Discrete Wavelet Transform), and optimized through an adversarial game combining visual and adversarial losses, yielding an adversarial latent representation. Ultimately, the adversarial latent code is reconstructed into an adversarial image leveraging the powerful generative capabilities of the diffusion model, effectively defending against deepfake manipulations. Moreover, we design a multi-level loss that combines pixel-level and semantic-level constraints to enhance the visual quality and defense efficacy of the adversarial image. To evaluate the efficacy of our method, we conduct experiments on two representative deepfake tasks: attribute editing and face reenactment. Experimental results demonstrate that our method achieves superior performance over existing benchmarks in both visual quality and defensive performance.

Keywords Security-Oriented Causal Intervention, stable diffusion model, Visual Security and Adversarial Defense, Visual Reasoning

1 Introduction

In recent years, the rapid advancement of generative techniques for image synthesis and editing, particularly Generative Adversarial Networks (GANs) [1], has greatly enhanced the capability of consumer electronics, including smartphones, smart cameras, and social media platforms, to produce and manipulate high-quality facial images. Yet these very capabilities also poses emerging risks to consumer-level security, digital trust, and user identity protection across modern devices. As generative models become increasingly embedded in consumer electronics, they can be exploited to deceive facial recognition systems [2–4] integrated into such products, or to enable malicious actors to fabricate realistic videos of public figures using widely accessible tools. Thus, developing effective defense mechanisms against deepfakes that are suitable for deployment in consumer electronics has become a critical research imperative.

Existing defense paradigms against deepfakes are broadly divided into two categories: passive detection and proactive defense. Passive detection methods [5–7] typically utilize binary classifiers to identify artifacts introduced during the image manipulation process, thereby distinguishing authentic from manipulated facial images. In contrast, proactive defense methods [8–10] aim to protect facial images prior to the creation of a deepfake, ensuring that the fabricated images cannot meet the malicious intent of the adversary. Current proactive defense strategies predominantly rely on injecting ℓ_p-norm constrained adversarial perturbations into pixel space to dis-

• Chengsheng Yuan, Youqiang Cao and Yuchen Wu are with the Engineering Research Center of Digital Forensics, Ministry of Education, School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China. Email: yuancs@nuist.edu.cn; caoyq@nuist.edu.cn; 18434286148@163.com
• Fengjun Xiao is with Zhejiang Informatization Development Institute, Hangzhou Dianzi University, Hangzhou 310018, China. Email: bhxfj@126.com (Corresponding author)
• Xinting Li is with the School of Foreign Languages, National University of Defense Technology, Nanjing 210039, China. Email:lixt@tju.edu.cn (Corresponding author)
• Zhihua Xia is with the College of Cyber Security, Jinan University, Guangzhou 510632, China. Email: xiazhihua@jnu.edu.cn
Manuscript received: 2025-12-04; revised: 2026-01-03; accepted: 2026-01-16

rupt deepfake generation. These perturbation-based methods, however, often introduce human-perceptible artifacts that compromise visual quality. Thus, it is essential to explore more effective approaches that maintain high visual quality in facial images while providing powerful defense capabilities.

In this paper, we propose DiffDefend, a novel proactive defense framework that introduces low-frequency perturbations in the diffusion latent space to counter deepfake generation. Unlike existing methods that directly inject perturbations into pixel space—often resulting in visible artifacts and limited robustness—our approach leverages the latent space of diffusion models [11] to achieve two key advantages. First, the inherent progressive denoising mechanism naturally suppresses high-frequency noise while ensuring that adversarial examples remain consistent with the underlying data distribution, thereby significantly improving imperceptibility. Second, the iterative refinement process guided by semantic constraints enables enhanced visual fidelity and better structural coherence compared to GAN-based defense approaches. Capitalizing on these advantages, we specifically perturb the low-frequency subbands of the latent codes, which are extracted via discrete wavelet transform (DWT). Low-frequency perturbations offer superior imperceptibility to human observers and demonstrate stronger robustness against common image transformations-such as compression and filtering-compared to high-frequency or full-spectrum perturbations that are more vulnerable to such degradations. To jointly optimize stealth and robustness, we introduce a multi-level constraint that simultaneously enforces pixel-level naturalness and semantic-level consistency. This ensures that the reconstructed adversarial images maintain photorealistic quality while effectively disrupting deepfake generation. As a result, our method enables the release of a protected surrogate image-reconstructed from the optimized latent code-that preserves visual fidelity while degrading the performance of deepfake models. Compared to pixel-space adversarial perturbation methods, our approach produces reconstructed images with significantly better visual quality, as evidenced in Fig. 1.

The main contributions of this paper are summarized as follows:

- To enhance the imperceptibility and robustness of adversarial images, this paper proposes a novel proactive defense method against deepfakes by optimizing the latent space of diffusion models. Specifically, adversarial perturbations are injected into the low-frequency subband of latent codes via discrete wavelet transform (DWT), preserving visual fidelity while resisting common image transformations.
- To further enhance both the visual fidelity and defensive efficacy of adversarial images, this paper formulates multi-level visual and adversarial constraints that exploit information at both the pixel and semantic feature levels for the optimization of latent codes.
- Extensive experimental results demonstrate that the proposed framework outperforms existing methods in terms of both visual quality and defense capability.

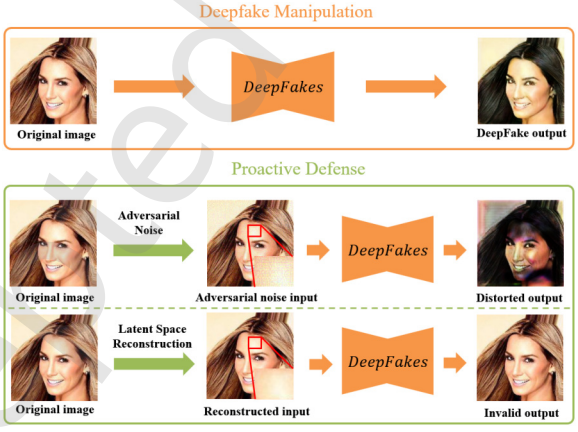


Fig. 1 Deepfake manipulation and proactive defense. Different defense methods yield varying effects on original and manipulated images. Introducing adversarial noise directly into the image pixel space frequently leads to the inclusion of artifacts in the original image, with the objective of producing distorted manipulated versions. In contrast, latent space reconstruction methods aim to produce manipulated images that are visually more natural and closely resemble the original images.

2 Related Works

2.1 Deepfake Models

In recent years, Generative Adversarial Networks (GANs) [1] and Diffusion Models [11] have achieved remarkable success in the fields of image synthesis and manipulation, facilitating the generation of highly realistic and superior-quality visual content. Leveraging these advancements, DeepFake technology has emerged, allowing the creation of fabricated facial images or videos that pose significant threats to personal privacy, political security, and societal trust. Broadly categorized, attribute editing and face reenactment represent two prevalent types of DeepFakes, both of which exploit GAN or diffusion model architectures for different manipulation tasks. Attribute editing involves fine-grained modification of facial features using models like StarGAN and DiffAE [12, 13], which can alter attributes such as hair color, age and gender while preserving identity. This technique has practical applications in areas like virtual avatars and aesthetic enhancement

but also presents severe misuse risks, such as identity forgery and unauthorized image manipulation. Face reenactment, or expression swapping, centers on transferring facial expressions from a source image to a target image. Methodologies like GANimation and Face-Adapter [14, 15] demonstrate proficiency in generating realistic expression changes, enabling applications in animation and entertainment. However, their misuse for creating deceptive or malicious content has emerged as a growing security concern.

This paper primarily addresses these two categories of DeepFakes, highlighting their threats to privacy and security. Given the advancing sophistication of both GAN-based and diffusion-based manipulation techniques, developing robust defense mechanisms capable of countering attribute editing and face reenactment attacks has become increasingly critical.

2.2 Deepfake Proactive Defense

Currently, the majority of proactive defense strategies seek to thwart DeepFake generation by directly injecting adversarial perturbations, which are constrained by the ℓ_p-norm, onto the image surface [8, 10, 16]. These approaches predominantly revolve around introducing pixel-level noise to degrade the performance of DeepFake models during image manipulation. Furthermore, certain investigations [9, 17] explore grey-box and black-box scenarios, where the internal architecture of the target model is partially or entirely unknown. Other research [18, 19] explores the utilization of universal adversarial perturbations, aiming to create perturbations that can generalize effectively across various inputs. Although these methods demonstrate certain effectiveness in disrupting DeepFake generation, they are accompanied by several notable limitations. Specifically, the injection of adversarial noise directly into the pixel space can significantly degrade the visual quality of the original image, introducing high-frequency artifacts or unnatural distortions that are readily perceptible to the human eye. Moreover, these adversarial artifacts are especially susceptible to a variety of post-processing methods, including smoothing, filtering, or even re-encoding, which can effectively mitigate the intended perturbations [20]. As a result, the robustness and stealthiness of these defense strategies are often compromised, especially when faced with adaptive attackers capable of implementing countermeasures.

In contrast, other studies [21, 22] have investigated the feasibility of searching for adversarial latent codes within the latent space of GANs. Instead of directly applying perturbations to the pixel space, these methodologies aspire to manipulate the latent representations of images to produce adversarial effects. By functioning within the latent space, such approaches can potentially achieve higher stealthiness, as perturbations are more aligned with the underlying data distribution and less likely to introduce perceptible noise artifacts. However, these latent-space-based methods are constrained by the inherent limitations of GANs' generative capabilities. Specifically, manipulations carried out within the latent space of GANs may unintentionally modify the semantic content of the original image, resulting in unintended modifications in facial structure, expression, texture, or other high-level attributes. Although these semantic shifts are often subtle, they can nonetheless compromise the visual fidelity of the generated adversarial examples, making them unsuitable for applications where preserving the original content is essential.

Therefore, while these latent-space-based approaches offer certain benefits in terms of stealthiness, they remain limited by the architectural constraints of GANs and their vulnerability to semantic deterioration.

2.3 Stable Diffusion Models

Recent years have witnessed growing interest in stable diffusion models [11], which have demonstrated remarkable performance across diverse generative tasks. As a category of probabilistic generative models, they synthesize high-quality samples through a Markov chain that progressively recovers data from noise. Unlike GANs that directly map random noise to data via a single generator pass, diffusion models employ an iterative refinement mechanism. This multi-step denoising procedure enhances training stability and output quality, making them particularly suitable for high-dimensional generation tasks in image, audio, and video domains. The underlying framework consists of a forward process that gradually injects noise into data until it becomes pure Gaussian noise, and a reverse process that iteratively denoises the signal to reconstruct realistic samples.

After extensive training on large-scale datasets, stable diffusion models can generate high-quality images from randomly sampled noise. Moreover, they are highly versatile and can produce specific images guided by textual prompts, enabling diverse applications such as text-to-image generation, style transfer, and artistic rendering. Due to their outstanding performance, diffusion models have been successfully extended to other domains, including image inpainting [23, 24], image super-resolution [25], and real image editing [26, 27]. Their inherent iterative refinement process allows them to achieve impressive results even in challenging tasks that requires high levels of detail and precision. When compared to GANs, diffusion models demonstrate superior ability in producing perceptually accurate and high-quality natural images. Their

structured generation process effectively mitigates issues such as mode collapse and training instability commonly encountered in GANs, thereby offering a more reliable framework for high-fidelity image synthesis. Given these advantages, this paper introduces diffusion models into the realm of proactive defense against DeepFakes. By leveraging their robust generation process, diffusion models can enhance the visual quality of protected facial images while maintaining semantic integrity. Additionally, their ability to produce high-quality output with minimal perceptual distortion aligns well with the imperceptibility requirements in proactive defense mechanisms.

Recent Advancements in Diffusion-Based Defense. Concurrently with the evolution of deepfake generators, defense mechanisms have also rapidly evolved. Notably, diffusion models have been repurposed not only for generation but also as powerful tools for adversarial purificationand robust feature learning. For instance, Wei et al.[28] leveraged a diffusion model to restore images corrupted by adversarial patches, demonstrating the model's inherent robustness to localized, physically-realizable attacks. Similarly, Lu et al.[29] proposed a dual-stage framework that employs a diffusion model for both detection and purification of adversarial examples in remote sensing, highlighting its versatility. These works underscore the potential of diffusion models as a backbone for constructing next-generation defense systems. However, they primarily focus on reactive purification in pixel space after an attack occurs. In contrast, our work explores a proactive defense paradigm within the latent space of diffusion models, aiming to prevent manipulation at its origin by exploiting the frequency-domain properties of latent representations.

In conclusion, diffusion models present a promising alternative to GAN-based approaches, offering improved robustness and visual fidelity in protecting facial images from malicious manipulation.

3 Methodology

3.1 Problem Definition

Given an original facial image x_{ori}, the goal of proactive defense against deepfakes involving invalid attacks is to introduce perturbations to x_{ori} such that the resulting adversarial image x'_{adv} maintains visual consistency with x_{ori}, while simultaneously degrading the generative capabilities of the deepfake model $\mathcal{F}$. More specifically, the adversarial image x'_{adv} should ensure that the fake image x'_{fake}, generated by the deepfake model $\mathcal{F}$ from x'_{adv}, closely resembles the original image x_{ori}:

$$\arg\min_{x'_{\text{adv}}} \mathcal{L}\left(\mathcal{F}\left(x'_{\text{adv}}\right), x_{\text{ori}}\right) \quad s.t. \quad d\left(x'_{\text{adv}}, x_{\text{ori}}\right) \leqslant \eta \quad (1)$$

where $\mathcal{L}(\cdot)$ denotes the loss function employed to assess the discrepancy between two images, while $d(\cdot)$ stands for the distance metric that quantifies the variation between the original and adversarial images. The symbol η represents the maximum permissible error.

Unlike previous methods that enforce ℓ_p-norm constraints on pixel values for pixel-based attacks, this paper introduces perturbations within the latent space of the diffusion model. By leveraging the inherent properties of the diffusion model, the attack ensures that the adversarial perturbations maintain a visually plausible and natural appearance.

3.2 Preliminaries

Rationale for Low-Frequency Perturbation:The rationale behind perturbing low-frequency components is twofold: The rationale behind selectively perturbing the low-frequency subband stems from its critical role in representing the global structural and semantic content of an image, while exhibiting inherent robustness to common image transformations. In the frequency domain derived via DWT, the LLLL subband encapsulates the coarse-scale information—such as facial contours, lighting distribution, and overall composition—which is perceptually dominant and resistant to alterations induced by operations like JPEG compression, Gaussian blurring, or resizing. In contrast, high-frequency subbands (LHLH, HLHL, HHHH) capture fine-grained details and textures that are both visually sensitive and easily degraded by such post-processing steps. By confining adversarial perturbations to the low-frequency domain, we ensure that the introduced modifications are structurally coherent with the original image, thereby enhancing imperceptibility, while simultaneously maintaining robustness against a wide range of practical image manipulations that typically preserve low-frequency components. This strategic focus aligns with the dual objective of our framework: to achieve strong defensive efficacy without compromising visual fidelity.

Denoising Diffusion Probabilistic Models (DDPM):[30] DDPM are a type of generative model that generate images by progressively denoising an initial Gaussian noise vector. The DDPM framework comprises two main processes: a forward and a denoising. The fundamental idea behind the forward process of the diffusion model is to progressively add noise, ultimately transforming the image data into pure noise. Given an image x_0, it is blended with noise, and at the t-th step of this process, the image is generated through an additive noise operation:

$$x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1-\alpha_t} \cdot \epsilon_t \quad (2)$$

where $\epsilon_t \sim \mathcal{N}(0,1)$ denotes noise sampled from a standard normal distribution, and α_t is a time-dependent decay factor. In this way, as t increases, the image gradually amasses more noise.

The denoising process of DDPM begins with a noisy image x_t and iteratively reduces the noise until the original image x_0 is recovered. For DDPM, the reverse process can be formulated as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z \quad (3)$$

where $\epsilon_\theta(x_t, t)$ denotes the noise predicted by the model, σ_t is a hyperparameter (typically dependent on time t in DDPM), and z denotes the random noise sampled from a standard normal distribution. By assigning $\sigma_t = 0$, we obtain a deterministic sampling process, which forms the basis of DDIM(Denoising Diffusion Implicit Models)[31].

To invert a latent representation of an image, it is necessary to ensure that the sampling process is deterministic, which is precisely what DDIM achieves. Contrary to the comprehensive denoising procedure in DDPM, DDIM employs a non-Markovian approach for the inversion process, which is encapsulated by the following formula:

$$x_t = \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_{t-1}}}x_{t-1} + \sqrt{\alpha_t}\left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1}\right) \epsilon_\theta(x_{t-1}, t-1) \quad (4)$$

Our approach leverages the popular and openly accessible Stable Diffusion Model [11], wherein the forward diffusion process is implemented on the latent image encoding $z = E(x)$. Following the completion of the reverse diffusion process, the image is reconstructed through the image decoder, resulting in $x = D(z)$.

3.3 Main Framework

Fig. 2 illustrates our three-stage deepfake proactive defense framework, DiffDefend, comprising inversion, optimization, and reconstruction stages. We use the widely-used open-source Stable Diffusion model [11], which has undergone extensive training on a vast text-image dataset. The objective of proactive deepfake defense is to disrupt the deepfake model by introducing perturbations to the image, which can be regarded as a unique form of image manipulation. Drawing inspiration from recent diffusion-based editing methods [26, 27, 32], our framework utilizes DDIM inversion techniques.

Inversion. During the inversion stage, the input image x_{ori} is initially encoded into a primary latent code z_0 using the encoder E of an autoencoder. Following this, a reverse deterministic sampling procedure that spans t steps is employed to revert the initial latent code z_0 to the final latent code z_t, which can reconstruct the original image x_{ori} with exceptional fidelity. The entire process can be mathematically represented by the following formula:

$$z_0 = E(x_{\text{ori}}) \quad (5)$$

$$z_t = \text{Inverse}\,(z_{t-1}) = \underbrace{\text{Inverse} \circ \cdots \circ \text{Inverse}}_{t}\,(z_0) \quad (6)$$

where Inverse$(\cdot)$ represents the DDIM Inversion operation.

Optimization. In the optimization stage, the latent code z_t derived from the inversion stage can faithfully reconstruct the original image. However, the reconstructed image from z_t lacks adversarial properties, rendering it ineffective for attacking the deepfake model. Consequently, it is necessary to adversarially optimize z_t to obtain the adversarial latent code z_t'.

Most real-world image processing operations, such as JPEG compression and filtering, tend to suppress high-frequency signals while preserving low-frequency information. Therefore, to enhance robustness against such common transformations and preserve visual fidelity, we apply a discrete wavelet transform (DWT) to decompose the latent representation into frequency subbands and inject adversarial perturbations specifically into the low-frequency subband. Given the latent code $z_t \in \mathbb{R}^{C\times H\times W}$ obtained after DDIM inversion, we first perform a single-level 2D DWT on each channel to decompose it into four frequency subbands:

$$LL, LH, HL, HH = DWT(z_t) \quad (7)$$

where LL denotes the low-frequency (approximation) subband that captures the most structurally stable components of the latent representation, and LH, HL, HH respectively represent horizontal, vertical, and diagonal high-frequency subbands.

Next, we inject an adversarial perturbation δ_{LL} only into the low-frequency subband:

$$LL' = LL + \delta_{LL} \quad (8)$$

This operation preserves the original high-frequency content while manipulating only the structurally stable components of the latent representation.

Finally, we reconstruct the adversarial latent code z_t' by applying the inverse wavelet transform to the modified low-frequency subband LL' and the untouched high-frequency subbands:

$$z_t' = IDWT(LL', LH, HL, HH) \quad (9)$$

This strategy ensures that the adversarial perturbation is embedded in the most robust and perceptually stable region

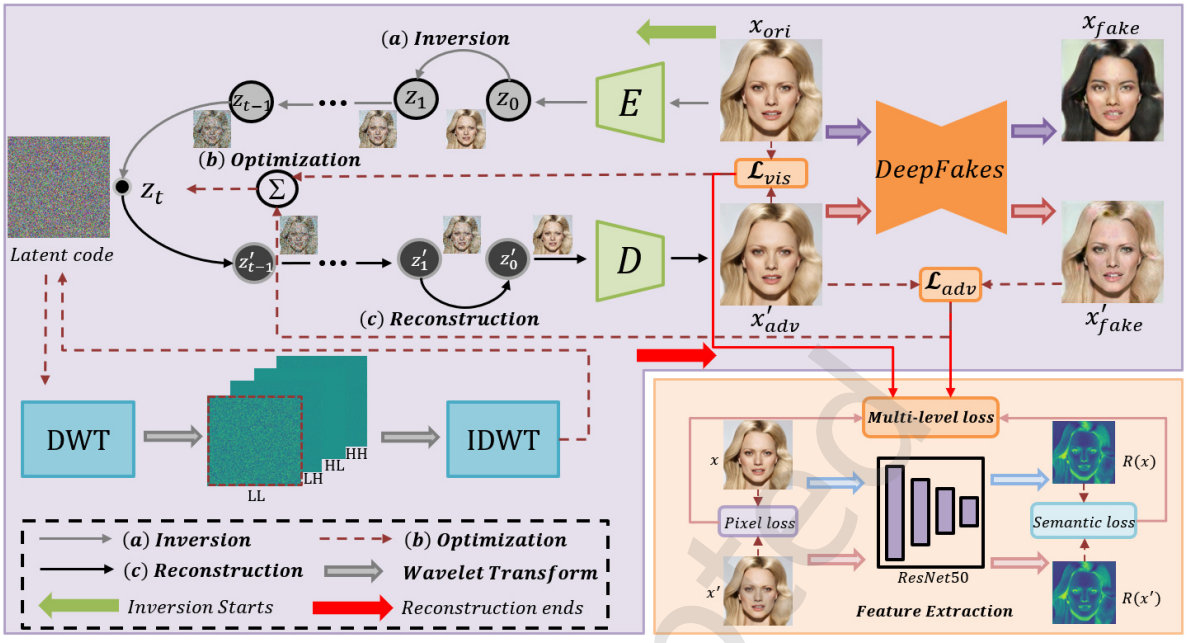


Fig. 2 The proposed DiffDefend framework. It comprises three stages: inversion, optimization, and reconstruction. It utilizes Stable Diffusion and DDIM inversion techniques to map clean face images into latent codes within the latent space. The low-frequency subband of these latent codes undergo adversarial optimization, guided by a meticulously crafted multi-level visual loss and adversarial loss. Ultimately, the diffusion model reconstructs adversarial images that efficiently safeguard against deepfake models while maintaining exceptional visual quality.

of the latent space, making it more resilient to real-world post-processing such as compression and filtering.

Earlier approaches often depended exclusively on pixel-level losses to gauge image disparities, thereby somewhat neglecting structural, textural, and perceptual coherence. Thus, to ensure that the adversarial latent code z_t' generates an adversarial image x'_{adv} that not only exhibits robust adversarial attributes but also upholds superior visual quality, we have devised a multi-level visual loss $\mathcal{L}_{\text{vis}}$ that integrates both pixel-level and semantic-level assessments, in conjunction with an adversarial loss $\mathcal{L}_{\text{adv}}$. The multi-level loss measures the similarity between input image pairs at both the pixel and semantic levels, and their joint constraints ensure the image pairs remain close in distance. To extract image features, we utilize a pre-trained ResNet-50 network [33], and we measure feature discrepancies between images using cosine distance. Meanwhile, pixel differences are quantified using mean squared error (MSE). The mathematical formulations for these two loss functions are provided below:

$$\begin{aligned}\mathcal{L}_{\text{vis}} = \alpha\,&\text{MSE}\left(x_{\text{ori}}, x'_{\text{adv}}\right)\\ &+ (1-\alpha)\ \text{Cos}\left(R\left(x_{\text{ori}}\right), R\left(x'_{\text{adv}}\right)\right)\end{aligned} \tag{10}$$

$$\begin{aligned}\mathcal{L}_{\text{adv}} = \beta\,&\text{MSE}\left(x'_{\text{adv}}, F\left(x'_{\text{adv}}\right)\right)\\ &+ (1-\beta)\ \text{Cos}\left(R\left(x'_{\text{adv}}\right), R\left(F\left(x'_{\text{adv}}\right)\right)\right)\end{aligned} \tag{11}$$

where α, β are utilized to regulate the weights of the pixel-level loss and semantic-level loss, respectively, R represents the feature extraction network, and F signifies the deepfake model. The ultimate objective of the optimization phase is to guarantee that the adversarial latent code z_t', after optimization, generates an adversarial image x'_{adv} that remains visual consistency with the original image while maximizing disruption to the deepfake model. This ensures that the generated fake image x'_{fake} closely resembles the original, encapsulating the comprehensive strategy of our optimization process, which can be formulated as:

$$\arg\min_{\delta_{LL}} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vis}} + \lambda \mathcal{L}_{\text{adv}} \tag{12}$$

where λ is a hyperparameter used to adjust the weights between different loss functions.

Reconstruction. In the reconstruction stage, the adversarial latent code z_t' acquired from the optimization stage, undergoes a t steps denoising process, culminating in the adversarial initial latent code z_0', which corresponds to the original initial latent code z_0. Following this, the decoder D of an autoencoder is employed to transform the latent code into the adversarial image x'_{adv}:

$$z_0' = \underbrace{\text{Denoise} \circ \cdots \circ \text{Denoise}}_{t}\left(z_t'\right) \tag{13}$$

$$x'_{\text{adv}} = D\left(z_0'\right) \tag{14}$$

The adversarial image ultimately obtained through these

Algorithm 1 DiffDefend Framework

Input: Original image x_{ori}, Encoder E, Decoder D, Feature extraction network R, DeepFake model F, Inversion and Denoising steps t, Iteration number T, Hyperparameters λ, Learning rate η.

Output: Adversarial image x'_{adv}

Inversion Stage:

$z_0 \leftarrow E(x_{\text{ori}})$;

$z_t \leftarrow \text{Inverse}(z_0)$ via t steps of DDIM Inversion;

Optimization Stage:

$LL, LH, HL, HH \leftarrow \text{DWT}(z_t)$;

Initialize $\delta_{LL} \leftarrow 0$;

for $i = 1$ **to** T **do**

$LL' \leftarrow LL + \delta_{LL}$;

$z'_t \leftarrow \text{IDWT}(LL', LH, HL, HH)$;

Compute visual loss $\mathcal{L}_{\text{vis}}$ with Eq. (10);

Compute adversarial loss $\mathcal{L}_{\text{adv}}$ with Eq. (11);

$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vis}} + \lambda \mathcal{L}_{\text{adv}}$;

$\delta_{LL} \leftarrow \delta_{LL} - \eta \nabla_{\delta_{LL}} \mathcal{L}_{\text{total}}$;

end for

Reconstruction Stage:

$z'_0 \leftarrow \text{Denoise}(z'_t)$ via t steps of DDIM Denoising;

$x'_{\text{adv}} \leftarrow D(z'_0)$;

return x'_{adv}

three stages not only effectively launches attacks on DeepFake models, enabling proactive defense, but also retains high visual quality, achieving a harmonious balance between robust defense capabilities and visual fidelity. The detailed procedures of the DiffDefend framework are outlined in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Dataset: Drawing inspiration from previous methods [21, 22], we utilize the CelebA-HQ [34] and FFHQ [35] dataset to evaluate the efficacy of our defense framework. CelebA-HQ contains 30,000 high-resolution face images at 1024×1024 resolution, while FFHQ includes 70,000 high-resolution images of faces. For our experimental configuration, we initially resize these images to 256×256 pixels and subsequently select 1,000 facial images at random, ensuring they belong to diverse identities, for the purpose of assessment.

Implementation Details: In the experiments, we evaluate the performance of the proposed defense framework on two distinct DeepFake tasks: attribute editing and face reenactment. To this end, we select four representative DeepFake models: GAN-based StarGAN [12] and diffusion-based DiffAE [13] for attribute editing, as well as GAN-based GANimation [14] and diffusion-based Face-Adapter [15] for face reenactment. Specifically, StarGAN edits hair color across three categories (black, blonde, brown), while DiffAE modifies gender and age. For reenactment, GANimation and Face-Adapter each control two facial action units to manipulate expressions.

We utilize DDIM as the sampler for Stable Diffusion, configuring it with 50 sampling steps. To optimize the latent code z_t, we employ the Adam optimizer with a learning rate of 5e-2 and set the number of iterations to 100. The hyperparameters α, β, and λ are assigned values of 0.5, 0.5, and 10, respectively. All experimental procedures are performed on a single NVIDIA GeForce RTX 3090 GPU.

Evaluation Metrics: To thoroughly evaluate the defense capabilities of our framework against various types of DeepFake attacks, we employ a range of evaluation metrics. For facial attribute editing, we utilize a facial recognition model [36] to extract facial features from both the input and output of each attribute editing operation, and then calculate the cosine similarity between the extracted feature vectors. Furthermore, we assess the effectiveness of the defense by computing the mean squared error (MSE) between the input and output images generated by StarGAN and DiffAE. In the case of face reenactment, we determine the Euclidean distance between the feature vectors extracted by a facial landmark detection model [37] from each facial expression operation image. Additionally, we evaluate the defense performance by calculating the mean absolute error (MAE) between the input and output images of GANimation and Face-Adapter. To assess the visual quality of the protected facial images, we use several metrics including the Structural Similarity Index Measure (SSIM) [38], Peak Signal-to-Noise Ratio (PSNR) [39], Learned Perceptual Image Patch Similarity (LPIPS) [40], and Mean Squared Error (MSE).

Comparisons: Given that latent space reconstruction as a defense strategy against DeepFakes is an emerging area with scarce research, we primarily benchmark our method against LAE [21] and LOFT [22], both of which leverage GAN latent spaces. As a reference for comparison, we also include several representative methods that introduce perturbations in the pixel space, namely AdvNoise [8], FOUND [19], and KGAD [16].

4.2 Defense Effectiveness Evaluation

In this section, we evaluate the effectiveness of the proposed DiffDefend framework against two different types of Deep-

Table 1 The quantification results of different defense methods against StarGAN and DiffAE on CelebA-HQ and FFHQ. The best results in each group are highlighted in bold.

Dataset	Method	StarGAN				DiffAE		
		Black↑	Blonde↑	Brown↑	MSE↓	Gender↑	Age↑	MSE↓
CelebA-HQ	Original	0.756	0.740	0.775	0.045	0.524	0.573	0.063
	AdvNoise [8]	0.773	0.759	0.792	0.031	0.592	0.635	0.042
	FOUND [19]	0.778	0.764	0.796	0.029	0.607	0.648	0.038
	KGAD [16]	0.883	0.871	0.902	0.016	0.721	0.757	0.024
	LAE [21]	0.770	0.745	0.794	0.027	0.713	0.736	0.025
	LOFT [22]	0.926	0.925	0.928	**0.003**	0.843	0.826	0.021
	Ours	**0.948**	**0.949**	**0.956**	0.006	**0.874**	**0.883**	**0.016**
FFHQ	Original	0.738	0.742	0.757	0.043	0.516	0.568	0.065
	AdvNoise [8]	0.759	0.761	0.775	0.035	0.615	0.645	0.046
	FOUND [19]	0.768	0.765	0.780	0.031	0.624	0.653	0.041
	KGAD [16]	0.876	0.871	0.885	0.018	0.732	0.761	0.029
	LAE [21]	0.774	0.758	0.767	0.026	0.698	0.710	0.032
	LOFT [22]	0.921	0.923	0.919	0.007	0.823	0.832	0.024
	Ours	**0.934**	**0.936**	**0.940**	**0.007**	**0.872**	**0.876**	**0.013**

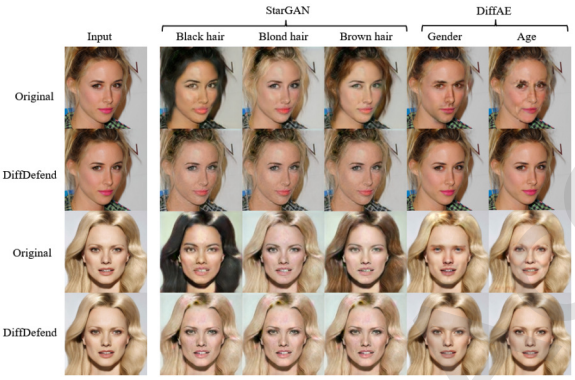


Fig. 3 Visual examples of DiffDefend applied to StarGAN and DiffAE attribute manipulation. The first and third rows show attribute manipulations on the original images, including alterations in hair color (black, blond, brown), gender, and age. The second and fourth rows show attribute manipulations on face images protected by the DiffDefend method. It is evident that DiffDefend successfully protects the face images from manipulation, preserving their original characteristics.

Fakes. Defense performance is quantified by measuring the discrepancy between the original face image and the manipulated output produced by the deepfake model when applied to our protected image. A smaller discrepancy indicates stronger defense capability, as it reflects the attacker's reduced success in altering facial content.

In the facial attribute editing task (Fig. 3), both StarGAN and DiffAE successfully manipulated attributes such as hair color, age, and gender in original facial images. However, when applied to images protected by our method, these models failed to produce the intended attribute changes. This result confirms that our approach effectively preserves facial attributes against malicious edits while maintaining high visual similarity to the original images. The quantitative

Table 2 The quantification results of different defense methods on CelebA-HQ and FFHQ against GANimation and Face-Adapter. The best results in each group are highlighted in bold.

Dataset	Method	GANimation			Face-Adapter		
		AU1↓	AU2↓	MAE↓	AU3↓	AU4↓	MAE↓
CelebA-HQ	Original	3.82	3.67	0.085	4.57	4.67	0.092
	AdvNoise [8]	1.65	1.58	0.016	1.75	1.68	0.035
	FOUND [19]	1.59	1.52	0.015	1.70	1.62	0.033
	KGAD [16]	1.57	1.50	0.014	1.68	1.59	0.032
	LAE [21]	2.58	2.51	0.020	3.09	2.74	0.038
	LOFT [22]	1.63	1.60	0.015	1.69	1.53	0.032
	Ours	**1.42**	**1.37**	**0.009**	**1.47**	**1.41**	**0.026**
FFHQ	Original	3.95	3.82	0.087	4.59	4.71	0.087
	AdvNoise [8]	1.78	1.62	0.018	1.79	1.72	0.033
	FOUND [19]	1.74	1.58	0.017	1.73	1.65	0.032
	KGAD [16]	1.69	1.55	0.016	1.71	1.61	0.031
	LAE [21]	2.63	2.58	0.019	3.21	2.68	0.036
	LOFT [22]	1.67	1.59	0.016	1.71	1.57	0.033
	Ours	**1.49**	**1.39**	**0.009**	**1.46**	**1.33**	**0.027**

findings are presented in Table 1. The first row provides reference results obtained by feeding original images directly into StarGAN and DiffAE. The table lists the cosine similarity between the feature vectors extracted from the input and output images for each attribute editing operation (black hair, blond hair, brown hair, gender, and age). A higher similarity score indicates that the input and output images from StarGAN and DiffAE are closer in terms of their features, reflecting a stronger defense capability. Furthermore, we measure the defense effectiveness by computing the mean squared error (MSE) between the input and output images for each of the five attribute editing operations. A lower MSE indicates superior defense performance. As shown in Table 1, our method exhibits superior defense ability across both StarGAN and DiffAE models on the CelebA-HQ and FFHQ datasets,

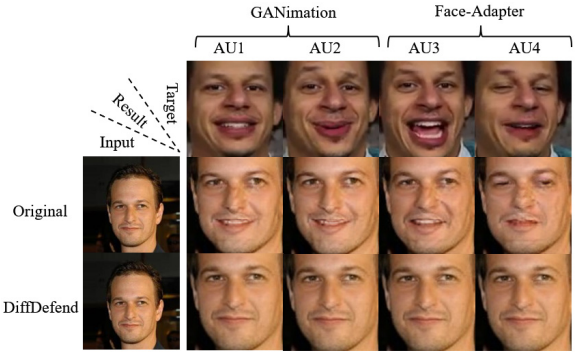


Fig. 4 Visual examples of DiffDefend defending against GANimation and Face-Adapter. The first row displays four target expressions. The second row illustrates the expression manipulations applied to the original images. The third row presents the expression manipulations on face images protected by DiffDefend. It is clear that DiffDefend effectively prevents the face images from undergoing expression changes, maintaining the integrity of the original facial features.

consistently obtaining the highest cosine similarity scores across all five attribute operations compared to baseline approaches. Although the MSE for StarGAN is slightly higher than the LOFT method by 0.003, our method still exhibits the best overall defense performance.

For the facial reenactment task, we evaluated our method using four images of the target face exhibiting different facial expressions, labeled as AU1 to AU4. As illustrated in Fig. 4, GANimation and Face-Adapter successfully edits the expression of the original image, but its performance significantly drops when attempting to edit the protected face images generated by our method, indicating that GANimation and Face-Adapter is incapable of successfully altering the facial expressions of these images. The numerical results are presented in Table 2, which presents the Euclidean distance between the feature vectors extracted from the input and output images for each AU operation. A lower value indicates fewer changes. Additionally, since facial expression editing typically affects sparse regions of the face, Table 2 also shows the Mean Absolute Error (MAE) between the input and output images of the four AU operations in GANimation and Face-Adapter. A lower MAE indicates better defense performance. As evident from Table 2, our method surpasses the comparison methods for all four AU operations across two datasets, demonstrating robust defense capabilities.

4.3 Visual Quality Evaluation

The goal of proactive defense against deepfake not only requires the protected facial image to resist deepfake models but also demands that the protected facial image visually resembles the original image. Fig. 5 presents visualizations

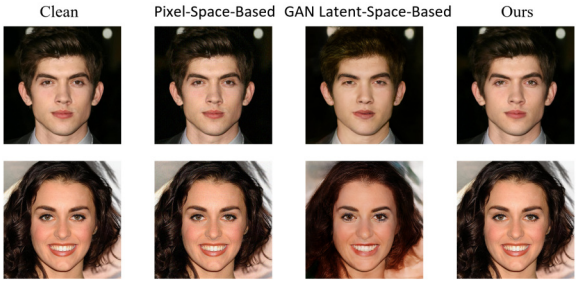


Fig. 5 The visualization of different proactive defense methods applied to protected face images. The first column displays the original clean face images. The second column shows results from a pixel-space-based method, which introduces visible noise artifacts. The third column presents outputs from a GAN latent-space-based approach, exhibiting slight but noticeable facial distortions. The fourth column illustrates our method, which neither introduces extra noise nor alters the face, providing a faithful reconstruction of the original face. Please zoom in for a better view.

Table 3 The quantification results of different defense methods on the visual quality for attribute editing and face reenactment. The best results are highlighted in bold.

Task	Method	SSIM↑	PSNR↑	LPIPS↓	MSE↓
Attribute Editing	AdvNoise [8]	0.8152	27.83	0.2713	0.0165
	FOUND [19]	0.8281	28.42	0.2450	0.0138
	KGAD [16]	0.8324	28.75	0.2331	0.0127
	LAE [21]	0.6200	21.19	0.3570	0.0360
	LOFT [22]	0.7860	28.26	0.1680	0.0070
	Ours	**0.8524**	**31.04**	**0.0965**	**0.0009**
Face Reenact -ment	AdvNoise [8]	0.8261	27.94	0.2805	0.0158
	FOUND [19]	0.8390	28.60	0.2510	0.0132
	KGAD [16]	0.8433	28.93	0.2394	0.0120
	LAE [21]	0.6360	22.91	0.2670	0.0250
	LOFT [22]	0.7720	27.70	0.1620	0.0080
	Ours	**0.9027**	**31.59**	**0.0690**	**0.0010**

of protected images generated through various methods. Pixel-space-based methods directly apply perturbations to the image surface, whereas GAN latent-space-based methods introduce perturbations within the GAN latent space. As Fig. 5 illustrates, pixel-space methods result in a noticeable layer of noise on the surface of the protected facial image. Although GAN latent-space methods avoid additional noise, they slightly alter facial features during reconstruction, leading to a loss of fidelity in the protected image compared to the original. Conversely, our method achieves unparalleled reconstruction quality: the protected image remains free of extraneous noise and retains a high degree of fidelity to the original, thereby ensuring both visual coherence and robustness.

Our proposed method, like LAE [21] and LOFT [22], involves latent space reconstruction of images. However, while LAE and LOFT utilize the latent space of GANs, our approach exploits the latent space of diffusion models.

Despite this distinction, all three methods aim to achieve faithful and realistic reconstruction of the original image, ensuring perceptual similarity between the reconstructed and input images. To quantify the quality of the reconstructed images, we utilize four visual quality metrics: SSIM [38], PSNR [39], LPIPS [40], and MSE.

As shown in Table 3, our method outperforms the comparison methods across all four visual quality evaluation metrics for both types of DeepFake tasks. This highlights the superior imperceptibility of our defense method, as well as the enhanced visual quality of the images generated by the diffusion model. Specifically, these images exhibit greater clarity, realism, and perceptual fidelity compared to those generated by GANs. These results underscore the efficacy of our approach in maintaining high visual standards while achieving powerful defense against DeepFake generation.

Table 4 The quantitative results of DiffDefend under different image processing operations.

Processing	Parameters	MSE↓	MAE↓
None	–	0.006	0.009
JPEG Compression	30	0.022	0.030
	50	0.018	0.025
	70	0.013	0.018
	90	0.011	0.013
Gaussian Blur	0.3	0.017	0.027
	0.4	0.021	0.028
	0.5	0.025	0.031
	0.6	0.027	0.033
Gaussian Noise	0.01	0.009	0.013
	0.03	0.015	0.022
	0.05	0.021	0.029
	0.07	0.026	0.035
Resizing	3/4	0.014	0.020
	5/4	0.011	0.016
	3/2	0.008	0.012
	7/4	0.013	0.016

4.4 Robustness to Image Processing Operations

In real-world deployment, facial images are frequently subject to various post-processing operations, making robustness against such transformations essential for practical viability. To evaluate this aspect, we apply a series of common image processing techniques-including JPEG compression, Gaussian blur, Gaussian noise, and resizing-to the protected images generated by DiffDefend. As summarized in Table 4, our method demonstrates strong robustness across these transformations. Although defensive performance experiences a slight degradation under stronger processing intensities, the framework retains sufficient capability to effectively disrupt deepfake generation.

This robustness primarily benefits from our proposed strategy of applying discrete wavelet transform (DWT) to the latent code and injecting adversarial perturbations into its low-frequency subband. Since most practical image transformations predominantly affect high-frequency content while preserving low-frequency structures, perturbations embedded in this domain demonstrate greater survivability under various processing conditions, thereby ensuring both effective defense and high visual fidelity.

Table 5 Transferability for defending against different deepfake models (MSE↓)

Test \ Train	Original (StarGAN)	DiffDefend (StarGAN)
StarGAN	0.045	0.006
AttGAN	0.063	0.013
AGGAN	0.042	0.029
HiSD	0.051	0.034

4.5 Transferability Evaluation

The transferability of protective adversarial perturbations is crucial for real-world deployment, especially in black-box scenarios where only a single deepfake generator is accessible to the defender. In such settings, the generalization capability of adversarial examples across diverse manipulation models becomes essential. To evaluate this property, we randomly select 300 images from the CelebA-HQ dataset and generate adversarial perturbations against StarGAN using our method. These protected images are subsequently applied to other unseen manipulation models-including AttGAN, AGGAN, and HiSD-for transferability assessment.

As shown in Table 5, our method achieves strong defensive performance against StarGAN, reducing the MSE from 0.045 to 0.006, while also demonstrating notable transferability to unseen models. Specifically, the MSE decreases from 0.063 to 0.013 on AttGAN, from 0.042 to 0.029 on AGGAN, and from 0.051 to 0.034 on HisD-even though no dedicated transferability module is incorporated. These results indicate that our latent-space perturbation strategy generalizes effectively across diverse deepfake generators, offering reliable protection even in challenging black-box scenarios.

4.6 Ablation Study

Different loss functions are crucial for both image reconstruction quality and defense performance. Table 6 presents the ablation study results on attribute editing and face reenactment task, evaluating the impact of combining pixel loss and semantic loss on defense performance and image quality. The findings reveal that utilizing only pixel loss yields

Table 6 Ablation Study on the Impact of Pixel Loss and Semantic Loss. The best results are highlighted in bold.

Pixel loss	Semantic loss	Attribute Editing				Face Reenactment			
		MSE↓	SSIM↑	PSNR↑	LPIPS↓	MAE↓	SSIM↑	PSNR↑	LPIPS↓
✓	✗	0.012	0.8354	28.31	0.1253	0.019	0.8594	29.86	0.1049
✗	✓	0.018	0.8026	28.04	0.1329	0.023	0.8433	30.03	0.1295
✓	✓	**0.006**	**0.8524**	**31.04**	**0.0965**	**0.009**	**0.9027**	**31.59**	**0.0690**

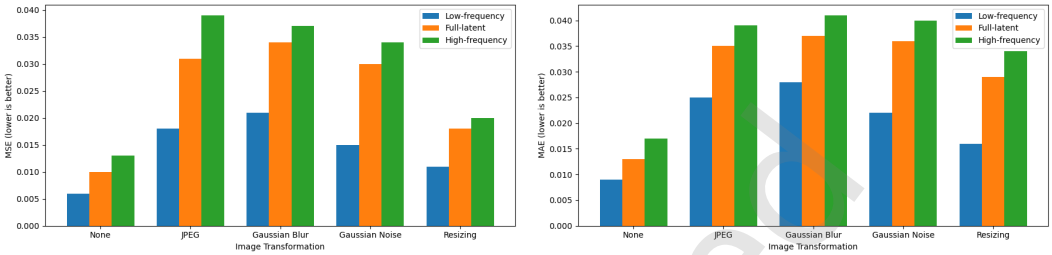


Fig. 6 Robustness comparison under different image transformations using three perturbation strategies: low-frequency perturbation, full-latent perturbation, and high-frequency perturbation. (a) MSE results. (b) MAE results. Lower values indicate better defense performance under transformations.

Table 7 Sensitivity Analysis of α and β on Defense Performance.

α	β	SSIM↑	PSNR↑	LPIPS↓	MSE↓	MAE↓
0.1	0.9	0.8263	30.32	**0.0927**	0.009	0.012
0.3	0.7	0.8375	29.79	0.0931	0.010	0.012
0.5	0.5	**0.8524**	**31.04**	0.0965	**0.006**	**0.009**
0.7	0.3	0.8524	30.92	0.1192	0.012	0.015
0.9	0.1	0.8521	30.86	0.1246	0.012	0.014

relatively high SSIM and PSNR values; however, the LPIPS metric remains suboptimal, and the defense performance is constrained. When only feature loss is applied, all metrics degrade significantly, particularly PSNR and LPIPS, suggesting that relying solely on semantic loss is insufficient for effectively defending against attacks. In contrast, the combined use of pixel loss and semantic loss substantially improves both defense performance and image quality, with particularly noticeable enhancements in LPIPS and PSNR. Overall, the results clearly demonstrate that the synergistic combination of pixel loss and semantic loss is essential for achieving optimal defense performance and preserving high visual quality. By simultaneously addressing low-level pixel differences and high-level perceptual features, our method effectively balances fidelity and robustness, leading to superior performance against a variety of deepfake models. This finding further underscores the importance of multi-level optimization in enhancing proactive defense strategies.

To investigate the impact of the weighting parameters α and β on defense performance and visual quality, we conducted a sensitivity analysis, with results shown in Table 7. Here, α and β control the balance between the pixel-level and semantic-level components in the visual and adversarial losses, respectively. As observed, when $\alpha = 0.5$ and $\beta = 0.5$, the framework achieves the best overall performance: SSIM reaches 0.8524, PSNR is 31.04, and both MSE and MAE attain their minimum values, indicating that high visual fidelity is maintained while effectively defending against deepfake attacks. Lower α values tend to degrade pixel-level fidelity, whereas higher α values may underemphasize semantic-level constraints, potentially reducing adversarial effectiveness. Overall, this experiment demonstrates that appropriately setting α and β achieves a favorable trade-off between visual quality and defense capability, providing empirical guidance for the multi-level loss design.

To further verify the effectiveness of the proposed perturbation strategy, we compare three strategies: low-frequency, full-latent, and high-frequency perturbations under various image transformations, including JPEG compression, Gaussian blur, Gaussian noise, and resizing. As shown in Fig. 6, the low-frequency perturbation consistently achieves the lowest MSE and MAE values, demonstrating its superiority in preserving defense performance under both lossy compression and geometric distortions. In contrast, the high-frequency perturbation yields the highest errors, suggesting that modifications in the high-frequency domain are highly sensitive to common transformations. The full-latent perturbation exhibits intermediate performance but still lags behind the low-frequency approach. These results validate the necessity of restricting perturbations to the low-frequency latent components, which not only enhance robustness against real-world transformations but also maintain imperceptible visual quality.

5 Conclusion

In this paper, we propose DiffDefend, a novel three-stage proactive defense framework against deepfake attacks. This paper enhances defense stealth and preserves visual fidelity by injecting adversarially optimized perturbations into the low-frequency components of diffusion latent codes. Unlike pixel-space perturbation methods, our framework leverages the strong generative prior of diffusion models, where operating in the low-frequency subspace ensures robustness to common image transformations while minimizing perceptible artifacts. Extensive experimental results demonstrate that DiffDefend achieves strong performance across two representative deepfake tasks-facial attribute editing and face reenactment-effectively mitigating attacks while maintaining high visual consistency with the original images, demonstrating both robust defensive capability and broad practical applicability.

While the current framework has demonstrated excellent performance, we recognize that deepfake attack and defense constitute an ongoing and evolving technological contest. To ensure the long-term effectiveness of the defense system, this paper further explores the adaptive evolution potential of the DiffDefend framework. In the future, the framework could employ online learning or incremental learning mechanisms to dynamically fine-tune the perturbation generation module using newly emerging adversarial samples, thereby countering continuously advancing generative models. Concurrently, it would be valuable to explore content-aware dynamic defense strategies that adaptively adjust perturbation patterns based on the semantic context of the image, achieving an intelligent balance between defensive strength and visual quality. Furthermore, researching how to effectively integrate this framework with other lightweight defense paradigms—such as digital watermarking and feature-space disruption—to construct a multimodal, redundant, and collaborative defense system represents a crucial direction for enhancing its generalization capability against unknown attacks. These adaptive strategies aim to transform DiffDefend from an outstanding static defense solution into an intelligent defense ecosystem capable of continuous learning, dynamic adjustment, and multi-strategy coordination, thereby maintaining enduring and robust protective value in a rapidly changing threat landscape.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grants U22B2062 and U23B2023; by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2023B1515020041.

Reference

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, *Advances in neural information processing systems*, vol. 27, 2014.

[2] S. R. Arashloo, Unseen face presentation attack detection using sparse multiple kernel Fisher null-space, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4084–4095, 2020.

[3] H. Zhang, B. Chen, J. Wang, and G. Zhao, A local perturbation generation method for GAN-generated face anti-forensics, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 661–676, 2022.

[4] L. Qin, M. Wang, C. Deng, K. Wang, X. Chen, J. Hu, and W. Deng, SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation, *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[5] J. Wu, B. Zhang, Z. Li, G. Pang, Z. Teng, and J. Fan, Interactive two-stream network across modalities for deepfake detection, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6418–6430, 2023.

[6] F. Gu, Y. Dai, J. Fei, and X. Chen, Deepfake detection and localisation based on illumination inconsistency, *International Journal of Autonomous and Adaptive Communications Systems*, vol. 17, no. 4, pp. 352–368, 2024.

[7] Y. Yu, X. Liu, R. Ni, S. Yang, Y. Zhao, and A. C. Kot, Pvass-mdd: predictive visual-audio alignment self-supervision for multimodal deepfake detection, *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[8] N. Ruiz, S. A. Bargal, and S. 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*, 2020, pp. 236–251.

[9] Q. Huang, J. Zhang, W. Zhou, W. Zhang, and N. Yu, Initiative defense against facial manipulation, in *Proc. AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1619–1627.

[10] C.-Y. Yeh, H.-W. Chen, H.-H. Shuai, D.-N. Yang, and M.-S. Chen, Attack as the best defense: Nullifying image-to-image translation gans via limit-aware adversarial attack, in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16188–16197.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[12] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, Stargan: Unified generative adversarial networks for multi-

domain image-to-image translation, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[13] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, Diffusion autoencoders: Toward a meaningful and decodable representation, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10619–10629.

[14] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, Ganimation: Anatomically-aware facial animation from a single image, in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 818–833.

[15] Y. Han, J. Zhu, K. He, X. Chen, Y. Ge, W. Li, X. Li, J. Zhang, C. Wang, and Y. Liu, Face-adapter for pre-trained diffusion models with fine-grained id and attribute control, in *European Conference on Computer Vision*, 2024, pp. 20–36.

[16] D. Zhou, Z. Su, D. Liu, T. Liu, N. Wang, and X. Gao, A Knowledge-guided Adversarial Defense for Resisting Malicious Visual Manipulation, *IEEE Transactions on Dependable and Secure Computing*, 2025.

[17] J. Dong, Y. Wang, J. Lai, and X. Xie, Restricted black-box adversarial attack against deepfake face swapping, *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2596–2608, 2023.

[18] H. Huang, Y. Wang, Z. Chen, Y. Zhang, Y. Li, Z. Tang, W. Chu, J. Chen, W. Lin, and K.-K. Ma, Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes, in *Proc. AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 989–997.

[19] L. Tang, D. Ye, Z. Lu, Y. Zhang, and C. Chen, Feature extraction matters more: An effective and efficient universal deepfake disruptor, *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 2, pp. 1–22, 2024.

[20] Z. Chen, L. Xie, S. Pang, Y. He, and B. Zhang, Magdr: Mask-guided detection and reconstruction for defending deepfakes, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9014–9023.

[21] Z. He, W. Wang, W. Guan, J. Dong, and T. Tan, Defeating deepfakes via adversarial visual reconstruction, in *Proc. 30th ACM International Conference on Multimedia*, 2022, pp. 2464–2472.

[22] S. Zeng, W. Wang, F. Huang, and Y. Fang, LOFT: Latent Space Optimization and Generator Fine-Tuning for Defending Against Deepfakes, in *Proc. ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4750–4754.

[23] W. Li, X. Yu, K. Zhou, Y. Song, Z. Lin, and J. Jia, SDM: Spatial diffusion model for large hole image inpainting, *arXiv preprint arXiv:2212.02963*, vol. 1, 2022.

[24] S. Xie, Z. Zhang, Z. Lin, T. Hinz, and K. Zhang, Smartbrush: Text and shape guided object inpainting with diffusion model, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22428–22437.

[25] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, Image super-resolution via iterative refinement, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[26] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, Zero-shot image-to-image translation, in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[27] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, Null-text inversion for editing real images using guided diffusion models, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6038–6047.

[28] Xingxing Wei, Caixin Kang, Yinpeng Dong, Zhengyi Wang, Shouwei Ruan, Yubo Chen, and Hang Su, Real-World Adversarial Defense Against Patch Attacks Based on Diffusion Model,*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 12, pp. 11124-11140, 2025.

[29] Z. Lu, H. Sun, L. Lei, Y. Xu, Y. Sun, and G. Kuang, DiffDual-AD: Diffusion-Based Dual-Stage Adversarial Defense Framework in Remote Sensing With Denoiser Constraint, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1-20, Art no. 5622720, 2025.

[30] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[31] J. Song, C. Meng, and S. Ermon, Denoising diffusion implicit models, *arXiv preprint arXiv:2010.02502*, 2020.

[32] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, Diffedit: Diffusion-based semantic image editing with mask guidance, *arXiv preprint arXiv:2210.11427*, 2022.

[33] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[34] T. Karras, Progressive Growing of GANs for Improved Quality, Stability, and Variation, *arXiv preprint arXiv:1710.10196*, 2017.

[35] T. Karras, S. Laine, and T. Aila, A style-based generator architecture for generative adversarial networks, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[36] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[37] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, Face alignment across large poses: A 3d solution, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 146–155.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[39] A. Almohammad and G. Ghinea, Stego image quality and the

reliability of PSNR, in *Proc. 2010 2nd International Conference on Image Processing Theory, Tools and Applications*, 2010, pp. 215–220.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[41] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye, Implicit identity driven deepfake face swapping detection, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4490–4499.

[42] Y. Huang, F. Juefei-Xu, Q. Guo, Y. Liu, and G. Pu, Dodging deepfake detection via implicit spatial-domain notch filtering, *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[43] Vu Tuan Truong, Luan Ba Dang, and Long Bao Le, Attacks and Defenses for Generative Diffusion Models: A Comprehensive Survey,*ACM Computing Surveys*, vol. 57, no. 8, Article 216, pp. 1-44, 2025.

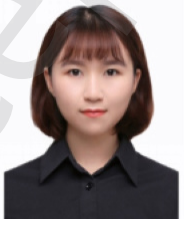
Author biography

Chengsheng Yuan received his PhD degrees at the School of Computer and software from Nanjing University of Information Science and Technology, China, in 2019. He is currently an associate professor with School of Computer Science, Nanjing University of Information Science and Technology, China. His current research interests include Multimedia Security, Artificial Intelligence Security and Information Hiding.

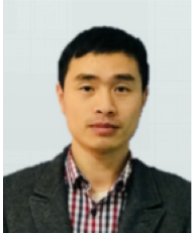
Youqiang Cao is currently a Master's candidate at the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China, and is affiliated with the Engineering Research Center of Digital Forensics, Ministry of Education. His research interests include computer vision, generative models, and information security.

Yuchen Wu is a Master's candidate at the School of Computer Science, Nanjing University of Information Science and Technology, and is affiliated with the Engineering Research Center of Digital Forensics, Ministry of Education, in Nanjing, China. Her research interests include deep learning, image processing, and privacy protection.

Fengjun Xiao is an assistant professor at the Zhejiang Informatization Development Institute, Hangzhou Dianzi University, Hangzhou, China. He received his Ph.D. from Beijing University of Aeronautics and Astronautics. His research interests include data governance, digital platform governance, and information security.



Xinting Li is an Associate Professor at the School of Foreign Languages, National University of Defense Technology, Nanjing, China. Her research interests include computational linguistics, cross-cultural communication, and technical translation.



Zhihua Xia is a Professor at the College of Cyber Security, Jinan University, Guangzhou, China. He received his Ph.D. in Computer Science and Technology from Hunan University. His research interests include adversarial machine learning, digital forensics, and computer vision.