Robust, Generalizable Proactive Face-Swapping Defense via Semantic Gradient Divergence

Seung-hyeok Back, **Do Hyun Ki**, **Juwan Kim** and **Seok Bong Yoo***
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
sbyoo@jnu.ac.kr

Abstract

The rapid progress of identity-feature-based face-swapping technology has raised concerns about impersonation and privacy violations. Although proactive defenses aim to block identity extraction at the source, existing methods suffer from perceptible visual artifacts, poor generalization across diverse deepfake models, and vulnerability to post-processing techniques (e.g., diffusion purification, image compression, and transformations). This work proposes a robust, generalizable proactive face-swapping defense via semantic gradient divergence (SGD-Guard) to address these challenges. It introduces an integrated feature gallery that uses CLIP features and a generalized identity feature, obtained by iteratively refining heterogeneous identity features into a homogeneous representation. This framework facilitates our semantic distortion attack by leveraging consensus weighting to target specific facial attributes within a CLIP-identity joint embedding space, disrupting deepfake generation while preserving visual fidelity. Furthermore, to ensure robustness against purification and post-processing, this method incorporates a module that prioritizes critical transformations by exploiting directional discrepancies. Comprehensive experiments demonstrate that the method effectively defends against diverse face-swapping models with high cross-model transferability.

1 Introduction

Face-swapping technology (a subset of deepfakes) substitutes the facial identity of a subject in visual media with that of another. Recent progress in identity-feature-based methods has enabled the generation of swaps for unseen target images without requiring supplementary data or retraining. Although these innovations facilitate legitimate uses in industry (e.g., entertainment, gaming, and digital avatar design), they pose significant ethical risks, including identity theft, the spread of disinformation, and reputational damage.

* Corresponding author

Recent research has focused on two paradigms for risk mitigation: deepfake detection and proactive defense. Although detection methods [Tao *et al.*, 2025] identify manipulated content post hoc and cannot curb its spread, proactive defense protects users by disrupting identity feature extraction at the source. However, most existing proactive approaches [Komkov and Petiushko, 2021; Sun *et al.*, 2024; Jeong *et al.*, 2025] are tailored to a single deepfake backbone and thus generalize poorly across face-swapping models (Fig. 1(a)). Moreover, in pursuing stronger protection without patches or explicit noise optimization, they often introduce perceptible artifacts that degrade image fidelity (Fig. 1(b)). Even defenses with less perceptible perturbations can be weakened by post-processing such as diffusion purification, JPEG compression, and transformations (Fig. 1(c)), leaving deepfakes perceptually close to the source identity and facial privacy vulnerable to exploitation.

This paper proposes a proactive face-swapping defense framework capable of generating natural-looking protected images that are robust against adaptive attacks and input transformations to address these limitations while demonstrating generalized performance across diverse deepfake generation backbones. The framework integrates key strategies to achieve these goals. First, it computes gradients on diffusion-purified images to iteratively purify adversarial noise during updates, constraining perturbations to remain imperceptible on the natural image manifold while resisting adaptive diffusion-based purification. Next, a feature gallery is created for the protection mechanism. Applying representative backbones of deepfake generation models as surrogate models, we semantically refine the extracted heterogeneous identity cues into a homogeneous identity feature.

This work proposes a proactive defense method that safeguards facial images by introducing distortions in deepfake generation output using a feature gallery of integrated features. The objective is to update perturbations that allow rapid distortion by identifying vector directions corresponding to crucial facial components in the entangled identity space. This work proposes a novel joint embedding space by semantically combining CLIP features and identity features. In this space, the representation is steered by consensus weighting toward the direction that induces the strongest distortion away from the protected image's original attributes. Based on these objectives, the resulting deepfakes undergo signifi-

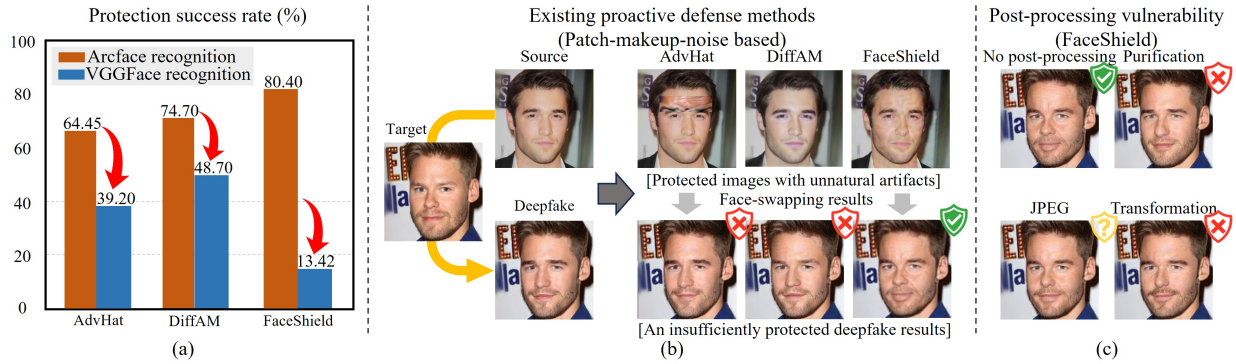

Figure 1: (a) Existing proactive defenses lack robustness across the backbones of deepfake generation models. (b) Limitations of proactive defense methods: Top: protected images; Bottom: corresponding face-swapping results. (c) Post-processing attacks neutralize proactive defenses.

cant distortion to the extent that the source cannot be inferred. This work analyzes the feature shift in the identity space caused by applying these transformations to the purified image to ensure robustness against post-processing operations. Optimizing perturbations to counteract this shift and restoring features to their intended distorted state achieves strong resilience against purification attacks. The SGD-Guard approach demonstrates exceptional robustness against adaptive attacks by providing an effective defense with minimal visual artifacts while achieving efficient image generation with low computational latency. Our source code and appendix are available at https://github.com/BACKAI/SGD-Guard. The following are the principal contributions of this work:

- This work proposes a semantic distortion (SD) attack in a CLIP-identity joint embedding space (CI-JES), leveraging consensus weighting to target facial attributes and disrupting deepfake generation while preserving visual fidelity. By aligning high-level semantics with identity representations, the distortions transfer across domains.
- This work presents semantic direction expectation over transformation (EOT), replacing the standard uniform expectation with direction-adaptive weighting based on semantic shifts in the CI-JES. It prioritizes critical transformations and prevents gradient dilution for efficient and robust optimization.
- This work proposes a semantic iterative identity refinement (SIIR) via channelwise disagreement to refine heterogeneous identity embeddings into a homogeneous representation. It thereby produces an identity feature that generalizes across diverse deepfake generation backbones.

2 Related Work

2.1 Face-Swapping Deepfake

Face swapping transfers the identity of a source face onto a target image while preserving the target attributes (e.g., the expression, illumination, and pose). Early learning-based methods [Korshunova *et al.*, 2017] often required subject-specific training for each identity pair, limiting generalization to unseen identities. Recent subject-agnostic frameworks have enabled efficient identity transfer without fine-tuning. Regardless of the generative architecture, whether based on convolutional neural network encoder–decoder models [Wang *et al.*, 2021], generators based on the generative adversarial network [Chen *et al.*, 2020; Rosberg *et al.*, 2023], or emerging diffusion models [Zhao *et al.*, 2023; Kim *et al.*, 2025; Baliah *et al.*, 2025], the core mechanism includes feature injection and fusion. Source embeddings are inserted into the latent feature maps of the target, where a fusion module combines source identity cues with the structural attributes of the target. This work aims to disrupt identity transfers by perturbing the source embeddings before fusion, inducing semantic divergence in the synthesized output.

2.2 Proactive Face-Swapping Defense

Proactive face-swapping defense methods modify facial images so that edits become null or visually odd. A line of work relies on adversarial watermarking, where small, structured perturbations are embedded into facial images to interfere with downstream generators. Cross-model processes [Huang *et al.*, 2022] and watermark-embedding networks [Zhang *et al.*, 2024] follow this paradigm but typically assume white-box access, limiting scalability.

Universal defenses have been proposed to relax this assumption, including transferable adversarial generators [Dong *et al.*, 2023] and compression-aware perturbations [Wang *et al.*, 2022; Qu *et al.*, 2024]. However, these methods target reconstruction-based encoder-decoder processes and offer limited protection against identity-driven face swapping in an identity-embedding space.

Other approaches disrupt identity extraction by attacking face-recognition components [Hu *et al.*, 2024] or applying local modifications [Ma *et al.*, 2023], attribute editing [Jia *et al.*, 2022], and makeup-based perturbations [Hu *et al.*, 2022; Sun *et al.*, 2024]. Although effective, these methods often require stronger perturbations or visible changes for robustness, degrading visual fidelity.

Recent methods mitigate such artifacts [Wang *et al.*, 2025; Jeong *et al.*, 2025] but may remain vulnerable to post-processing or depend on specific identity extractors.

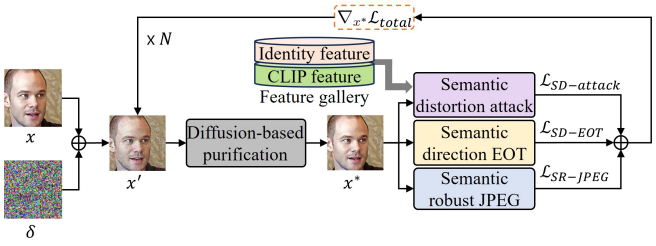

Figure 2: Architecture of the SGD-Guard framework.

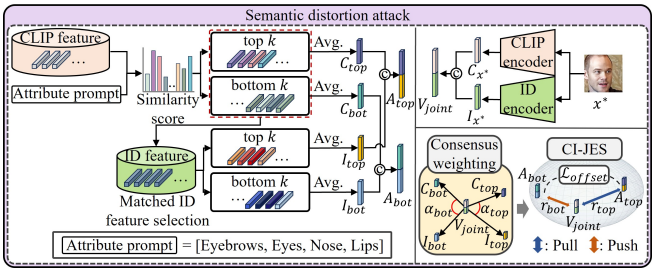

Figure 3: Illustration of semantic distortion attack.

This work proposes an adversarial watermarking framework driven by semantic gradient divergence to address these limitations, offering transferable protection robust to complex post-processing while preserving high visual fidelity.

3 Proposed Method

3.1 Overview

This work proposes a proactive face-swapping defense that facilitates high visual fidelity and generalized robustness across diverse deepfake backbones. In Figure 2, the framework iteratively optimizes the perturbation δ via a diffusion-guided strategy to resist diffusion-based adaptive attacks [Xue *et al.*, 2023]. Integrating one-step diffusion purification [Lei *et al.*, 2025] imperceptible yet resilient noise while accelerating the optimization process. Prior to the perturbation update, unified identity features are synthesized from backbone outputs and are integrated with CLIP to create a stable reference gallery. This work applies gallery features and four identity prompts to estimate directions that maximize SD in deepfake outputs. To ensure robustness, this approach estimates transformation-induced latent shifts offline across the dataset and neutralizes them during optimization to prevent post-processing degradation.

3.2 Semantic Distortion Attack

Proactive deepfake defenses often fail to protect synthesized outputs because key identity-relevant facial attributes [Lestriandoko *et al.*, 2022; Lee *et al.*, 2023] remain intact, offering limited protection. Thus, this work assigns larger weights to distortion directions that strongly degrade identity-critical attributes, making the original image unidentifiable. Accordingly, we select four local identity-critical attributes (eyebrows, eyes, nose, and lips) as manipulation anchors, since broader cues such as pose or style are less identity-specific and may dilute the semantic distortion gradient. Further discussion and experiments are provided in *Appendix A*. We then construct the prompt set $t_j{}_{j=1}^{4}$, where each t_j denotes the CLIP text feature for the jth attribute. Figure 3 presents a cosine-similarity assessment between these prompts and the CLIP image features from a prebuilt feature gallery, yielding a set of similarity scores S:

$$S = \{\cos(C_i, t_j)\,,\ i \in \{1, \ldots, N\}, j \in \{1, \ldots, 4\}\}\,, \quad (1)$$

where C_i denotes the CLIP image feature corresponding to the ith sample in the feature gallery.

From S, this method ranks the gallery samples for each attribute j and selects the top-k and bottom-k subsets. To retrieve samples efficiently from S, this work uses FAISS [Douze *et al.*, 2024] and averages the CLIP features in each subset to obtain C_{top} and C_{bot}, and retrieves identity features from the same samples to calculate I_{top} and I_{bot}. This work concatenates the corresponding CLIP and identity centroids to form the joint anchors A_{top} and A_{bot} for each attribute, producing four anchor pairs. Integrating features extracted from these distinct latent spaces creates CI-JES that enables the concurrent exploitation of semantic and identity-related representations. Unlike isolated-space defenses, the framework synergizes high-level semantics and identity embeddings.

Next, this work computes the joint representation of the purified sample x^* and encodes x^* using a CLIP encoder E_{clip} and an identity encoder E_{id} to obtain C_{x^*} and I_{x^*}, which are combined to form V_{joint}. The relative alignment toward the top and bottom anchors for each attribute is calculated to determine the distortion direction for V_{joint}, which is initiated by defining vectors between the features of x^* and each anchor in the CLIP and identity spaces:

$$\tilde{C}_v^j = C_v^j - C_{x^*}, \quad \tilde{I}_v^j = I_v^j - I_{x^*}, \quad v \in \{\text{top}, \text{bot}\}, \quad (2)$$

Because CLIP and identity embeddings reside in heterogeneous spaces, we first map their directional vectors into a shared comparison space using lightweight projection heads, followed by ℓ_2 normalization. Next, the cosine agreement for the top and bottom pairs in each space is computed as follows:

$$\alpha_{\text{top}}^j = \cos\big(\tilde{C}_{\text{top}}^j, \tilde{I}_{\text{top}}^j\big), \quad \alpha_{\text{bot}}^j = \cos\big(\tilde{C}_{\text{bot}}^j, \tilde{I}_{\text{bot}}^j\big). \quad (3)$$

Then, the weight r for each anchor is calculated based on the alignment degree of the top and bottom pairs:

$$r_v^j = \frac{\exp(\alpha_v^j)}{\exp(\alpha_{top}^j) + \exp(\alpha_{bot}^j)}, \quad (4)$$

The consensus weight r assigns relative importance to the distortion direction of V_{joint}. Calculating the similarity between V_{joint} and the attribute anchors $(A_{\text{top}}^j, A_{\text{bot}}^j)_{j=1}^{4}$ using cosine similarity applies an offset loss that pulls the representation toward directions with low similarity and away from those with high similarity. The r parameter weights each direction to modulate the distortion intensity adaptively toward the direction with higher semantic consensus in each space. The offset loss is defined as follows:

$$\mathcal{L}_{\text{offset}}^j = \max\Big(0,\ r_{\text{bot}}^j \cos(V_{\text{joint}}, A_{\text{bot}}^j) - r_{\text{top}}^j \cos(V_{\text{joint}}, A_{\text{top}}^j)\Big), \quad (5)$$

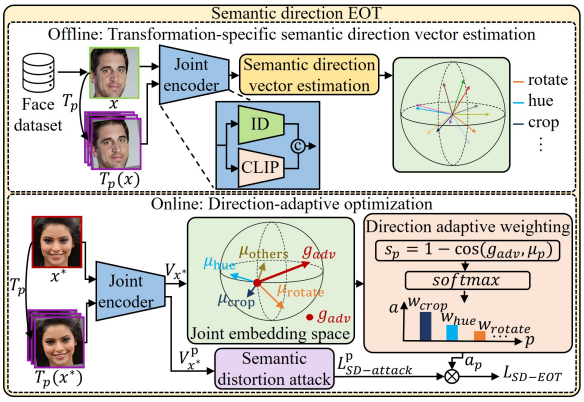

Figure 4: Illustration of semantic direction EOT.

For each attribute, we compute the cosine agreement between the top and bottom anchors in each feature space to obtain an alignment weight w_j that reflects attribute-wise divergence. This weight emphasizes attributes that contribute most to the distortion. These weights scale the corresponding offset losses, yielding the final SD attack loss as follows:

$$\mathcal{L}_{\text{SD-attack}} = \sum_{j=1}^{4} w_j \, \mathcal{L}^{j}_{\text{offset}}. \tag{6}$$

3.3 Semantic Direction EOT

Expectation over transformation (EOT) [Yuan *et al.*, 2022] improves robustness to photometric and geometric variations in face-swapping processes and can enhance black-box transferability. However, uniform expectations may dilute informative gradients by averaging them with many weak signals, slowing optimization. In Figure 4, Semantic direction EOT (SD-EOT) replaces the uniform expectation with direction-adaptive weighting, precomputing transformation-specific semantic direction vectors in the CI-JES, and assigning stable weights based on their directional conflict with the current attack direction for more robust optimization.

Offline Semantic Direction Estimation. For each transformation in a fixed bank, a semantic direction vector is estimated to capture the expected shift induced by the transformation in the CI-JES. Following prior EOT settings, a fixed set of differentiable photometric and geometric transformations $\{T_p\}_{p=1}^{P}$ is adopted. For each transformation T_p, the CI-JES shift vector ΔV_x^p for a face image x is defined by:

$$\Delta V_x^p = E_{\text{joint}}(T_p(x)) - E_{\text{joint}}(x), \tag{7}$$

where $E_{\text{joint}}(\cdot)$ extracts an identity embedding (e.g., ArcFace) and a CLIP image embedding from the input image and concatenates them to form the joint representation. The semantic direction vector μ_p of T_p is defined by:

$$\mu_p = \mathbb{E}_{x\sim\mathcal{D}}\left[\Delta V_x^p\right]. \tag{8}$$

where $\mathcal{D}$ represents a face dataset (e.g., FFHQ), and $\mathbb{E}_{x\sim\mathcal{D}}[\cdot]$ denotes the expectation over samples from $\mathcal{D}$. The direction vectors $\mu_p{}_{p=1}^{P}$ are computed once offline and shipped with the defense, requiring no additional data at deployment time.

Online direction-adaptive optimization. During online optimization, the loss terms for each transformation are adaptively weighted to focus on those that most impede the current attack instance. Given the current adversarial image x^*, the attack direction in the CI-JES, g_{adv}, is defined as follows:

$$g_{\text{adv}} = -\frac{\partial \mathcal{L}_{\text{SD-attack}}}{\partial E_{\text{joint}}(x^*)}, \tag{9}$$

where $\mathcal{L}_{\text{SD-attack}}$ denotes the semantic distortion attack loss.

The directional conflict score s_p is defined as follows to quantify how a transformation T_p counteracts the current attack direction:

$$s_p = 1 - \frac{g_{\text{adv}} \cdot \mu_p}{\|g_{\text{adv}}\|_2 \, \|\mu_p\|_2}, \tag{10}$$

where μ_p denotes the precomputed semantic direction of T_p, $\cdot$ represents the dot product, and $\|\cdot\|_2$ indicates the ℓ_2 norm.

The direction-adaptive weight a_p is obtained by applying a softmax normalization over $\{s_p\}_{p=1}^{P}$:

$$a_p = \frac{\exp(s_p)}{\sum_{q=1}^{P} \exp(s_q)}, \tag{11}$$

The SD-EOT objective is defined using the weighted expectation of the attack loss over the transformation set:

$$\mathcal{L}_{\text{SD-EOT}} = \sum_{p=1}^{P} a_p \, \mathcal{L}^{(p)}_{\text{SD-attack}}. \tag{12}$$

where $\mathcal{L}^{(p)}_{\text{SD-attack}}$ denotes the semantic distortion attack loss evaluated on $T_p(x^*)$.

3.4 Semantic Robust JPEG

The JPEG compression is common in real-world face-swapping pipelines. Its nondifferentiable quantization suppresses high-frequency perturbations and weakens defense performance. Existing defenses use differentiable JPEG surrogates [Qu *et al.*, 2024] or low-frequency constraints [Jeong *et al.*, 2025]; however, the former may deviate from real JPEG behavior, and the latter can limit semantic disruption. We adopt BPDA [Athalye *et al.*, 2018] with a differentiable surrogate, evaluating the forward objective on real JPEG outputs and approximating gradients through the surrogate.

The real JPEG operator is denoted by $J_{\text{real}}(x) = \text{JPEG}(x, q)$ and its differentiable surrogate is denoted by $J_{\text{diff}}(x) = \text{JPEG}_{\text{diff}}(x, q)$, where the quality factor q is randomly sampled during optimization. In the forward pass, the semantic distortion loss is evaluated on the real JPEG output:

$$\mathcal{L}_{\text{SR-JPEG}} = \mathcal{L}^{(\text{real})}_{\text{SD-attack}}, \tag{13}$$

where $\mathcal{L}^{(\text{real})}_{\text{SD-attack}}$ denotes the semantic distortion attack loss evaluated on $J_{\text{real}}(x^*)$.

In the backward pass, BPDA computes gradients using $J_{\text{diff}}(x^*)$ while maintaining the forward evaluation on $J_{\text{real}}(x^*)$.

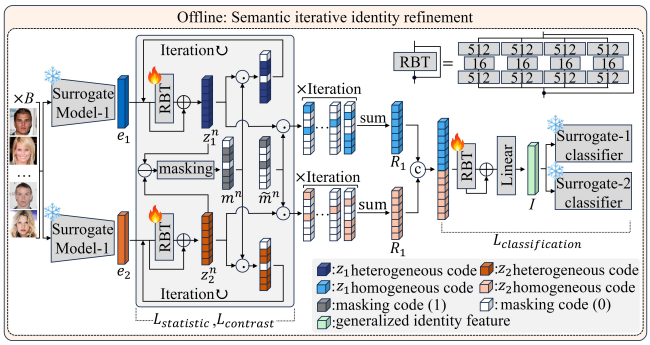

Figure 5: Illustration of semantic iterative identity refinement.

3.5 Total Loss Function

This work integrates the three proposed semantic-driven losses into a unified objective function, defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SD-attack}} + \mathcal{L}_{\text{SD-EOT}} + \mathcal{L}_{\text{SR-JPEG}}. \tag{14}$$

Finally, the perturbation θ is updated using the total loss with a learning rate η according to the following equation:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta \mathcal{L}_{total}. \tag{15}$$

3.6 Generalized Feature Gallery Construction

This work preconstructs a feature gallery to facilitate the perturbation update. The gallery comprises CLIP and identity features for each image. The CLIP features were extracted using E_{clip}. Prior gallery-based defenses rely on a single identity extractor [Lee *et al.*, 2025], limiting generalization under extractor shifts. Therefore, the identity features were extracted using the SIIR.

Offline Semantic Iterative Identity Refinement. This method extracts an identity feature that remains consistent across identity extractors for proactive defense under extractor shifts. Inspired by cross-model compatibility [Meng *et al.*, 2021], we align embeddings from surrogate extractors used in deepfake pipelines and iteratively consolidate cross-extractor homogeneous components while suppressing extractor-specific variation. As shown in Figure 5, embeddings $e_1, e_2 \in \mathbb{R}^{512}$ are obtained from two frozen surrogates, ArcFace and FaceNet [Deng *et al.*, 2019; Schroff *et al.*, 2015], which cover widely used identity spaces. Two RBT modules [Wang *et al.*, 2020] refine them along separate paths to preserve identity cues and facilitate alignment, producing z_1^n and z_2^n at iteration step n.

Channelwise disagreement d_c^n and a binary heterogeneous mask m_c^n are computed at each iteration step:

$$d_c^n = \left| |z_1^n[c]| - |z_2^n[c]| \right|, \quad m_c^n = \begin{cases} 1, & \text{if } d_c^n > H(d^n), \\ 0, & \text{otherwise}, \end{cases} \tag{16}$$

where $H(d^n)$ denotes the harmonic mean of the channelwise disagreement values $\{d_c^n\}_{c=1}^{512}$. This adaptive threshold is recomputed at each iteration to adapt to the changing distribution of channelwise disagreements. Channels selected by m_c^n are treated as heterogeneous codes and refined in subsequent iterations, whereas channels selected by $\tilde{m}_c^n = 1 - m_c^n$ are treated as homogeneous codes and accumulated across iterations to form the homogeneous code accumulation embeddings R_1 and R_2. The refinement stops when the mask no longer changes between consecutive iterations ($m^n = m^{n-1}$), after which R_1 and R_2 are concatenated and the result is fed into an RBT module followed by a linear projection to obtain the generalized identity feature $I \in \mathbb{R}^{512}$.

This work defines two alignment objectives on the last iteration embeddings z_1^n and z_2^n obtained from the two surrogate paths. First, a statistical alignment loss matches the mean and covariance computed over the minibatch embeddings of z_1^n and z_2^n, where $\bar{z}_1, \Sigma_1$ and $\bar{z}_2, \Sigma_2$ denote the corresponding mean and covariance, respectively:

$$\mathcal{L}_{\text{statistic}} = \left\| \bar{z}_1 - \bar{z}_2 \right\|_2^2 + \left\| \Sigma_1 - \Sigma_2 \right\|_F^2, \tag{17}$$

Second, contrastive loss pulls the embeddings of the same gallery image closer while separating different gallery images in a minibatch $\mathcal{B}$:

$$\mathcal{L}_{\text{contrast}} = -\mathbb{E}_{u \in \mathcal{B}} \log \frac{\exp\left(\cos(z_{1,u}^n, z_{2,u}^n)\right)}{\sum_{b \in \mathcal{B}} \exp\left(\cos(z_{1,u}^n, z_{2,b}^n)\right)}, \tag{18}$$

In our setting, we use one gallery image per identity and assign a unique identity index as its label. Here, $z_{1,u}^n$ and $z_{2,u}^n$ denote the last iteration embeddings of the u-th gallery image obtained from the two surrogate paths. $\mathcal{L}_{\text{contrast}}$ encourages cross-path consistency while preserving identity separability.

Finally, the classifier heads of the two surrogate identity extractors are frozen, and I is input, minimizing $\mathcal{L}_{\text{classification}}$ so that both heads predict the identity label for I. Further, $\mathcal{L}_{\text{classification}}$ is defined as the average of two cross-entropy loss functions from the prediction of each head with respect to the identity label. We optimize the gallery objective $\mathcal{L}_{\text{gallery}}$ to learn a robust generalized identity feature I, defined as:

$$\mathcal{L}_{\text{gallery}} = \mathcal{L}_{\text{statistic}} + \mathcal{L}_{\text{contrast}} + \mathcal{L}_{\text{classification}}. \tag{19}$$

4 Experiment

4.1 Experimental Setup

Datasets and Face-swapping Models. This work evaluates the method on two high-quality face datasets, CelebA-HQ [Karras *et al.*, 2017] and VGGFace2-HQ [Chen *et al.*, 2023]. The CelebA-HQ dataset contains 30K celebrity images, and VGGFace2-HQ comprises 9,630 identities with about 1.3M facial images, spanning diverse identities, poses, expressions, and illumination. The quantitative evaluation used 1,000 source images, generated 1,000 swaps per set via random source–target pairing, and averaged results across 10 sets to reduce pairing bias. This large-scale random pairing enables statistically grounded evaluation by averaging results over diverse identity combinations and reducing sensitivity to specific pair selections. This work compares the approach against face-swapping models: SimSwap [Chen *et al.*, 2020], FaceDancer [Rosberg *et al.*, 2023], DiffSwap [Zhao *et al.*, 2023], DiffFace [Kim *et al.*, 2025], REFace [Baliah *et al.*, 2025].

Methods	SimSwap (ACM MM'20)			FaceDancer (WACV'23)			DiffSwap (CVPR'23)			DiffFace (PR'25)			REFace (WACV'25)		
	PSNR‡ ↓	SSIM‡ ↓	DSR↑	PSNR‡ ↓	SSIM‡ ↓	DSR↑	PSNR‡ ↓	SSIM‡ ↓	DSR↑	PSNR‡ ↓	SSIM‡ ↓	DSR↑	PSNR‡ ↓	SSIM‡ ↓	DSR↑
CelebA-HQ dataset															
Anti-Forgery	26.47	0.823	40.2	25.62	0.812	47.4	24.77	0.806	52.1	27.05	0.833	57.4	26.64	0.818	40.3
CMUA	25.34	0.810	48.3	24.58	0.808	49.6	25.10	0.811	48.7	24.10	0.807	49.7	25.48	0.811	48.1
DF-RAP	24.10	0.807	58.4	25.06	0.810	54.3	24.83	0.809	59.3	24.13	0.806	55.2	24.35	0.808	53.5
DiffAM	23.77	0.805	69.0	24.62	0.809	68.1	23.54	0.807	66.8	24.97	0.807	67.7	23.51	0.803	70.6
FaceShield	22.53	0.801	83.6	21.13	0.781	82.2	22.47	0.803	82.5	21.94	0.791	78.3	22.94	0.801	81.5
SCOL	20.54	0.757	**95.1**	20.98	0.776	**95.0**	21.79	0.781	86.4	21.68	0.784	82.5	22.78	0.804	86.7
SGD-Guard (Ours)	**20.41**	**0.739**	94.3	**20.79**	**0.760**	94.1	**21.05**	**0.745**	**89.5**	**21.37**	**0.743**	**88.4**	**20.03**	**0.759**	**92.3**
VGGFace2-HQ dataset															
Anti-Forgery	27.10	0.826	49.0	26.54	0.843	46.5	25.65	0.812	56.1	27.18	0.824	55.3	27.45	0.831	54.6
CMUA	25.98	0.812	48.4	25.31	0.809	49.0	25.23	0.808	53.2	24.85	0.807	51.5	25.32	0.810	53.5
DF-RAP	24.73	0.805	47.9	25.01	0.807	52.5	25.25	0.809	51.3	24.56	0.810	52.2	23.83	0.806	61.0
DiffAM	24.12	0.808	69.5	24.43	0.806	70.0	24.29	0.807	70.1	24.33	0.807	69.1	23.75	0.806	64.2
FaceShield	26.53	0.815	56.2	25.33	0.808	48.2	25.07	0.805	55.8	25.11	0.810	46.7	26.74	0.816	44.3
SCOL	20.93	0.774	**94.5**	20.12	0.775	**96.0**	21.86	0.792	85.3	22.48	0.785	83.7	23.62	0.802	87.3
SGD-Guard (Ours)	**20.85**	**0.768**	93.7	**20.05**	**0.762**	94.6	**21.28**	**0.768**	**90.6**	**22.27**	**0.746**	**89.8**	**20.14**	**0.771**	**91.5**

Table 1: Quantitative baseline comparison for face-swapping models (deepfake evaluation).

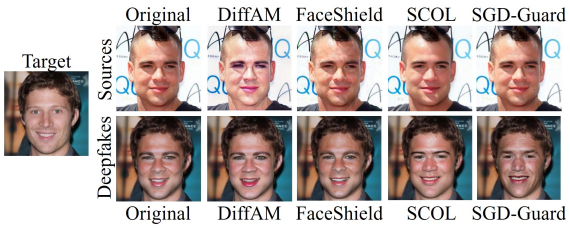

Figure 6: Visualization of protected images and the corresponding REFace results.

Methods	Publication	CelebA-HQ dataset			VGGFace2-HQ dataset		
		PSNR† ↑	SSIM† ↑	ISM↓	PSNR† ↑	SSIM† ↑	ISM↓
Anti-Forgery	IJCAI'22	34.91	0.930	0.428	34.73	0.931	0.418
CMUA-Watermark	AAAI'22	33.18	0.919	0.354	33.38	0.929	0.431
DF-RAP	TIFS'24	33.41	0.922	0.427	34.23	0.934	0.420
DiffAM	CVPR'24	17.66	0.658	0.238	17.75	0.625	0.255
FaceShield	ICCV'25	32.63	0.913	0.228	31.58	0.920	0.248
SCOL	ACM MM'25	32.52	0.924	0.163	32.01	0.921	0.158
SGD-Guard	–	**35.84**	**0.945**	**0.148**	**35.37**	**0.946**	**0.145**

Table 2: Quantitative comparison of original and protected images (source evaluation).

Figure 7: Protection evaluation of face-swapping models.

Evaluation Metrics. This work assesses visual fidelity and defense effectiveness using five metrics. The peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) metrics quantify visual similarity; higher values indicate better fidelity for original–protected pairs, while lower values on the deepfake outputs indicate stronger disruption. † for original-protected comparisons and ‡ for the deepfake images comparisons. The identity score matching (ISM) [Van Le *et al.*, 2023] measures identity similarity between the original and protected images, and defense success rate (DSR) and protection success rate (PSR) [Qu *et al.*, 2024; Lyu *et al.*, 2023] report the rates at which deepfakes fail to preserve the source identity and protected sources are not recognized as the original identity, respectively. In all tables, the best results are in bold, and the second-best are underlined.

Baselines. This work compares SGD-Guard with six proactive defenses: Anti-Forgery [Wang *et al.*, 2022], CMUA-Watermark [Huang *et al.*, 2022], DF-RAP [Qu *et al.*, 2024], DiffAM [Sun *et al.*, 2024], FaceShield [Jeong *et al.*, 2025], and SCOL [Lee *et al.*, 2025]. The DiffAM is makeup-based. Anti-forgery, CMUA-Watermark, DF-RAP, and FaceShield perturb protected images, whereas SCOL performs fusion in the latent space and injects adversarial watermark-like perturbation. For all noise-based methods, ϵ is set to 0.05 [Jeong *et al.*, 2025] under the ℓ_∞ norm for fair comparison. For optimization or generation based baselines, we follow the hyperparameters and protocols specified in the respective papers.

Implementation Details. The feature gallery stores paired generalized identity features and CLIP features. This work adopts FaRL as the CLIP image encoder E_{clip} for face analysis [Zheng *et al.*, 2022]. The gallery is precomputed offline and occupies only 72 MB, as it stores compressed feature embeddings rather than raw image data, enabling easy reproducibility and deployment. As described in Section 3.2, we set the retrieval parameter to $k = 10$ throughout. For semantic-direction EOT, the direction vectors are estimated using the full FFHQ dataset. Details on k are provided in *Appendix B*, and the algorithms for offline feature gallery building and online SGD-Guard are summarized in *Appendix D*.

4.2 Main Results

Evaluation of Deepfake Results. Table 1 compares proactive defenses on five face-swapping models across two datasets. Anti-Forgery and CMUA achieve low DSR, while DF-RAP, DiffAM, and FaceShield show high variance. In contrast, SGD-Guard is the most consistent, reaching top DSR in most cases and the lowest PSNR and SSIM on deepfake outputs. Figure 6 qualitatively compares SGD-Guard with DiffAM, FaceShield, and SCOL, which are used as state-of-the-art baselines in later experiments, and also evaluates protection against ReFace. DiffAM often causes large facial changes, FaceShield adds visible noise, and SCOL can blur protected images. In contrast, SGD-Guard better disrupts identity cues and preserves natural visual quality. Additional visual results are included in *Appendix C*.

Models	Protected success rate↑			
	DiffAM	FaceShield	SCOL	SGD-Guard
ArcFace	74.70	80.40	84.6	**91.4**
FaceNet	24.20	24.80	59.2	**79.2**
VGGFace	48.70	13.42	44.4	**59.2**
SphereFace	31.20	14.00	32.7	**58.4**

Table 3: Generalization evaluation of protection methods on face recognition models using the protection success rate.

Models	Purification		JPEG		Crop & rotate	
	PSNR‡ ↓	DSR↑	PSNR‡ ↓	DSR↑	PSNR‡ ↓	DSR↑
DiffAM	25.25	67.7	24.02	73.5	23.87	72.7
FaceShield	27.42	45.1	23.85	74.0	23.25	77.8
SCOL	25.45	67.5	23.12	79.3	23.06	83.2
SGD-Guard (Ours)	**22.43**	**83.9**	**20.52**	**88.6**	**20.15**	**90.1**

Table 4: Protection under post-processing attacks using REFace on CelebA-HQ.

Evaluation of Protected Images. Table 2 compares the protected-image quality and identity suppression. Anti-forgery preserves PSNR and SSIM but yields a high ISM value, indicating weak identity protection, whereas DiffAM lowers the ISM value at the cost of severe visual degradation. In contrast, SGD-Guard achieves the best PSNR and SSIM with the lowest ISM value, offering substantial identity disruption without sacrificing visual fidelity.

Generalizability. Figure 7 compares DiffSwap, DiffFace, and ReFace under face-swap model shifts. The first row shows swaps from source images, and the second shows swaps from protected sources. Across all models, SGD-Guard consistently degrades synthesized outputs, indicating robust protection under model changes. Table 3 evaluates identity extractor shifts using four face verifiers. Methods optimized for ArcFace achieve high PSR on it but drop by over 45%p on average on other verifiers. In contrast, SGD-Guard leverages ArcFace and FaceNet surrogates, achieving the highest PSR while limiting the average gap to 26%p on VGGFace and SphereFace. These results suggest transfer beyond the surrogate pair, although they do not guarantee coverage of proprietary identity extractors whose embedding distributions substantially differ from the evaluated verifiers.

Robustness to Post-processing Attacks. Table 4 evaluates robustness to standard post-processing on CelebA-HQ using REFace. We apply diffusion purification [Nie *et al.*, 2022], JPEG compression with random quality ($q \sim \mathcal{U}[1, 70]$), and crop/rotate transforms, since vulnerability varies by transformation. Following [Yuan *et al.*, 2022], images are randomly cropped to retain at least 90% area and rotated up to 15°. Across all settings, SGD-Guard is the most robust, achieving the highest DSR and lowest PSNR, indicating disruption that persists after post-processing. Additional evaluations under challenging conditions (e.g. real-world extreme poses and various resolutions), with qualitative robustness results, are provided in *Appendix C*.

Ablation Study. Table 5 presents an ablation study on the distortion and generalization performance of SD-attack and SIIR using REFace on the VGGFace2-HQ dataset. SD-attack achieves a low SSIM regardless of its integration with the SIIR module, signifying substantial distortion in the protected images. When utilized alongside the SIIR model, it demonstrates high DSR across various backbones and exhibits robust generalization performance.

SD-attack	SIIR	SSIM‡ ↓	ArcFace	FaceNet	VGGFace	SphereFace
			DSR↑			
✓	✓	0.771	91.5	86.7	77.1	75.2
✓		0.771	91.5	52.9	49.3	44.5
		1.000	15.4	13.6	12.9	12.1

Table 5: Ablation study of SD-attack and SIIR assessing protection efficacy and generalization using REFace on VGGFace2-HQ.

SD-EOT	SR-JPEG	Clean		Purification		JPEG		Crop & rotate	
		SSIM‡ ↓	DSR↑	SSIM‡ ↓	DSR↑	SSIM‡ ↓	DSR↑	SSIM‡ ↓	DSR↑
✓	✓	0.771	91.5	0.777	83.2	0.775	87.1	0.773	90.3
✓		0.772	90.7	0.783	82.9	0.794	80.5	0.774	89.2
	✓	0.773	90.3	0.780	83.3	0.776	86.8	0.795	80.4
		0.771	91.5	0.784	81.2	0.819	57.6	0.807	70.2

Table 6: Ablation study of SD-EOT and SR-JPEG in SGD-Guard using REFace on VGGFace2-HQ.

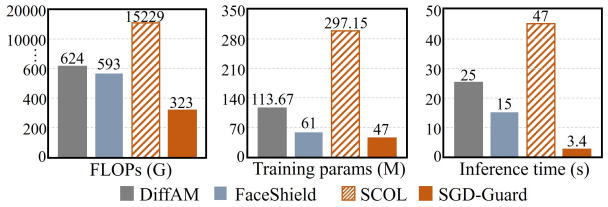

Figure 8: Complexity comparison of proactive defense methods.

Table 6 ablates the robustness of SD-EOT and SR-JPEG to post-processing under the same setting as Table 5. Using only SD-EOT achieves robustness to crop-and-rotate transformations, while adding SR-JPEG strengthens robustness to JPEG compression. Combining SD-EOT and SR-JPEG yields the most reliable protection, maintaining consistently low SSIM and high DSR values under various post-processing variants.

Complexity. Figure 8 shows SGD-Guard is the most efficient defense in FLOPs, parameters, and inference time. DiffAM is costly due to diffusion inversion and repeated U-Net passes, while FaceShield increases cost by optimizing diffusion cross-attention. In contrast, SGD-Guard freezes CLIP/identity encoders and optimizes only a compact pixel-space perturbation with similarity-based objectives. It uses a single-step latent diffusion purifier with LoRA adapters [Lei *et al.*, 2025] and reduces inference time by parallelizing computation and precomputing SIIR and SD-EOT directions offline. A discussion of limitations is provided in *Appendix E*.

5 Conclusion

This study proposes SGD-Guard, a proactive defense strategy that provides robust, generalized protection against face-swapping deepfakes. By integrating facial attribute and identity spaces, SGD-Guard identifies distortion directions and learns perturbations that steer representations toward a protected state. Furthermore, diffusion-based purification and post-processing techniques prevent feature shifts in the latent space to counteract attacks intended to neutralize perturbations, ensuring resilience against image transformations. The proposed method achieves generalized defense across diverse model backbones in a time-efficient manner.

Acknowledgements

This work was supported by the IITP grant funded by the Korea government (MSIT) (RS-2022-00156287, RS-2023-00256629, RS-2024-00437718, RS-2026-25619158).

Contribution Statement

Seung-hyeok Back and Do Hyun Ki contributed equally to this work.

References

[Athalye *et al.*, 2018] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[Baliah *et al.*, 2025] Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1062–1071. IEEE, 2025.

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

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

[Deng *et al.*, 2019] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[Dong *et al.*, 2023] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 18:2596–2608, 2023.

[Douze *et al.*, 2024] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

[Hu *et al.*, 2022] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15014–15023, 2022.

[Hu *et al.*, 2024] Cong Hu, Yuanbo Li, Zhenhua Feng, and Xiaojun Wu. Toward transferable attack via adversarial diffusion in face recognition. *IEEE Transactions on Information Forensics and Security*, 19:5506–5519, 2024.

[Huang *et al.*, 2022] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 989–997, 2022.

[Jeong *et al.*, 2025] Jaehwan Jeong, Sumin In, Sieun Kim, Hannie Shin, Jongheon Jeong, Sang Ho Yoon, Jaewook Chung, and Sangpil Kim. Faceshield: Defending facial image against deepfake threats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10364–10374, 2025.

[Jia *et al.*, 2022] Shuai Jia, Bangjie Yin, Taiping Yao, Shouhong Ding, Chunhua Shen, Xiaokang Yang, and Chao Ma. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. *Advances in Neural Information Processing Systems*, 35:34136–34147, 2022.

[Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

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

[Komkov and Petiushko, 2021] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th international conference on pattern recognition (ICPR)*, pages 819–826. IEEE, 2021.

[Korshunova *et al.*, 2017] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017.

[Lee *et al.*, 2023] Isack Lee, Eungi Lee, and Seok Bong Yoo. Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1536–1546, 2023.

[Lee *et al.*, 2025] Eungi Lee, Jae Hyun Yoon, and Seok Bong Yoo. Scol: Style code orchestration in latent space for proactive face-swapping defense. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11472–11481, 2025.

[Lei *et al.*, 2025] Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24331–24340, 2025.

[Lestriandoko *et al.*, 2022] Nova Hadi Lestriandoko, Raymond Veldhuis, and Luuk Spreeuwers. The contribution

of different face parts to deep face recognition. *Frontiers in Computer Science*, 4:958629, 2022.

[Lyu *et al.*, 2023] Yueming Lyu, Yue Jiang, Ziwen He, Bo Peng, Yunfan Liu, and Jing Dong. 3d-aware adversarial makeup generation for facial privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13438–13453, 2023.

[Ma *et al.*, 2023] Haotian Ma, Ke Xu, Xinghao Jiang, Zeyu Zhao, and Tanfeng Sun. Transferable black-box attack against face recognition with spatial mutable adversarial patch. *IEEE Transactions on Information Forensics and Security*, 18:5636–5650, 2023.

[Meng *et al.*, 2021] Qiang Meng, Chixiang Zhang, Xiaoqiang Xu, and Feng Zhou. Learning compatible embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9939–9948, 2021.

[Nie *et al.*, 2022] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[Qu *et al.*, 2024] Zuomin Qu, Zuping Xi, Wei Lu, Xiangyang Luo, Qian Wang, and Bin Li. Df-rap: A robust adversarial perturbation for defending against deepfakes in real-world social network scenarios. *IEEE Transactions on Information Forensics and Security*, 19:3943–3957, 2024.

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

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Sun *et al.*, 2024] Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, and Yongdong Zhang. Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24584–24594, 2024.

[Tao *et al.*, 2025] Renshuai Tao, Ziheng Qin, Yifu Ding, Chuangchuang Tan, Jiakai Wang, and Wei Wang. Unlocking the potential of lightweight quantized models for deepfake detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 520–528, 2025.

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

[Wang *et al.*, 2020] Chien-Yi Wang, Ya-Liang Chang, Shang-Ta Yang, Dong Chen, and Shang-Hong Lai. Unified representation learning for cross model compatibility. *arXiv preprint arXiv:2008.04821*, 2020.

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

[Wang *et al.*, 2022] Run Wang, Ziheng Huang, Zhikai Chen, Li Liu, Jing Chen, and Lina Wang. Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 761–767, 2022.

[Wang *et al.*, 2025] Tianyi Wang, Shuaicheng Niu, Harry Cheng, Xiao Zhang, and Yinglong Wang. Nullswap: Proactive identity cloaking against deepfake face swapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9945–9954, 2025.

[Xue *et al.*, 2023] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. *Advances in Neural Information Processing Systems*, 36:2894–2921, 2023.

[Yuan *et al.*, 2022] Zheng Yuan, Jie Zhang, and Shiguang Shan. Adaptive image transformations for transfer-based adversarial attack. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022.

[Zhang *et al.*, 2024] Yunming Zhang, Dengpan Ye, Caiyun Xie, Long Tang, Xin Liao, Ziyi Liu, Chuanxi Chen, and Jiacheng Deng. Dual defense: Adversarial, traceable, and invisible robust watermarking against face swapping. *IEEE Transactions on Information Forensics and Security*, 19:4628–4641, 2024.

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

[Zheng *et al.*, 2022] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709, 2022.