Adversarial Robust Safeguard for Evading Deep Facial Manipulation

Jiazhi Guan[1], **Yi Zhao**[2*], **Zhuoer Xu**[3], **Changhua Meng**[3], **Ke Xu**[1,4], **Youjian Zhao**[1,4*]
[1]DCST, BNRist, Tsinghua University
[2]Beijing Institute of Technology
[3]Ant Group
[4]Zhongguancun Laboratory
{guanjz20@mails., zhaoyoujian@}tsinghua.edu.cn, yizhao.tsinghua@gmail.com

Abstract

The non-consensual exploitation of facial manipulation has emerged as a pressing societal concern. In tandem with the identification of such fake content, recent research endeavors have advocated countering manipulation techniques through proactive interventions, specifically the incorporation of adversarial noise to impede the manipulation in advance. Nevertheless, with insufficient consideration of robustness, we show that current methods falter in providing protection after simple perturbations, *e.g.*, blur. In addition, traditional optimization-based methods face limitations in scalability as they struggle to accommodate the substantial expansion of data volume, a consequence of the time-intensive iterative pipeline. To solve these challenges, we propose a learning-based model, Adversarial Robust Safeguard (ARS), to generate desirable protection noise in a single forward process, concurrently exhibiting a heightened resistance against prevalent perturbations. Specifically, our method involves a two-way protection design, characterized by a basic protection component responsible for generating efficacious noise features, coupled with robust protection for further enhancement. In robust protection, we first fuse image features with spatially duplicated noise embedding, thereby accounting for inherent information redundancy. Subsequently, a combination comprising a differentiable perturbation module and an adversarial network is devised to simulate potential information degradation during the training process. To evaluate it, we conduct experiments on four manipulation methods and compare recent works comprehensively. The results of our method exhibit good visual effects with pronounced robustness against varied perturbations at different levels.

Introduction

Human-central entertainment is booming with rapid advancements in artificial intelligence applications. One of the most important topics, facial manipulation, is now being able to deliver photo-realistic outcomes given simply a target facial image. Although these techniques are proposed for enriching people's living world, they also induce malicious usages such as non-consensual portrait violation, even producing nasty porn content. Concerns from society promote many countermeasures in academic groups.

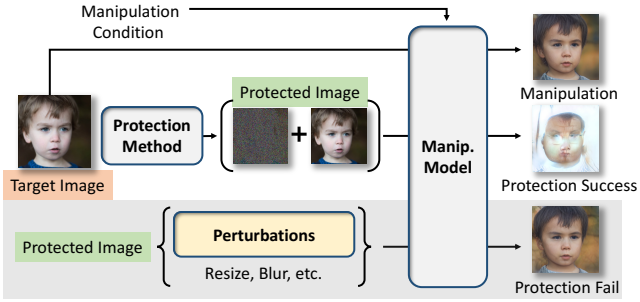

Figure 1: Current protection methods can protect images from malicious manipulations by adding adversarial noise, as we see disrupted outcomes of the manipulation (manip.) model. Nevertheless, the performance is largely compromised following simple perturbations, as the target is successfully manipulated with a desirable style in the last row.

Fake detection has been broadly studied in recent years (Li et al. 2020; Sun et al. 2021a; Xu et al. 2023; Huang et al. 2023). Identifying the existing fake content cannot erase the possible negative impact that has already been done. Therefore, an additional avenue of priors (Ruiz, Bargal, and Sclaroff 2020; Huang et al. 2022) delves into a proactive solution, leveraging adversarial noise to pre-protect facial contents. Despite promising results reported that injected noise is able to disrupt the manipulation process, there are still some issues to be solved. 1) Optimization-based methods (*e.g.*, (Ruiz, Bargal, and Sclaroff 2020)) require from-scratch optimization for each image independently, which costs considerable time and computational resources. 2) The manipulation methods are sensitive to subtle changes introduced by the injected noise, while the injected noise itself is also susceptible to minor alterations. There is still no comprehensive evaluation regarding the robustness of the injected noise against manipulations when fronting commonly existing perturbations on the Internet. This problem is intuitively demonstrated in Fig. 1. Our preliminary experiment finds the vulnerability of current studies in this aspect.

With a spirit to tackle these challenges, in this paper, we develop a learning-based protection framework named Adversarial Robust Safeguard (ARS). ARS is designed to achieve highly efficient and perturbation-resistant pre-

*Corresponding author.

protection, in the context of countering deep facial manipulations. One of the pivotal concepts central to our approach involves the design of a neural network to generate input-specific protection noises. These noises are subsequently injected into the original input with the primary objective of disrupting the outcomes of manipulations. Thus, different from optimization-based methods (Ruiz, Bargal, and Sclaroff 2020; Yeh et al. 2020), the entire protection could be done in a single-forward process after one-time training of our model.

On the other hand, the robustness of noises is emphasized in our method from two folds. Initially, we propose a two-way protection framework. The first one, denoted as Basic Protection, is designed to learn and generate an effective noise feature. And second one, termed Robust Protection, is aimed at producing a noise signal that ensures robustness. We learn the information replication measure from (Wang et al. 2021; Guan et al. 2023a), which is inspired by Shannon's capacity theorem that redundancy could improve robustness. Different from direct duplication of the input in previous works, we spatially repeat the noise embedding from the Basic Protection. Thus the protective information is injected individually and spread throughout the spatial dimensions, ensuring proper information integrity from being recovered after certain lossy perturbations. Second, from the training aspect, a combination of a differentiable perturbation module and an adversarial net is proposed to simulate possible information loss in training. Similar to related applications in (Zhu et al. 2018), including the perturbation module can enhance information preservation. We adopt the same idea to power better resistance to perturbations of our protection noises. In addition, an adversarial net is proposed to play a role against our protection model, which aims to erase the injected information. As a result, our model can benefit from an adversarial training process, achieving more robust protection.

We conduct experiments on four manipulation methods and compare recent works comprehensively. From the first aspect of protection visual effect, our method reports objectively better performance in several metrics and shows clearly better disruption effect visually. Regarding robustness evaluation, we conduct comparisons with six types of perturbations at five different levels. Our method consistently reports the best performance both quantitatively and qualitatively. Our contributions can be summarized as:

- We develop a learning-based two-way model achieving high efficiency and perturbation-resist pre-protection against deep facial manipulations.
- By introducing a differentiable perturbation module and an adversarial training strategy, our model is learned to generate more robust protection noises that better tolerate possible information loss fronting perturbations.
- We conduct comprehensive experiments in different aspects. Comparisons with previous studies show that our method achieves better protection both qualitatively and quantitatively.
- To the best of our knowledge, we are the first to benchmark the robustness evaluation of adversarial protection methods against facial manipulations. Considering the ubiquitous post-processing in multimedia platforms, in-depth studies in this aspect could greatly promote the real-world applications of related protection methods.

Related Works

Facial Manipulation

In light of the rapid progress in generative methods, recent research has demonstrated the capability to attain vivid facial manipulations. StyleGAN (Karras et al. 2020) represents a milestone in image synthesis, capable of being trained to produce high-fidelity images given only random initialization. Drawing upon this foundation, plenty of studies have been devised to realize controlled manipulations (Yang et al. 2021; Guan et al. 2023b). For instance, built upon a well-trained StyleGAN serving as the decoder, (Richardson et al. 2021) attains desirable manipulations through the employment of diverse encoding strategies. An even more powerful application involves identity swapping. (Chen et al. 2020) is able to seamlessly transform facial identities between two provided images, all the while ensuring the preservation of the original attributes with high fidelity. The pressing issue of non-consensual exploitation of these manipulation methods has garnered significant societal attention. It is imperative to underscore the critical importance of developing countermeasures.

Facial Manipulation Detection

Detection of facial manipulation, as a passive defense measure against abuses, has been widely studied in recent years. Most works (Sun et al. 2021b; Qian et al. 2020; Li et al. 2020; Guan et al. 2023c) are formulated as a binary classification problem, with the objective of identifying pre-existing fake contents. While powerful deep neural networks could provide a good understanding of the differences between real faces and the generated fake ones, recent studies (Sun et al. 2021a; Luo et al. 2021; Guan et al. 2022; Dong et al. 2022; Yao et al. 2023; Yan et al. 2023; Dong et al. 2023) find it hard to keep a consistent performance from the training set to untapped manipulations. The weak generalization ability encourages researchers to dig deeper into this problem. Even if we could identify fakes with precision, we cannot erase the negative impact that could already be done. Thus, an additional avenue of priors delves into a proactive solution.

Adversarial Attack against Facial Manipulation

Adversarial attack (Goodfellow, Shlens, and Szegedy 2014) is initially studied in the context of a classification problem, that a deep network can be fooled with small pixel-level changes. It has penetrated into many fields such as emotion recognition (Zhao et al. 2021b), network intrusion detection (Chen et al. 2023), speech recognition (Guo et al. 2022), and edge computing (Zhao et al. 2021a). Recent researchers have adopted a similar idea and turned it into a proactive defensive method against facial manipulations. (Ruiz, Bargal, and Sclaroff 2020) first conduct experiments with well-established baselines. Their results prove a fruitful avenue

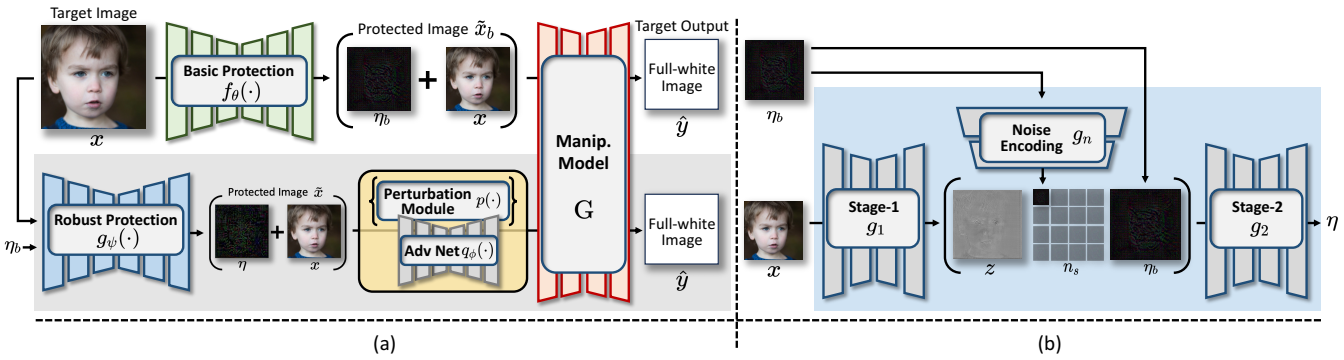

Figure 2: (a): Overview of the proposed Adversarial Robust Safeguard (ARS). Our two-way protection model consists of the Basic Protection (green) and the Robust Protection (blue). Manipulation (Manip.) Model (red), Perturbation Module (yellow), and Adv Net (yellow) are exclusively employed during the training phase, where the yellow part is especially leveraged to simulate instances of information loss after the protection. (b): Illustration of the Robust Protection.

for proactive defense against facial manipulations. (Huang et al. 2022) adopt the idea of universal adversarial perturbation and devise an effective method against multiple manipulations simultaneously. (He et al. 2022) propose to protect images in the latent space of StyleGAN (Karras et al. 2020) and disable manipulations via reconstruction. Tiny changes in the latent space could disable manipulations, but also introduce minor changes to the original identity. (Aneja, Markhasin, and Nießner 2022) is a close study of our motivations, which focuses on designing a learning-based protection model. Their method is able to protect images through one single forward process. While these methods can achieve successful protection, the performance when fronting various kinds of perturbations is still unclear. Considering the ubiquitous post-processing in multimedia platforms, our method is distinct from previous works especially designs in robust protection, for better support of practical scenarios.

Method

Preliminary on Adversarial Attack

In this paper, our focus lies on safeguarding images by employing attacks on facial manipulation models. For this reason, we refrain from explicitly distinguishing between the terms "attack" and "protect" within the following context. Both terms are applicable in describing our objective, where the "attack" pertains to the manipulation model, and the "protect" pertains to the target image we aim to safeguard.

Given a target image $x \in \mathbb{R}^{H\times W\times 3}$ and a well-trained manipulation model G. The original manipulated output $y = \mathrm{G}(x, c)$, where c is an optional conditional input depending on G. We aim to find a desirable noise $\eta \in \mathbb{R}^{H\times W\times 3}$, which is additive to x as $\tilde{x} = \mathrm{clip}(x + \eta)$, where $\tilde{x}$ is the protected image and clip function ensures a validate result of addition. Thus the disrupted/protected manipulation output $\tilde{y}$ is defined as $\tilde{y} = \mathrm{G}(\tilde{x}, c)$.

A desirable noise η should be human-neglectable to the visual effect of x and be able to considerably disrupt the generative results of G. Learning from adversarial attack methods (Kurakin, Goodfellow, and Bengio 2018), a FGSM-based baseline could be defined as:

$$\eta = \epsilon\,\mathrm{sign}(\nabla_x J(\mathrm{G}(x, c))), \tag{1}$$

where ϵ describes that each pixel of x can be changed no more than ϵ, and J is a loss function defining the protection goal. When J is L_2-norm constraint, maximum $J(\mathrm{G}(x, c))$ will lead to a protected output $\tilde{y}$ to be less similar to the original manipulation in y, achieving our goal of protection.

To provide clarity in our following descriptions, in this paper, the term "noise" refers to the outcome η in our objectives. Conversely, the term "perturbation" is employed to signify potential post-processing that occurs subsequent to the completion of our protection.

Adversarial Robust Safeguard

After defining the problem, we elaborate the Adversarial Robust Safeguard (ARS) in this section. The overview of the proposed method is illustrated in Fig. 2. ARS follows a learning-based manner, where we generate desirable noise η for each target image x independently in only one forward process. The two-way design is outlined at the upper and lower pipelines in Fig. 2 (a), respectively.

Basic Protection. In the upper one, the basic protection function $f_\theta(\cdot)$ with trainable parameters θ will generate basic noise $\eta_b = f_\theta(x)$. Different from the maximum-disruption optimization protocol adopted in (Yeh et al. 2020; Huang et al. 2022), we set a protection target similar to (Aneja, Markhasin, and Nießner 2022). Thus f_θ is optimized as:

$$\min_\theta \|\mathrm{G}(\tilde{x}_b) - \hat{y}\|_2 + \lambda \|\eta_b\|_2\,, \tag{2}$$

$$\begin{aligned} \tilde{x}_b &= \mathrm{clip}(x + \eta_b) = \mathrm{clip}(x + f_\theta(x)), \\ s.t. \quad & \|\eta_b\|_\infty < \epsilon, \end{aligned} \tag{3}$$

where $\hat{y}$ is the protection target that is simply set as full-white image in the same shape of x, λ is set to 10, the last $\|\eta_b\|_2$ is a regular term preventing too large the noise, and L_∞-constraint is implemented by clip function as well.

To optimize Eq. (2), we could have a desirable noise η_b that is suitable for target image x to nullify the manipulation of G. Naturally, sensitive information for attacking the manipulation model is also extracted in η_b. We thus later leverage this critical information in the next stage for further enhancement.

Robust Protection. In the second way of robust protection, we also introduce a protection function $g_\psi(\cdot)$ with trainable parameters ψ. Details of its design are illustrated in Fig. 2 (b). The main idea of this part is to enhance information redundancy for robust considerations. Therefore, our protection could be more effective even after perturbations.

The robust protection $g_\psi(\cdot)$ is split it into three components, where the stage-1, stage-2, and noise encoding are denoted as g_1, g_2, and g_n, respectively. The stage-1 takes the target image x as input, for feature extraction of the specific input:

$$z = g_1(x), \quad z \in \mathbb{R}^{H\times W\times C}, \tag{4}$$

where C is a hyper-parameter. The noise encoding function g_n is responsible for compressing the spatial dimension of η_b while keeping useful information for our protection purposes:

$$n = g_n(\eta_b), \quad n \in \mathbb{R}^{1\times 1\times C}. \tag{5}$$

Then, we spatially repeat the compressed feature n by $H \times W$ times to be $n_s \in \mathbb{R}^{H\times W\times C}$. At last, we concatenate the above feature in channel dimension as input to get the final protection:

$$\begin{aligned} \eta &= g_2(\mathrm{cat}(z, n_s, \eta_b)), \\ \tilde{x} &= \mathrm{clip}(\mathrm{x} + \eta), \\ s.t. \quad &\|\eta\|_\infty < \epsilon. \end{aligned} \tag{6}$$

Therefore, even if the protected $\tilde{x}$ is processed to lose certain information at a specific sub-corner, our method can ensure proper information integrity from being recovered. The whole g_ψ is optimized end-to-end with the following parts, which we will introduce later.

Perturbation Module and Adv Net. Both the two components are adopted to enhance the robustness of the protection noise η. We denote the perturbation module as function $p(\cdot)$, which is consisted of four differentiable functions, including p_{drop}, p_{resize}, p_{blur}, and p_{jpeg}, to simulate possible perturbations. For p_{drop}, we define it as:

$$p_{drop}(\tilde{x}, x) = \tilde{x} \cdot m + x \cdot (1 - m), \tag{7}$$

where $m \in \{0,1\}^{\mathbb{R}^{H\times W\times 1}}$ here is a mask that each value is randomly assigned independently. Therefore, p_{drop} could simulate pixel-level lossy perturbations. Additionally, p_{resize} is simply implemented using bilinear interpolation. We randomly downscale and then upscale $\tilde{x}$ back. Moreover, p_{blur} is implemented using convolutional operation with predefined blur kernels. Lastly, p_{jpeg} aims to simulate information loss after JPEG compression. Since there is a lossy quantization step in the compression process, which is non-differentiable, we cannot directly adopt a standard JPEG compression in our training. Learning from an approximate method (Zhu et al. 2018), we randomly mask out part of the high-frequency coefficients after the DCT-transform step of the compression process. During training, we apply the introduced functions with a probability of 0.5. Therefore, different combinations of these functions will simulate a variety of situations.

Despite $p(\cdot)$ can be helpful in dealing with certain perturbations, we cannot exhaustively enumerate all possible scenarios in training. Therefore, we propose the Adv Net, denoted as $q_\phi(\cdot)$ with trainable parameters ϕ, to nullify our protection. q_ϕ is optimized by:

$$\min_\phi \|\mathrm{G}(q_\phi(\tilde{x})) - y\|_2, \tag{8}$$

recalling $y = \mathrm{G}(x, c)$ is the original manipulation result. Our model is trained to bypass any information loss that could be caused in q_ϕ as:

$$\min_\psi \|\mathrm{G}(q_\phi(\tilde{x})) - \hat{y}\|_2. \tag{9}$$

As a result, our model can benefit from the above adversarial process, generating more robust protection noise.

To summarize, our robust protection function g_ψ is end-to-end optimized by:

$$\min_\psi \|\mathrm{G}(\tilde{x}) - \hat{y}\|_2 + \lambda_1 \|\mathrm{G}(q_\phi(\tilde{x})) - \hat{y}\|_2 + \lambda_2 \|\eta\|_2, \tag{10}$$

where $\lambda_1 = \lambda_2 = 10$ and $\|\eta\|_2$ is also a regular term as Eq. (2).

Experiments

Set Up

Implementations. Image inputs are resized to 256×256 and normalized in $[-1, 1]$ for all our experiments. The functions including f_θ, g_1, g_2, are implemented using U-net (Ronneberger, Fischer, and Brox 2015) architecture. And noise encoding function g_n is implemented by stacking several Conv-BN-ReLU layers. ϵ is set to 0.1 by default. The channel C for noise encoding is 128. We use Adam optimizer with a learning rate of 10^{-4} in training. And our model converges at around 30 epochs.

Manipulation Methods. We conduct comprehensive experiments on several powerful facial manipulation methods. 1) **pSp-mix** (Richardson et al. 2021) achieves style translation of facial images by mixing latent codes of different sources. 2) **pSp-recon** (Richardson et al. 2021) uses the same architecture as pSp-mix, while being leveraged to reconstruct a given facial image. 3) **SimSwap** (Chen et al. 2020) is a state-of-the-art face swap method that we utilized for simulating abuses of identity replacement. 4) **StyleClip** (Patashnik et al.

Manip. Method	Method	MSE ↑	LPIPS ↑	SSIM ↓	PSNR ↓
pSp-mix	FGSM	0.0433	0.1286	0.6381	19.9703
	PGD	0.3081	0.2945	0.4691	11.3420
	Disrupt	0.1879	0.2375	0.5249	13.5339
	TAFIM	0.0158	0.0978	0.7045	24.4298
	ARS	**1.3218**	**0.8007**	**0.3602**	**5.7422**

Table 1: Performance against pSp-mix.

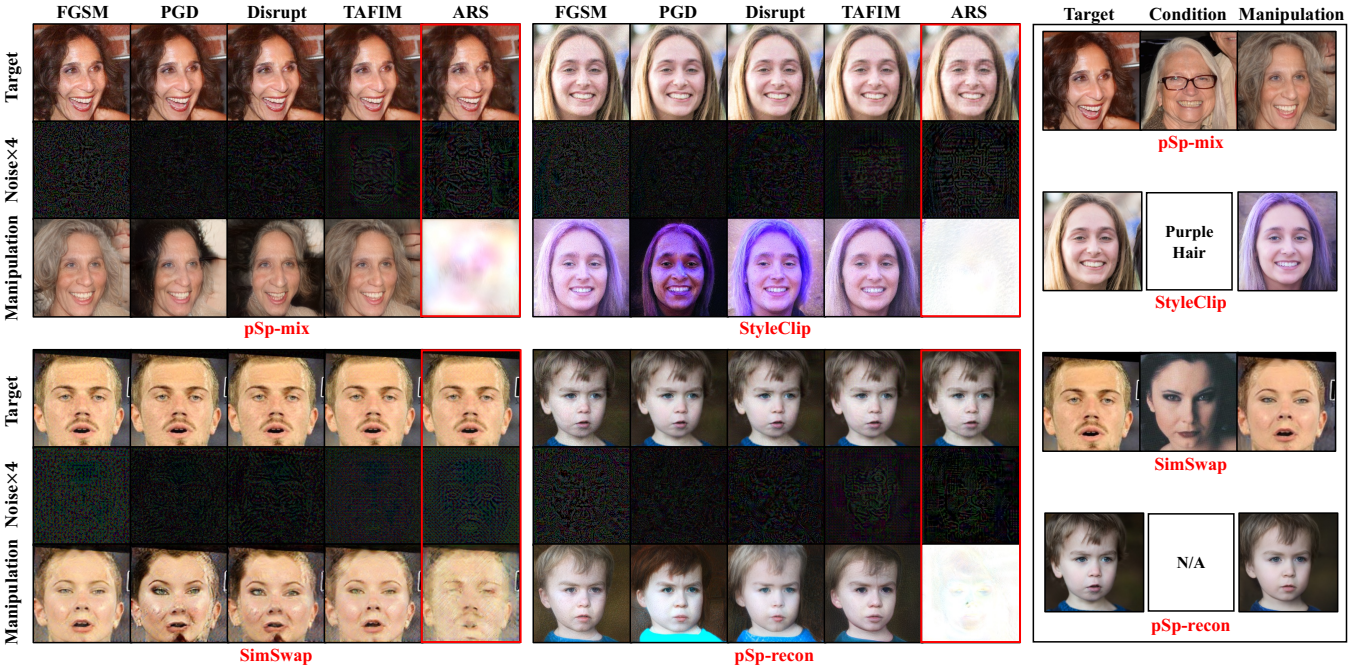

Figure 3: Qualitative comparisons. We present the protected results of four manipulation methods, alongside the original manipulation results shown in the rightmost column, for reference. Our results are in red boxes.

2021) is a recent study focusing on text-driven manipulation. For all the manipulation methods, we collect the open-source codes and model weights from their official implementations.

Datasets. Since the above four manipulation methods are trained on different datasets, we also conduct experiments using the data aligned with their implementations. For pSp-mix, pSp-recon, and StyleClip, we randomly select 10000 images from FFHQ (Karras, Laine, and Aila 2019) in training, 2000 images for validation, and 2000 images for testing. For SimSwap, we also introduce a custom subset of VGGFace2 (Cao et al. 2018) including 10000, 2000, and 2000 images for train, validation, and test, respectively.

Baselines. We conduct comparisons with several well-established baselines. 1) **FGSM** (Kurakin, Goodfellow, and Bengio 2016) is one of the most well-known methods in adversarial training. 2) **PGD** (Madry et al. 2017) is also adopted to attack the manipulation methods in an iterative way. 3) **Disrupt** (Ruiz, Bargal, and Sclaroff 2020) is one of the earliest works investigating adversarial attacks for protection against facial manipulations. We adopt the best model of their proposals, which involves a spread-spectrum evasion of blur defenses. 4) **TAFIM** (Aneja, Markhasin, and Nießner 2022) is a recent learning-based method. We retrain their model with our experimental conditions for fair comparisons.

Metrics. Following previous works (Ruiz, Bargal, and Sclaroff 2020; Huang et al. 2022; Aneja, Markhasin, and Nießner 2022), we use metrics including Mean Square Error (**MSE**), Peak Signal-to-Noise Ratio (**PSNR**), Learned Perceptual Image Patch Similarity (**LPIPS**) (Zhang et al. 2018), Structural Similarity (**SSIM**), Success Rate (**SR**) (Ruiz, Bargal, and Sclaroff 2020), and Masked Success Rate ($\mathbf{SR}_{mask}$) (Huang et al. 2022) for evaluations. MSE, PSNR, LPIPS, and SSIM are calculated between the original manipulated outcomes and the possible disrupted output after protection. Therefore, larger MSE and LPIPS indicate better protection ability. While regarding PSNR and SSIM, the smaller the number, the better the protection. SR and SR_{mask} are proposed to represent the successful ratio of protection. When the output of manipulation methods could be considerably disrupted, the protection is regarded as a successful case. Please find details of SR and SR_{mask} in (Ruiz, Bargal, and Sclaroff 2020; Huang et al. 2022).

Results

Visual Comparisons. We first present intuitive visual comparisons in Fig. 3. From the figure, single-step FGSM cannot work well with a small $\epsilon = 0.1$. Iterative-based methods, PGD and Disrupt, perform better in disrupting the manipulations. Although TAFIM is also trained with a white target image, it cannot work well with a small ϵ. Additionally, our learning-based ARS is trained in a targeted-protection way, thus the protected manipulation should be close to our predefined full-white image. A comparison between all manipulation methods shows that SimSwap is the most challenging to defend against. For quantitative comparisons, we tabulate detailed results against pSp-mix in Table 1. Our method consistently outperforms baselines.

Robust Comparisons. Next, we conduct comparisons with perturbations included. When the perturbation is included, we find some baselines almost lost protection ability with

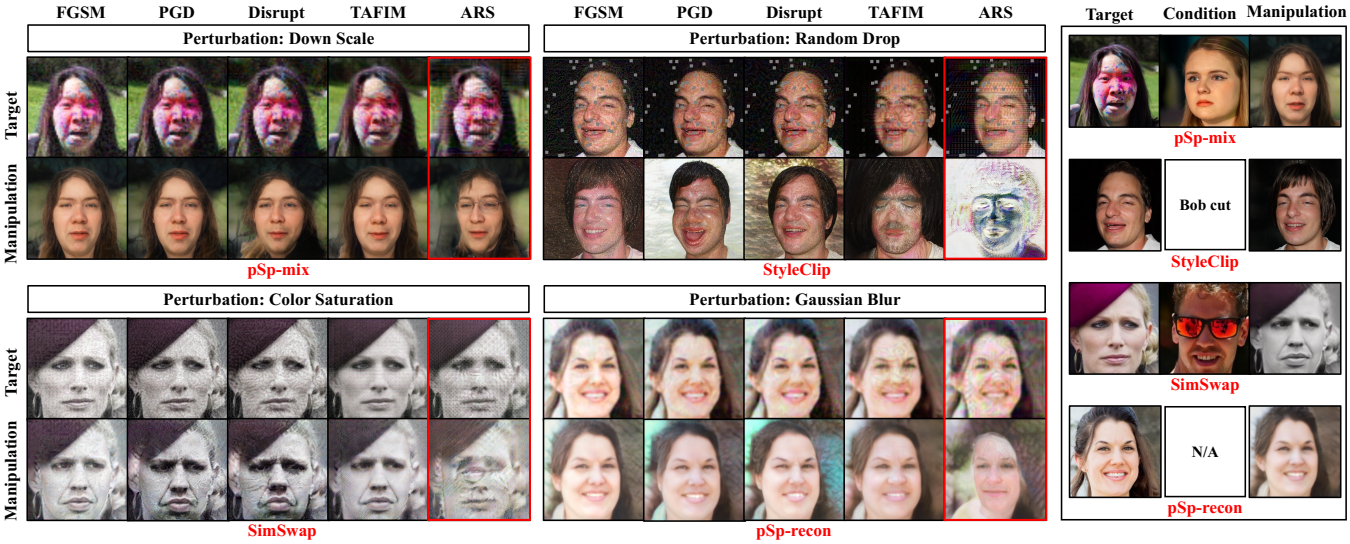

Figure 4: Qualitative comparisons with perturbations. We present the protected results of four manipulation methods, alongside the original manipulation results (w/ perturbations) shown in the rightmost column, for reference. Our results are in red boxes.

Method	FGSM	PGD	Disrupt	TAFIM	ARS
$\epsilon = 0.1$	0.04	0.48	0.41	0.02	**0.53**
$\epsilon = 0.3$	0.31	0.81	0.86	0.64	**0.96**

Table 2: Success Rate evaluated on pSp-mix averaged from all kinds of perturbations at the highest level.

Method	FGSM	PGD	Disrupt	TAFIM	ARS
pSp-mix↦StyleClip	0.31	0.30	0.30	0.19	**0.77**
StyleClip↦pSp-mix	0.17	0.16	0.21	0.13	**0.92**

Table 3: Success Rate evaluated between pSp-mix and StyleClip averaged from all perturbations at the highest level.

a small $\epsilon = 0.1$ (see Success Rate in Table 2). A greater ϵ than 0.3 would seriously impact the visual quality, we thus set $\epsilon = 0.3$ as an appropriate value for a better demonstration. We first demonstrate the protection in Fig. 4. From the results, our method can perform successful protection and disrupt the manipulations considerably. Moreover, for quantitative comparisons, we show results with pSp-mix at each level averaged from all perturbations in Fig. 5 for intuitive demonstration. From the results, we see that our method significantly outperforms all baselines in visual metrics including LPIPS and PSNR. Regarding SR and SR_{mask}, Disrupt and our method both report a saturated successful ratio when lower-level perturbations are included. When a higher level of perturbation is confronted, our method achieves better protection. In addition, we tabulate performance evaluated against pSp-mix at the highest level for each kind of perturbation in Table 6. Comparing different perturbations, "Gaussian Blur" happens to be the most powerful one. In spite of the incorporation of a blur module within the Disrupt for the sake of robustness, our approach, distinguished by its specially tailored robust protection, significantly outperforms all baselines. Note the included perturbation will also have an impact on the original manipulated result, *e.g.*, the instances in rightmost column of Fig. 4. For details of the introduced perturbations, please find them in the appendix.

Transferable Protection. In addition to the white-box setting, we further perform transfer experiments and tabulate the comparisons in Table 3. The success rate is evaluated with the protection model trained using one manipulation method and tested on another one. The results show our method holds a good transferable protection ability.

Speed Comparisons. Running speed is also a key property in applications. We tabulate the comparisons in Table 4. PGD and Disrupt spend the most time on protecting one image due to the time-intensive iterative pipeline. FGSM can yield results within a singular forward-backward process, but its protection effect is not satisfactory. While TAFIM and our method require only one forward process to generate protection noises, which costs the least amount of time.

Method	FGSM	PGD	Disrupt	TAFIM	ARS
Time Cost (s/image)	0.65	2.79	2.84	**0.46**	**0.46**

Table 4: Speed comparisons. We run methods against pSp-mix with the same experimental environment and report the performance averaged from 5 independent runnings.

Ablation Study

In this section, we conduct ablations on key designs of our approach. Here we first explain these alternative designs. a) "Basic" denotes the model without the proposed second way in Fig. 2, *i.e.*, the Robust Protection, the Perturbation Module, and Adv Net are removed; b) "Basic w/ Robust" removes the Robust Protection but keeps the Perturbation

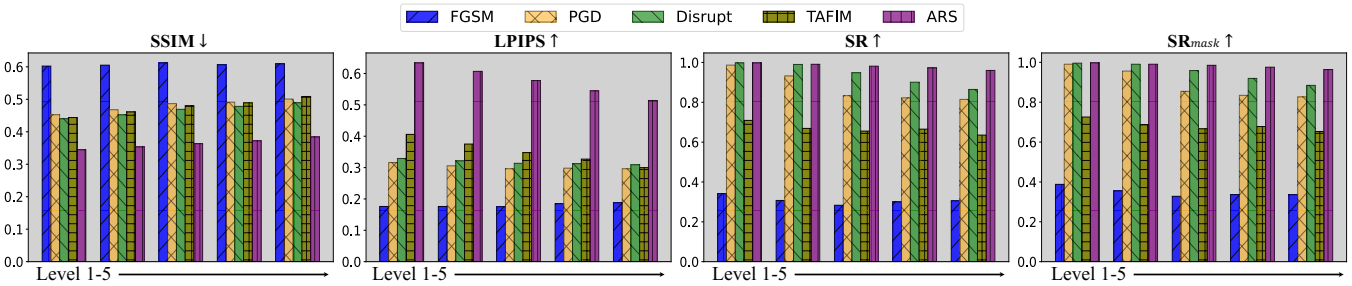

Figure 5: Robust evaluation. Protection against pSp-mix averaged from all kinds of perturbations at each level.

Model	MSE ↑	PSNR ↓	SR ↑	SR$_{mask}$ ↑
Basic	0.2221	13.6986	0.8741	0.8866
Basic w/ Robust	0.6304	10.7555	0.9300	0.9340
non-lossy ARS	**0.7944**	**10.2761**	0.8930	0.9130
ARS w/o Adv	0.6145	10.8189	0.9290	0.9390
ARS	0.5035	11.9364	**0.9600**	**0.9680**

Table 5: Ablations on pSp-mix. We compare several alternative designs with randomly chosen perturbations at random levels. All the experiments are initialized with the same random seed to ensure fairness.

Module and Adv Net in training. In this model, the output of the Basic Protection can still benefit from the simulation of information loss in training, but lost information redundancy provided in Robust Protection; c) "non-lossy ARS" indicates that we remove both Perturbation Module and Adv Net; d) "ARS w/o Adv" indicates that we remove Adv Net and the adversarial training process but keep the Perturbation Module; e) "ARS" denotes the whole method we proposed.

We compare these models with randomly chosen perturbations at random levels. All the experiments are initialized with the same random seed to ensure fairness. The Results are tabulated in Table 5. "Basic" could be regarded as a simple learning-based baseline. Comparing "Basic" and "Basic w/ Robust", we see the proposed Perturbation Module and Adv Net can greatly promote the protection performance of the learning-based method. "non-lossy ARS" achieves the best performance in MSE and PSNR, but has poorer successful protection rate (SR and SR$_{mask}$). This indicates this model is able to disrupt the manipulation more significantly but fails in tough cases with severe perturbations. In addition, comparing "non-lossy ARS" and "ARS w/o Adv", we see that simply including the Perturbation Module for certain types of information loss can benefit the successful rate. Further, including the Adv Net and adversarial training process in "ARS" will slightly impact the visual performance, but greatly improve the success rate.

Conclusion

In this paper, we present a learning-based method, abbreviated ARS, to pre-protect facial images from being manipulated without consensus. Prominent attributes of our proposed approach encompass its good efficiency and notable robustness. Compared with related baselines, our method outperforms them both qualitatively and quantitatively. Especially, we first benchmark the robustness evaluation of adversarial protection methods against four kinds of facial manipulations. Considering the ubiquitous post-processing in multimedia platforms, in-depth studies in this aspect could greatly promote the applications of related pre-protection methods in the future.

Noise Type	Method	MSE ↑	PSNR ↓	SR ↑	SR$_{mask}$ ↑
Color Saturation	FGSM	0.0419	20.3315	0.2700	0.3120
	PGD	0.1228	15.4775	0.9867	0.9833
	Disrupt	0.1036	16.2793	0.9300	0.9333
	TAFIM	0.0766	17.6059	0.7850	0.7960
	ARS	**0.1720**	**14.2258**	**0.9960**	**0.9950**
Color Contrast	FGSM	0.0335	21.4004	0.1610	0.1640
	PGD	0.1235	15.4683	0.9900	0.9867
	Disrupt	0.1066	16.1490	0.9433	0.9433
	TAFIM	0.0906	17.7739	0.6100	0.6210
	ARS	**1.1085**	**6.6215**	**1.0000**	**1.0000**
Dandom Drop	FGSM	0.0541	19.0030	0.5020	0.5530
	PGD	0.1511	14.4892	**1.0000**	**1.0000**
	Disrupt	0.1272	15.2732	0.9933	0.9933
	TAFIM	0.1066	16.5021	0.8780	0.9020
	ARS	**0.3711**	**11.2601**	**1.0000**	**1.0000**
Gaussian Blur	FGSM	0.0238	22.9322	0.0590	0.0670
	PGD	0.0320	21.6120	0.1400	0.1867
	Disrupt	0.0534	19.3434	0.4633	0.5433
	TAFIM	0.0231	23.2245	0.0660	0.0810
	ARS	**0.0759**	**17.5743**	**0.8040**	**0.8300**
Down Scale	FGSM	0.0586	18.9503	0.5400	0.5620
	PGD	0.0758	17.6559	0.7700	0.8033
	Disrupt	0.0856	17.0568	0.8600	0.8967
	TAFIM	0.0562	19.0650	0.5160	0.5530
	ARS	**0.0973**	**16.3995**	**0.9590**	**0.9580**
JPEG Compression	FGSM	0.0426	20.2552	0.2990	0.3570
	PGD	0.1549	14.4008	**1.0000**	**1.0000**
	Disrupt	0.1459	14.6833	0.9967	0.9967
	TAFIM	0.2899	13.0587	0.9570	0.9650
	ARS	**0.6911**	**8.4913**	**1.0000**	**1.0000**

Table 6: Robust evaluation. Protection against pSp-mix at the highest level for each kind of perturbation.

Acknowledgments

This work was in part supported by National Natural Science Foundation of China with No. 62394322, No. 61932016, No. 62132011 and No. 62202258, National Science Foundation for Distinguished Young Scholars of China with No. 61825204, Beijing Outstanding Young Scientist Program with No. BJJWZYJH01201910003011, China Postdoctoral Science Foundation with No. 2021M701894, China National Postdoctoral Program for Innovative Talents, Shuimu Tsinghua Scholar Program, Ant Group, and the Beijing National Research Center for Information Science and Technology (BNRist) key projects.

References

Aneja, S.; Markhasin, L.; and Nießner, M. 2022. TAFIM: Targeted Adversarial Attacks against Facial Image Manipulations. In *European Conference on Computer Vision*, 58–75. Springer.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Chen, J.; Zhao, Y.; Li, Q.; Feng, X.; and Xu, K. 2023. FedDef: Defense Against Gradient Leakage in Federated Learning-Based Network Intrusion Detection Systems. *IEEE Transactions on Information Forensics and Security (TIFS)*, 18: 4561–4576.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2003–2011.

Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H.; and Ge, Z. 2023. Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3994–4004.

Dong, S.; Wang, J.; Liang, J.; Fan, H.; and Ji, R. 2022. Explaining Deepfake Detection by Analysing Image Matching. *arXiv preprint arXiv:2207.09679*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guan, J.; Hu, T.; Zhou, H.; Guo, Z.; Deng, L.; Quan, C.; Ding, E.; and Zhao, Y. 2023a. Building an Invisible Shield for Your Portrait against Deepfakes. *arXiv preprint arXiv:2305.12881*.

Guan, J.; Zhang, Z.; Zhou, H.; Hu, T.; Wang, K.; He, D.; Feng, H.; Liu, J.; Ding, E.; Liu, Z.; et al. 2023b. StyleSync: High-Fidelity Generalized and Personalized Lip Sync in Style-based Generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1505–1515.

Guan, J.; Zhou, H.; Guo, Z.; Hu, T.; Deng, L.; Quan, C.; Fang, M.; and Zhao, Y. 2023c. Dual-Modality Co-Learning for Unveiling Deepfake in Spatio-Temporal Space. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 85–94.

Guan, J.; Zhou, H.; Hong, Z.; Ding, E.; Wang, J.; Quan, C.; and Zhao, Y. 2022. Delving into sequential patches for deepfake detection. *Advances in Neural Information Processing Systems*, 35: 4517–4530.

Guo, H.; Wang, Y.; Ivanov, N.; Xiao, L.; and Yan, Q. 2022. SPECPATCH: Human-In-The-Loop Adversarial Audio Spectrogram Patch Attack on Speech Recognition. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, 1353–1366.

He, Z.; Wang, W.; Guan, W.; Dong, J.; and Tan, T. 2022. Defeating DeepFakes via Adversarial Visual Reconstruction. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2464–2472.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4490–4499.

Huang, H.; Wang, Y.; Chen, Z.; Zhang, Y.; Li, Y.; Tang, Z.; Chu, W.; Chen, J.; Lin, W.; and Ma, K.-K. 2022. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 989–997.

Karras, T.; Laine, S.; and 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*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5001–5010.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16317–16326.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2085–2094.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, 86–103. Springer.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2287–2296.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Ruiz, N.; Bargal, S. A.; and Sclaroff, S. 2020. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European Conference on Computer Vision*, 236–251. Springer.

Sun, K.; Liu, H.; Ye, Q.; Liu, J.; Gao, Y.; Shao, L.; and Ji, R. 2021a. Domain general face forgery detection by learning to weight. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2638–2646.

Sun, K.; Yao, T.; Chen, S.; Ding, S.; Ji, R.; et al. 2021b. Dual Contrastive Learning for General Face Forgery Detection. *arXiv preprint arXiv:2112.13522*.

Wang, R.; Juefei-Xu, F.; Luo, M.; Liu, Y.; and Wang, L. 2021. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3546–3555.

Xu, Y.; Liang, J.; Jia, G.; Yang, Z.; Zhang, Y.; and He, R. 2023. TALL: Thumbnail Layout for Deepfake Video Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22658–22668.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023. UCF: Uncovering Common Features for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22412–22423.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 672–681.

Yao, K.; Wang, J.; Diao, B.; and Li, C. 2023. Towards Understanding the Generalization of Deepfake Detectors from a Game-Theoretical View. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2031–2041.

Yeh, C.-Y.; Chen, H.-W.; Tsai, S.-L.; and 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*, 53–62.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhao, Y.; Xu, K.; Wang, H.; Li, B.; and Jia, R. 2021a. Stability-Based Analysis and Defense against Backdoor Attacks on Edge Computing Services. *IEEE Network*, 35(1): 163–169.

Zhao, Y.; Xu, K.; Wang, H.; Li, B.; Qiao, M.; and Shi, H. 2021b. MEC-Enabled Hierarchical Emotion Recognition and Perturbation-Aware Defense in Smart Cities. *IEEE Internet of Things Journal*, 8(23): 16933–16945.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, 657–672.