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

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

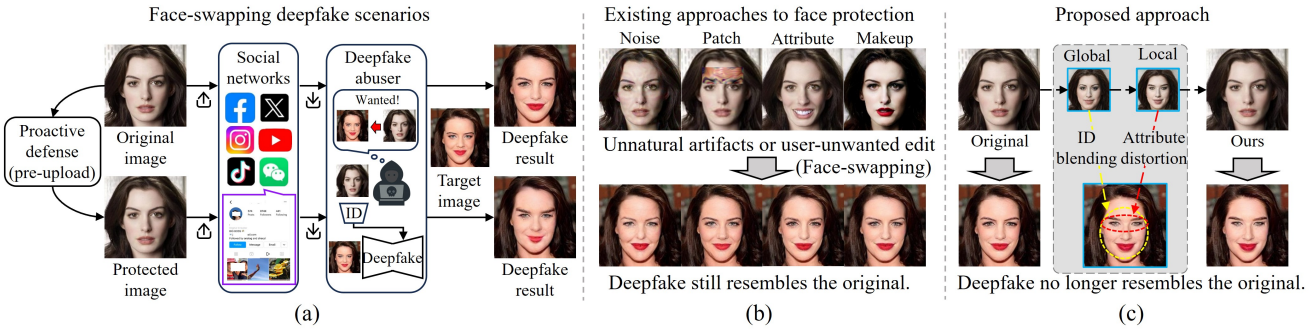

Figure 1. (a) Illustration of how proactive defense prevents unauthorized deepfake generation from publicly shared facial images. (b) Limitations of existing face protection methods. Top: protected images; bottom: corresponding face-swapping results. (c) Proposed defense combining global identity blending and local attribute distortion.

Abstract

Face-swapping deepfakes allow realistic identity transfer, which can serve creative purposes but increases the risk of identity abuse. A proactive defense aims to prevent deepfake creation by obstructing identity feature extraction from input images, essential for identity-driven face-swapping. Existing proactive defense approaches aim to protect faces by hindering accurate identity feature extraction, but tend to introduce visible artifacts and fail to degrade the visual quality of the face-swapping deepfakes. We propose a proactive face-swapping defense using identity blending and attribute distortion (DeepProtect) that integrates global identity fusion in the latent space and local prompt-driven adversarial watermarking to address these problems. We dilute distinct identity representations by channel-wise blending of multiple identities in the latent space and optimizing the generator for visual consistency. The proposed approach distorts facial components in the identity space, directly influencing how faces are reconstructed in deepfakes. Our approach applies semantic directions derived from user-provided text prompts to embed imperceptible adversarial watermarks that selectively distort facial attributes, affecting the visual fidelity of deepfake results. The proposed method hinders face-swapping deepfakes while preserving the perceptual quality of the protected images, offering a robust and practical solution for facial privacy protection. The experimental results reveal that DeepProtect effectively defends against face-swapping deepfakes while preserving visual consistency. This code is available at https://github.com/BACKAI/DeepProtect.

1. Introduction

Face-swapping is a type of deepfake in which the facial identity of one person is replaced. Recent advances in face-swapping deepfakes have introduced identity feature-based approaches [4, 5, 27, 36, 38, 51, 53], enabling generation for unseen target images without additional data or training. While these advancements enable applications in film production, digital avatars, and gaming, they also raise serious concerns over unauthorized impersonation, misinformation, and defamation.

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

Recent research has focused on two defense paradigms to mitigate these risks: deepfake detection and proactive defense. While detection methods [25, 26, 28, 39, 41, 47] identify manipulated content after distribution, they often fail to prevent early spreading. In contrast, proactive defenses [7, 16, 33, 40] block identity feature extraction at the source, offering continuous offline (pre-upload) protection without requiring real-time processing. Figure 1(a) contrasts two scenarios: (1) uploading the original facial image and (2) uploading a proactively protected facial image. In the former, attackers can extract identity features for impersonation. In the latter, identity protection significantly reduces this risk before image sharing.

As shown in Fig. 1(b), existing proactive defenses based on noise [16], patch [33], attribute manipulation [7], makeup [40] produce visible artifacts or fail to preserve the original appearance. Although these methods reduce identity similarity by shifting the feature representation globally, the modified representation often corresponds to a visually similar individual. The resulting deepfakes remain perceptually close to the original, risking facial privacy.

To overcome the limitations, we propose a unified defense framework, as presented in Fig. 1(c), which protects the source image while maintaining visual consistency. The proposed framework combines two critical strategies: a generation-based approach that globally dilutes identity representation and a local, attribute-based defense that distorts critical facial components. This identity blending weakens the overall identity, enhancing the effectiveness of localized distortions. Building on the $\mathcal{W}^+$ space which captures identity and appearance information [20], we introduce a strategic style vector fusion to achieve identity blending. The generator is further optimized using an identity lock loss to ensure that the diluted identity maintains visual alignment with the original input. Additionally, we apply partial fine-tuning to the generator via low-rank adaptation (LoRA) [13], which improves efficiency of optimization and significantly reduces computational overhead.

Next, we introduce the attribute distortion method as an adversarial watermarking strategy to disrupt face-swapping deepfakes. The goal is to selectively distort specific facial attributes according to user-defined prompts (e.g., eyes, as shown in Fig. 1(b)). This is achieved by identifying vector directions corresponding to key facial components within the entangled identity space. Using these identified vector directions, the original identity feature is encouraged to diverge from its original attribute state. Then, imperceptible adversarial watermarks are embedded into the identity blended image, selectively distorting these local components. The proposed method offers several advantages. First, it effectively defends against face-swapping attacks while maintaining user satisfaction via minimal visual interference. Second, by targeting user-specified local regions, the distortions enable the deepfake results to incorporate semantic cues, which can be employed for downstream tasks such as post-hoc detection and traceability. We summarize the contributions below:

- We propose a proactive defense framework that dilutes facial identity while preserving appearance realism. Unlike conventional GAN-based manipulation, our approach reverses the usual paradigm by altering identity in the $\mathcal{W}^+$ space without compromising visual fidelity.
- We introduce an adversarial watermarking technique that targets identity-critical facial components, enabling user-guided, localized disruption of deepfakes.
- We develop a lightweight identity-lock optimization using LoRA in StyleGAN middle layers, reducing training parameters and computation for practical deployment.

2. Related Work

2.1. Face-swapping Deepfakes

Face-swapping aims to replace a face in an image. Traditional face-swapping models including non-identity-driven ones [23] require separate training for source-target pair data, limiting generalization to unseen identities [3]. Recent subject-agnostic methods [5, 27, 36, 51, 53] remove the dependence on identity-specific decoders by directly embedding source identity features into target representations, enabling more scalable face-swapping via learning shared identity spaces. This work aims to disrupt the identity transfer process by altering the extracted identity cues before fusing them into the target image, intentionally inducing semantic divergence in the synthesized output.

2.2. Proactive Deepfake Defense

Proactive defense against deepfakes has gained increased attention, with adversarial attacks emerging to disrupt synthetic content generation. Watermark-based methods have been widely explored to embed adversarial perturbations into facial images, disrupting deepfake generation [15, 32, 50]. However, they are typically built on white-box assumptions, limiting their applicability in real-world. Recent model-agnostic defenses [7, 11, 33] are mainly optimized for encoder-decoder architectures and remain ineffective against identity-driven face-swapping. Transferable noise-based [8, 12, 16, 45] and patch-based [22, 29, 46] methods often produce visible artifacts. Editing-based approaches [14, 17, 40, 49] aim to improve realism but still introduce undesirable changes for users. A recent effort has increasingly explored latent-space identity obfuscation to weaken identity cues [24]. However, this method requires an exhaustive two-stage, full-StyleGAN optimization to obfuscate identity and align visual appearance, resulting in substantial computational overhead. Moreover, it is limited to global identity manipulation and does not explicitly

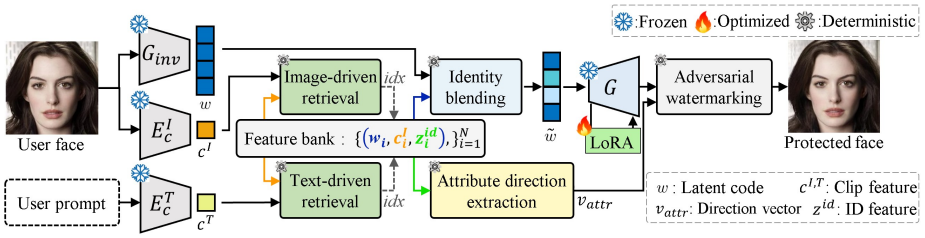

Figure 2. Illustration of DeepProtect. The process applies global identity blending via style vector fusion, followed by localized attribute distortion using prompt-driven adversarial perturbations along semantic directions.

protect local regions that are critical for identity recognition. In contrast, we introduce a text-guided adversarial watermarking framework that enables efficient, prompt-guided perturbations through partial adaptation of identity-critical layers. Our method effectively degrades deepfake outputs while maintaining superior visual fidelity.

3. Method

3.1. Overview

We propose a proactive defense against face-swapping deepfakes by altering identity cues and injecting prompt-guided, attribute-specific distortion watermarks. As illustrated in Fig. 2, the method comprises two stages: identity blending to dilute identity information and adversarial watermarking to distort attributes in generated deepfakes. First, an input image is encoded into a latent code (w) and a CLIP feature (c^I) [34]. Using c^I, latent codes of visually similar samples are retrieved from a feature bank and blended with w on a per-style-channel basis to generate $\tilde{w}$. To maintain visual consistency, only identity-relevant generator layers are fine-tuned via an identity lock loss. Concurrently, we estimate a text-driven attribute direction (v_{attr}) in the identity space. An adversarial watermark is then embedded into the optimized image along this direction to induce targeted attribute distortions in the deepfake result. At inference, optimization is performed with an identity encoder as a surrogate, while deepfake models remain black-box.

3.2. Identity Blending

This process begins by obtaining the latent code w of the input image using a GAN inversion model [43] (Fig. 3) to dilute the identity of the input face without compromising visual consistency. The latent representation is embedded in the extended latent space $\mathcal{W}^+$ [1], permitting layerwise latent codes for fine-grained image control. Then, the CLIP embedding space retrieves latent codes of visually similar but identity-different references from a feature bank. The feature bank is built with three types of features: identity features, CLIP image features c^I, and latent codes. By comparing c^I of the original input image with the retrieved CLIP

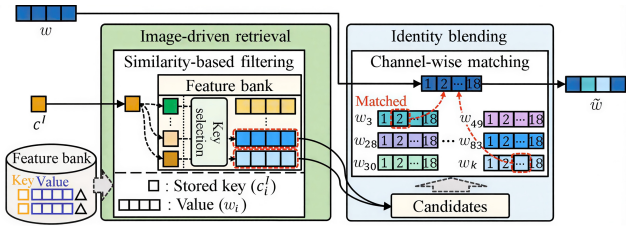

Figure 3. Identity blending process.

features c_i^I, this approach constructs a candidate set $\mathcal{C}$ of latent codes that preserves the appearance similarity to the input while enabling identity dilution:

$$\mathcal{C} = \left\{ w_i \;\middle|\; \cos(c^I, c_i^I) \geq \tau,\ \forall i \in \{1, \ldots, N\} \right\}, \tag{1}$$

where c^I and c_i^I denote the CLIP image embeddings of the input and the i-th sample in the feature bank, respectively, and N denotes the number of image-derived feature entries in the feature bank. This approach constructs $\mathcal{C}$ by retrieving the latent codes w_i corresponding to samples whose CLIP features have cosine similarity to the input above the threshold τ. The rationale behind using CLIP-based retrieval is illustrated in Sec. 6.1 of the Supplement Material (SM) due to space constraints.

To prevent unauthorized identity extraction, the model performs identity blending by modifying specific style vectors of w in the $\mathcal{W}^+$ space. Each w consists of 18 style vectors of 512 dimensions each, among which the middle layers are known to capture identity-critical information [2]. We adopt layers three through seven as representative middle layers. To dilute the original identity, each style vector in these layers is replaced with the most similar counterpart from the candidate set $\mathcal{C}$, based on the cosine similarity:

$$\tilde{w}[l] = \arg\max_{w_j \in \mathcal{C}} \cos\left(w[l],\ w_j[l]\right), \tag{2}$$

where $w[l]$ denotes the style vector in layer l in w, and j represents the index of w in $\mathcal{C}$. The remaining layers retain the original. This channel-wise substitution fuses features from semantically similar identities, obscuring identity-specific characteristics without converging to any single identity.

Selective modification of identity-critical mid-level layers reduces the optimization overhead, improving the practical applicability. The $\tilde{w}$ produces facial images with diluted identity features, weakening the association with any specific individual's characteristics. Our approach performs channel-wise fusion of style components and transfers CLIP-space semantic cues into the identity space, effectively diluting the original identity to prevent misuse. Such relaxation alleviates rigid identity constraints and enables more pronounced, yet coherent, attribute distortions.

3.3. Generator Optimization

The generator is optimized to preserve visual consistency with the input image while applying identity blending. Replacing the original latent code with a blended one disrupts the identity consistency, the basis of the proposed method. Building on the prior optimization-based approach [35] that minimizes pixel-level and perceptual losses, our method further introduces an identity-lock constraint that enforces alignment with the diluted identity. This encourages the generator to retain the intended identity obfuscation during optimization. Hence, this work defines an identity-lock loss $\mathcal{L}_{\text{id-lock}}$ that measures the cosine similarity between the identity features of the initial generator output and those of the optimized result. Given a blended latent code $\tilde{w}$, the identity-lock loss is defined as follows:

$$\mathcal{L}_{\text{id-lock}} = 1 - \cos\left(E_{id}(G_{\text{init}}(\tilde{w})), E_{id}(G(\tilde{w}))\right), \quad (3)$$

where $E_{id}(\cdot)$ is a pretrained identity encoder [6], and G_{init} and G denote the initial and fine-tuned generators. The ID-lock loss (Eq. 3) penalizes deviation from the diluted identity features of the blended latent code, preserving the diluted identity representation and preventing overfitting from reintroducing original identity cues.

This approach replaces only the middle-style vectors of w to dilute identity, which are known to encode identity information. Accordingly, only the middle layers of StyleGAN are fine-tuned, sufficing to maintain consistency with the input image while significantly reducing the number of trainable parameters. The coarse style layers and affine transforms in the fine layers remain frozen; thus their outputs remain constant during training. This method caches these outputs at initialization and reuses them throughout the process, improving optimization efficiency.

Despite the partial tuning, the model still requires approximately 11M trainable parameters. We further reduce this overhead by adopting LoRA. The model applies LoRA to the affine transform module, generating style vectors from the latent code and modulating the convolutional weights. The original weight matrix $\theta_0 \in \mathbb{R}^{512\times b}$, where b is the number of output channels in the affine transformation, is decomposed into two smaller matrices: $\theta_1 \in \mathbb{R}^{512\times r}$ and $\theta_2 \in \mathbb{R}^{r\times b}$, where the LoRA ranks $r \ll \min(512, b)$.

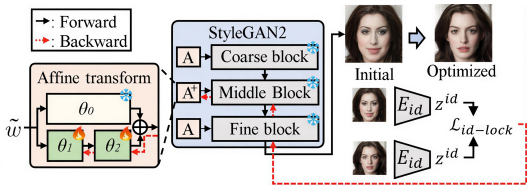

Figure 4. Partial tuning of generator.

During optimization, θ_0 is frozen while only θ_1 and θ_2 are updated, following the tuning strategy shown in Fig. 4. The objective is defined over the LoRA parameters:

$$\theta^* = \underset{\theta=\{\theta_1,\theta_2\}}{\arg\min}\ \mathcal{L}_2 + \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{id-lock}}\mathcal{L}_{\text{id-lock}}, \quad (4)$$

where θ denotes the LoRA parameters of the generator and $\lambda_{\text{id-lock}}$ balances the identity-lock loss against the reconstruction losses ($\mathcal{L}_2$ and $\mathcal{L}_{\text{LPIPS}}$) [35].

3.4. Attribute Distortion

We design a controllable attribute distortion in the identity embedding space. This design is motivated by the observation that selectively perturbing identity-critical regions (e.g., eyebrows, eyes, nose, and lips) can substantially distort face-swapped outputs, as qualitatively demonstrated in Sec. 7 of the SM. Figure 5 illustrates that the proposed method enables users to define target facial features for alteration using text prompts. These prompts are associated with pertinent visual features through the application of the joint multimodal embedding space of CLIP. We employ facial representation learning (FaRL) [52], a CLIP-aligned model optimized for facial representation embedding tasks, to enhance the fidelity of facial attribute representations.

Text-guided Attribute Vector Retrieval. Although the identity embedding space primarily captures holistic identity information, it involves the entanglement of facial components without explicit separation. We aim to find semantically meaningful directions corresponding to specific facial components without additional training, even within an entangled space. We propose a retrieval-based approach to achieve this goal, applying the joint image–text embedding space of CLIP. Given a user-defined text prompt (e.g., 'lips'), the CLIP text encoder E_c^T extracts the text feature $c^T = E_c^T(I)$ and computes the dot product between the embedding text feature and the CLIP features of facial images stored in the feature bank:

$$s_i = c_i^I \cdot c^T, \quad \forall i \in \{1, ..., N\}. \quad (5)$$

The alignment scores, $S = \{s_i\}_{i=1}^N$, indicate how strongly the queried attribute is semantically expressed in each image. To efficiently retrieve samples under large N, we use approximate nearest neighbor search (FAISS [9]) to rank c^I

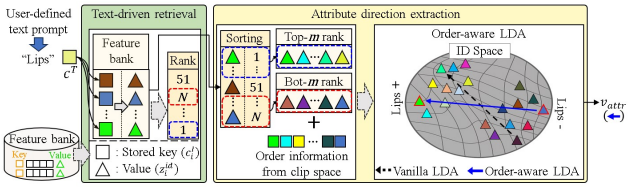

Figure 5. Text-guided discovery of attribute vectors.

by S and obtain the identity feature sets $\mathcal{Z}_k^{id}$ associated with strong and weak correlations to the target attribute.

This approach enables the extraction of attribute-specific directions directly in the entangled identity space, without relying on explicit supervision or additional model training. This approach selects the top-m (strongest) and bottom-m (weakest) identity features based on the ranking indices $\mathcal{R}_k$:

$$\mathcal{Z}_k^{id} = \left\{ z_i^{id} \mid i \in \text{argsort}(S)[\mathcal{R}_k] \right\}, \quad k \in \{\text{top}, \text{bot}\}, \tag{6}$$

where $\mathcal{R}_{\text{top}} = [:m]$ and $\mathcal{R}_{\text{bot}} = [-m:]$, and m denote the number of samples selected from each end of the ranking.

Attribute Direction Extraction in the Identity Space. The goal is to disrupt identity-based face-swapping models by manipulating identity features. To this end, we estimate a direction in the identity space that captures variations in the targeted attribute. We propose an order-aware linear discriminant analysis (LDA) objective that separates identity features with and without strong attribute expressions, while preserving their relative semantic order, inspired by LDA [42]. Given the identity features from the top and bottom groups identified via CLIP-based ordering, this approach optimizes the following objective:

$$v_{attr} = \arg\max_{v} \frac{v^\top S_B v}{v^\top (S_W + \lambda_R R) v}, \tag{7}$$

where S_B and S_W denote the between-class and within-class scatter matrices, calculated from $\mathcal{Z}_k^{id}$. The S_B reflects inter-group covariance, encouraging projection v to maximize separation, while S_W captures intra-group variance, promoting compactness. To preserve semantic strength ordering, we introduce the order-aware regularization R:

$$R = \sum_{i,j \in \{1,\ldots,2m\}} |\text{rank}_{\text{CLIP}}(i) - \text{rank}_{\text{CLIP}}(j)| \cdot (z_i^{id} - z_j^{id}). \tag{8}$$

This regularization preserves the relative ordering derived from CLIP scores within the identity space. The rank difference $|\text{rank}_{\text{CLIP}}(i) - \text{rank}_{\text{CLIP}}(j)|$ encourages the preservation of the relative ordering observed in the CLIP space along the projection direction in the identity space. The expression $(z_i^{id} - z_j^{id})$ measures differences in the identity feature space and encourages the arrangement of the identity space to align with the semantic ranking derived from the CLIP space. In Eq. (7), λ_R balances the separation of top and bottom groups with the preservation of semantic order. We solve this optimization using the LSQR algorithm [31], which efficiently handles the generalized Rayleigh quotient form. Our order-aware LDA, a conceptual advance over conventional LDA, enforces the semantic ordering inherited from CLIP features while computing discriminative boundaries in identity space. LSQR then yields the projection vector v_{attr} that maximizes group separation along this order-preserving direction, serving as the manipulation vector for targeted attribute distortion.

Adversarial Watermarking. We propose an adversarial watermark, W_{attr}, which distorts deepfake output by amplifying or suppressing specific facial components. The target direction is automatically determined as $v_{target} = -\text{sign}(p) \cdot v_{attr}$, where p is the projection of the identity feature z^{id} onto the attribute direction v_{attr}, encouraging deviation from the original attribute state. The watermark is iteratively updated via a gradient-based method to maximize displacement along v_{target}. At each step t, the adversarial image is $I_t = I + W_{attr}$, and its identity feature z_t is extracted. The optimization objective is defined as follows:

$$\mathcal{L}(z_t, v_{target}) = z_t \cdot v_{target}, \tag{9}$$

encouraging alignment with the distortion direction. The watermark is updated using a sign-based gradient step:

$$W_{attr} \leftarrow W_{attr} + \alpha \cdot \text{sign}\left(\nabla_{W_{attr}} \mathcal{L}\right), \tag{10}$$

where α denotes step size, and W_{attr} is projected onto the ℓ_∞-ball to ensure imperceptibility:

$$W_{attr} \leftarrow \max(\min(W_{attr}, \epsilon), -\epsilon), \tag{11}$$

where ϵ bounds the perturbation magnitude. Leveraging global identity blending and optimization, our method creates a favorable condition for embedding adversarial watermark. As a result, the watermark induces a semantic shift in the identity space without introducing perceptible artifacts.

4. Experiment

4.1. Experimental Setup

Datasets and face-swapping models. We evaluate on CelebA-HQ [19] and VGGFace2-HQ [4], comprising 30K celebrities and 9,630 identities with 1.3M facial images, respectively. Both datasets are known for their diversity, enabling generalized evaluation. CelebA-HQ is used for cross-validation against the fixed VGGFace2-HQ bank, and no bank rebuilding is needed for practical use. For quantitative evaluations, we used 1,000 sources and 10,000 generated images with random identities. For comparison, we used state-of-the-art (SOTA) face-swapping models, including SimSwap [3], FaceDancer [36], BlendFace [38], FaceSwapper [27] and DiffFace [21].

Evaluation metrics. We evaluated visual fidelity and defense effectiveness using four metrics: PSNR, SSIM, identity score matching (ISM) [44], and defense success rate (DSR) [33]. PSNR and SSIM assess visual similarity—higher for original-protected pairs and lower for their deepfakes. For clarity, † denotes comparisons between original and protected images, and ‡ indicates those between their deepfakes. ISM measures identity similarity between original and protected images, while DSR quantifies identity disruption in generated deepfakes. We also introduce the preservation-disruption score (PDS), defined as the SSIM difference between original-protected pairs and their deepfakes, to capture the trade-off between visual fidelity and identity disruption. Higher PDS values indicate stronger disruption with minimal visual degradation.

Baselines. We benchmark DeepProtect against seven publicly available SOTA proactive defenses, covering both adversarial noise-based and makeup-based approaches. CMUA-Watermark [15], DF-RAP [33], and FaceShield [16] employ adversarial noises, whereas AMT-GAN [14], CLIP2Protect [37] and DiffAM [40], WDP [10] adopt makeup-based strategies. We evaluate two variants of our method: an attribute-only version to assess the standalone effectiveness of distortion, and the full DeepProtect combining identity blending and attribute manipulation. The attribute-only variant is lightweight and easily deployable. We apply only proactive manipulation without prior knowledge to ensure fair comparisons. Unless otherwise specified, our method defends the image based on the 'nose' text. For all noise-based methods, including DeepProtect, the perturbation budget (ϵ) is constrained to a ℓ_∞ bound of 0.02 to preserve the visual fidelity of the source image. Further justification for the constraint is provided in Sec. 7.2 of SM. Results are averaged over five runs to ensure statistical robustness and reduce variance.

Implementation details. We use FaRL [52], a CLIP-style model trained on LAION-Face20M and optimized for the facial domain, ensuring that retrieved features effectively reflect identity components. The feature bank is constructed from 4,605 distinct identities in VGGFace2-HQ, creating a total of $N = 4,605$ samples, each associated with pre-extracted latent codes, CLIP features, and identity features. The order-aware LDA method groups the features into top and bottom sets with $m = 30$. StyleGAN2 is pretrained on FFHQ [18], optimized with a learning rate of 3×10^{-4} and LoRA with rank $r = 8$. None of models have seen CelebA-HQ or VGGFace2-HQ during training, ensuring unbiased evaluation under real-world conditions. We use pretrained E4E and StyleGAN2 without additional training. In the experiments, we set τ=0.75, $\lambda_{\text{id-lock}}$=0.1, and λ_R=1. Details on thresholds are provided in Sec. 6 of the SM.

Table 1. Quantitative comparison of original and protected images (source evaluation).

Method	Publication	CelebA-HQ			VGGFace2-HQ		
		PSNR† ↑	SSIM† ↑	ISM↓	PSNR† ↑	SSIM† ↑	ISM↓
CMUA-Watermark [15]	AAAI'22	35.14	0.925	0.438	35.06	0.922	0.431
DF-RAP [33]	TIFS'24	35.44	0.939	0.427	35.21	0.934	0.420
FaceShield [16]	ICCV'25	32.63	0.933	0.228	31.58	0.930	0.268
AMT-GAN [14]	CVPR'22	24.95	0.859	0.307	24.56	0.853	0.303
CLIP2Protect [37]	CVPR'23	18.07	0.660	0.249	17.84	0.653	0.247
DiffAM [40]	CVPR'24	17.96	0.608	0.244	17.89	0.633	0.240
WDP [10]	CVPR'25	25.35	0.773	0.289	25.12	0.747	0.273
DeepProtect (Attribute)	Ours	**35.89**	**0.944**	0.419	**35.72**	**0.941**	0.411
DeepProtect (Combined)		32.02	0.902	**0.201**	31.57	0.904	**0.198**

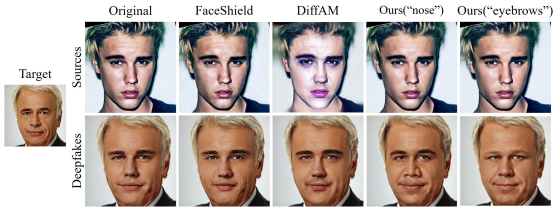

Figure 6. Visualization of protected images and corresponding SimSwap results.

4.2. Baseline Comparison

Evaluation of protected images. Table 1 reports visual consistency (PSNR, SSIM) and identity protection (ISM) of protected images. Noise-based methods yield high PSNR/SSIM due to strict perturbation constraints, but protection is ineffective against deepfake generation. DeepProtect (Attribute-only) addresses this limitation through targeted attribute distortion, achieving a better balance between visual fidelity and identity masking. Makeup-based methods (AMT-GAN, CLIP2Protect, DiffAM) offer stronger identity disruption but at the cost of noticeable quality degradation. In contrast, DeepProtect (Combined) achieves the lowest ISM while maintaining visual quality, demonstrating effective identity protection without compromising fidelity.

Deepfake output evaluation. Table 2 presents a quantitative comparison of proactive defenses against face-swapping deepfakes. Compared to noise-based methods, DeepProtect (Attribute-only) achieves higher performance by focusing on facial attribute representations that more strongly affect deepfake generation. DeepProtect (Combined) further boosts defense by integrating identity blending and attribute distortion, achieving the highest DSR while simultaneously reducing PSNR and SSIM, indicating greater perceptual divergence from the original. These results demonstrate that the proposed method offers comprehensive protection across various face-swapping models.

Qualitative comparison. Figure 6 compares DeepProtect with two SOTA proactive defenses: FaceShield and DiffAM. The first row shows protected images, and the second

Table 2. Quantitative comparison of baselines for face-swapping models (deepfake evaluation).

Method	SimSwap (ACM MM'20) [3]			FaceDancer (WACV'23) [36]			BlendFace (ICCV'23) [38]			FaceSwapper (TPAMI'24) [27]			DiffFace (PR'25) [21]		
	PSNR‡ ↓	SSIM‡ ↓	DSR↑	PSNR‡ ↓	SSIM‡ ↓	DSR↑	PSNR‡ ↓	SSIM‡ ↓	DSR↑	PSNR‡ ↓	SSIM‡ ↓	DSR↑	PSNR‡ ↓	SSIM‡ ↓	DSR↑
CelebA-HQ dataset															
CMUA-Watermark	27.10	0.851	41.0	26.84	0.850	41.2	27.31	0.876	35.1	27.19	0.843	43.5	28.55	0.868	37.1
DF-RAP	26.73	0.840	42.7	26.55	0.841	42.5	27.20	0.872	36.9	26.94	0.838	44.1	28.40	0.863	38.0
FaceShield	22.78	0.752	81.6	23.91	0.764	84.1	24.62	0.814	72.3	20.76	0.783	82.5	32.05	0.811	83.1
AMT-GAN	23.96	0.802	67.2	24.80	0.818	63.5	25.16	0.825	60.9	24.79	0.819	63.3	23.64	0.790	73.9
CLIP2Protect	23.39	0.781	79.4	24.13	0.809	78.9	24.83	0.810	73.7	24.22	0.805	79.6	23.11	0.766	85.4
DiffAM	23.11	0.779	79.7	23.95	0.784	80.8	24.07	0.804	74.0	24.37	0.809	79.3	23.05	0.763	85.8
WDP	29.04	0.900	38.6	30.11	0.925	32.7	33.60	0.974	24.1	31.26	0.952	26.6	28.84	0.932	28.4
DeepProtect (Attribute)	26.01	0.824	60.5	25.95	0.820	60.4	26.38	0.832	50.9	25.08	0.825	60.5	25.88	0.801	67.4
DeepProtect (Combined)	**21.09**	**0.710**	**94.8**	**21.75**	**0.712**	**94.3**	**22.91**	**0.730**	**92.2**	**20.18**	**0.711**	**94.7**	**21.02**	**0.704**	**96.7**
VGGFace2-HQ dataset															
CMUA-Watermark	27.07	0.852	40.8	26.81	0.848	41.5	27.29	0.877	35.0	27.13	0.842	43.6	28.58	0.869	37.0
DF-RAP	26.71	0.843	42.4	26.52	0.838	42.8	27.23	0.873	36.6	26.91	0.835	44.2	28.43	0.864	37.8
FaceShield	22.72	0.748	81.3	23.45	0.761	84.4	24.41	0.818	72.8	17.92	0.784	82.9	29.20	0.814	82.6
AMT-GAN	23.89	0.805	66.8	24.78	0.817	63.8	25.19	0.826	60.5	24.76	0.816	63.7	23.61	0.793	73.5
CLIP2Protect	23.37	0.784	79.0	24.11	0.806	79.3	24.86	0.812	73.3	24.19	0.804	80.0	23.03	0.767	85.0
DiffAM	23.09	0.778	79.8	23.93	0.781	81.2	24.14	0.807	73.6	24.34	0.806	79.7	23.05	0.766	85.2
WDP	27.98	0.878	36.2	29.13	0.903	36.4	33.74	0.977	24.1	31.23	0.954	27.0	28.61	0.924	28.7
DeepProtect (Attribute)	25.98	0.823	60.5	25.92	0.819	60.8	26.41	0.835	50.5	25.05	0.823	60.9	25.85	0.804	67.1
DeepProtect (Combined)	**21.07**	**0.708**	**95.0**	**21.73**	**0.711**	**94.5**	**22.89**	**0.729**	**92.5**	**17.85**	**0.710**	**94.9**	**21.00**	**0.703**	**96.5**

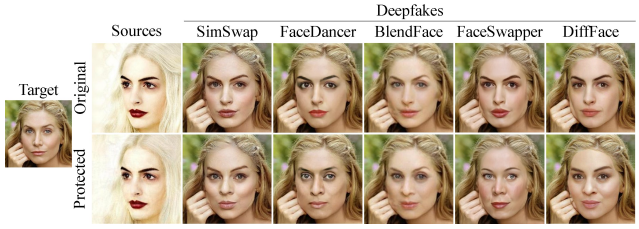

Figure 7. Evaluating protection across face-swapping models on the CelebA-HQ dataset.

Figure 8. Text-to-video deepfake results from original vs. protected images, showing protection beyond face-swapping.

presents face-swapped results using SimSwap. FaceShield adds adversarial noise, but this fails to provide natural protection in flat regions, and its subtle perturbations fail to affect deepfakes. DiffAM plausibly distorts deepfakes but visibly alters the source, harming visual fidelity. In contrast, DeepProtect uses identity blending to dilute identity cues and attribute distortion guided by prompts (e.g., 'nose', 'eyebrows') to induce localized changes. For instance, nose prompts enhance prominence, while eyebrow prompts reduce expression. These effects reflect v_{attr}, which automatically determines the optimal distortion direction. With the same perturbation budget as FaceShield, DeepProtect achieves stronger protection without compromising visual fidelity. Due to space constraints, the expanded results for Figs. 6 and 7 are provided in Sec. 8.4 of the SM.

4.3. Comprehensive Evaluation

Generalizability. Figure 7 presents the visual results for five deepfake models using a 'nose' prompt. The first row displays face-swapped outputs from the original sources, and the second row presents those from protected sources. The proposed method induces distortions around the nose region, degrading deepfake outputs across various models and demonstrating strong generalization. We further investigate how identity information is affected by our method; results on identity retrieval are provided in Sec. 9 of the SM.

Table 3. Human evaluation results on CelebA-HQ, based on a five-point MOS scale for visual fidelity and identity disruption.

Method	Source MOS↑	Deepfake MOS↓
DiffAM [40]	3.45	3.20
FaceShield [16]	4.7	3.82
DeepProtect	4.7	1.35

Beyond face-swapping, DeepProtect can defend against a broader range of deepfakes. As shown in Fig. 8, when a source is used in a SOTA diffusion-based deepfake generation model [48], DeepProtect preserves privacy by ensuring a clear visual difference in identity between videos generated from original and protected images.

Human evaluation. We conducted a user study with 10 participants to evaluate visual fidelity and identity protection using a five-point mean opinion score (MOS) scale (1: very poor, 5: excellent) on 100 images (20 protected, 80 deepfakes). Source MOS (higher is better) measures the visual fidelity of protected images, while deepfake MOS

Table 4. Ablation study of DeepProtect using SimSwap-generated deepfakes on CelebA-HQ.

Identity blending	Generator optimization	Attribute distortion	SSIM† ↑	ISM↓	PDS↑
✓			0.584	0.241	-0.191
✓	✓		0.915	0.249	0.136
		✓	0.944	0.419	0.120
✓		✓	0.571	0.194	-0.196
✓	✓	✓	0.902	0.201	0.192

Table 5. Complexity comparison of proactive deepfake defenses.

Method	FLOPs (G) ↓	Training params (M)↓	Inference time (s)↓
CLIP2Protect	171	30.37	23
DiffAM	624	113.67	25
FaceShield	593	0.00	15
DeepProtect	126 (72.65+45.12+8.25)	0.04	12 (1+10+1)

(lower is better) indicates the degree of identity disruption in generated deepfakes. As shown in Table 3, DeepProtect achieves the highest source MOS (4.7) and the lowest deepfake MOS (1.35), significantly outperforming DiffAM (3.45/3.20) and FaceShield (4.7/3.82) in preserving image quality and disrupting identity in deepfakes.

Ablation study. Table 4 shows an ablation study on SimSwap evaluating DeepProtect components. Applying only identity blending disrupts identity (ISM: 0.241) but reduces visual fidelity (SSIM: 0.584) due to noticeable appearance shifts. Adding generator optimization improves visual fidelity while maintaining identity disruption. The fourth row displays substantial identity disruption but reduced visual fidelity due to the excluded generator optimization. Combining all components achieves the highest identity disruption, deepfake degradation, and visual fidelity. Section 8 in the SM provides additional qualitative examples illustrating the effect of each component on deepfake outputs, and Sec. 9 of the SM details the effects of identity blending.

Computational complexity. Noise-based methods are excluded as they fail under the imperceptibility constraint. DiffAM, as a diffusion-based method, requires high FLOPs, many parameters, 25 seconds inference, and over 4 hours to train a new style. CLIP2Protect uses the CLIP encoder repeatedly, resulting in higher FLOPs and 23 seconds optimization time. FaceShield is training free, but heavy with diffusion UNet, CLIP, and detectors, so FLOPs are high. DeepProtect needs only a single CLIP encoding and partial generator tuning, lowering overhead. Identity blending takes under 1 second, generator tuning about 10 seconds until LPIPS converges, and attribute distortion under 1 second. LoRA-based tuning reduces trainable parameters for efficient optimization and inference.

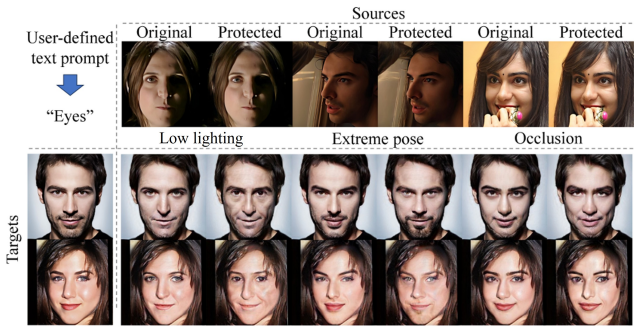

Figure 9. Deepfake outputs from original vs. DeepProtect-protected images under real-world challenging scenarios.

Table 6. Quantitative evaluation of DeepProtect's robustness under post-processing and adaptive attack conditions.

Post-processing	-	Gaussian noise ($\sigma = 0.08$)	Gaussian blur ($\sigma = 1$, kernel size 7×7)	JPEG compression (QF=25)	Adaptive Attack
DSR	94.8	93.7	93.5	93.5	93.2

4.4. Real-world Scenarios

To evaluate the robustness of DeepProtect, we tested it on images under challenging real-world conditions (Fig. 9). The VGGFace2-HQ images include low lighting, diverse facial angles, and occlusions, demonstrating effective mitigation of face-swapping deepfakes.

We report quantitative results under various post-processing conditions in Table 6. Experiments are conducted using SimSwap on CelebA-HQ, reporting the DSR of DeepProtect under each transformation. Despite common degradations such as Gaussian noise, blur, and JPEG compression, DeepProtect maintains over 93% DSR, including 93.2% under adaptive purification attacks [30]. This robustness demonstrates that identity protection through style-vector fusion is not only robust to low-level perturbations but also resistant to adaptive purification-based attacks that attempt to reverse our protection. Additional experiments are provided in Sec. 8.3 of the SM, pseudocodes in Sec. 10, and discussions on applicability, limitations, and future work in Sec. 11.

5. Conclusion

We propose a method to protect users against face-swapping deepfakes by performing identity blending in the $\mathcal{W}^+$ space to shield identity features and generating adversarial watermarks to distort user-specified attributes via text prompts. This approach safeguards users and mitigates the risk of unintended impersonation by enabling users to control attribute manipulation directly in deepfake outputs. Furthermore, this method preserves the visual fidelity of protected images while defending against face-swapping deepfakes.

Acknowledgements

This work was supported by the IITP grant funded by the Korea government (MSIT) (RS-2022-00156287, RS-2023-00256629, RS-2024-00437718) and by the ETRI grant funded by the Korean government (26ZK1100).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 3

[2] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8001–8010, 2023. 3

[3] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 2, 5, 7

[4] Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. Simswap++: Towards faster and high-quality identity swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):576–592, 2023. 1, 5

[5] Kaiwen Cui, Rongliang Wu, Fangneng Zhan, and Shijian Lu. Face transformer: Towards high fidelity and accurate face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 668–677, 2023. 1, 2

[6] 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. 4

[7] 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. 2

[8] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7714–7722, 2019. 2

[9] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024. 4

[10] Salar et al. Enhancing facial privacy protection via weakening diffusion purification. In *CVPR*, 2025. 6

[11] Jiazhi Guan, Yi Zhao, Zhuoer Xu, Changhua Meng, Ke Xu, and Youjian Zhao. Adversarial robust safeguard for evading deep facial manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 118–126, 2024. 2

[12] Cong Hu, Yuanbo Li, Zhenhua Feng, and Xiaojun Wu. Towards transferable attack via adversarial diffusion in face recognition. *IEEE Transactions on Information Forensics and Security*, 2024. 2

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

[14] 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. 2, 6

[15] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI conference on artificial intelligence*, pages 989–997, 2022. 2, 6

[16] 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. 2, 6, 7

[17] 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. 2

[18] Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019. 6

[19] 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. 5

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[21] 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. 5, 7

[22] 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. 2

[23] 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. 2

[24] 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 In-*

ternational Conference on Multimedia, pages 11472–11481, 2025. 2

[25] Eun-Gi Lee, Isack Lee, and Seok-Bong Yoo. Cluecatcher: Catching domain-wise independent clues for deepfake detection. *Mathematics*, 11(18):3952, 2023. 2

[26] Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. Freqblender: Enhancing deepfake detection by blending frequency knowledge. In *Advances in Neural Information Processing Systems*, pages 44965–44988. Curran Associates, Inc., 2024. 2

[27] Qi Li, Weining Wang, Chengzhong Xu, Zhenan Sun, and Ming-Hsuan Yang. Learning disentangled representation for one-shot progressive face swapping. *IEEE transactions on pattern analysis and machine intelligence*, 2024. 1, 2, 5, 7

[28] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16815–16825, 2024. 2

[29] 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. 2

[30] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827. PMLR, 2022. 8

[31] Christopher C Paige and Michael A Saunders. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1): 43–71, 1982. 5

[32] Tong Qiao, Bin Zhao, Ran Shi, Meng Han, Mahmoud Hassaballah, Florent Retraint, and Xiangyang Luo. Scalable universal adversarial watermark defending against facial forgery. *IEEE Transactions on Information Forensics and Security*, 2024. 2

[33] Zuomin Qu, Zuping Xi, Wei Lu, Xiangyang Luo, Qian Wang, and Bin Li. Df-rap: A robust adversarial perturbation for defending against deepfakes in real-world social network scenarios. *IEEE Transactions on Information Forensics and Security*, 2024. 2, 6

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

[35] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. 4

[36] 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. 1, 2, 5, 7

[37] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20595–20605, 2023. 6

[38] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7634–7644, 2023. 1, 5, 7

[39] Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. In *Advances in Neural Information Processing Systems*, pages 101474–101497. Curran Associates, Inc., 2024. 2

[40] 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. 2, 6, 7

[41] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2

[42] Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. Linear discriminant analysis: A detailed tutorial. *AI communications*, 30(2):169–190, 2017. 5

[43] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. 3

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

[45] 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. 2

[46] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11845–11854, 2021. 2

[47] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024. 2

[48] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decompo-

sition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025. 7

[49] Qian Zhang, Qing Guo, Ruijun Gao, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. Adversarial relighting against face recognition. *IEEE Transactions on Information Forensics and Security*, 2024. 2

[50] 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*, 2024. 2

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

[52] 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. 4, 6

[53] Yixuan Zhu, Wenliang Zhao, Yansong Tang, Yongming Rao, Jie Zhou, and Jiwen Lu. Stableswap: stable face swapping in a shared and controllable latent space. *IEEE Transactions on Multimedia*, 2024. 1, 2