Anti-Avatar: Protect Against Unauthorized 3D Head Avatar Generation via Dual-Space Divergence

Lingzhuang Meng[1], Mingwen Shao[2*], Xiang Lv[1], Mengyao Wu[1], Yuanjian Qiao[3], Jie Zhang[1],

[1]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software,
Qingdao Institute of Software, College of Computer Science and Technology,
China University of Petroleum (East China), China
[2] Artificial Intelligence Research Institute, Shenzhen University of Advanced Technology, China
[3] College of Computer Science (College of Software), Inner Mongolia University, China
lzhmeng1688@163.com, smw278@126.com, lvxiang1997@126.com, z24070063@s.upc.edu.cn,
yjqiao58@163.com, zjedu1225@126.com

Abstract

Head avatar generation is facilitated to construct high-fidelity 3D virtual personas from a single portrait, but it also raises the risk of unauthorized personal avatars generation. Recent 2D portrait protection methods actively prevent malicious image generation by perturbing the identity features. However, there are two key limitations when directly applied to prevent 3D head avatar generation: **1)** These methods neglect the inherent 3D geometric structure of portrait, thus failing to disrupt the modeling of 3D shapes or poses. **2)** They focus only on identity offset and are unable to interfere with the overall appearance, resulting in excessive preservation of facial characteristics. To overcomes these limitations, we propose a 3D defense framework termed **Anti-Avatar**, tailored to protect against unauthorized 3D head avatar generation from a single portrait. Specifically, Anti-Avatar consists of two key designs: Geometric Disruption and Perceptual Confusion. The former disrupts the precise reconstruction of 3D structure by interfering with the estimation of geometric parameters, thus affecting the structural accuracy of the 3D avatar. Collaboratively, the latter confuses image features by dispersing attention distribution, thereby hindering the effective perception of portrait appearance. Benefiting from the above dual-space divergence in geometry and perception, the avatars generated by our protected portraits exhibit substantial discrepancies from the originals. Extensive experiments show that our Anti-Avatar outperforms 2D methods in protection performance and effectively resists reconstruction and manipulation by state-of-the-art 3D head avatar generation methods.

Introduction

Recently, 3D head avatar generation (Wang et al. 2025c; Ren et al. 2025) has made remarkable progress, enabling the creation of photo-realistic 3D virtual representations and allowing for flexible driving. This advancement provides opportunities for various applications (Lyu et al. 2025; Jiang et al. 2025), such as virtual reality, gaming, and online communication. Nevertheless, the rapid progress also brings substantial privacy and security concerns, particularly the unauthorized reconstruction and manipulation of personal avatars, leading to identity theft and the spread of false information. (Huang, Wu, and Wang 2025; Gan et al. 2025).

Early approaches (Feng et al. 2021; Ma et al. 2023; Deng et al. 2024) in the field of 3D head avatar reconstruction primarily rely on 3D Morphable Model (3DMM) (Li et al. 2017), using parameter fitting to estimate facial parameters, such as shape, expression, pose, and texture, thereby constructing a complete 3D head structure through 3D mapping. To achieve more detailed and realistic avatar generation, some methods introduce Neural Radiance Fields (Gafni et al. 2021; Hong et al. 2022) and 3D Gaussian Splatting (3DGS) (Qian et al. 2024; Wang et al. 2025b; Xu et al. 2024) for rendering, achieving high-fidelity results and enabling flexible viewpoint control. Along this research trajectory, recent approaches (Chu et al. 2024; Chu and Harada 2024; He et al. 2025) combine 3D geometric structure with facial representation to drive a neural 3D reconstruction module, significantly improving the generalization ability of the task and achieving real-time reconstruction. Despite these significant advances, existing research rarely consider the risks posed by unauthorized avatar generation, raising serious concerns about privacy breaches and identity theft.

To protect portraits from malicious 2D editing and generation, existing methods (Shih et al. 2025; Choi et al. 2025) focus on actively interfering with the generation process of diffusion models by introducing adversarial perturbations into images. These methods can be broadly categorized into two categories. One type of methods (Liang et al. 2023; Zheng, Liang, and Wu 2025) focus on directly interfering with the denoising process of diffusion models, aiming to hinder the gradual image refinement during the generation process, thereby preventing high-quality or semantically consistent image synthesis The other type of methods manipulates the attention distribution (Jeon et al. 2025; Lo et al. 2024) or identity feature (Song et al. 2025b; Wang et al. 2025a) in diffusion model to hinder editing in specific regions or suppress the retention of identity-related information. However, the above 2D approaches exhibit two key limitations when applied to the defense against 3D head avatar generation. On the one hand, they disregard the facial geometric structure

*Corresponding author.

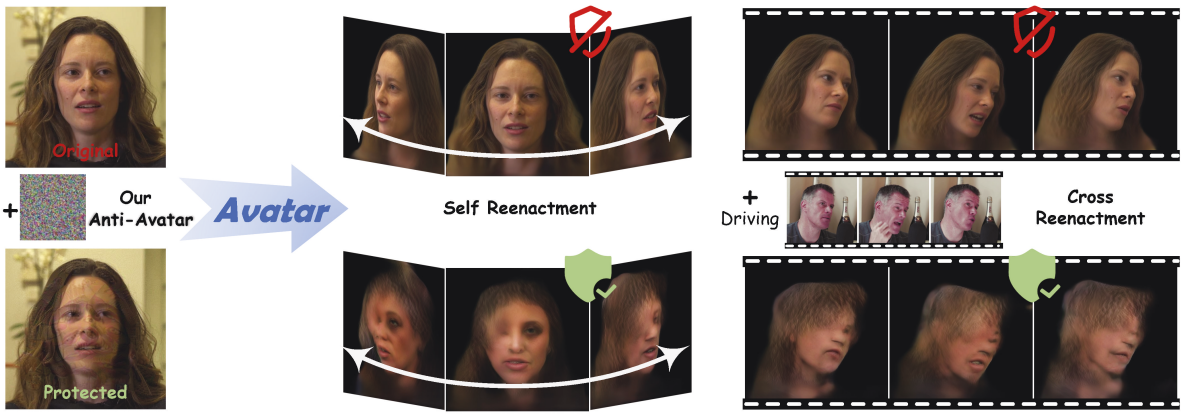

Figure 1: Comparison of 3D avatars reconstructed before and after protection. **Top row:** Unprotected portraits can be used to generate highly realistic and identity-consistent 3D virtual humans without authorization in both self-reenactment and cross-reenactment. **Bottom row:** Portraits protected by our Anti-Avatar fail to produce visually plausible or identity-preserving avatars, thereby effectively preventing unauthorized 3D head avatar generation.

inherent in portraits, thus failing to disrupt the 3D geometric shapes and poses during the generation process. On the other hand, they lack perturbation of facial appearance, leading to the overly preserved portrait features in the generated avatars and thereby limiting the protective effect.

To address the aforementioned limitations, we propose a novel 3D defense framework, known as Anti-Avatar, for preventing the unauthorized 3D head avatar generation from portraits, as shown in Figure 1. Specifically, our Anti-Avatar is built upon two complementary modules: Geometric Disruption and Perceptual Confusion. The former induces imprecise facial geometry by perturbing parameter estimation of 3D geometric in the FLAME (Li et al. 2017), thereby producing unnatural deviations in shape and pose of avatars. The latter confuses facial characteristics by dispersing self-attention distribution during feature extraction, thus deviating the extracted appearance features from the original portrait. This dual-space discrepancy design ensures that the generated avatars differ significantly from the original in both geometric structure and appearance perception, thereby effectively preventing unauthorized reconstruction and manipulation. Extensive experiments demonstrate that our Anti-Avatar outperforms 2D defense methods and offers more targeted protection against various state-of-the-art (SOTA) 3D head avatar generation schemes.

Our main contributions are summarized as follows:

- We propose Anti-Avatar, a novel defense framework against 3D avatar reconstruction. To the best of our knowledge, this is the first dedicated framework to protect portraits from malicious avatar generation.
- We design Geometric Disruption that perturbs the 3D parameter estimation of facial geometry to damage the reconstructed facial 3D structure.
- To prevent retention of the original appearance, we devise Perceptual Confusion that distracts attention during feature extraction for hindering perception of portraits.
- Experimental results show that our Anti-Avatar can effectively defend against unauthorized reconstruction and manipulation, and outperforms existing 2D defense methods in both qualitative and quantitative results.

Related Works

Single-image 3D Head Avatar Generation. 3D head avatar generation techniques have achieved notable progress and have been widely applied in fields such as virtual reality (Wang et al. 2025c). Traditional approaches (Zhang et al. 2025a; Ren et al. 2025; Qian et al. 2024) typically rely on multi-view images or video input to reconstruct accurate head geometry and appearance, but they exhibit limited generalization capabilities when faced with unseen identities or novel scenes. To ameliorate this dilemma, recent research focus on constructing animatable 3D head avatars from a single portrait (Feng et al. 2021; Ma et al. 2023; Gerogiannis et al. 2025). They combine powerful facial foundation models with identity-guided optimization to achieve identity preservation and expressive control. Specifically, GPAvatar (Chu et al. 2024) and GAGAvatar (Chu and Harada 2024) employ dual-lifting strategies and point-based expression field to enable real-time and high-fidelity reenactment. The latest frameworks, Morphable Diffusion (Chen et al. 2024), LAM (He et al. 2025) and SEGA (Guo et al. 2025) combine generative diffusion models (Zhou et al. 2025; Zhang et al. 2025b) with 3D structural priors to further achieve full-view virtual character synthesis. However, these advances also raise significant concerns regarding unauthorized misuse, and there is still no specific work to defend against the 3D head avatar generation. To address this pressing issue, we propose Anti-Avatar, a novel protection framework that degrades reconstruction quality in both geometric and perceptual spaces to defend against unauthorized 3D avatar reconstruction and malicious manipulation.

Defending Against Malicious Image Generation. To resist unauthorized 2D image generation and editing, an active protection mechanism (Liang et al. 2023; Ahn et al. 2025) based on adversarial perturbations (Meng et al. 2024) has

been proposed to hinder the generation process of diffusion models. For example, AdvDM (Liang et al. 2023) effectively disrupts the denoising step by maximizing the diffusion loss, while ACE (Zheng, Liang, and Wu 2025) encourages outputs that resemble predefined target images, thereby causing the generated content to deviate from the original intent. Furthermore, AdvPaint (Jeon et al. 2025) and SemanticAttack (Lo et al. 2024) actively disrupt the self-attention and cross-attention mechanisms in diffusion models, making it difficult to accurately locate target regions, further providing superior protection against local editing and semantic manipulation. In addition, IDProtector (Song et al. 2025b) and FaceLock (Wang et al. 2025a) maximize the identity feature discrepancy between the protected and original images, effectively preventing the generation of consistent identity images. Despite excellent performance in defending against malicious 2D image generation, the aforementioned methods ignore the intrinsic geometric structure and appearance features of facial images, resulting in limited effectiveness against 3D head avatar generation. In this paper, we deliberately design Geometric Disruption and Perceptual Confusion modules to mislead the models in constructing facial structure and perceiving portrait appearance, thereby effectively defending against 3D head avatar generation.

Method

Preliminary

3D Head Avatar Reenactment. Given a source image I_{src} and a driving image I_{drv}, the goal of self-reenactment is to generate a 3D head avatar from I_{src}, while preserving its identity, expression, and pose. In contrast, cross-reenactment constructs a 3D head avatar that retains the identity of I_{src} while adopting expression and pose of I_{drv}.

Existing framework of 3D head avatar generation uses a FLAME estimator (Li et al. 2017) to offline estimate the 3D structural parameters of both portraits, as follows:

$$(s_{src}, e_{src}, p_{src}) = \Phi(I_{src}), \tag{1}$$

$$(s_{drv}, e_{drv}, p_{drv}) = \Phi(I_{drv}), \tag{2}$$

where Φ denotes the FLAME parameter regression network, s represents the shape parameter, e denotes the expression parameter, and p refers to the pose parameter.

Simultaneously, appearance features are extracted from the source image I_{src}, as follows:

$$F_{src} = E(I_{src}), \tag{3}$$

where E is the appearance encoder, and F_{src} represents the high-dimensional appearance features of the source image.

Then, for self-reenactment, we use the source parameters to generate structural through a generator $\mathcal{G}$:

$$G_{self} = \mathcal{G}(s_{src}, e_{src}, p_{src}). \tag{4}$$

For cross-reenactment, structural are generated based on the source identity parameter s_{src} and the expression and pose parameters (e_{drv}, p_{drv}) of the driving image:

$$G_{cross} = \mathcal{G}(s_{src}, e_{drv}, p_{drv}). \tag{5}$$

At last, combining structure and appearance features, customized mapping and rendering functions are used to generate the target avatar of self and cross reenactment:

$$\mathcal{V}_{self} = R(F_{src}, G_{self}), \tag{6}$$

$$\mathcal{V}_{cross} = R(F_{src}, G_{cross}), \tag{7}$$

where R represents the mapping and rendering function, which can be neural rendering or 3DGS rendering.

Prevent Malicious Image Generation. Given an original image x and a generate model $\mathcal{D}$, the objective is to optimize an imperceptible perturbation δ such that the the protected image $(x+\delta)$ can hinder unauthorized generation or editing as much as possible. The general form can be written as:

$$\delta \coloneqq \arg\min_{\delta} \mathcal{L}_{adv}(x+\delta; \mathcal{D}) \quad \text{s.t. } \|\delta\|_p \leq \epsilon, \tag{8}$$

where $\mathcal{L}_{adv}$ denotes the adversarial loss, quantifying the generation or editing effectiveness under different protection objectives. The perturbation δ is restricted to the range of $(-\epsilon, \epsilon)$ by the L_p norm.

To prevent image generation with retained identities, $\mathcal{L}_{adv}$ is defined as the similarity between two features, as follows:

$$\mathcal{L}_{adv} = sim\left(f(x+\delta), f(x)\right), \tag{9}$$

where f represents the identity feature encoder, such as ArcFace (Deng et al. 2019) or CLIP (Radford et al. 2021), and sim denotes cosine similarity.

Problem Definition

For a portrait $\mathbf{I}$, the existing avatar generation model $\mathcal{V}$ can construct a head avatar $\mathcal{V}(\mathbf{I})$ from $\mathbf{I}$. In this paper, we propose a 3D defense framework aiming to prevent unauthorized generation of 3D head avatar, which can be defined as an optimization problem.

Definition 1 (Preventing Unauthorized 3D Avatar Generation). Given a portrait $\mathbf{I}$, there exists a perturbation δ such that the 3D avatar $\mathcal{V}(\mathbf{I}+\delta)$ generated by the protected portrait $(\mathbf{I}+\delta)$ has no distinguishable natural characteristics. The perturbation is given by the following objectives:

$$\begin{aligned} \delta \coloneqq \arg\max_{\delta} & \mathcal{L}_{dist}(\mathcal{V}(\mathbf{I}+\delta), \mathcal{V}(\mathbf{I})), \\ \text{s.t. } & \|\delta\|_p \leq \epsilon, \end{aligned} \tag{10}$$

where $\mathcal{L}_{dist}$ denotes the distance function used to measure the differences between avatars before and after protection.

However, directly measuring the difference between two digital avatars is challenging, as conventional image-based metrics often fail to capture variations in facial structure and identity attributes. To overcome this limitation, we instead assess discrepancies in geometric G and appearance feature F representations, which are the principal factors influencing 3D avatar synthesis, as demonstrated in Eqs. (6) and (7). Building on this insight, we define our objective as follows:

$$\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}_{geo}(G', G) + \lambda \mathcal{L}_{per}(F', F), \tag{11}$$

where G' and F' denote the geometric and appearance feature obtained from the protected portraits, respectively. $\mathcal{L}_{geo}$ and $\mathcal{L}_{per}$ denote the distance in geometric and perceptual

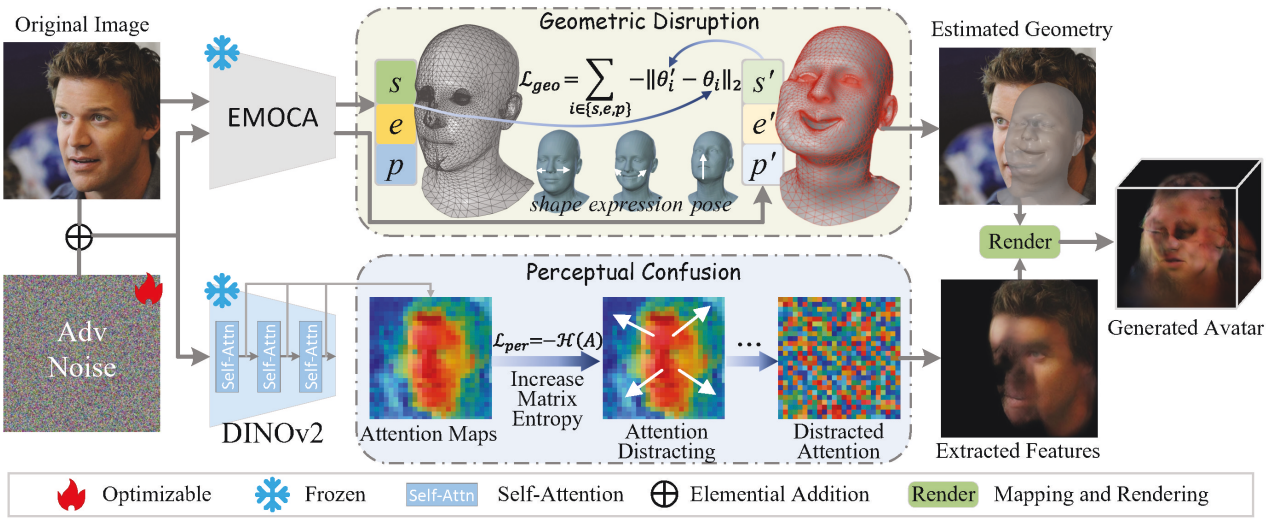

Figure 2: Overview of our Anti-Avatar. We introduce adversarial noise to the original image and employ two complementary modules to optimize the noise. The **Geometric Disruption** disrupts 3D facial structure by interfering with the estimation of facial geometric parameters, while the **Perceptual Confusion** maximizes the entropy of attention maps to distract feature extraction. Through joint optimization, the avatars generated from the protected portraits exhibit dual-space deviations in both geometric and appearance spaces, thereby effectively preventing unauthorized 3D avatar reconstruction.

space, λ is a weight used to balance this two distance. Through this objective, we can reduce the geometric and appearance quality of reconstructed head avatars, thereby resisting unauthorized avatar generation.

However, the above strategy assumes that malicious generation has already taken place, which is precisely the scenario we seek to prevent. Therefore, we shift our focus toward a more proactive strategy: directly maximizing the intrinsic geometric and perceptual discrepancies within the portrait, thereby disrupting the reconstruction at source:

$$\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}_{geo}(\Phi(\mathbf{I}+\delta), \Phi(\mathbf{I})) + \lambda \mathcal{L}_{per}(E(\mathbf{I}+\delta), E(\mathbf{I})). \tag{12}$$

Following the above analysis, we propose Anti-Avatar, which disrupts the generation process by perturbing dual-space of geometry and perception, as illustrated in Figure 2. Specifically, Anti-Avatar is implemented through two complementary modules: the Geometric Disruption and the Perceptual Confusion, which are described in detail as follows.

Geometric Disruption

To disrupt the structural accuracy of reconstructed avatars, we design a Geometric Disruption to interfere with the estimation of key geometric parameters. Specifically, the geometric structure of the reconstructed 3D head is essentially encoded by three key parameters representing facial geometry, as shown in Eqs. (4) and (5). For a given input facial image, the geometric parameters θ are estimated via the parameter estimation model, as follows:

$$\theta_i = \Phi(\mathbf{I}), \quad i \in \{s, e, p\}. \tag{13}$$

Based on this dependency, we directly attack the estimation model of geometric parameters, providing a reasonable means to disrupt the 3D structure reconstruction process. Accordingly, we define the geometric disruption loss as the distance between the predicted geometric parameters of the original and protected portraits:

$$\mathcal{L}_{geo} = \sum_{i\in\{s,e,p\}} -\omega_i \|\theta_i' - \theta_i\|_2, \tag{14}$$

where θ_i' represent the geometric parameters predicted from the protected portraits and ω_i is a weighting coefficient controlling the contribution of each parameter. By maximizing this discrepancy, the perturbations induce substantial changes in the estimated geometric structure, thereby disrupting self-reenactment and cross-reenactment rendering.

Perceptual Confusion

To interfere with the appearance quality of the generated 3D avatars, we design a matrix entropy-based loss function to disrupt attention during feature extraction. From an information-theoretic perspective, matrix entropy (Skean et al. 2023; Song et al. 2025a) is an effective metric for measuring the richness of feature representations and the similarity structure among samples. Given a positive definite Gram matrix $K \in \mathbb{R}^{d\times d}$, its matrix entropy is defined as:

$$\mathcal{H}(K) = -tr\left(\frac{1}{d}K \log \frac{1}{d}K\right) = -\sum_{i=1}^{d} \frac{\rho_i}{d} \log \frac{\rho_i}{d}, \tag{15}$$

where tr denotes the matrix trace and ρ_i is the i-th eigenvalue of matrix K. A higher matrix entropy indicates that the relationships between samples in the feature space tend toward a uniform distribution, with no obvious clustering.

3D head avatar generation methods often rely on concentrated attention patterns to capture key appearance features

Figure 3: Comparison of avatar reconstruction results using different defense methods, including AdvPaint (Jeon et al. 2025), ACE (Zheng, Liang, and Wu 2025), and FaceLock (Wang et al. 2025a). **The top row** presents results on the CelebA-HQ dataset (Karras et al. 2018), and **the bottom row** presents results on the VFHQ dataset (Xie et al. 2022).

to achieve high-quality reconstruction and realistic identity retention. Inspired by this, we propose Perceptual Confusion to optimize the attention matrix entropy:

$$\mathcal{L}_{per} = -\mathcal{H}(A) = tr\left(\frac{1}{N} A \log \frac{1}{N} A\right), \tag{16}$$

where $A \in \mathbb{R}^{N \times N}$ is the attention matrix generated during feature extraction. By maximizing the matrix entropy $\mathcal{H}(A)$, this module encourages the attention distribution of the model to tend toward uniformity, thereby weakening its ability to accurately capture specific appearance details.

The Perception Confusion not only disrupts the ability of the model to locate and extract meaningful facial information, but also distorts identity features, significantly enhancing the robustness of our defense.

Optimization

We combine the geometric disruption and perceptual confusion objectives into a unified total loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{geo} + \lambda \mathcal{L}_{per}. \tag{17}$$

The adversarial perturbation δ is obtained by solving the following constrained optimization problem:

$$\delta^{(k+1)} = Proj_{\|\delta\|_p \leq \epsilon}\left[\delta^{(k)} - \alpha \cdot \nabla_\delta \mathcal{L}_{total}\right], \tag{18}$$

where k indicates the iteration step, α is the update step size, and $Proj$ represents the projecting function that controls the visual imperceptibility of the perturbation.

Through the coordinated effect of these modules, our Anti-Avatar forces the generation process to deviate simultaneously in the dual spaces of geometry and perception, thereby substantially hindering accurate and realistic 3D avatar reconstruction from protected portraits.

Experiments

Experimental Setup

Datasets and Models. We conduct experiments on widely used facial image and video datasets, including CelebA-HQ (Karras et al. 2018) and VFHQ (Xie et al. 2022). These datasets provide high-quality portraits covering diverse identities, expressions, and poses, which are helpful for comprehensively evaluating virtual character reconstruction and preservation. In addition, we employ several SOTA single-image 3D head avatar generation methods, including GPAvatar (Chu et al. 2024), GAGAvatar (Chu and Harada 2024), and LAM (He et al. 2025), with all settings using the default parameters, and GAGAvatar is default model.

Evaluation Metrics. We employ several metrics to evaluate the effectiveness of the proposed method. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are used to evaluate the difference in image quality between the reconstructed images before and after protection. Identity Similarity Metric (ISM) quantifies the degree of consistency in facial identity, while CLIP similarity measures high-level feature differences between images. In addition, Fréchet Inception Distance (FID) is utilized to

Method	Self Reenactment						Cross Reenactment					
	PSNR↓	SSIM↓	ISM↓	CLIP↓	FID↑	FDR↓	PSNR↓	SSIM↓	ISM↓	CLIP↓	FID↑	FDR↓
ACE	17.91	0.7461	0.6574	0.9106	50.85	1.0000	17.98	0.7573	0.6426	0.9033	52.27	0.9973
AdvPaint	14.63	0.6690	0.2923	0.7775	113.13	0.9786	15.41	0.6865	0.3954	0.7833	112.65	0.9824
FaceLock	20.85	0.7832	0.2516	0.9186	36.14	1.0000	21.56	0.7948	0.3131	0.9172	38.51	0.9973
Anti-Avatar (Ours)	14.17	0.5459	0.2175	0.7554	116.84	0.8328	14.58	0.5683	0.1897	0.7528	115.55	0.9394

Method	Self Reenactment						Cross Reenactment					
	PSNR↓	SSIM↓	ISM↓	CLIP↓	FID↑	FDR↓	PSNR↓	SSIM↓	ISM↓	CLIP↓	FID↑	FDR↓
ACE	17.17	0.7499	0.6393	0.8994	61.41	1.0000	16.68	0.7716	0.6061	0.9083	45.12	0.9992
AdvPaint	15.95	0.6683	0.3731	0.8225	125.12	0.9751	16.85	0.6881	0.3541	0.8277	111.95	0.9501
FaceLock	20.46	0.7928	0.4386	0.9186	42.77	1.0000	20.93	0.8128	0.4227	0.9287	23.30	0.9999
Anti-Avatar (Ours)	15.58	0.6696	0.3433	0.8089	97.37	0.9384	16.02	0.6987	0.337	0.8189	112.28	0.9452

Table 1: Quantitative comparison of defense methods, including AdvPaint, ACE, FaceLock, and Anti-Avatar. **The top table** presents results on the CelebA-HQ dataset, and **the bottom table** presents results on the VFHQ dataset. ↑: higher is better, ↓: lower is better. Red indicates optimal and orange indicates suboptimal.

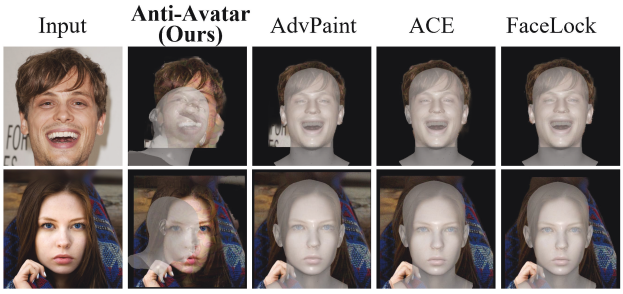

Figure 4: Visualization of estimated 3D geometric structures under different defense methods. Our Anti-Avatar introduces significant geometric disruption, effectively hindering the faithful reconstruction of 3D shapes.

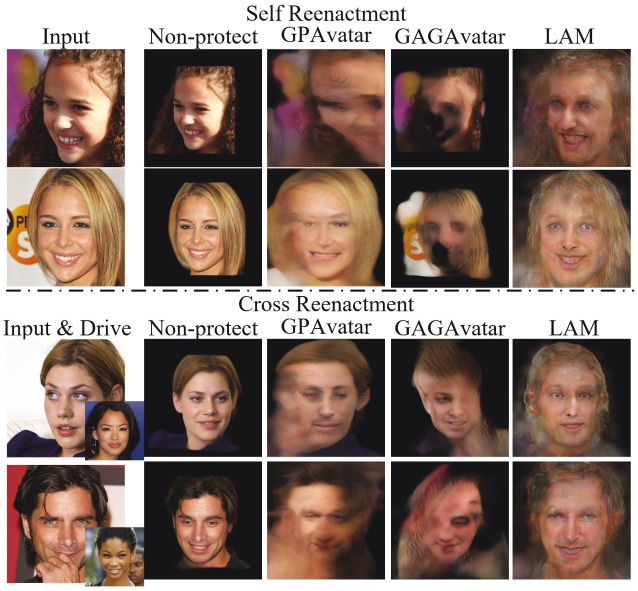

Figure 5: Qualitative comparison of the generalization performance across multiple 3D avatar generation pipelines, including GPAvatar, GAGAvatar, and LAM.

evaluate the overall fidelity and realism of rendered avatars, and Face Detection Rate (FDR) is reported to assess the recognizability of faces in the protected portraits.

Implementation Details. We utilize EMOCA (Daněček, Black, and Bolkart 2022) as the parameter estimation network and DINOv2 (Oquab et al. 2024) as the feature extraction network, which are both most commonly used in existing head avatar generation. The perturbation budget ϵ is set to $8/255$, and the number of iteration $K = 120$. The hyperparameter $\lambda = 0.01$ for loss balancing and the weights $\omega_{s,e,p} = \{0.01, 0.01, 1\}$ in the geometric perturbation. All experiments are conducted using NVIDIA GPU RTX 4090.

Comparative Evaluation

We present comparison of avatar reconstruction results on different defense methods, including AdvPaint, ACE, FaceLock, and our Anti-Avatar, on the CelebA-HQ and VFHQ datasets. As shown in Figure 3, although AdvPaint, ACE, and FaceLock introduce slight distortions or identity shifts, the reconstructed avatars generally retain similarity to the original appearance. In contrast, our Anti-Avatar produces more intense facial distortions and significantly disrupt the structural integrity and identity consistency of the generated avatars. Furthermore, our method demonstrates outstanding performance in both self-reenactment and cross-reenactment settings, underscoring its ability to resist unauthorized 3D head avatar generation and manipulation.

As summarized in Table 1, our Anti-Avatar achieves the lowest ISM and CLIP among all methods, indicating a substantial reduction in identity similarity between the reconstructed and original avatars. Meanwhile, it obtains competitive values for PSNR, SSIM, and FID, demonstrating that the overall visual quality and perceived realism of the avatars

Method	Self Reenactment					
	PSNR↓	SSIM↓	ISM↓	CLIP↓	FID↑	FDR↓
original	Inf	1.0	1.0	1.0	0.0	1.0
GPAvatar	15.76	0.6511	0.3378	0.8962	112.75	0.8524
GAGAvatar	14.17	0.5459	0.2175	0.7554	116.84	0.8328
LAM	16.78	0.7340	0.5636	0.9656	44.56	0.9547
Method	Cross Reenactment					
	PSNR↓	SSIM↓	ISM↓	CLIP↓	FID↑	FDR↓
original	Inf	1.0	1.0	1.0	0.0	1.0
GPAvatar	15.70	0.4204	0.3652	0.8101	117.87	0.9640
GAGAvatar	14.58	0.5683	0.1897	0.7528	115.55	0.9394
LAM	16.94	0.7444	0.5398	0.9856	45.55	0.9834

Table 2: Quantitative evaluation of generalization on GPAvatar, GAGAvatar, and LAM. Using the CelebA-HQ dataset. ↑: higher is better, ↓: lower is better.

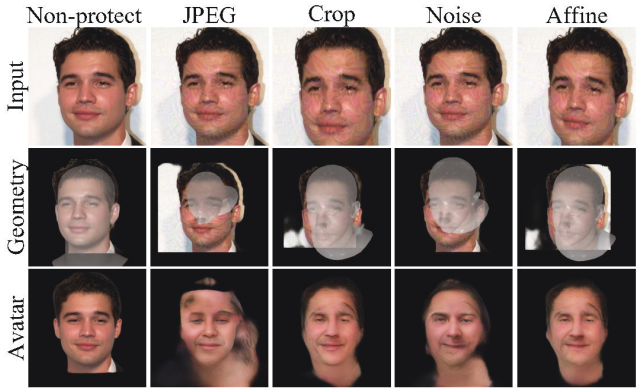

Figure 6: Robustness of our Anti-Avatar. The protected portrait remains resistant even after JPEG compression, image cropping, noise addition, and affine transformation.

are compromised. These results suggest that our method effectively disrupts unauthorized 3D head avatar synthesis by reducing identity retention and image quality.

Furthermore, Figure 4 visualizes the estimated geometric structure of the protected portrait. It is evident that existing methods fail to interfere with the geometric structure, resulting in the avatar closely resembles the original structure. In contrast, our Anti-Avatar significantly disrupts the estimation of geometric parameters, validating its unique superiority in preventing unauthorized 3D head avatar generation.

Generalization Evaluation

We validate the generalizability on GPAvatar, GAGAvatar, and LAM. As illustrated in Figure 5, consistent and effective performance is achieved across all evaluated 3D avatar generation frameworks. Meanwhile, our method yields significantly lower reconstruction metrics compared to standard avatar construction, as presented in Table 2. This is because our Anti-Avatar disrupts the geometric and perceptual features of the portrait at its source, essentially breaking the underlying information of the portrait rather than relying on vulnerabilities in specific model frameworks. Therefore, our

Figure 7: Ablation study on of key components. "w/o $\mathcal{L}_{geo}$" denotes not using the geometric disruption loss, and "w/o $\mathcal{L}_{per}$" denotes not using the perceptual confusion loss.

method demonstrates broad applicability and transferability across various 3D virtual character synthesis scenarios.

Robustness Analysis

We further investigate the robustness of our method against complex scenarios, including JPEG compression, image cropping, image noise, and affine transformations. As shown in Figure 6, our method maintain strong performance even when images are subjected to operations such as compression. This robustness can be attributed to the image cropping and segmentation operations integrated into the EMOCA module, which enhance the ability to focus perturbations on local facial regions and learn robust geometric features, thereby enabling Anti-Avatar to effectively resist various scene-level image manipulations.

Ablation Study

The ablation results in the Figure 7 indicate that without the geometric distortion loss $\mathcal{L}_{geo}$, the geometric structure of the avatar are preserved, resulting in minimal protection. Without the perceptual confusion loss $\mathcal{L}_{per}$, more appearance details are retained and reconstructed. The full model containing both losses effectively suppresses geometric structure and appearance quality, achieving optimal performance.

Conclusion

In this paper, we propose Anti-Avatar, a novel 3D defense framework dedicated to protecting portraits from unauthorized 3D head avatar generation. In contrast to existing methods that solely perturb 2D identity features, our Anti-Avatar simultaneously disrupts both 3D structure and appearance of reconstructed avatars, offering more effective and targeted prevention. Specifically, we elaborate complementary Geometric Disruption and Perceptual Confusion modules that interfere with the estimation of geometric parameters and the extraction of appearance features, thereby amplifying reconstruction discrepancies in the dual-space of geometry and perception. Extensive experiments demonstrate that our Anti-Avatar substantially reduces the realism and identity of reconstructed avatars, delivering generalized protection against various single-image 3D head avatar reconstruction and manipulation frameworks.

Acknowledgments

The authors are very indebted to the anonymous referees for their critical comments and suggestions for the improvement of this paper. This work was supported by the National Key Research and Development Program of China (2021YFA1000102), National Natural Science Foundation of China (Nos. 62376285, 61673396), Natural Science Foundation of Shandong Province (No: ZR2022MF260).

References

Ahn, N.; Yoo, K.; Ahn, W.; Kim, D.; and Nam, S.-H. 2025. Nearly Zero-Cost Protection Against Mimicry by Personalized Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28801–28810.

Chen, X.; Mihajlovic, M.; Wang, S.; Prokudin, S.; and Tang, S. 2024. Morphable Diffusion: 3D-Consistent Diffusion for Single-image Avatar Creation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10359–10370.

Choi, J. S.; Lee, K.; Jeong, J.; Xie, S.; Shin, J.; and Lee, K. 2025. DiffusionGuard: A Robust Defense Against Malicious Diffusion-based Image Editing. In *International Conference on Learning Representations*.

Chu, X.; and Harada, T. 2024. Generalizable and Animatable Gaussian Head Avatar. In *Advances in Neural Information Processing Systems*, volume 1838, 57642–57670.

Chu, X.; Li, Y.; Zeng, A.; Yang, T.; Lin, L.; Liu, Y.; and Harada, T. 2024. GPAvatar: Generalizable and Precise Head Avatar from Image(s). In *International Conference on Learning Representations*.

Daněček, R.; Black, M.; and Bolkart, T. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *EEE/CVF Conference on Computer Vision and Pattern Recognition*, 20279–20290.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4685–4694.

Deng, Y.; Wang, D.; Ren, X.; Chen, X.; and Wang, B. 2024. Portrait4D: Learning One-Shot 4D Head Avatar Synthesis Using Synthetic Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7119–7130.

Feng, Y.; Feng, H.; BLACK, M. J.; and Bolkart, T. 2021. Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. *ACM Transactions on Graphics*, 40(4): 1–13.

Gafni, G.; Thies, J.; Zollhöfer, M.; and Nießner, M. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8645–8654.

Gan, Y.; Miao, J.; Wang, Y.; and Yang, Y. 2025. Silence Is Golden: Leveraging Adversarial Examples to Nullify Audio Control in LDM-based Talking-Head Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13434–13444.

Gerogiannis, D.; Papantoniou, F. P.; Potamias, R. A.; Lattas, A.; and Zafeiriou, S. 2025. Arc2Avatar: Generating Expressive 3D Avatars from a Single Image via ID Guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10770–10782.

Guo, C.; Su, Z.; Wang, J.; Li, S.; Chang, X.; Li, Z.; Zhao, Y.; Wang, G.; and Huang, R. 2025. SEGA: Drivable 3D Gaussian Head Avatar from a Single Image. *arXiv preprint arXiv.2504.14373*.

He, Y.; Gu, X.; Ye, X.; Xu, C.; Zhao, Z.; Dong, Y.; Yuan, W.; Dong, Z.; and Bo, L. 2025. LAM: Large Avatar Model for One-shot Animatable Gaussian Head. In *SIGGRAPH Conference*.

Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; and Zhang, J. 2022. HeadNeRF: A Realtime NeRF-based Parametric Head Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20342–20352.

Huang, H.; Wu, Y.; and Wang, Q. 2025. ROBIN: Robust and Invisible Watermarks for Diffusion Models with Adversarial Optimization. In *Advances in Neural Information Processing Systems*, volume 37, 3937–3963.

Jeon, J.; Kim, W. J.; Ha, S.; Son, S.; and Yoon, S.-e. 2025. AdvPaint: Protecting Images from Inpainting Manipulation via Adversarial Attention Disruption. In *International Conference on Learning Representations*.

Jiang, J.; Lin, G.; Rong, Z.; Liang, C.; Zhu, Y.; Yang, J.; and Zhong, T. 2025. MobilePortrait: Real-Time One-Shot Neural Head Avatars on Mobile Devices. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15920–15929.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Transactions on Graphics*, 36(6): 194:1–194:17.

Liang, C.; Wu, X.; Hua, Y.; Zhang, J.; Xue, Y.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples. In *International Conference on Machine Learning*, 20763–20786.

Lo, L.; Yeo, C. Y.; Shuai, H.-H.; and Cheng, W.-H. 2024. Distraction Is All You Need: Memory-Efficient Image Immunization against Diffusion-Based Image Editing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24462–24471.

Lyu, W.; Zhou, Y.; Yang, M.-H.; and Shu, Z. 2025. FaceLift: Single Image to 3D Head with View Generation and GS-LRM. In *International Conference on Computer Vision*, 12691–12701.

Ma, Z.; Zhu, X.; Qi, G.; Lei, Z.; and Zhang, L. 2023. OTAvatar: One-Shot Talking Face Avatar with Controllable Tri-Plane Rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16910.

Meng, L.; Shao, M.; Wang, F.; Qiao, Y.; and Xu, Z. 2024. Advancing Few-Shot Black-Box Attack With Alternating Training. *IEEE Transactions on Reliability*, 73(3): 1544–1558.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.; Li, S.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Qian, S.; Kirschstein, T.; Schoneveld, L.; Davoli, D.; Giebenhain, S.; and Nießner, M. 2024. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20299–20309.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I.; Marina, M.; and Tong, Z. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of Machine Learning Research*.

Ren, H.; Duan, W.; Li, W.; Liu, Y.; Guo, Y.; Huang, S.; Zhang, J.; and Huang, H. 2025. EGAvatar: Efficient GAN Inversion for Generalizable Head Avatar from Few-shot Images. *IEEE Transactions on Visualization and Computer Graphics*, 1–14.

Shih, C.; Peng, L.; Liao, J.; Chu, E.; Chou, C.-F.; and Chen, J.-C. 2025. Pixel Is Not A Barrier: An Effective Evasion Attack for Pixel-Domain Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6905–6913.

Skean, O.; Osorio, J. K. H.; Brockmeier, A. J.; and Giraldo, L. G. S. 2023. DiME: Maximizing Mutual Information by a Difference of Matrix-Based Entropies. *arXiv preprint arXiv.2301.08164*.

Song, K.; Tan, Z.; Zou, B.; Chen, J.; Ma, H.; and Huang, W. 2025a. Exploring Information-Theoretic Metrics Associated with Neural Collapse in Supervised Training. *arXiv preprint arXiv.2409.16767*.

Song, Y.; Yang, P.; Ci, H.; and Shou, M. Z. 2025b. IDProtector: An Adversarial Noise Encoder to Protect Against ID-Preserving Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3019–3028.

Wang, H.; Zhang, Y.; Bai, R.; Zhao, Y.; Liu, S.; and Tu, Z. 2025a. Edit Away and My Face Will Not Stay: Personal Biometric Defense against Malicious Generative Editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23806–23816.

Wang, J.; Xie, J.; Li, X.; Xu, F.; Pun, C.-M.; and Gao, H. 2025b. GaussianHead: High-Fidelity Head Avatars With Learnable Gaussian Derivation. *IEEE Transactions on Visualization and Computer Graphics*, 31(7): 4141–4154.

Wang, Y.; Wang, X.; Yi, R.; Fan, Y.; Hu, J.; Zhu, J.; and Ma, L. 2025c. 3D Gaussian Head Avatars with Expressive Dynamic Appearances by Compact Tensorial Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21117–21126.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. VFHQ: A High-Quality Dataset and Benchmark for Video Face Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 656–665.

Xu, Y.; Chen, B.; Li, Z.; Zhang, H.; Wang, L.; Zheng, Z.; and Liu, Y. 2024. Gaussian Head Avatar: Ultra High-Fidelity Head Avatar via Dynamic Gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Zhang, D.; Liu, Y.; Lin, L.; Zhu, Y.; Chen, K.; Qin, M.; Li, Y.; and Wang, H. 2025a. HRAvatar: High-Quality and Relightable Gaussian Head Avatar. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26285–26296.

Zhang, J.; Wu, Z.; Liang, Z.; Gong, Y.; Hu, D.; Yao, Y.; Cao, X.; and Zhu, H. 2025b. FATE: Full-head Gaussian Avatar with Textural Editing from Monocular Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5535–5545.

Zheng, B.; Liang, C.; and Wu, X. 2025. Targeted Attack Improves Protection against Unauthorized Diffusion Customization. In *International Conference on Learning Representations*.

Zhou, Z.; Ma, F.; Fan, H.; and Chua, T. 2025. Zero-1-to-A: Zero-Shot One Image to Animatable Head Avatars Using Video Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15941–15952.