International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02259-5

Anti-Fake Vaccine: Safeguarding Privacy Against Face Swapping via Visual-Semantic Dual Degradation

Jingzhi Li[1,5] · Changjiang Luo[1] · Hua Zhang[1] · Yang Cao[2] · Xin Liao[3] · Xiaochun Cao[4]

Received: 30 August 2023 / Accepted: 26 September 2024

Abstract

Deepfake techniques pose a significant threat to personal privacy and social security. To mitigate these risks, various defensive techniques have been introduced, including passive methods through fake detection and proactive methods through adding invisible perturbations. Recent proactive methods mainly focus on face manipulation but perform poorly against face swapping, as face swapping involves the more complex process of identity information transfer. To address this issue, we develop a novel privacy-preserving framework, named *Anti-Fake Vaccine*, to protect the facial images against the malicious face swapping. This new proactive technique dynamically fuses visual corruption and content misdirection, significantly enhancing protection performance. Specifically, we first formulate constraints from two distinct perspectives: visual quality and identity semantics. The visual perceptual constraint targets image quality degradation in the visual space, while the identity similarity constraint induces erroneous alterations in the semantic space. We then introduce a multi-objective optimization solution to effectively balance the allocation of adversarial perturbations generated according to these constraints. To further improving performance, we develop an additive perturbation strategy to discover the shared adversarial perturbations across diverse face swapping models. Extensive experiments on the CelebA-HQ and FFHQ datasets demonstrate that our method exhibits superior generalization capabilities across diverse face swapping models, including commercial ones.

Keywords Face privacy · Face swapping · Generative adversarial perturbation

1 Introduction

Personal privacy and social security face a significant threat from powerful Deepfake techniques (Juefei-Xu et al., 2022; Thambawita et al., 2021). One prominent application of Deepfakes is face swapping (Nirkin et al., 2019; Rosberg et al., 2023; Jiang et al., 2023), which plays a crucial role and has wide applications in the film industry and computer games. Face swapping aims to create a deceptive image that combines the identity of a source facial image with the attributes of a target facial image, such as the background, pose, and expression. Existing technologies can produce remarkably realistic and natural synthetic results, as shown in the first and third rows of Fig. 1. As this technology becomes more accessible, malicious actors can easily and quickly create unauthorized fake photos, videos, and GIFs of individuals, often without extensive expertise. The presence of a victim's identity in fake videos could mislead the public, causing significant damage to the victim's reputation. Additionally, face swapping can effectively fool many facial liveness APIs, posing further security risks (Teotia et al., 2022). Thus, it is critical for individuals to have tools to safeguard their facial images shared on social media to prevent potential exploitation by attackers.

In response to the aforementioned concerns caused by Deepfakes, existing research has mainly focused on *pas-*

Communicated by Segio Escalera.

✉ Jingzhi Li
lijingzhi@iie.ac.cn

[1] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[2] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

[3] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

[4] School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China

[5] Key Laboratory of Education Informatization for Nationalities (Yunnan Normal University), Ministry of Education, Kunming, China

Published online: 01 November 2024

sive defense techniques (Frank & Holz, 2021; Ilyas et al., 2023; Kim et al., 2021). Researchers have proposed various impressive methods that have achieved considerably high accuracy in distinguishing forged facial images or videos from real ones. However, these methods primarily serve as the ex-post forensic tools, and cannot provide reliable privacy protection when individuals share their facial data. In recent years, *proactive defense techniques* have emerged that add adversarial perturbations to original facial image to defend against Deepfakes (Ruiz et al., 2020; Wang et al., 2022; Huang et al., 2022). These methods mostly draw inspiration from previous adversarial attacks (e.g., FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018)) to generate imperceptible adversarial perturbations through gradient-based or optimization-based strategies. When Deepfakes are applied to facial images with added perturbations, the results often exhibit visually noticeable artifacts.

Essentially, these proactive defense methods create adversarial examples to defend against deepfake models by maximizing the pixel-level distance between the perturbed output and the original output. However, their performance against face swapping models could be quite unsatisfactory: 1) Insufficient protection. Face manipulation directly modifies the facial region in the source image, and existing methods can effectively counter this approach (Huang et al., 2021, 2022). However, face swapping involves the transfer of identity information, a more complicated process where existing methods struggle (Wang et al., 2022). 2) Weak generalization. Current approaches are often trained on specific deepfake models, resulting in poor performance when attackers change deepfake models (Ruiz et al., 2020).

When analyzing these vulnerabilities, a significant insight arises: existing methods primarily capture adversarial gradients that rely solely on pixel-level visual corruption, overlooking the content-level identity transfer inherent in face swapping. Inspired by this, we aim to enhance performance through a dual degradation mechanism targeting both visual quality and identity semantics. To this end, we propose *Anti-Fake Vaccine*, a face privacy preserving framework that helps individuals protect their images from unauthorized face swapping by generating erroneous image content. Our method forces the deepfake models, whether used or not used in generating protective noise, to incorporate deceptive visual elements when creating forgeries. For example, a user who wants to share personal photos on social media or the public web can add small, imperceptible alterations to their photos before uploading them. If these photos are collected by a third-party attacker and used to produce deepfakes, the images with injected protective noise would result in the model producing erroneous swapped results. Figure 1 illustrates our proposed face protection method. In summary, the introduced *Anti-Fake Vaccine* aims to achieve three primary goals: imperceptible perturbations to human eyes, significant reduced degradation performance of face swapping, and effective against unseen face swapping models without requiring prior knowledge.

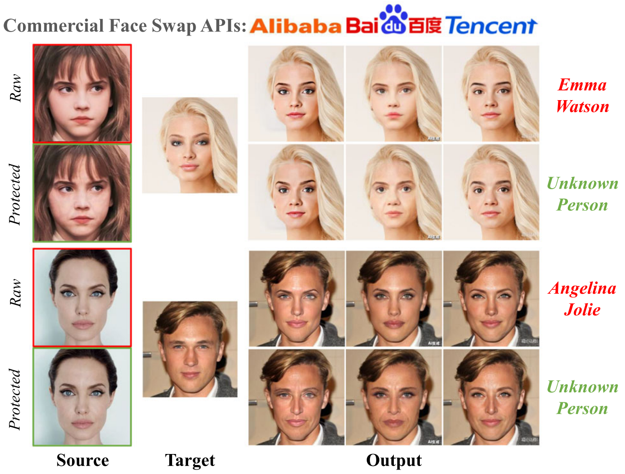

Fig. 1 *Anti-Fake Vaccine* protects user privacy by injecting adversarial perturbations into their online photos, effectively countering the threats posed by malicious Face Swaps. Different from the prior methods that rely on pixel-level distance regularization to generate the perturbations, our method introduces a dual degradation mechanism targeting both visual appearance and identity semantics. This mechanism significantly improve the protection performance. The red boxes denote the raw image online, while the green boxes represent the protected images, which are injected with protective noise imperceptible to the human eye. In this figure, We employ third-party systems from Alibaba, Baidu and Tencent to perform the face swapping (color figure online)

Specifically, we built a generator to create the protective noise that meets the above goals. The generator is optimized using three key techniques. First, we formulate constraints from two perspectives: visual quality and identity semantics. The visual perceptual constraint operates in the visual space, inducing noticeable degradation in the resulting outputs. The identity similarity constraint operates in the semantic space, preventing the reconstruction of individuals' identities during face swapping. Second, to achieve the optimality for these two regularization terms, we introduce a multi-objective optimization solution that adaptively adjusts weight parameters at each iteration. Finally, to further improve the performance against unseen deepfake models, we iteratively optimize the perturbation generator using adversarial gradients from multiple face swapping models. This technique enables us to create a compatible ensemble of adversarial attacks.

We conduct comprehensive experiments against six face swapping methods and three commercial APIs in a black-box setting. The experimental results demonstrate the effectiveness of our proposed method in terms of privacy protection, image utility, and robustness against the third parties. Furthermore, a detailed ablation analysis is provided to justify the motivations behind our design.

Our main contributions can be summarized as follows.

Table 1 A list of variables and their descriptions

Variable	Description
$I_{org}, I_{tar} \in \mathbb{R}^{W \times H \times C}$	The source face and the target face
$I^{sw} \in \mathbb{R}^{W \times H \times C}$	The swapped face
$I_{prv} \in \mathbb{R}^{W \times H \times C}$	The protected source face
$v_a \in \mathbb{R}^{W \times H \times C}$	The protective noise *Anti-Fake Vaccine*
$\varepsilon \in [1, \infty)$	The maximum magnitude of the perturbation
I^{sw}_{prv},$I^{sw}_{org} \in \mathbb{R}^{W \times H \times C}$	The original swapped face and the protected swapped face
λ_a,$\lambda_{id} \in (0, 1)$	The weight of two loss objectives
$\eta \in (0, 1)$	The content boundary
θ	The model parameters
$\nabla_\theta \mathcal{L}_a(\theta)$,$\nabla_\theta \mathcal{L}_{id}(\theta)$	The gradient of two loss objectives with respect of θ
m_a,$m_{id} \in (0, 1)$	The boundary constraint for two loss objectives

- We develop a novel privacy-preserving framework, named *Anti-Fake Vaccine*, to counter malicious face swapping attempts on facial images. This framework exhibits a strong generalization to unseen models and even commercial APIs, while preserving the utility of the original image through imperceptible pixel-level alterations.
- We introduce a dual degradation mechanism that dynamically combines visual corruption and content misdirection using a multi-objective optimization solution, thereby maximizing privacy protection and generalization capabilities.
- To further enhance generalization, we design an additive perturbation strategy based on meta-learning to integrate shared adversarial perturbations from two face swapping models.

The remaining sections of the paper are organized as follows: Section 2 provides a comprehensive review of related work. Section 3 introduces our privacy-preserving model, detailing the key components and methodologies. Section 4 presents the experimental results from extensive evaluations on different datasets. Section 5 offers the conclusions.

2 Related Work

2.1 DeepFake Creation

In recent years, deepfake techniques have made remarkable progress, posing a potential threat to personal privacy and social security. These methods could be broadly categorized into two groups based on their intended forgery goals: face manipulation and face swapping.

Face manipulation focuses on modifying facial attributes such as age, expression, and hairstyle. The attribute editing is commonly achieved using popular Generative Adversarial Networks (GANs) like StyleGAN (Tov et al., 2021) and InterfaceGAN (Shen et al., 2022). By incorporating attribute labels into the training process, these GAN-based models enable precise control over characteristics like hairstyle, age, gender, and other facial attributes (He et al., 2019; Wu et al., 2019; Li et al., 2021; Ververas & Zafeiriou, 2020). Another approach within face manipulation is face reenactment, which involves analyzing the driving face to capture the facial expressions and movements, and mapping them onto the source face (Thies et al., 2016; Wiles et al., 2018; Tripathy et al., 2020). Face2Face (Thies et al., 2016) is textcolorbluewell-known tool for this purpose, textcolorbluecapable of generating high-quality facial reenactment videos by aligning the facial features and expressions between the driving and source faces.

Face swapping, a prominent framework within DeepFake techniques, involves swapping the identity between a source and a target facial image (Nirkin et al., 2019; Jiang et al., 2023; Li et al., 2020; Liu et al., 2023b; Suwajanakorn et al., 2017). Autoencoders and conditional GAN architectures are commonly employed as fundamental components, enabling the generation of convincingly realistic results. These methods are primarily identity-oriented, merging the identity information from the source face with the target face images or videos, and utilizing GANs to produce visually coherent face-swapped content. These methods can be categorized into two types based on their operation mode: target-oriented face swapping and source-oriented face swapping.

The target-oriented face swapping methods use different encoders to extract identity embeddings and attribute embeddings, then gradually inject the identity embeddings into the attribute embeddings to generate swapped faces. FaceShifter (Li et al., 2020) designs a two-stage network to merge target non-identity attributes with source identity and refine the output. SimSwap (Chen et al., 2020) employs weak feature matching loss to maintain target attributes and achieve high-fidelity swapped faces. InfoSwap (Gao et al., 2021) uses an information bottleneck to disentangle iden-

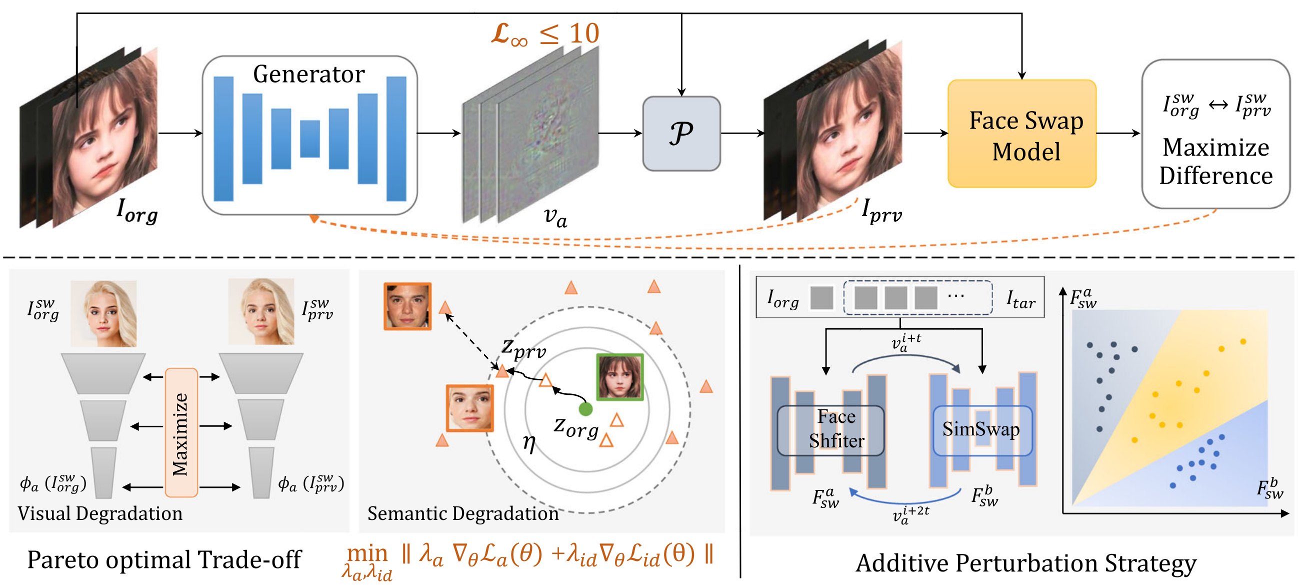

Fig. 2 The overview of the proposed *Anti-Fake Vaccine* framework. The upper part shows the pipeline for generating the protected images. The generator, using the ResNet architecture, produces the adversarial perturbations that are scaled to comply with a norm constraint. These perturbations are then added to the original image and clipped to create the protected image. The orange dashed line represents the gradient flow. The lower part depicts the key components involved in the optimization process (color figure online)

tity information. SmoothSwap (Kim et al., 2022) focuses on identity space smoothness and applies a supervised contrastive loss to improve source identity preservation, albeit at the expense of target attribute fidelity. MegaFS (Zhu et al., 2021) leverages StyleGAN to produce high-resolution images, while FSLSD (Xu et al., 2022) applies it for structural disentanglement. BlendFace (Shiohara et al., 2023) focuses on mitigating attribute leakage issues in face swapping. FaceDancer (Rosberg et al., 2023) proposes a one-stage network that adaptively fuses attribute features and identity features without requiring any additional face segmentation process. SelfSwapper (Lee et al., 2024) implements a self-supervised training framework aimed at improving the separation of identity and non-identity features.

The source-oriented face swapping methods incorporate face reenactment models to warp the source image to match the target's pose and expression, and then blend it with the target background to swap faces. DeepFaceLab (Liu et al., 2023b) combines person-specific reenacted faces with mask and color correction, leading to issues in preserving facial shape and texture. Naruniec et al. (2020) achieve high-resolution swapping results within this framework, while Otto et al. (2022) enhance it using a differentiable 3D network. FSGAN (Nirkin et al., 2019) integrates face swapping and reenactment through a two-stage process, followed by a face inpainting network to blend the reenacted source with target images. E4S (Liu et al., 2023a) adopts a reenact model to align the pose and expression between the target face and the swapped face.

In these two types of deepfakes, face swapping mainly transfers the identity of the victim into a target image with an unknown scene, raising significant privacy or security concerns to the individuals. In this work, we focus on privacy-preserving strategies against face-swapping techniques.

2.2 DeepFake Passive Defense

Detecting whether a facial image is real or fake by identifying subtle differences is a straightforward approach to defending against Deepfakes. Efforts in image or visual deepfake detection have focused on identifying specific artifacts introduced by face forgery. Studies have shown that these subtle differences can be observed in forgery both the spatial (Frank & Holz, 2021; Wang et al., 2021) and frequency domain (Qian et al., 2020). Zhao et al. (2021) introduce a multi-attentional deepfake detection network, framing deepfake detection as a fine-grained classification problem. Li et al. (2021) propose a discriminative feature learning framework to capture frequency cues, enabling the learning of more distinctive features with reduced optimization difficulty. Recent works have explored intramodal inconsistency by comparing source and forged images (Kim et al., 2021; Zhao et al., 2021). Additionally, there have been efforts in multimodal deepfake detection. Cheng et al. (2022) approach deepfake detection from a voice-face matching perspective, leveraging the intrinsic correlation between facial and audio cues. Ilyas et al. (2023) propose an unified audio-visual detection framework based on modality dissonance and dense hierarchical features.

Although some detection-based methods achieve considerably high accuracy in ex-post forensics, they cannot be applied to prevent the widespread dissemination of Deepfakes before they cause significant damage.

2.3 DeepFake Proactive Defense

Countermeasures designed to prevent the manipulation of facial images have received significant attention in the past two years (Ruiz et al., 2020; Huang et al., 2022; Ruiz et al., 2023; Yang et al., 2021; Li et al., 2022). A similar study on face image protection involves designing protective noise to resist face recognition systems (Deb et al., 2020; Yang et al., 2021; Sun et al., 2024). Yang et al. (2021) develop a method that utilizes differentiable random image transformations to generate adversarial faces with noticeable artifacts. Ruiz et al. (2020) proposed a gradient-based method to attack facial manipulation systems successfully in white-box settings, protecting faces from being manipulated by specific models. Yeh et al. introduce two types of adversarial attacks against image translation GANs by utilizing a specially designed adversarial loss function and gradient optimization (Yeh et al., 2020). Huang et al. (2021) propose an defense framework against face manipulation in black-box settings. This work produces the perturbations by training a surrogate model, however, it may face the challenge of finding a suitable surrogate model when it is applied to defend against face swapping. Recently, Wang et al. (2022) propose perceptual-aware adversarial perturbations operating on the Lab color space for both types of DeepFake.

However, existing proactive defense techniques provide insufficient protection against face swapping and lack generalization ability across Deepfake models. In this work, we examine the challenges of preserving privacy against third-party face swapping models and even commercial models in dealing with our *Anti-Fake Vaccine*. Our goal is to develop stealthy and robust imperceptible protective noise that are effective against unseen models, providing strong protection for the shared facial images on social media.

3 Methodology

3.1 Problem Settings

The goal of face swapping is to transfer the identity of the source face to a target image while preserving other information unchanged such as background, illumination, and pose. There are two paradigms for face swapping. The first involves gradually infuse identity codes from the source face into the attribute codes of the target face using various frameworks to produce the swapped face (Li et al., 2020; Gao et al., 2021; Chen et al., 2020). The second involves manipulating the source face with the target face's attributes, such as pose and expression, and then blending it with the remaining information from the target image to obtain the final swapped face (Nirkin et al. 2019; Liu et al. 2023a, 2023b). Table 1 presents the variables used in this paper along with their descriptions. We formulate these two paradigms in a general form:

$$I^{sw} = F_{sw}(I_{org}, I_{tar}), \tag{1}$$

where I_{org} and I_{tar} are the source face and the target face, and F_{sw} represent the different face swapping models.

Our privacy-preserving goal is to generate protective noise v_a that are added to the source face I_{org} to disrupt the generation of the swapped face. Ideally, $F_{sw}(I_{org}, I_{tar})$ and $F_{sw}((I_{org} + v_a), I_{tar})$ should exhibit a noticeable difference in appearance to human eyes. v_a can be obtained by solving a constrained optimization problem:

$$\begin{aligned} &\underset{v_a}{max} \mathcal{L}(F_{sw}((I_{org} + v_a), I_{tar}), F_{sw}(I_{org}, I_{tar})), \\ &s.t. \parallel v_a \parallel_p \leq \varepsilon, \end{aligned} \tag{2}$$

where ε is the maximum magnitude of the perturbation, and $\mathcal{L}(.)$ is a loss function for measuring the difference between the original swapped face and the protected swapped face. This paper defines this difference along two dimensions: visual quality at the pixel level and identity semantics at the content level. The perturbation v_a is constrained by the ℓ_p-norm ($p \in 0, 2, \infty$).

The *Anti-Fake Vaccine* should satisfy the following properties.

- Imperceptible perturbations to human eyes. Facial images shared on social media platforms carry the significant social value, thus the visual consistency should be preserved.
- Significantly reduced performance of face swapping. To defend against malicious face swapping, we should provide the sufficient protection against the transfer of the victim's identity.
- Effective against unseen face swapping models. Since we cannot know the prior information about face swapping models in real scenarios, we need to ensure the generalization of our method.

3.2 Overall Framework

Figure 2 shows the architecture for generating the protected facial image that satisfy the above properties. The framework

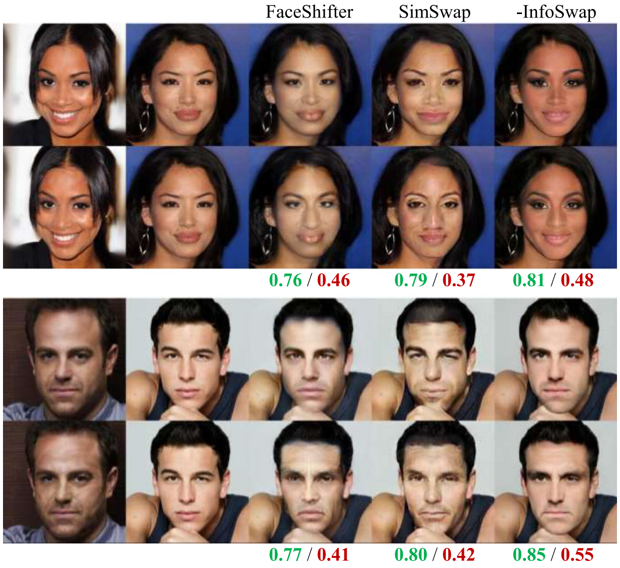

Fig. 3 The visual results of our method. Green annotations indicate the cosine distance between the original swapped faces and original faces, while red annotations indicate the distance between the protected swapped faces and original faces (color figure online)

mainly consists of a perturbation generator G_v and face swapping model F_{sw}. Given an original facial image I_{org}, our goal is to generate the protected facial image I_{prv}, which causes the black-box face swapping model F_{sw} to produce an incorrect output. To begin, I_{org} is passed through the generator G_v to create the perturbation v_a, which is scaled to have a fixed norm. Then, the protected facial image can be obtained by:

$$I_{prv} = clip(min(I_{org} + \varepsilon, max(v_a, I_{org} - \varepsilon))), \tag{3}$$

where ε is the maximum magnitude of the perturbation, and the operation $clip$ limits the I_{prv} in the range $[I_{org} - \varepsilon, I_{org} + \varepsilon]$. Then, the protected image I_{prv} as well as the original image I_{org} are fed to the face swapping model F_{sw} to obtain the swapped faces I_{prv}^{sw} and I_{org}^{sw}, respectively.

Subsequently, we optimize the generator G_v using the objective loss function given in Eq. (2). The objective loss function consists of two parts: the visual perceptual constraint $\mathcal{L}_a$ and the identity similarity constraint $\mathcal{L}_{id}$. Equation (2) can be expressed in its minimized form as follows:

$$\underset{v_a}{min} \lambda_a \mathcal{L}_a + \lambda_{id} \mathcal{L}_{id}, s.t. \parallel v_a \parallel_p \leq \varepsilon, \tag{4}$$

where λ_a, λ_{id} are the trade-off parameters. The visual perceptual constraint $\mathcal{L}_a$ operates in the visual space, inducing the noticeable degradation in the resulting Deepfakes. The identity similarity constraint $\mathcal{L}_{id}$ operates in the semantic space, preventing the reconstruction of individuals' identities during face swapping. To ensure the optimality of the two regularization terms and thus enhance protection performance, we introduce a multi-objective optimization solution based on Lagrangian multipliers. This solution dynamically adjusts weight parameters during each iteration. Furthermore, considering the differences between the two face swapping paradigms, we design an additive perturbation strategy to create a compatible ensemble of adversarial perturbations.

To illustrate the impact of protective noise on face swapping, we present an intuitive visual comparison in Fig. 3. The first column shows the source images before and after adding perturbations. The second column shows the target images. The remaining columns display the swapped results generated by three models using the perturbed and original source images. Additionally, we employed a face recognition model (ArcFace) to evaluate the identity similarity between the source face and the swapped faces. The closer the distance is to 1, the more similar the faces are. The results reveal that the protective noise we introduced has little visual impact on the source images, but can significantly affect the face-swapping results, including for unseen models.

3.3 Visual-Semantic Dual Degradation

Existing methods (Ruiz et al., 2020; Huang et al., 2022; Yang et al., 2021) that employ pixel distance regularization facilitate the notable visual quality degradation in deepfake-generated content. However, the protection brought by this degradation is limited. In contrast, we design a dual degradation of visual quality and identity semantics, which enhances content protection in source images, and improves the generalization to various face swapping models. Specifically, a visual perceptual constraint term is designed to achieve image quality degradation in the visual space, while an identity similarity constraint induces image content changes in the semantic space.

3.3.1 Visual Perceptual Constraint

The protected facial images are intended to safeguard individuals' real appearance from unauthorized face swapping. To this end, prior approaches (Ruiz et al., 2020; Yang et al., 2021) typically use the pixel-level distance regularization to introduce significant distortions on the outputs. However, pixel regularization is sensitive to cluttered background and the scale of face images. Therefore, we introduce the perceptual similarity between the protected swapped face I_{prv}^{sw} and the original swapped face I_{org}^{sw} as the regularization. Specifically, we adopt a perceptual model $\mathcal{A}$ to compute the feature inconsistency (Zhang et al., 2018). The perceptual regularization $\mathcal{L}_a$ between two images is defined as:

$$\mathcal{L}_a = -\sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \parallel \omega_l \odot \left(f_a^l\left(I_{prv}^{sw}\right) - f_a^l\left(I_{org}^{sw}\right)\right) \parallel_2, \tag{5}$$

where $f_a^l(.)$ represents the normalized features of the two images. First, extract feature stack from the l_{th} layer of network $\mathcal{A}$, and then unit normalize these features along the channel dimension. H and W denote the height and width of the feature maps. ω_l is used to scale the channel-wise activations. By optimizing this regularization, the perturbations created by the generator G_v could make a noticeable modifications in the resulting outputs.

3.3.2 Identity Similarity Constraint

The visual perceptual constraint is an unsupervised technique that aims to maximize the pixel distance between the protected swapped face and the original swapped face. This constraint introduces visual artifacts, such as black spots, into the swapped result, but has little impact on the image content. Furthermore, this attack mechanism generates model-specific adversarial perturbations, meaning that the perturbation performs poorly in defending against third-party deepfake models. To improve the generalization of our model, an intuitive way is to explore the common vulnerabilities of models operating on the source face (the protected subject). Given that both face-swapping paradigms core on extracting the identity embedding of the source face, we believe that content changes can lead to more general defense capabilities. Thus, we propose an identity similarity constraint to prevent the reconstruction of individuals' identities within the created Deepfakes. In other words, the attacker can only obtain a misleading identity code, which significantly deviates from the accurate code.

Specifically, as illustrated in Fig. 2, this regularization attempts to move the embedding of the protected face beyond the content boundary of the original face embedding. A pre-trained identity encoder $\mathcal{ID}$ is used to perform this regularization. In the identity semantic space, maximizing the distance between two identity features is the direct and effective approach. However, in our task, we need to account for the constraint that optimizing this loss should not exceed the maximum magnitude of the perturbation (imperceptible). And some experimental results have shown that a certain content boundary can effectively protect face identity. Inspire by this observation, we set a content boundary to achieve the image content alterations with reduced costs. Note that the identity code contains various semantic contents, such as the gender, age, etc. The identity similarity regularization is defined as:

$$\begin{aligned}\mathcal{L}_{id} = &\parallel \Delta_{cos}\left[f_{id}(I_{prv}), f_{id}(I_{org}) \right] - \eta \parallel_1 \\ &+ \parallel \Delta_{cos}\left[f_{id}(I_{prv}^{sw}), f_{id}(I_{org}) \right] - \eta \parallel_1,\end{aligned} \tag{6}$$

where Δ_{cos} is the cosine distance, $f_{id}(.)$ is the normalized image feature after applying the encoder $\mathcal{ID}$, and $\parallel . \parallel_1$ is L_1 norm distance. Our loss function limits the maximization of face embedding distance, providing two benefits: (i) it guarantees that the image content is altered only to the extent necessary to effectively protect the face identity, and (ii) it allows us to control the visual impact by adjusting this boundary.

3.4 Multi-objective Optimization

In our task, the generator G_v needs to optimize two objectives simultaneously: the visual perceptual constraint $\mathcal{L}_a$ described in Sect. 3.3.1 and the identity similarity constraint $\mathcal{L}_{id}$ described in Sect. 3.3.2. The visual perceptual constraint aims to introduce the distortions at the pixel level to the swapped face. And the identity similarity constraint focuses on erasing identity-related information of the swapped face at the semantic level. The two regularization terms are constrained by the set perturbation intensity, making the simultaneous optimization of these objectives challenging. Typically, the optimal solution for one objective is sub-optimal for the other, a fact that has been validated in the experiments section. Consequently, this represents a multi-objective optimization problem. In the overall objective loss function, we aggregate the two objectives with different weights.

The Pareto optimality is an important concept in multi-objective optimization. Given a system which aims to minimize a series of objective functions, Pareto optimality is a state when it is impossible to improve one objective without hurting other objectives. Our goal is to achieve the Pareto optimality. Existing gradient-based multi-objective optimization algorithms (Sener & Koltun, 2018; Lin et al., 2019) have proven that any solution that satisfies the Karush-Kuhn-Tucker (KKT) conditions (Kuhn et al., 2013) can achieve the Pareto optimality. The KKT conditions for the parameters of our model can be expressed as:

$$\begin{aligned}&\lambda_a + \lambda_{id} = 1, \exists \lambda_a \geq m_a, \lambda_{id} \geq m_{id}, \\ &\lambda_a \nabla_\theta \mathcal{L}_a(\theta) + \lambda_{id} \nabla_\theta \mathcal{L}_{id}(\theta) = 0,\end{aligned} \tag{7}$$

where λ_a and λ_{id} are the trade-off parameters, m_a and m_{id} are the lower bounds of the importance of the two constraints, $\nabla_\theta \mathcal{L}_a(\theta)$ and $\nabla_\theta \mathcal{L}_{id}(\theta)$ are the gradients of two constraints. And these conditions can be transformed into the following minimization form:

$$\min_{\lambda_a,\lambda_{id}} \parallel \lambda_a \nabla_\theta \mathcal{L}_a(\theta) + \lambda_{id} \nabla_\theta \mathcal{L}_{id}(\theta) \parallel_2, \\ s.t. \lambda_a + \lambda_{id} = 1, \lambda_a \geq m_a, \lambda_{id} \geq m_{id}. \tag{8}$$

The solution to this minimization problem is either 0, which satisfies the KKT conditions, or gives the descent direction that improves all tasks (Sener & Koltun, 2018). That is to say, the solutions to λ_a and λ_{id} under these conditions achieve the Pareto optimality.

We adopt the method proposed in Lin et al. (2019) to solve this constrained optimization problem. Similar to the traditional solution to the constrained optimization problems, this method first solve the optimization problem only subject to the equality constraint, and then introduce a projection procedure to produce a valid solution with all the constraints. Before solving this problem, we reformulate the optimization problem as:

$$\min_{\hat{\lambda}_a,\hat{\lambda}_{id}} \parallel (\hat{\lambda}_a + m_a) \nabla_\theta \mathcal{L}_a(\theta) + (\hat{\lambda}_{id} + m_{id}) \nabla_\theta \mathcal{L}_{id}(\theta) \parallel_2, \\ s.t. \hat{\lambda}_a + \hat{\lambda}_{id} = 1 - m_a - m_{id}, \hat{\lambda}_a \geq 0, \hat{\lambda}_{id} \geq 0, \tag{9}$$

where $\hat{\lambda}_a = \lambda_a - m_a$, $\hat{\lambda}_{id} = \lambda_{id} - m_{id}$. This is equivalent to Eq. (8). First, we employ the Lagrange multipliers method to solve the equality constrained optimization problem in Eq. (9). According to the theorem in Lin et al. (2019), the solution to this problem is:

$$\begin{bmatrix} \hat{\lambda}_a^* \\ \hat{\lambda}_{id}^* \\ \lambda \end{bmatrix} = \left(H H^T \right)^{-1} H \begin{bmatrix} -G G^T M \\ 1 - E^T M \end{bmatrix}, \tag{10}$$

where $G = [\nabla \mathcal{L}_a, \nabla \mathcal{L}_{id}]^T$, $E = [1, 1]^T$, $M = [m_a, m_{id}]^T$, H is $\begin{bmatrix} GG^T & E \\ E^T & 0 \end{bmatrix}$, and λ is the Lagrange multiplier. Then, we perform a projection step to obtain a valid solution that satisfies the inequality constraints in Eq. (9). This solutions is given by solving a non-negative least squares problem:

$$\min_{\tilde{\lambda}_a,\tilde{\lambda}_{id}} \parallel \left(\tilde{\lambda}_a - \hat{\lambda}_a^*\right) + \left(\tilde{\lambda}_{id} - \hat{\lambda}_{id}^*\right) \parallel_2, \\ s.t. \tilde{\lambda}_a + \tilde{\lambda}_{id} = 1, \tilde{\lambda}_a \geq 0, \tilde{\lambda}_{id} \geq 0, \tag{11}$$

where the $\tilde{\lambda}_a$ and $\tilde{\lambda}_{id}$ are the valid solution to Eq. (9). Finally, the solution to λ_a and λ_{id} that achieves the Pareto optimality in our multi-objective optimization is given by

$$\lambda_a = \tilde{\lambda}_a + m_a, \lambda_{id} = \tilde{\lambda}_{id} + m_{id}. \tag{12}$$

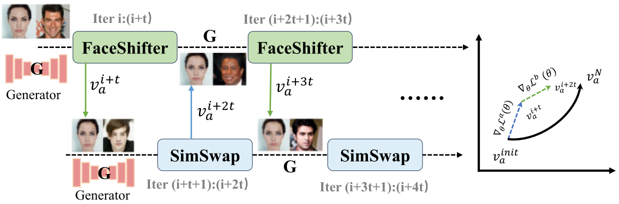

Fig. 4 The procedure of additive perturbation strategy. we calculate gradients from two face swapping models in order and use them to update the perturbations. In each iteration, we change the target face. Over multiple iterations, the perturbations accumulate gradients from both domains, resulting in a stronger protection (color figure online)

3.5 Additive Perturbation Strategy

To further enhance performance against unseen deepfake models, we develop an additive perturbation strategy that merges the shared adversarial perturbations from different face swapping models. This is inspired by meta-learning (Shao et al., 2020; Yin et al., 2023), which aims to train a model capable of adapting to new tasks with minimal training iterations and data. Similar to the classical meta-learning framework, we iteratively optimize the perturbation generator G_v using adversarial gradients from two face swapping models. Additionally, in each iteration, we utilize distinct target facial images to improve the generalization across various scenarios. Our additive perturbation strategy could collect the advantages of both models and integrate them in a compatible way.

The overall process of our additive perturbation strategy is illustrated in Fig. 4. Let F_{sw}^a and F_{sw}^b denote the two different face swapping models, respectively. In each round of iterations, we alternately select the two face swapping models and compute the gradients based on the overall loss $\mathcal{L}(\theta, I_{org}, I_{tar})$. The final performance is not affected by the order of the face swapping models. During the experimental process, we observed that the number of iterations before the alternate optimization of these two models significantly impacts the results. Consequently, we set an iteration step size t, leading to the generation of t optimization parameters, denoted as $\{\theta^1, ..., \theta^t\}$. At first, the optimization of the first face swapping model F_{sw}^a is calculated by:

$$\theta^i \leftarrow \theta^{i-1} - \gamma \cdot \nabla_\theta \mathcal{L}^a \left(\theta^{i-1}, I_{org}, I_{tar} \right), \tag{13}$$

where γ is the updating hyper-parameter, and $\mathcal{L}^a$ represents the overall loss on F_{sw}^a. After t iterations, we update the perturbation generator G_v and generate the perturbations v_a^{i+t}. Subsequently, we apply the perturbations v_a^{i+t} to the original facial image I_{org} to create the preserved facial image I_{prv}, which serves as the initialization for the other face swapping model F_{sw}^b. We use a new target facial image to update the perturbation generator:

Algorithm 1 Our proposed *Anti-Fake Vaccine*

```
Require: Original image $I_{org} \in \mathbb{R}^{H\times W\times C}$; Target image $I_{tar} \in \mathbb{R}^{H\times W\times C}$; Two face swapping models $F_{sw}^{a}$, $F_{sw}^{b}$; Perturbation generator $G_v$;
Ensure: Protected image $I_{prv}$.
  Initialize model parameter $\theta$; trade-off parameters $\lambda_a$, $\lambda_{id}$.
  for $i \in N$ do
    Select one face swapping model $F_{sw}^{a}$;
    for $i \in t$ do
      Calculate the appearance perceptual loss $\mathcal{L}_a$ via Eq.5;
      Calculate the identity semantic loss $\mathcal{L}_{id}$ via Eq.6;
      Compute $\nabla_{\theta^{i-1}}\mathcal{L}_a(\theta^{i-1})$ and $\nabla_{\theta^{i-1}}\mathcal{L}_{id}(\theta^{i-1})$ via Eq.8;
      Compute $\tilde{\lambda}_a$ and $\tilde{\lambda}_{id}$ via Eq.10 and Eq.11;
      Update $\lambda_a$ and $\lambda_{id}$ via Eq.12;
      Update $\theta^{i}$ via Eq.13;
      Obtain $v_a^{i}$ through perturbation generator $G_v$;
      Update $I_{prv}^{i} = I_{org} + v_a^{i}$;
    end for
    Select another face swapping model $F_{sw}^{b}$ to replace $F_{sw}^{a}$
    Obtain $\theta^{i+t+1}$ via Eq.14;
    Update $I_{prv}^{i+t+1} = I_{org} + v_a^{i+t+1}$;
  end for
  return $I_{prv}^{N}$
```

$$\theta^{i+t+1} \leftarrow \theta^{i+t} - \gamma \cdot \nabla_{\theta}\mathcal{L}^{b}\left(\theta^{i+t}, I_{org}, I_{tar}^{new}\right), \tag{14}$$

where I_{tar}^{new} denotes the new target facial image, and $\mathcal{L}^{b}$ represents the overall loss on F_{sw}^{b}.

After several rounds of iterations, the computed gradients from different face swapping models and diverse target facial images are aggregated to stabilize the current optimizations. Our strategy enhances optimization stability and promotes the generalization by collecting gradients across different models and target face images. The complete algorithm of *Anti-Fake Vaccine* is presented in Algorithm 1.

4 Experiments

In this section, we perform a series of experiments to assess the effectiveness of our proposed framework. These experiments are divided into three distinct parts, each aimed at analyzing various facets of our task. Firstly, we validate the effectiveness of privacy protection using six face swapping models and three commercial APIs. Secondly, we analyze the utility of protected images in relation to realism, visual quality, and resistance to common image manipulation. Thirdly, we investigate the contributions of different components through ablation studies.

4.1 Experimental Setup

Datasets: 1) CelebA-HQ (Lee et al., 2020) is a high-resolution facial image dataset, which is widely employed in the recent DeepFake studies. It contains 30, 000 images selected from the CelebA with different demographics such as age, gender, and ethnicity. The size of each image is 1024 × 1024. We randomly select 2, 000 images for quantitative evaluations, and the remaining images are used as the training set. 2) FFHQ (Karras et al., 2019) is another high-quality facial image dataset, crawled from Flickr and automatically aligned and cropped using dlib. It contains 70,000 images with considerable variation in terms of age, ethnicity and image background. The size of each image is also 1024 × 1024. We randomly select 2000 images of different identities from the test set.

Face Swapping Models: To validate the transferability of the protected images generated by *Anti-Fake Vaccine*, we evaluated our proposed method on a combination of various offline and online face swapping models. Specifically, we choose: 1) six well-known face swapping models from recent research: FaceShifter (Li et al., 2020), InfoSwap (Gao et al., 2021), SimSwap (Chen et al., 2020), E4S (Liu et al., 2023a), MegaFS (Zhu et al., 2021), FaceDancer (Rosberg et al., 2023); 2) three widely used online commercial face swapping systems: Alibaba,[1] Baidu,[2] Tencent.[3] For all evaluations, we use FaceShifter (Li et al., 2020) and SimSwap (Chen et al., 2020) to generate the protected facial images and evaluate the protection performance on the other models. All parameters of the offline models are obtained from the original literature. We provide a unified summary of the training datasets and network architectures for the six models in Table 2.

Comparison Methods: We design three groups of typical task-related baselines, which have been widely applied in the literature. The DeepFake-based adversarial methods include: Disrupting (Ruiz et al., 2020), which generates adversarial perturbations using a gradient-based technique PGD; Anti-forgery (Wang et al., 2022) which generates perceptual-aware adversarial perturbations operating in the Lab color space; and Initiative (Huang et al., 2021) which produces perturbations through a generative model trained with an adversarial mechanism. The transfer-based adversarial methods include: Regional homogeneity (Li et al., 2020) and Transfer (Nakka & Salzmann, 2021). These methods are designed to generate strongly transferable adversarial examples. The face recognition(FR)-based adversarial methods include: Advfaces (Deb et al., 2020) and TIP-IM (Yang et al., 2021). These methods are designed to generate adversarial perturbations against face recognition model.

[1] https://vision.aliyun.com/facebody.

[2] https://ai.baidu.com/tech/face/merge.

[3] https://cloud.tencent.com/product/facefusion.

Table 2 A list of face swapping models we used in our experiments

Model name	Training dataset	Network architecture
FaceShifter Li et al. (2020)	CelebA-HQ, FFHQ, VGGFace	Encoder-Decoder
SimSwap Chen et al. (2020)	VGGFace2	GANs
InfoSwap Gao et al. (2021)	CelebA-HQ, FFHQ	Encoder-Decoder
E4S Liu et al. (2023a)	CelebAMask-HQ, FFHQ	StyleGAN
MegaFS Zhu et al. (2021)	CelebA, CelebA-HQ, FFHQ	StyleGAN2
FaceDancer Rosberg et al. (2023)	VGGFace2, LS3D-W	Conditional GAN

4.2 Evaluation Metrics

In our experiments, we design two types of metrics to evaluate the privacy protection and utility of data, respectively.

Privacy Metrics: To evaluate the degree of visual damage caused by protective noise to the face-swapping model, we adopt two visual quality evaluation metrics: PSNR and LPIPS. PSNR calculates differences based on the raw pixel values between two images, while LPIPS is a perceptual metric that quantifies the human-perceived similarity between two images. The average PSNR and LPIPS are used to quantify the similarity between the original swapped image and the injected swapped image. A lower PSNR score denotes significant corruption introduced, while a higher LPIPS score indicates substantial distortion. Additionally, we design two protection success rates (PSRs) to report successfully protected facial images, i.e., the ratio of swapped faces that are successfully destroyed. PSR_1 indicates the proportion of injected swapped images whose distortion exceeds a certain threshold, specifically when the L_2 distance is greater than 0.05. PSR_2 indicates the proportion of injected swapped images whose identities are incorrectly identified by the face recognition (FR) model. The threshold for the FR model is the cosine similarity score at the 0.001 FPR (False Positive Rate) level.

Utility Metrics: To evaluate the imperceptibility of the source face images before and after adding perturbations, we compare the RMSE, PSNR, SSIM, and LPIPS between them, which represent different levels of visual perception. RMSE directly measures the average error between two images, focusing on the magnitude of differences. PSNR uses a logarithmic scale to assess image quality relative to the maximum signal strength, making it useful for comparing distortion and noise levels across images. SSIM evaluates image quality by analyzing alterations in structural information, luminance, and contrast due to distortions. In contrast, LPIPS assesses visual similarity by comparing features extracted through a convolutional neural network, providing a measure that better reflects human visual perception. A higher SSIM score and a lower LPIPS score indicate that the protected image is closer to the original image, while a higher PSNR score and a lower RMSE score signify better visual quality of the protected image.

4.3 Implementation Details

In *Anti-Fake Vaccine*, we design the architecture of G_v according to the traditional generative adverisal network (Poursaeed et al., 2018). We adopt the network proposed in Zhang et al. (2018) as the perceptual model $\mathcal{A}$ to perform the regularization $\mathcal{L}_a$. The regularization $\mathcal{L}_{id}$ is performed using the identity encoder $\mathcal{ID}$ based on ArcFace (Deng et al., 2019). We trained this model for 100 epochs on the selected datasets with batch size 20, achieving an accuracy of 96.8% on CelebA-HQ and 91.5% on VGGFace2-HQ. In the training process, the learning rate of G_v is set to 0.0004. Empirically, the parameter of Eq. (6) is set to $\eta = 0.3$. Similar to Poursaeed et al. (2018), we set the L_2-norm threshold as 2000 and the L_∞-norm threshold as 10. To achieve the desirable outputs, we set the number of iterations for each face swapping model to be 300, and the total number of iterations to be 2400.

4.4 Evaluation on Privacy Protection

In this task, we conduct two experiments: defense against offline models and defense against online models. We train the *Anti-Fake Vaccine* framework with two face swapping models to generate the protected faces and test them in black-box settings. We compute the privacy metrics for the original swapped faces and the protected swapped faces.

4.4.1 Defense Against Offline Models

We present the protected results produced by our method and seven comparison methods in Tables 3 and 4, evaluated on the CelebA-HQ and FFHQ datasets, respectively. The corresponding visualization results are displayed in Fig. 5. According to the quantitative results, our method outperforms almost all face swapping models, whether they were used in training or not. This indicates strong generalization in preserving facial appearance. The visualization results further affirm its superiority. Our method effectively disrupts the

Table 3 Comparison of the proposed method with existing methods against six face swapping models on the CelebA-HQ dataset

Methods	FS models	Faceshifter Li et al. (2020)				Simswap Chen et al. (2020)				-InfoSwap Gao et al. (2021)			
	metrics	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑
DF-Based	Disrupting Ruiz et al. (2020)	33.40	0.04	0.0	0.24	**26.27**	**0.14**	0.19	**0.93**	32.43	0.06	0.0	0.25
	Anti-forgery Wang et al. (2022)	44.47	0.01	0.0	0.08	42.02	0.01	0.0	0.04	36.36	0.04	0.0	0.10
	Initiative Huang et al. (2021)	37.03	0.02	0.03	0.06	39.33	0.04	0.0	0.01	35.19	0.03	0.05	0.08
Transfer-based	Regional Li et al. (2020)	**29.10**	**0.07**	**0.35**	0.71	30.06	0.10	**0.35**	0.45	29.38	**0.10**	**0.33**	0.70
	Transfer Nakka and Salzmann (2021)	32.24	0.04	0.0	0.14	33.81	0.05	0.0	0.04	31.81	0.03	0.01	0.13
FR-based	Advfaces Deb et al. (2020)	31.89	0.04	0.0	0.49	30.90	0.07	0.0	0.16	31.05	0.04	0.01	0.46
	TIP-IM Yang et al. (2021)	29.16	0.05	0.05	**0.84**	29.91	0.11	0.0	0.46	**28.91**	0.08	0.05	**0.85**
	Anti-Fake Vaccine	**28.26**	**0.13**	**0.92**	**0.95**	**24.94**	**0.17**	**1.0**	**1.0**	**28.87**	**0.16**	**0.92**	**0.92**

Methods	FS models	-E4S Liu et al. (2023a)				-MegaFS Zhu et al. (2021)				-FaceDancer Rosberg et al. (2023)			
	metrics	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑
DF-Based	Disrupting Ruiz et al. (2020)	32.31	0.04	0.0	0.47	30.93	0.08	0.12	0.47	33.01	0.03	0.0	0.30
	Anti-forgery Wang et al. (2022)	35.08	0.03	0.0	0.32	40.92	0.02	0.01	0.27	44.20	0.02	0.0	0.10
	Initiative Huang et al. (2021)	31.23	0.07	**0.58**	0.33	28.67	0.11	**0.72**	0.31	38.93	0.02	0.0	0.10
Transfer-based	Regional Li et al. (2020)	31.52	**0.09**	0.4	0.75	32.36	**0.12**	0.65	0.71	**28.87**	0.05	**0.22**	**0.75**
	Transfer Nakka and Salzmann (2021)	**28.87**	0.05	0.03	0.36	27.63	0.08	0.24	0.39	33.55	0.04	0.0	0.16
FR-based	Advfaces Deb et al. (2020)	**30.06**	0.04	0.04	0.71	**27.46**	0.08	0.26	**0.79**	30.17	0.04	0.0	0.67
	TIP-IM Yang et al. (2021)	30.78	0.04	0.0	**0.78**	27.83	0.09	0.26	0.76	29.31	**0.06**	0.03	0.67
	Anti-Fake Vaccine	30.68	**0.14**	**0.62**	**0.92**	**26.31**	**0.22**	**0.83**	**0.90**	**27.72**	**0.16**	**0.97**	**0.96**

Bold and underlined values indicate the best performance, while bolded values indicate the second-best performance
'-' indicates the face swapping model that is not used in the training process

Table 4 Comparison of the proposed method with existing methods against six face swapping models on the FFHQ dataset

Methods	FS models	Faceshifter Li et al. (2020)				Simswap Chen et al. (2020)				-InfoSwap Gao et al. (2021)			
	metrics	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑
DF-Based	Disrupting Ruiz et al. (2020)	33.74	0.04	0.0	0.31	**26.50**	**0.16**	**0.35**	**0.95**	32.64	0.04	0.0	0.40
	Anti-forgery Wang et al. (2022)	44.06	0.01	0.0	0.12	43.71	0.02	0.0	0.02	36.39	0.02	0.0	0.10
	Initiative Huang et al. (2021)	37.04	0.02	0.01	0.09	38.39	0.04	0.0	0.04	34.59	0.03	0.06	0.16
Transfer-based	Regional Li et al. (2020)	29.39	**0.11**	**0.15**	0.67	29.23	0.08	0.23	0.55	29.12	**0.09**	**0.23**	0.76
	Transfer Nakka and Salzmann (2021)	32.39	0.05	0.0	0.17	34.19	0.04	0.0	0.04	31.81	0.05	0.01	0.27
FR-based	Advfaces Deb et al. (2020)	32.22	0.04	0.0	0.51	31.22	0.03	0.0	0.15	31.02	0.05	0.0	0.54
	TIP-IM Yang et al. (2021)	**29.25**	0.07	0.04	**0.82**	29.27	0.04	0.06	0.49	**28.44**	0.07	0.12	**0.84**
	Anti-FakeVaccine	**27.50**	**0.19**	**0.97**	**0.95**	**24.42**	**0.21**	**0.98**	**0.99**	**27.56**	**0.18**	**0.95**	**0.97**

Methods	FS models	-E4S Liu et al. (2023a)				-MegaFS Zhu et al. (2021)				-FaceDancer Rosberg et al. (2023)			
	metrics	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑	PSNR ↓	LPIPS↑	PSR_1↑	PSR_2↑
DF-Based	Disrupting (Ruiz et al., 2020)	31.56	0.05	0.05	0.49	31.71	0.07	0.06	0.46	32.78	0.05	0.0	0.40
	Anti-forgery Wang et al. (2022)	34.36	0.04	0.01	0.27	40.21	0.03	0.02	0.22	43.79	0.03	0.0	0.10
	Initiative Huang et al. (2021)	29.28	0.06	**0.72**	0.35	29.25	0.11	**0.63**	0.42	38.56	0.02	0.0	0.17
Transfer-based	Regional Li et al. (2020)	32.90	**0.09**	0.21	0.67	29.79	**0.12**	0.43	0.66	29.43	**0.06**	**0.23**	0.71
	Transfer Nakka and Salzmann (2021)	**28.56**	0.06	0.12	0.46	**27.12**	0.10	0.24	0.49	33.21	0.04	0.0	0.22
FR-based	Advfaces (Deb et al., 2020)	29.15	0.06	0.10	0.67	27.61	0.06	0.2	0.71	29.81	0.03	0.04	0.70
	TIP-IM Yang et al. (2021)	30.00	0.05	0.06	**0.69**	27.57	0.06	0.2	**0.78**	**28.82**	0.03	0.12	**0.75**
	Anti-FakeVaccine	**28.92**	**0.13**	**0.93**	**0.93**	**27.44**	**0.21**	**0.9**	**0.90**	**26.83**	**0.12**	**0.94**	**0.96**

Bold and underlined values indicate the best performance, while bolded values indicate the second-best performance
‘-‘ indicates the face swapping model that is not used in the training process

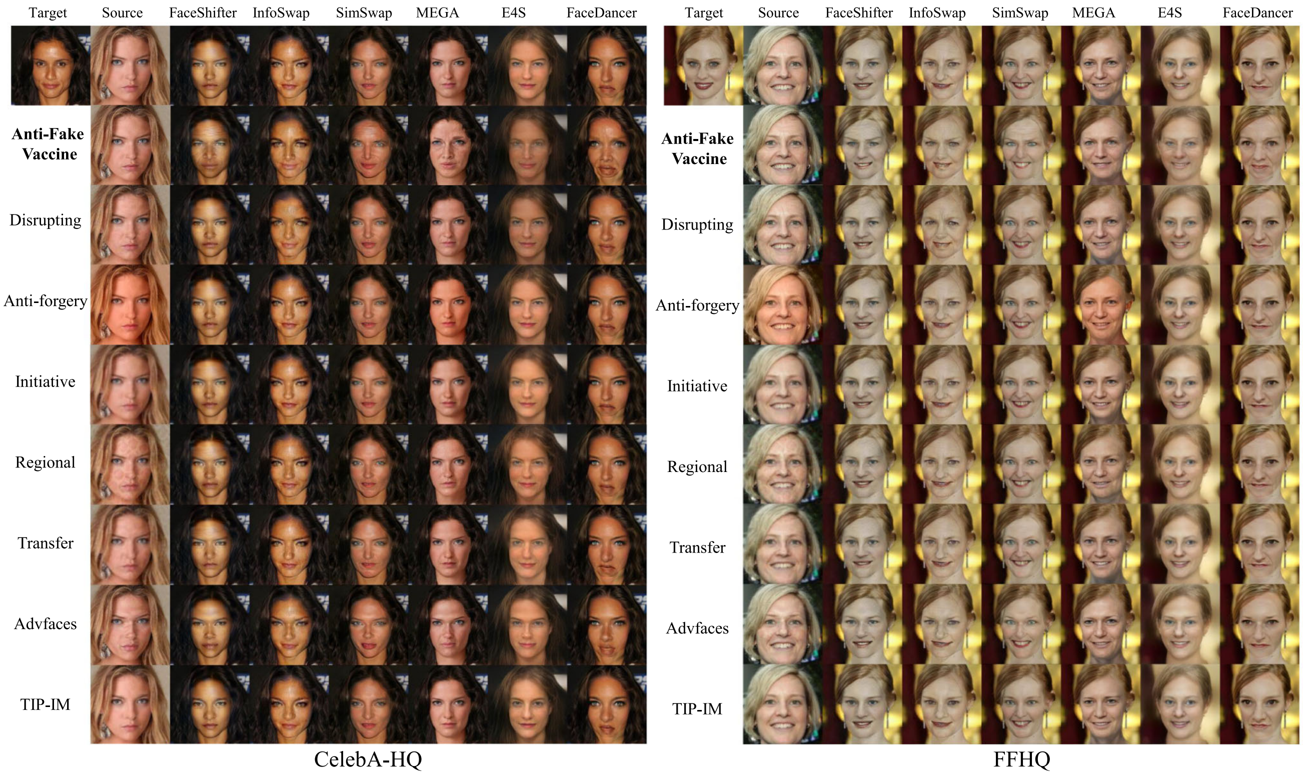

Fig. 5 Qualitative comparison with different methods on the CelebA-HQ and FFHQ datasets. The first column shows the target face. The second column shows the original face and the protected face produced by different methods. The remaining columns show the visual results of different methods against six face-swapping models (color figure online)

fake face synthesis process across various models, resulting in face swapping models obtaining inaccurate identity codes.

Comparison with DeepFake-based Adversarial Methods. Three DeepFake-based adversarial methods are implemented. From Tables 3 and 4, we observe that the DeepFake-based methods exhibit weak transferability on most target models since they are designed and tuned for specific DeepFake models. Specifically, we train the Disrupting method (Ruiz et al., 2020) on SimSwap (Chen et al., 2020). We observe that this method achieves the second-best performance on SimSwap (Chen et al., 2020) and performs poorly on the other face swapping models. For the Anti-forgery method (Wang et al., 2022), we directly use the model trained on the public tool Faceswap.[4] Obviously, the results show that this method has very limited defense against the six face swapping models. The Initiative method (Ruiz et al., 2020) utilizes a surrogate model of the face manipulation model for training, yet its performance remains unsatisfactory. The same conclusion can be drawn from Fig. 5 as well. Since these methods rely on specific models to generate adversarial perturbations, they exhibit weak transferability to unseen models. In contrast, our method incorporates the common features of FS models in terms of attack style and iterative optimization, enabling it to generate strong defense perturbations.

Comparison with Transfer-based Adversarial Methods. Since our privacy-preserving method utilizes adversarial example techniques, traditional adversarial methods are a feasible way. We select two adversarial methods designed to achieve strong transferability (Regional (Li et al., 2020) and Transfer (Nakka & Salzmann, 2021)) as the baselines. As observed from the results in Tables 3 and 4, *Anti-Fake Vaccine* outperforms all the transfer-based methods across various settings and evaluated models. The quantitative results show that the protected images produced by the Regional method (Li et al., 2020) exhibit a modest impact across diverse face swapping models compared with all the other methods. However, the visualization results 5 reveal that this method introduces a significant amount of distortion into the source image. Furthermore, our method still outperforms it on PSR metrics by more than 50% on average.

Comparison with FR-based Adversarial Methods. To illustrate the difference from face identity attacks, we also compare *Anti-Fake Vaccine* with two FR-based adversarial methods. From Tables 3 and 4, the results show that these methods are ineffective against the face swapping attacks.

[4] https://github.com/Oldpan/Faceswap-Deepfake-Pytorch.

While the FR-based methods show a certain defense effect on the PSR_2 metric, they perform poorly on the other three metrics. Similarly, as shown in Fig. 5, the swapped facial images obtained by these methods are indistinguishable from the original swapped facial images. Therefore, we can see that these methods are only effective for face recognition models and are not suitable for our task.

4.4.2 Defense Against Commercial Models

In the real world, the public commonly adopts commercial APIs to perform the face swap operations. To evaluate the effectiveness of our method, we employ three commercial face swapping systems. We use the same protected images generated by our method as input to various commercial APIs. Subsequently, we calculate the privacy metrics for the original swapped images and the protected swapped images.

As observed from the results of CelebA-HQ in Table 5, *Anti-Fake Vaccine* demonstrates strong defense performance against three commercial APIs. Specifically, our method achieves an impressive 63% of PSR_1 and 81% of PSR_2 on average, which are remarkably high in real-world scenarios. Additionally, the PSNR and LPIPS measurements indicate that our method substantially reduces the visual quality of the swapped images. Based on the visualization results presented in Fig. 1, it becomes evident that the swapped faces using our protected images exhibit distinct appearances compared to the original swapped faces. This demonstrates the success of our method in safeguarding individuals' face, significantly mitigating the impact of malicious face swapping.

Table 5 Privacy Protection performance of our method against three commercial FS models on the CelebA-HQ dataset

APIs	PSNR↓	LPIPS↑	PSR_1↑	PSR_2↑
Alibaba	28.85	0.11	0.81	0.75
Baidu	27.92	0.10	0.62	0.95
Tencent	29.66	0.13	0.50	0.74

4.5 Evaluation on Image Utility

To evaluate the utility of facial images for practical applications, we focus on two key aspects: visual quality and protect against the common image processing attacks. In the visual quality evaluation, we compute the utility metrics between the original images and protected images. For protection against the common image processing attacks, we calculate privacy metrics across different distraction attack scenarios.

As shown in the Table 6, the RMSE score indicates that our method has comparable or less variation at the pixel level than other methods, while the PSNR score demonstrates that our method generates images with good quality. Additionally, based on the SSIM and LPIPS scores, the images generated by our method exhibit a high degree of similarity to the original images in terms of human perception. These findings suggest that the images injected with our *Anti-Fake Vaccine* can be used normally. Similarly, Fig. 5 also shows our superior performance. In summary, our method generates more transferable and imperceptible protected faces.

In real scenarios, images uploaded on social media platforms could undergo various image processing procedures (e.g., compression, blurring, etc.). To validate the robustness of our method in these cases, we evaluate the effectiveness of our results against four common image processing techniques, including JPEG compression, Random noise, Rotation, and Gaussian blur. Specifically, we perform rotation using four angles and apply three kernels to perform Gaussian blur in the face area. Table 7 reports the PSR_1 results of our method in tackling various attack scenarios. The experimental results demonstrate that our method exhibits robustness and performs well under common image processing techniques.

Table 6 Visual quality performance of our method is compared with six baselines on the CelebA-HQ and FFHQ datasets

Dataset	CelebA-HQ				FFHQ			
metric	SSIM ↑	PSNR↑	RMSE ↓	LPIPS ↓	SSIM ↑	PSNR ↑	RMSE ↓	LPIPS ↓
Disrupting Ruiz et al. (2020)	0.92	**35.57**	**4.25**	0.14	0.89	**32.07**	**5.05**	0.16
Anti-forgery Wang et al. (2022)	**0.94**	35.50	4.36	**0.09**	**0.91**	31.60	6.78	**0.13**
Initiative Huang et al. (2021)	0.80	28.26	10.14	0.41	0.81	28.15	10.23	0.42
Regional Li et al. (2020)	0.55	24.00	16.18	0.59	0.55	23.81	16.54	0.61
Transfer Nakka and Salzmann (2021)	0.76	28.58	9.64	0.43	0.76	28.38	9.82	0.46
Advfaces Deb et al. (2020)	0.83	27.81	10.78	0.32	0.83	27.09	11.70	0.35
TIP-IM Yang et al. (2021)	0.87	31.14	7.14	0.24	0.86	30.89	7.33	0.23
Anti-Fake Vaccine	**0.96**	**35.63**	**4.24**	**0.05**	**0.92**	**32.58**	**6.07**	**0.11**

Bold and underlined values indicate the best performance, while bolded values indicate the second-best performance

Table 7 Privacy protection performance of our method under different common image processing attaks on the CelebA-HQ dataset

FS models	FaceShifter Li et al. (2020)				SimSwap Chen et al. (2020)				-InfoSwap Gao et al. (2021)			
image processing	Rescale	Random noise	Rotate	Gaussian	Rescale	Random noise	Rotate	Gaussian	Rescale	Random noise	Rotate	Gaussian
Disrupting Ruiz et al. (2020)	0.0	0.0	0.0	0.0	0.04	0.12	0.03	0.02	0.0	0.02	0.01	0.01
Anti-forgery Wang et al. (2022)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0
Initiative Huang et al. (2021)	0.06	0.04	0.04	0.02	0.03	0.02	0.01	0.03	0.10	0.14	0.15	0.08
Regional Li et al. (2020)	**0.26**	**0.31**	**0.28**	**0.32**	**0.28**	**0.3**	**0.23**	**0.25**	**0.25**	**0.28**	**0.21**	**0.16**
Transfer Nakka and Salzmann (2021)	0.0	0.02	0.01	0.0	0.0	0.02	0.01	0.0	0.01	0.01	0.03	0.01
Advfaces Deb et al. (2020)	0.0	0.04	0.01	0.02	0.0	0.02	0.01	0.03	0.01	0.13	0.03	0.07
TIP-IM Yang et al. (2021)	0.04	0.05	0.07	0.05	0.07	0.07	0.07	0.06	0.13	0.12	0.15	0.11
Anti-Fake Vaccine	**0.83**	**0.85**	**0.9**	**0.85**	**0.98**	**0.98**	**0.98**	**0.98**	**0.81**	**0.81**	**0.87**	**0.83**

Bold and underlined values indicate the best performance, while bolded values indicate the second-best performance

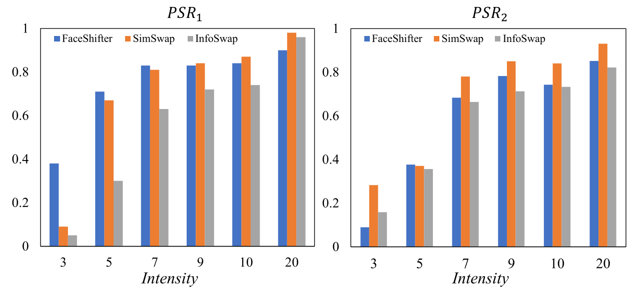

Fig. 6 The trend of the intensity of perturbations v_a and the protection success rate PSR_1 and PSR_2. The x-axis represents the intensity of v_a. The y-axis represents the protection success rate at the corresponding intensity against three face swapping models (color figure online)

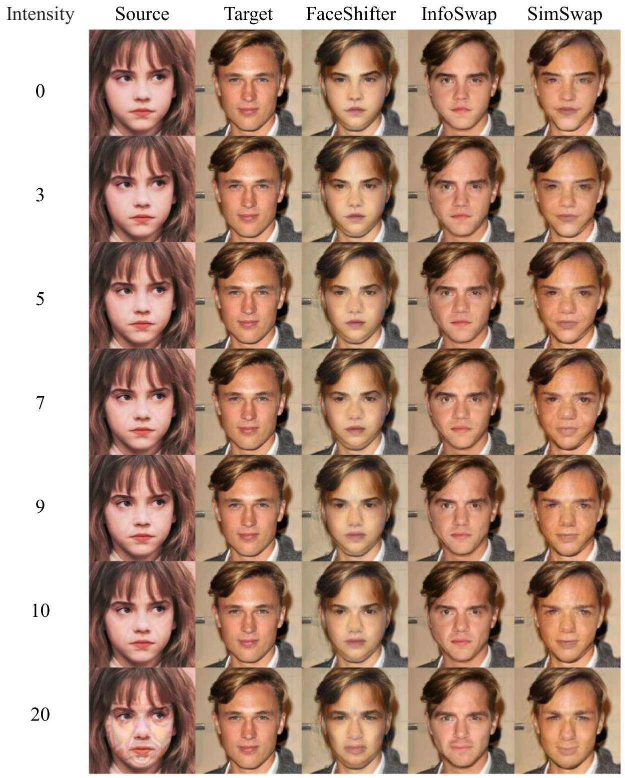

Fig. 7 The visual results controlled by the intensity of perturbations (color figure online)

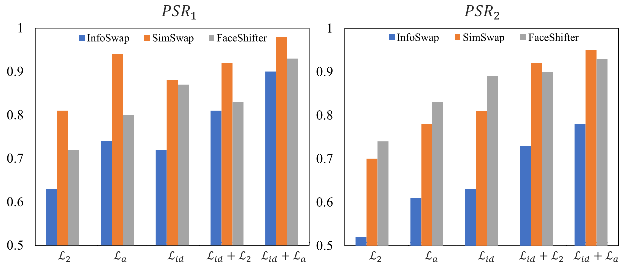

Fig. 8 The protection performance with different regularization constraints against three face swapping models (color figure online)

4.6 Ablation Study

To demonstrate the impact of different components of *Anti-Fake Vaccine*, we perform an ablation study on the CelebA-HQ dataset. Specifically, we focus on three components of the proposed framework: 1) the intensity of perturbations v_a, 2) two types of regularization constraints and multi-objective optimization method, 3) the additive perturbation strategy.

Firstly, we investigate the level of protection offered by our approach against face swapping under various perturbation magnitudes. Figure 6 shows the PSR values for the three face swapping models by adding perturbations from 0 to 20. The experimental results indicate that greater perturbations result in improved protection performance. However, Fig. 7 shows that the larger perturbations also introduce unnatural artifacts in the source data, which can compromise our image utility requirements. Thus, we set the maximum perturbation to 10 across all experiments.

Secondly, we analyze the privacy protection performance of different constraint tasks. For the visual perceptual constraint, $\mathcal{L}_2$ and $\mathcal{L}_a$ are loss functions that induce quality distortion in the swapped output. $\mathcal{L}_2$ refers to the pixel-level distance regularization commonly used by existing protection methods, while $\mathcal{L}_a$ is the similarity-aware regularization we proposed. Figure 8 shows that the visual perceptual regularization achieves better protection than the pixel-level distance regularization. Additionally, the combination of $\mathcal{L}_a$ and $\mathcal{L}_{id}$ achieves effective protection across three face swapping models. For the identity similarity constraint, we show the variation of visual effects under different boundary values of η. As shown in the Fig. 9, a smaller value of η, indicating a larger distance between the protected face identity and the original face identity, results in a more significant alteration in the content of the swapped facial image. Nevertheless, it is worth noting that a higher value can also yield the desired protection output. Hence, we design this parameter to achieve the protection effect while reducing the related cost. Additionally, users can choose different protection levels based on η.

For the multi-objective optimization term, we experimentally compare our proposed dynamically optimized weights with fixed weights. From Fig. 10a, we observe that employing dynamic optimization parameters leads to optimal performance in terms of both the strength and generalization of privacy protection. This is attributed to the fact that, in each iteration, we are able to attain an optimal balance between the two regularization terms.

Thirdly, we investigate the importance of the additive perturbation strategy. The objective of this module is to further improve the generalization of our method. In this experiment, we conduct individual optimizations using the three

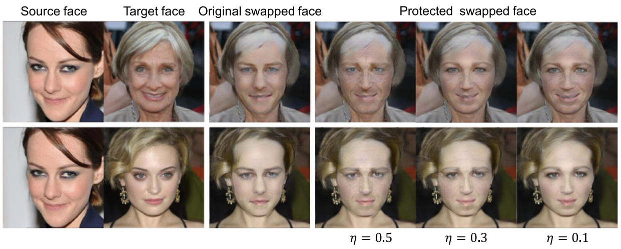

Fig. 9 The visual results controlled by identity boundaries η (color figure online)

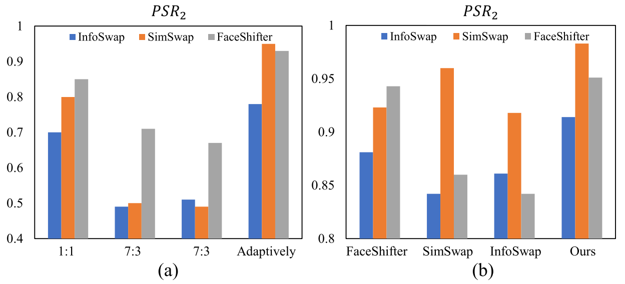

Fig. 10 The protection performance against three face swapping models: **a** PSR_2 obtained using different weighting values; **b** PSR_2 obtained using different gradient update method (color figure online)

face swapping models, and also employ optimization based on our proposed additive perturbation strategy. According to the quantitative results shown in Fig. 10b, the additive perturbation strategy improves both the protection strength and generalization.

5 Conclusion

In this paper, we develop a novel privacy-preserving framework *Anti-Fake Vaccine* for facial images to defend against the malicious face swapping. Our method exhibits strong generalization to unseen models while ensuring image utility. Unlike the prior defense methods employing pixel-level distance regularization, we introduce a dual degradation mechanism focusing on visual quality and identity semantics. During optimization, we propose a multi-objective optimal solution to achieve the utmost strength and generalization of privacy protection. Additionally, we design an additive perturbation strategy to integrate the shared adversarial perturbations from various face swapping models and target face images. Extensive experiments on the CelebA-HQ and FFHQ datasets demonstrate the effectiveness of *Anti-Fake Vaccine* in terms of privacy protection and image utility. Notably, our method shows the superior generalizability across various face swapping models, including commercial ones. In the future, our research will extend to privacy protection against other types of DeepFake models. Furthermore, we will explore the robustness of our privacy-preserving framework under more image processing attacks, including image reconstruction.

Acknowledgements This work was supported in part by the National Key R&D Program of China under (Grant No. 2021YFB3100800); in part by the National Natural Science Foundation of China (No. 62306308, 62025604, U22A2030, 62372448), in part by Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025), in part by the Foundation of Key Laboratory of Education Informatization for Nationalities (Yunnan Normal University), Ministry of Education (No. EIN2024B004).

Data Availibility The data that support the findings of this study are available from the authors upon reasonable request.

References

Chen, R., Chen, X., Ni, B., & Ge, Y., (2020) Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International conference on multimedia*, pp. 2003–2011.

Cheng, H., Guo, Y., Wang, T., Li, Q., Chang, X., & Nie, L. (2022). Voice-face homogeneity tells deepfake. arXiv preprint arXiv:2203.02195

Deb, D., Zhang, J., & Jain, A. K. (2020). Advfaces: Adversarial face synthesis. In *2020 IEEE international joint conference on biometrics (IJCB)*, pp. 1–10. IEEE.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, June.

Frank, J., & Holz, T. (2021). [RE] CNN-generated images are surprisingly easy to spot... for now. arXiv preprint arXiv:2104.02984

Gao, G., Huang, H., Fu, C., Li, Z., & He, R. (2021). Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3404–3413.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *Statistics, 1050*, 20.

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2019). Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing, 28*(11), 5464–5478.

Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., Chu, W., Chen, J., Lin, W., & Ma, K-K. (2022). Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 989–997.

Huang, Q., Zhang, J., Zhou, W., Zhang, W., & Nenghai, Y. (2021). Initiative defense against facial manipulation. In *Proceedings of the AAAI conference on artificial intelligence,* vol. 35, pp. 1619–1627.

Ilyas, H., Javed, A., & Malik, K. M. (2023). Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. *Applied Soft Computing, 136*, 110124.

Jiang, D., Song, D., Tong, R., & Tang, M. (2023). Styleipsb: Identity-preserving semantic basis of stylegan for high fidelity face swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 352–361.

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision, 130*(7), 1678–1734.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410.

Kim, J., Lee, J., & Zhang, B-T., (2022). Smooth-swap: A simple enhancement for face-swapping with smoothness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10779–10788.

Kim, M., Tariq, S., & Woo, S. S. (2021). Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1001–1012.

Kuhn, H. W., & Tucker, A. W. (2013). Nonlinear programming. In *Traces and emergence of nonlinear programming*, pp. 247–258. Springer.

Lee, J., Hyung, J., Jeong, S., & Choo, J. (2024). Selfswapper: Self-supervised face swapping via shape agnostic masked autoencoder. arXiv preprint arXiv:2402.07370

Lee, C. H., Liu, Z., Wu, L., & Luo, P. (2020). Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5549–5558.

Li, Y., Bai, S., Xie, C., Liao, Z., Shen, X., & Yuille, A. (2020). Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses. In *Computer Vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 795–813. Springer.

Li, L., Bao, J., Yang, H., Chen, D., & Wen, F. (2020). Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5074–5083.

Li, Z., Yu, N., Salem, A., Backes, M., Fritz, M., & Zhang, Y. (2022). Unganable: Defending against gan-based face manipulation. arXiv preprint arXiv:2210.00957

Li, X., Zhang, S., Hu, J., Cao, L., Hong, X., Mao, X., Huang, F., Wu, Y., & Ji, R. (2021). Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8639–8648.

Lin, X., Chen, H., Pei, C., Sun, F., Xiao, X., Sun, H., Zhang, Y., Ou, W., & Jiang, P. (2019). A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proceedings of the 13th ACM conference on recommender systems*, pp. 20–28.

Liu, Z., Li, M., Zhang, Y., Wang, C., Zhang, Q., Wang, J., & Nie, Y. (2023a). Fine-grained face swapping via regional gan inversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8578–8587.

Liu, K., Perov, I., Gao, D., Chervoniy, N., Zhou, W., & Zhang, W. (2023b). Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition, 141*, 109628.

Li, J., Xie, H., Lingyun, Y., Gao, X., & Zhang, Y. (2021). Discriminative feature mining based on frequency information and metric learning for face forgery detection. *IEEE Transactions on Knowledge and Data Engineering, 35*(12), 12167–12180.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International conference on learning representations*.

Nakka, K.K., & Salzmann, M. (2021). Learning transferable adversarial perturbations. In *NeurIPS*.

Naruniec, J., Helminger, L., Schroers, C., & Weber, R. M. (2020). High-resolution neural face swapping for visual effects. *Computer Graphics Forum, 39*, 173–184.

Nirkin, Y., Keller, Y., & Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7184–7193.

Otto, C., Naruniec, J., Helminger, L., Etterlin, T., Mignone, G., Chandran, P., Zoss, G., Schroers, C., Gross, M., Gotardo, P., et al. (2022). Learning dynamic 3d geometry and texture for video face swapping. *Computer Graphics Forum, 41*, 611–622.

Poursaeed, O., Katsman, I., Gao, B., & Belongie, S.(2018). Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4422–4431.

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer.

Rosberg, F., Aksoy, E. E., Alonso-Fernandez, F., & Englund, C. (2023) Facedancer: Pose- and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 3454–3463.

Ruiz, N., Bargal, S. A., & Sclaroff, S. (2020). Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *Computer Vision–ECCV 2020 workshops: Glasgow, UK, August 23–28, 2020, proceedings, Part IV 16*, pp. 236–251. Springer.

Ruiz, N., Bargal, S. A., Xie, C., & Sclaroff, S. (2023). Practical disruption of image translation deepfake networks. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 14478–14486.

Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems*, *31*.

Shao, R., Lan, X., & Yuen, P. C. (2020). Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 11974–11981.

Shen, Y., Yang, C., Tang, X., & Zhou, B. (2022). Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(4), 2004–2018.

Shiohara, K., Yang, X., & Taketomi, T. (2023). Blendface: Redesigning identity encoders for face-swapping. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7634–7644.

Sun, Y., Yu, L., Xie, H., Li, J., & Zhang, Y. (2024). Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24584–24594.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Transactions on Graphics (ToG), 36*(4), 1–13.

Teotia, D., Lapedriza, A., & Ostadabbas, S. (2022). Interpreting face inference models using hierarchical network dissection. *International Journal of Computer Vision, 130*(5), 1277–1292.

Thambawita, V., Isaksen, J. L., Hicks, S. A., Ghouse, J., Ahlberg, G., Linneberg, A., Grarup, N., Ellervik, C., Olesen, M. S., Hansen, T., et al. (2021). Deepfake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Scientific Reports, 11*(1), 21896.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387–2395.

Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., & Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG), 40*(4), 1–14.

Tripathy, S., Kannala, J., & Rahtu, E. (2020). Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3385–3394.

Ververas, E., & Zafeiriou, S. (2020). Slidergan: Synthesizing expressive face images by sliding 3d blendshape parameters. *International Journal of Computer Vision, 128*(10–11), 2629–2650.

Wang, R., Huang, Z., Chen, Z., Liu, L., Chen, J., & Wang, L. (2022). Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. arXiv preprint arXiv:2206.00477

Wang, R., Juefei-Xu, F., Luo, M., Liu, Y., & Wang, L. (2021). Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3546–3555.

Wiles, O., Koepke, A., & Zisserman, A. (2018). X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 670–686.

Wu, P-W., Lin, Y-J., Chang, C-H., Chang, E. Y , Liao, S-W. (2019). Relgan: Multi-domain image-to-image translation via relative attributes. In *Proceedings of the IEEE international conference on computer vision*, pp. 5914–5922.

Xu, Y., Deng, B., Wang, J., Jing, Y., Pan, J., & He, S. (2022). High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7642–7651.

Yang, C., Ding, L., Chen, Y., & Li, H. (2021). Defending against gan-based deepfake attacks via transformation-aware adversarial faces. In *2021 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE.

Yang, X., Dong, Y., Pang, T., Su, H., Zhu, J., Chen, Y., & Xue, H. (2021). Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 3897–3907, October.

Yeh, C. Y., Chen, H.-W., Tsai, S.-L., & Wang, S.-D. (2020). Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, pp. 53–62.

Yin, F., Zhang, Y., Wu, B., Feng, Y., Zhang, J., Fan, Y., & Yang, Y. (2023). Generalizable black-box adversarial attack with meta learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(3), 1804–1818.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595.

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2021). Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15023–15033.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194.

Zhu, Y., Li, Q., Wang, J., Xu, C-Z., & Sun, Z. (2021). One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4834–4844.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.