Initiative Defense against Facial Manipulation

Qidong Huang†, Jie Zhang†, Wenbo Zhou*, Weiming Zhang*, Nenghai Yu
University of Science and Technology in China
{hqd0037@mail., welbeckz@, zjzac@mail., zhangwm@, ynh@}ustc.edu.cn

Abstract

Benefiting from the development of generative adversarial networks (GAN), facial manipulation has achieved significant progress in both academia and industry recently. It inspires an increasing number of entertainment applications but also incurs severe threats to individual privacy and even political security meanwhile. To mitigate such risks, many countermeasures have been proposed. However, the great majority methods are designed in a passive manner, which is to detect whether the facial images or videos are tampered after their wide propagation. These detection-based methods have a fatal limitation, that is, they only work for ex-post forensics but can not prevent the engendering of malicious behavior.

To address the limitation, in this paper, we propose a novel framework of initiative defense to degrade the performance of facial manipulation models controlled by malicious users. The basic idea is to actively inject imperceptible venom into target facial data before manipulation. To this end, we first imitate the target manipulation model with a surrogate model, and then devise a poison perturbation generator to obtain the desired venom. An alternating training strategy are further leveraged to train both the surrogate model and the perturbation generator. Two typical facial manipulation tasks: face attribute editing and face reenactment, are considered in our initiative defense framework. Extensive experiments demonstrate the effectiveness and robustness of our framework in different settings. Finally, we hope this work can shed some light on initiative countermeasures against more adversarial scenarios.

Introduction

With the tremendous success of deep generative models (e.g. conditional-GAN), face manipulation has been an emerging topic in very recent years and a variety of methods (Thies et al. 2016; Korshunova et al. 2017; Nirkin, Keller, and Hassner 2019; Natsume, Yatagawa, and Morishima 2018; Wu et al. 2018) have been proposed. The face manipulation systems can change the target face with a different attribute, such as hairstyle or expression. As the manipulated results become more and more realistic, these techniques can easily be misused for malicious purposes such as violating individual privacy or even misleading political opinions. In details, the malicious users may edit the portrait image of any person without his/her permission. Moreover, the attacker is able to forge the expression (e.g. lip shape) of political figure's speech video, which might seriously mislead the public.

To alleviate the risks brought by malicious usage of face manipulation, various impressive countermeasures (Matern, Riess, and Stamminger 2019; Han et al. 2017; Afchar et al. 2018; Nguyen, Yamagishi, and Echizen 2019; Rössler et al. 2019; Li et al. 2020) have been proposed, most of which are ex-post detection-based methods used for forensics. Notwithstanding the considerably high accuracy in distinguishing forged facial images or video from the real ones, these methods are too passive and hard to eliminate the damages caused by malicious face manipulation, because the generation and wide spread of such forgeries have already become a fait accompli. How to proactively prevent such threats before its happening is a significant but seriously under-researched problem. Very recently, Ruiz *et al.*(Ruiz, Bargal, and Sclaroff 2020) propose a gradient-based method to attack facial manipulation models, which is regarded as the baseline method of this paper. However, its white-box assumption is not practical. The baseline method needs the inner information about the target models. Besides, it is invalid to some other types of manipulation tasks, like the real-time face reenactment.

In this work, we propose the concept of initiative defense against face manipulation, which is a totally fresh angle of defense. Specifically, we leverage a poison perturbation generator $\mathbf{PG}$ to generate invisible perturbation, and then superpose it onto the original face data before releasing the data to the public. The forger can only have access to the infected face data, and due to the existence of poison perturbation, the performance of face forgery models will be significantly degraded, whether infected data is used in inference stage or in training process. In other words, the infected face data must satisfy two ultimate objectives: (1) is visually consistent to the clean one and (2) can degrade the performance of the target model greatly.

To achieve these goals, we devise a two-stage training framework of initiative defense, which is generally appli-

*Corresponding authors, † Equal contribution.

cable to different facial manipulation tasks and different adversarial settings. It should be pointed out that the baseline method methoned above can be regarded as a special example of our framework. Here, we take the defense against attribute editing task as an example to elaborate our framework briefly. As shown in Figure 1, in the stage A, we train a surrogate model $\mathbf{SM}$ to imitate the target manipulation model's behavior, while we aim to obtain the corresponding perturbation generator $\mathbf{PG}$ in the stage B. In practice, it is hard to train a $\mathbf{PG}$ with the well-trained $\mathbf{SM}$ in the stage A. The main reason is that, the infected data produced by the initial $\mathbf{PG}$ will easily solve the objective (2) mentioned above, and make the training process trapped in the local optimum because of the non-convex property of DNNs. To this end, we propose an alternating training strategy to train $\mathbf{SM}$ and $\mathbf{PG}$ step by step. The whole training procedure is illustrated in Algorithm 1. Different from the popular adversarial training approach, in our framework, only the update of perturbation generator $\mathbf{PG}$ is influenced by the surrogate model $\mathbf{SM}$ while the $\mathbf{SM}$ is regularly updated in an isolated manner. For different tasks, we further leverage some task-specific strategies to enhance the defense effectiveness.

We conduct experiments on two typical types of face manipulation tasks: face attribute editing and face reenactment. Extensive experiments demonstrate the effectiveness of the proposed framework and the robustness in different adversarial settings, even in a black-box way. Furthermore, some ablation analysis is provided to justify the motivation of our design.

To summarize, the contributions of this paper are threefold as below:

- We introduce the concept of initiative defense against face manipulation, which is a new perspective of countermeasures. And we hope it can inspire more great works for this seriously under-researched field.
- We propose the two-stage training framework, which is generally applicable to different face manipulation tasks. Furthermore, we leverage an alternating training strategy to achieve the designed objectives, and some task-specific strategies to enhance the defense performance.
- Extensive experiments demonstrate the proposed framework can resist the manipulation from the forgers in different settings, including the black-box adversarial scenarios.

Related Work

Deep Face Forgery. Deep face forgery techniques have achieved tremendous progress in the past few years. According to different forgery goals, existing methods can be roughly divided into two categories: *face swapping* and *face manipulation*. Face swapping is a representative framework in deep face forgery which swaps the face between a source and a target person. In this class of methods (DeepFakes 2017; FaceSwap 2017; Bao et al. 2018; Afchar et al. 2018), auto-encoder and GANs are commonly used as backbones to generate realistic results. All these methods are "identity-changed" which focus on exchanging the identity face region and seldom concerns other attributes out of the face region. Face manipulation is a more common type of deep face forgery methods, this type of methods (Choi et al. 2018; Thies et al. 2016; Pumarola et al. 2019; Kim et al. 2018) includes many cases such as face attribute editing which modifies part of the facial attributes (e.g., hair style), and face reenactment which transfers the facial expression from one person to another. StarGAN (Choi et al. 2018) achieves facial attribute transfer among multiple domains by using a single generator. It is adequately trained on public face datasets and edit the target attribute in the inference stage. Face2face (Thies et al. 2016) is a typical approach of face reenactment. To transfer the facial expression from source to target, it requires a considerable amount of the target face data to train the generative model for each specific person.

Usually, the realism of results generated by face swapping are quite limit especially when source and target attribute are very different. However, face manipulation is easier to achieve much better visual quality because that it does not need to change the original face. Therefore, face manipulation has a wider range of application scenarios than face swapping currently. In this work, we mainly focus on facial manipulation to implement our initiative defense.

Existing Countermeasures against Facial Manipulation. As face manipulation causes great threaten to individual privacy even political security, it is of paramount importance to develop countermeasures against it. Previous works (Rössler et al. 2019; Matern, Riess, and Stamminger 2019; Li et al. 2020; Qian et al. 2020; Dang et al. 2020) are commonly designed in a passive manner. For instance, Rossler *et al.*(Rössler et al. 2019) introduce a simple but effective Xception Net as a binary classifier for face manipulation detection. Face X-ray (Li et al. 2020) considers the detection from a different view which focuses on the blending artifacts rather than the manipulation artifacts. The blending operator could not be avoided in the process of manipulation, thus face X-ray obtains better results than most works. However, although some detection-based methods achieve considerably high accuracy in ex-post forensics, they still can not eliminate the influences of malicious applications which had been caused in the wide spreading.

Recently, the countermeasures which aim to prevent malicious behavior before the manipulation have been studied. For example, Fawkes (Shan et al. 2020) protects faces from being stolen by recognition systems. But it can not prevent private face images from being edited by malicious users. Ruiz *et al.*(Ruiz, Bargal, and Sclaroff 2020) propose a gradient-based method to attack facial manipulation systems in white box successfully, which protects faces from being manipulated by some specific models. However, this method is time-consuming and laborious, which requires optimization for each single image in each iterations and performs badly in both gray-box and black-box scenes. To address these limitations, this paper is the first try to establish a complete framework of initiative defense against face manipulation, which aims to be more efficient and robust.

Method

In this section, we will elaborate the proposed framework of initiative defense in detail. Before that, the formal problem

definition will be introduced firstly.

Problem Definition. For facial manipulation tasks, we roughly divided them into two categories: *model-based manipulation* and *data-based manipulation*. The corresponding example tasks are facial attribute editing and face reenactment, respectively. Given a face image or video x, the former type is to modify x to forged example y with a well-trained manipulation model $\mathcal{M}$, while the latter one is to utilize the given face data x to train a new model $\mathcal{M}_x$, and feed some guidance information z (e.g. landmark) into the model to generate modified face y. Therefore, the target of facial manipulation tasks can be formulated by

$$y = \begin{cases} \mathcal{M}(x), & \text{model-based;} \\ \mathcal{M}_x(z), & \text{data-based.} \end{cases} \tag{1}$$

Before the forgers get access to the face data x, we can preprocess it to obtain the infected face data x' in an additive way, i.e.,

$$x' = x + \epsilon \cdot \delta_x, \qquad \|\delta_x\|_\infty \leq \tau. \tag{2}$$

Here, δ_x denotes the perturbation superposed on x. ϵ represents an adjustable intensity of δ_x, and τ is an threshold to constrain δ_x. With the infected data x', the forger will acquire the corresponding infected forged example y'. One of our objective is to guarantee the visual consistence between clean data x and infected data x'. The other objective is to destroy the generation of y, namely that, to maximize the distance $\mathcal{D}$ between the clean forged face y and the infected forged face y', i.e.,

$$\delta_x = \arg\max_{\delta} \mathcal{D}(y, y'). \tag{3}$$

Two-stage Training Framework of Initiative Defense. For the baseline method (Ruiz, Bargal, and Sclaroff 2020), its assumption that having the full access to the target manipulation model is not quite practical. To this end, we substitute a surrogate model $\mathbf{SM}$ for the target manipulation model $\mathbf{M}$, which is feasible because of the accessibility to the model type and training procedure of a specific manipulation task. Besides, we devise a poison perturbation generator $\mathbf{PG}$ to generate the $\delta_x = \mathbf{PG}(x)$, which will incur minor computational cost compared to the gradient-based method adopted by Ruiz *et al.*. We propose the two-stage training framework to train $\mathbf{SM}$ and $\mathbf{PG}$. Intuitively, we attempt to train the $\mathbf{PG}$ with a well-trained $\mathbf{SM}$, but it does not work because the objective function described in Eq.(3) will be fulfilled at the very beginning of the $\mathbf{PG}$'s training, which makes it trapped into the local optimum. Therefore, we propose the alternating training strategy to train both the $\mathbf{SM}$ and $\mathbf{PG}$ one by one from scratch. Concretely, in every training iteration, we first update the surrogate model $\mathbf{SM}$ regularly and get the corresponding infected forgery y in the stage A. And in the stage B, with the infected data x' calculated by the trained $\mathbf{PG}$ in the last iteration, we obtain the infected forgery y' based on Eq. (1). And then, a task-specific influence loss function $\mathcal{L}_{A\to B}$ will be designed to feedback the training influence from the stage A to the stage B. Both $\mathcal{L}_{A\to B}$ and the conventional loss objective $\mathcal{L}_B$ are leveraged to update the $\mathbf{PG}$ in the stage B. The pseudo code of two-stage training framework is illustrated in Algorithm 1.

Algorithm 1: Two-stage Training Framework

Input: number of iterations $maxiter$, original training dataset D.
Output: a perturbation generator PG_{best}.

// Initialization;
$SM_0, PG_0 \leftarrow RandomInit()$;
$PG_{bset} \leftarrow PG_0, maxdist \leftarrow -inf$;
for $i = 1$ **to** $maxiter$ **do**
  $x \sim SelectMiniBatch(D)$;
  // **Stage A**: train SM on clean data x regularly;
  $SM_i \leftarrow Update(SM_{i-1}, x, \mathcal{L}_A)$;
  // Get clean forgery y based on Eq.(1);
  $y \leftarrow GetForgery(x)$;
  // **Stage B**: obtain x' using last iteration's PG;
  $x' \leftarrow x + \epsilon \cdot PG_{i-1}(x)$;
  // Acquire the corresponding y' based on Eq.(1);
  $y' \leftarrow GetForgery(x')$;
  // Design task-specific influence loss $\mathcal{L}_{A\to B}$ to
  // maximize the distance between y and y';
  $\mathcal{L}_{A\to B} \leftarrow \max \mathcal{D}(y, y')$;
  // Train PG based on $\mathcal{L}_{A\to B}$ and $\mathcal{L}_B$;
  $PG_i \leftarrow Update(PG_{i-1}, x, \mathcal{L}_B, \mathcal{L}_{A\to B})$;
  // Compare distances and select the best one;
  if $maxdist < \mathcal{D}(y, y')$ **then**
    $maxdist \leftarrow \mathcal{D}(y, y')$;
    $PG_{best} \leftarrow PG_i$;
  end
end
return PG_{best}

Defending against Model-based Manipulation. As shown in Figure 1, we take the facial attribute editing task as example to elaborate the network architectures and loss functions adopted in our two-stage training framework in detail. We adopt an auto-encoder like networks with 6 residual blocks (" Res6") and UNet-128 (Ronneberger, Fischer, and Brox 2015) as the default architecture of surrogate model $\mathbf{SM}$ and perturbation generator $\mathbf{PG}$, respectively. Two discriminators $\mathbf{D_A}$ and $\mathbf{D_B}$ are introduced in both stages, which are common convolutional networks (7 convolutional layers appended with one dense layer). For $\mathbf{SM}$ in the stage A, we follow the settings in (Choi et al. 2018) to train it step by step. The objective loss function to train $\mathbf{PG}$ consists of two parts: the influence loss $\mathcal{L}_{A\to B}$ and the conventional training loss $\mathcal{L}_B$, i.e.,

$$\mathcal{L}_{PG} = \mathcal{L}_{A\to B} + \lambda \cdot \mathcal{L}_B, \tag{4}$$

where λ is the hyper parameter to balance these two loss terms. Below we will introduce the detailed formulation of $\mathcal{L}_{A\to B}$ and $\mathcal{L}_B$, respectively. Here, only the adversarial loss $\mathcal{L}_{adv}$ is considered in $\mathcal{L}_B$, which is enough to guarantee the visual quality between x and x', i.e.,

$$\begin{aligned} \mathcal{L}_{adv} = \mathbb{E}_x[D_B(x)] - \mathbb{E}_{x'}[D_B(x')] \\ -\lambda_1 \cdot \mathbb{E}_{\hat{x}}\Big[(\|\nabla_{\hat{x}} D_B(\hat{x})\|_2 - 1)^2\Big], \end{aligned} \tag{5}$$

where the last term leverages Wasserstein distance to penalize the gradient for training stability (Arjovsky, Chintala,

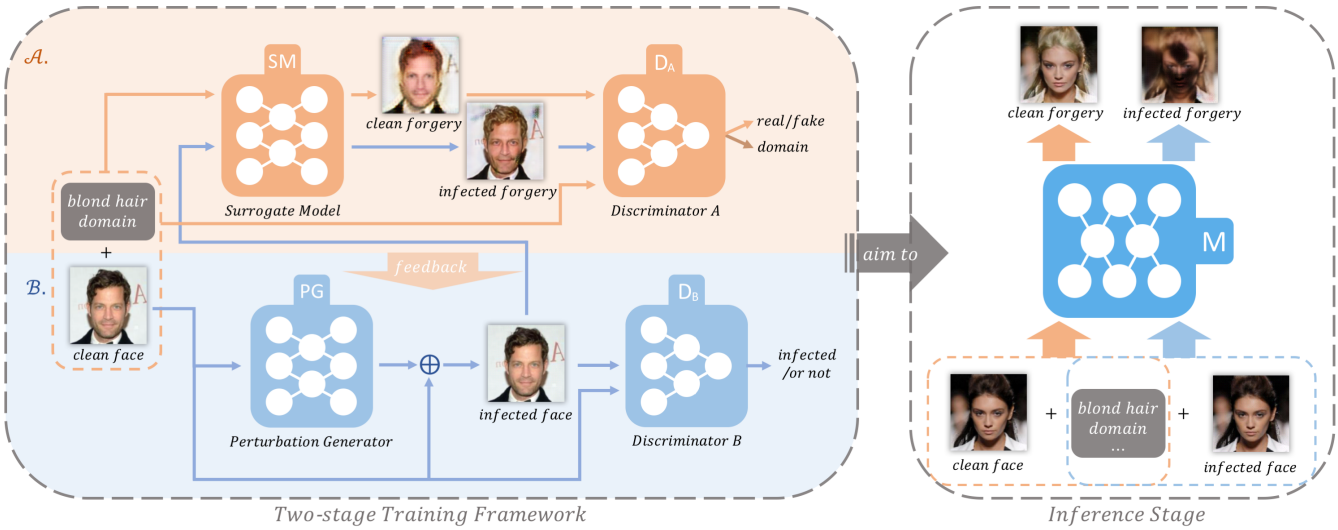

Figure 1: The overall pipeline to defend facial attribute editing. We set the blond hair as the target domain. The left part is two-stage training framework, and in the stage A, we train a surrogate model **SM** regularly on clean faces and domain labels to imitate the behavior of the target model **M**. In the stage B, a perturbation generator **PG** is trained on clean faces and constrained with the extra influence loss $\mathcal{L}_{A\to B}$ feedbacked from the stage A. During the inference stage, the infected faces generated by the well-trained **PG** will disrupt the target model **M** severely.

and Bottou 2017; Gulrajani et al. 2017) and λ_1 is the hyperparameters indicating the weights to the corresponding term. The responsibility of $\mathcal{L}_{A\to B}$ is to maximize the distance between the clean forgery y and infected forgery y'. In this task, $\mathcal{L}_{A\to B}$ is defined as the weighted sum of three terms, i.e.,

$$\mathcal{L}_{A\to B} = \lambda_2 \cdot \mathcal{L}_{bs} + \lambda_3 \cdot \mathcal{L}_{cyc} + \lambda_4 \cdot \mathcal{L}_{dom}. \tag{6}$$

where the first two terms are to degrade the visual quality of y' compared with y in diverse domains and the last term is to confuse the discriminator $\mathbf{D_A}$ to classify y' as fake outputs and the farthest domain from the one of y. In detail, the basic loss $\mathcal{L}_{bs}$ is used to maximize the pixel-level difference between y' and y, i.e.,

$$\mathcal{L}_{bs} = \mathbb{E}_{x,x',c}\Big[-\sum_j \mu_j \cdot \|SM(x,c_j) - SM(x',c_j)\|_1\Big], \tag{7}$$

where c_j denotes to a series of target domains derived from the original domain c of x, and $\|\cdot\|_1$ is L_1 norm distance. Here, μ_j is to balance the influence degree among different domains due to the discrepant area ratio in pixel-level, and we achieve it in an indirect way, i.e.,

$$\mu_j = \frac{\|x - SM(x,c_j)\|_1}{\sum_j \|x - SM(x,c_j)\|_1}. \tag{8}$$

Cycle-consistency loss is popular for many un-supervised tasks and we devise $\mathcal{L}_{cyc}$ to disrupt the consistency, which is described as

$$\mathcal{L}_{cyc} = \mathbb{E}_{x',c}\Big[-\sum_j \mu_j \cdot \|x' - SM(SM(x',c_j),c)\|_1\Big]. \tag{9}$$

To enlarge the difference of y and y' in high-dimensional level, we calculate the inverse domain c_{rj} for each c_j as the farthest domain, and minimum the confidence probability of y' from the perspective of $\mathbf{D_A}$ meanwhile. Therefore, the domain loss $\mathcal{L}_{dom}$ is defined as

$$\begin{aligned}\mathcal{L}_{dom} = \mathbb{E}_{x',c}\Big[\mathbb{E}_{c_j}[-\log D_A(c_{rj}|SM(x',c_j))]\Big] \\ +\mathbb{E}_{x',c}\Big[\mathbb{E}_{c_j}[D_A(SM(x',c_j))]\Big].\end{aligned} \tag{10}$$

Defending against Data-based Manipulation. Face reenactment is the classic type task of data-based manipulation and the defending against Face2Face is illustrated in Figure 2 for instance. Similarly, a surrogate model **SM** is learned conventionally according to the guidance by (Thies et al. 2016) in the stage A, which can be regarded as an image-to-image translation model. In the stage B, $\mathcal{L}_{adv}$ mentioned above is also applied in this task as $\mathcal{L}_B$ for **PG**'s training. The difference is the influence loss $\mathcal{L}_{A\to B}$ (in Eq.(4)) here designed to greatly weaken the ability of infected model $\mathbf{M}'$ compared with surrogate model **SM**, i.e.,

$$\mathcal{L}_{A\to B} = \mathbb{E}_{x,z}\Big[\|(SM_x(z) - x)\|_1 - \|(M'_{x'}(z) - x)\|_1\Big], \tag{11}$$

in which we simply use the $\|(SM_x(z) - x)\|_1$ to represent the reconstruction ability of the image translation model $\mathbf{SM_x}$. Intuitively, it is equivalent to a distance penalty to further corrupt the performance of the infected model $\mathbf{M'_{x'}}$. Besides, as shown in Figure 2, a temporary model **TM** is introduced to maintain the gradient information to **PG**, and we copy its parameters into an infected model $\mathbf{M}'$ to compute the influence loss $\mathcal{L}_{A\to B}$. Similar training mechanism is also leveraged in (Feng, Cai, and Zhou 2019). To simplify the framework, we assume that the perturbation will not influence the landmark of face data, which is perhaps a more rigorous assumption for this problem.

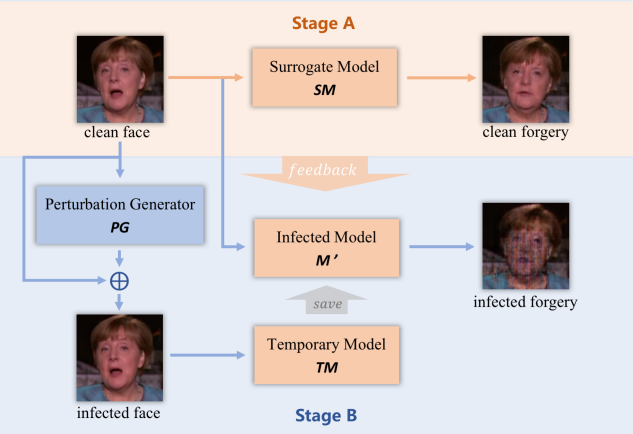

Figure 2: The brief pipeline for face reenactment task. We shall point out that we utilize a temporary model **TM** to maintain the gradients of **PG**, while **TM**'s parameters is copied into **M** to calculate the influence loss $\mathcal{L}_{A\to B}$.

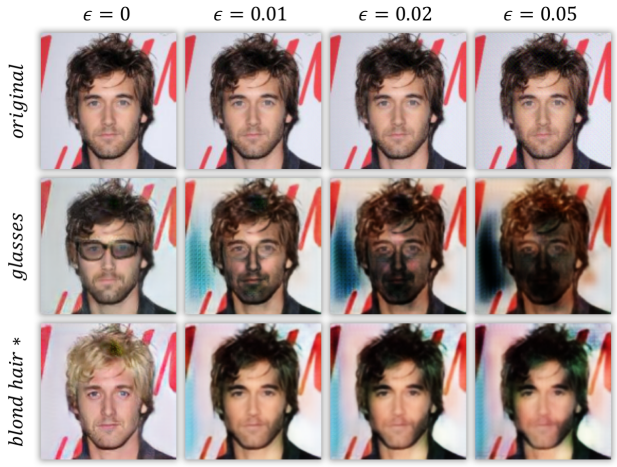

Figure 3: Some visual examples of defending against attribute editing under different perturbation intensities ϵ. The first row are the infected data and the last two rows are the infected forgery with different domains. $*$ denotes the wild domain not involved in the training phase of **PG**.

Task-specific Defense Enhancement. To further boost the defense ability, we also leverage some training skills to enhance the defense ability for specific task. For facial attribute editing task, we integrate the extra influence loss $\mathcal{L}'_{A\to B}$, calculated between the current **PG** and the **SM** in last iteration, into the original objective loss $\mathcal{L}_{PG}$ illustrated in Eq.(4), i.e.,

$$\mathcal{L}_{PG} = \mathcal{L}_{A\to B} + \mathcal{L}'_{A\to B} + \lambda \cdot \mathcal{L}_B, \tag{12}$$

which will absorb more influence information from the stage A and stabilize the optimization.

As for Face2Face task, we introduce the attention-based facial mask $m(x)$ of clean data x into Eq.(11), namely that, $\|(SM_x(z) - x) \cdot m(x)\|_1 - \|(M'_{x'}(z) - x) \cdot m(x)\|_1$, which assigns the facial area as 1.0 and the rest as 0.01. Guided by BiSeNet-based face matting method (Yu et al. 2018), $m(x)$ can enable the optimization to pay more attention to corrupted facial area.

Experiments

In this paper, we conduct experiments on two classical manipulation tasks: facial attribute editing and face reenactment. To demonstrate the effectiveness and robustness of our method, we first show the newly introduced initiative defense framework can greatly disrupt the malicious manipulation models while guaranteeing the visual quality of preprocessed face data. Then, we verify the proposed method is robust in different adversarial settings. Finally, some ablation studies are provided to justify the motivation of the training strategies we leverage and prove the feasibility of the extensions to the combined manipulation scenario.

Implementation Details. For facial attribute editing, we use face images from the CelebA dataset (Liu et al. 2015), which is further split into 100000 images for two-stage training, 100000 for training the target model and 2600 for inference stage. As for face reenactment, Face2Face is investigated as the representative forgery model, with the famous Ms Merkel's speech video as a face data example. We samples 1500 frames in this video and shuffle them randomly, which are utilized in the training phase as shown in Figure 2. Then we obtain the final infected model $\mathbf{M}'$ trained on the whole infected video (15000 frames), which is poisoned by the well-trained **PG**. And another 100 clean faces with expression landmark information are prepared for testing.

In the first scenario, the learning rate of both **PG** and the surrogate model are assigned as 0.0001 by default, and five randomly selected attribute domains are involved in the two-stage training framework. Empirically, we adopt $\lambda = 0.01$, $\lambda_1, \lambda_2 = 10$, $\lambda_3 = 2.5$, and $\lambda_4 = 1$ in the related loss function. For the second scenario, the initial learning rates of all models are equal to 0.0002 and $\lambda = 0.01$ by default.

Evaluation Metrics. To evaluate the visual quality, PSNR is used by default. To investigate the defending effect under each condition, we compute the average L_1 and L_2 norm distance, LPIPS (Zhang et al. 2018) and perceptual loss (Johnson, Alahi, and Fei-Fei 2016) between forgeries and original faces to measure the degradation degree of forgery generation. Specifically, in the first scenario (attribute editing), we regard the defense as success if the L_2 distance is bigger than 0.05. Based on it, the defense success rate (DSR) is further defined as the ratio of poisoned face images whose corresponding forgery is successfully corrupted. While in the second scenario (Face2Face), the corrupted area is more sparse than that in attribute editing task, and therefore, we replace the L_2 norm in the calculation of DSR by the sparse metric L_1 norm. In order to show the effectiveness of the proposed defense more intuitively, we adopt Local Binary Pattern (LBP) to depict the texture characteristics and identity information of the corresponding faces, which is often leveraged in some traditional face recognition algorithms (Guo, Zhang, and Zhang 2010; Zhang et al. 2010).

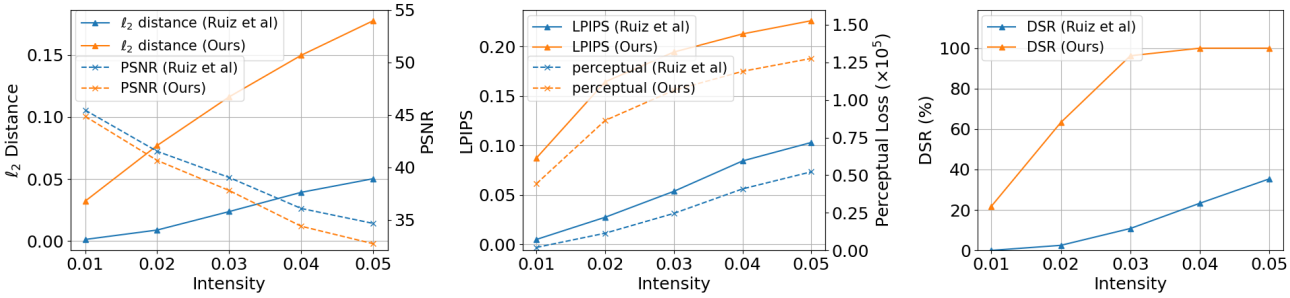

Figure 4: Quantitative comparison of defending against attribute editing between the baseline method (Ruiz, Bargal, and Sclaroff 2020) and our approach under different perturbation intensities ϵ. Left: PSNR for infected data and average L_2 for infected forgery; Middle: LPIPS and perceptual loss for infected forgery; Right: Defendse success rate (DSR).

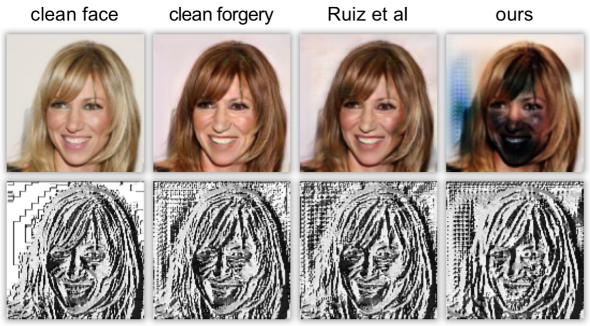

Figure 5: Visual comparison of defending against attribute editing between the baseline method (Ruiz, Bargal, and Sclaroff 2020) and our method in pixel-level (top) and LBP-level (bottom). ($\epsilon = 0.03$, in gray-box way.)

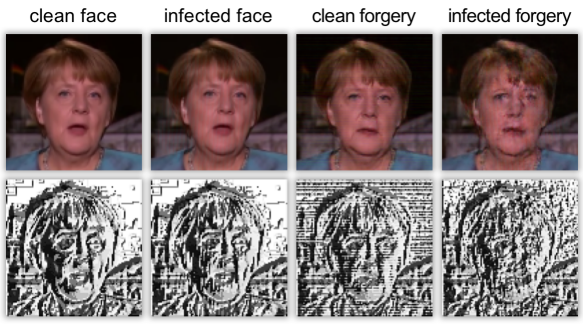

Figure 6: Some visual examples of defending against face reenactment ($\epsilon = 0.02$). The bottom row is the pertinent LBP features of the top row, and the forgeries are generated from the same expression landmark.

Effectiveness of The Proposed Initiative Defense. To demonstrate the effectiveness of the proposed initiative defense, we conduct the basic experiments in gray-box way, that is, we are clear about the network architecture and domain information (especially for facial attribute editing task) of the attacker's manipulation model $\mathbf{M}$, but its inner information such as parameters and gradients is still unknown for us. Then, we adopt the same network structure and attribute domain to train the surrogate model $\mathbf{SM}$.

For face attribute editing task, our method can greatly disrupt the malicious manipulation model while guaranteeing the visual quality of infected data. In addition, a control experiment on different perturbation intensities ϵ are considered. Some visual examples are shown in Figure 3, and we can observe the damaged scale on face forgery becomes larger as the increase of ϵ, which means the defending effect gets better. In other words, the forger can not manipulate our infected face to his/her desired domain, even if the threshold ϵ is set as 0.01, which represents very slight perturbation. Furthermore, we compare the effectiveness of our approach with a gradient-based method (Ruiz, Bargal, and Sclaroff 2020) proposed recently, which optimizes every single face image separately in white-box way. As shown in Figure 4 and Figure 5, we can achieve great incremental defense effectiveness while keeping a very comparable visual quality.

As for the experiments of defending against Face2Face, we obtain target manipulation model $\mathbf{M}$ and infected model $\mathbf{M}'$ by training with clean video and poisoned video data, and then evaluate them on the same face landmarks. In Figure 6, the output of the both visual quality and texture characteristic are corrupted significantly after data poisoning, thus the forger is unable to successfully manipulate the infected video protected by particular $\mathbf{PG}$. We shall point out that the baseline method is not applicable to data-based manipulation, and to the best of our knowledge, the proposed method is the first attempt to such adversarial scenario.

Robustness in Different Adversarial Settings. In this experiment, we consider more different adversarial settings aside from the gray-box setting. Specifically, four types of network architectures are utilized for training target manipulation model here: a vanilla convolutional network ("CNet"), two auto-encoder like networks with 9 and 16 residual blocks respectively ("Res6", "Res9"), an UNet-128 network ("UNet128"). For face attribute editing task, we also leverage two type choices of attribute domains to train $\mathbf{SM}$, namely that, the same domains "SD" (i.e. glasses) as training target model $\mathbf{M}$ or the different domains "DD" (i.e. blond hair). Besides the gray-box setting mentioned above, we regard all the other settings as black-box setting.

We can see in Table 1 and Table 2 that our initiative de-

Domain	Setting	Ruiz *et al.*		Ours		Ours†	
		DSR	L_2	DSR	L_2	DSR	L_2
SD	Res6*	35%	0.050	**100%**	**0.177**	95%	0.124
	Res9	4%	0.011	**90%**	**0.134**	85%	0.092
	CNet	54%	0.061	**93%**	**0.095**	67%	0.087
	UNet128	3%	0.014	**100%**	**0.129**	45%	0.052
DD	Res6	16%	0.031	**100%**	**0.139**	96%	0.127
	Res9	5%	0.017	**93%**	**0.110**	91%	0.108
	CNet	0%	0.008	**95%**	**0.101**	63%	0.087
	UNet128	46%	0.071	**100%**	**0.207**	89%	0.089

Table 1: Quantitive results of defending against attribute editing models of different settings ($\epsilon = 0.05$). Here, † denotes the results without alternating training strategy, and $*$ means the gray-box setting.

Setting	defense			defense†		
	DSR	L_1	L_2	DSR	L_1	L_2
UNet256*	**100%**	**0.057**	**0.006**	34%	0.049	0.004
CNet	**100%**	**0.165**	**0.064**	100%	0.135	0.060
Res6	**100%**	0.113	**0.025**	100%	**0.115**	0.024
Res9	**100%**	0.093	0.016	100%	**0.102**	**0.019**
UNet128	**94%**	**0.056**	**0.006**	61%	0.052	0.005

Table 2: Quantitive results of defending against face reenactment models of various network structures ($\epsilon = 0.02$). Here, † denotes the results without alternating training strategy. And $*$ means the gray-box setting.

fense of both tasks attains strong robustness in different adversarial settings. For face attribute editing task, the defense success rate (DSR) of the baseline method (Ruiz, Bargal, and Sclaroff 2020) degrade severely in most black-box settings, and even in gray-box setting, DSR is only 35%. On the other hand, our method's DSR is more than 93% in all cases, even tested on the unseen domain in the training phase, and some visual examples are also showcased in Figure 3.

For face reenactment task, we can still realize the robustness in all adversarial settings considered above, as shown in Table 2). It is worth noting that UNet performs especially well for face reenactment task by multi-scale skip connections. In comparison, it is poorer for the performance of other network structures we adopt such as "CNet", "Res6" and "Res9". In our experiment, we also regard those intrinsically worse manipulated results as successful defense. Therefore, the defense success rate (DSR) is all 100% in those settings even without the alternating training strategy.

Ablation Study

The Importance of Alternating Training Strategy (ATS). The goal of the alternating training strategy (ATS) is to evade falling into the undesired local optimum. According to the quantitative results shown in Table 1 and Table 2, the alternating training strategy can substantially improve the defense effectiveness. Visual results is displayed in Figure 7.

The Importance of Task-specific Defense Enhancement (TDE). As we can find in Figure 7, the more influence information absorbed in training procedure and the attention-based facial mask guidance are advantageous for corresponding tasks to enhance the defense ability. Without these, the distortion appeared on forged images might be much slighter in both defending scenarios.

Figure 7: Visual comparison results of two tasks trained with or without alternating training and defense enhancement.

Figure 8: Visual examples of the joint defense on Merkel's speech video (wild data for face attribute editing task) and choosing gray hair as target domain (black-box setting for face reenactment task).

Extensions to The Joint Scenarios. Considering that it is possible that the forger may extract one frame from the poisoned video (the second scenario) for attribute editing (the first scenario), we combine two training procedures by utilizing the first scenario's $\mathbf{PG}$ to process each frame of videos in advance. And then we continue to process these already infected faces with the second scenario's $\mathbf{PG}$. Based on perturbation overlaying, we can defend against the special scenes successfully while sacrificing some visual quality in this combined way, whose results can be seen in Figure 8.

Conclusion

In this paper, we newly introduce the initiative defense against face manipulation, aiming to protect face data from being forged. By poisoning the face data, the performance of manipulation model will be degraded. To achieve this goal, a general two-stage training framework is proposed to train a poison generator $\mathbf{PG}$. Besides, we further introduce the alternating training strategy and defense enhancement techniques into the training process. Two classic manipulation tasks, namely that, face attribute editing and face reenactment are considered. Extensive experiments demonstrate the effectiveness and robustness of our approach in different adversarial settings. Moreover, we conduct some ablation studies to prove the advantages of our design and the feasibility to the special scenarios. And we hope this work can spur more great works in this seriously under-researched filed.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047 ,U1636201, 62002334 and 62072421, Exploration Fund Project of University of Science and Technology of China under Grant YD3480002001, and by Fundamental Research Funds for the Central Universities under Grant WK2100000011. Jie Zhang is partially supported by the Fundamental Research Funds for the Central Universities WK5290000001.

Ethics Statement

With the popularity of online social platforms, people are keen to share their selfie pictures or videos. It means that plenty of human portraits are exposed in public without any protection, which provides an opportunity for malicious users to manipulate these face data. Facial manipulation may cause damage to a person's reputation, or even be used to defraud. More seriously, a manipulated political figure's video may bring threats to national security. Comparing to detection-based methods, our initiative defense framework can indeed protect against malicious usage of face data in advance rather than remedy the damages ex-post. This is effective to eliminate such risks to some extent. In general, our initiative defense will greatly inspire the relative researches in countermeasures against facial manipulation.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. IEEE.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. In *34th International Conference on Machine Learning (ICML)*, 214–223.

Bao, J.; Chen, D.; Wen, F.; Li, H.; and Hua, G. 2018. Towards open-set identity preserving face synthesis. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 6713–6722.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the detection of digital face manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5781–5790.

DeepFakes. 2017. DeepFakes. http://github.com/ondyari/FaceForensics/tree/master/dataset/. Accessed Sept. 30, 2019.

FaceSwap. 2017. FaceSwap. http://github.com/deepfakes/faceswap. Accessed July 4, 2019.

Feng, J.; Cai, Q.-Z.; and Zhou, Z.-H. 2019. Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder. In *Advances in Neural Information Processing Systems (NeurIPS)*, 11971–11981.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Guo, Z.; Zhang, L.; and Zhang, D. 2010. A Completed Modeling of Local Binary Pattern Operator for Texture Classification. *IEEE Transactions on Image Processing (TIP)* 19(6): 1657–1663.

Han, X.; Morariu, V.; Larry Davis, P. I.; et al. 2017. Two-stream neural networks for tampered face detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 19–27.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*.

Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; and Theobalt, C. 2018. Deep video portraits. *ACM Transactions on Graphics (TOG)* 37(4): 1–14.

Korshunova, I.; Shi, W.; Dambre, J.; and Theis, L. 2017. Fast face-swap using convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 3677–3685.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5001–5010.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*.

Matern, F.; Riess, C.; and Stamminger, M. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83–92. IEEE.

Natsume, R.; Yatagawa, T.; and Morishima, S. 2018. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447* .

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467* .

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. FSGAN: Subject agnostic face swapping and reenactment. In *IEEE international conference on computer vision (ICCV)*, 7184–7193.

Pumarola, A.; Agudo, A.; Martinez, A.; Sanfeliu, A.; and Moreno-Noguer, F. 2019. GANimation: One-Shot Anatomically Consistent Facial Animation. *International Journal of Computer Vision (IJCV)* .

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues. *arXiv preprint arXiv:2007.09355* .

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI* 234–241.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint arXiv:1901.08971* .

Ruiz, N.; Bargal, S. A.; and Sclaroff, S. 2020. Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems. *arXiv:2003.01279* .

Shan, S.; Wenger, E.; Zhang, J.; Li, H.; Zheng, H.; and Zhao, B. Y. 2020. Fawkes: Protecting Personal Privacy against Unauthorized Deep Learning Models. In *Proceedings of USENIX Security*.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Niebner, M. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, W.; Zhang, Y.; Li, C.; Qian, C.; and Change Loy, C. 2018. Reenactgan: Learning to reenact faces via boundary transfer. In *European conference on computer vision (ECCV)*, 603–619.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*, 334–349.

Zhang, B.; Gao, Y.; Zhao, S.; and Liu, J. 2010. Local Derivative Pattern Versus Local Binary Pattern: Face Recognition With High-Order Local Pattern Descriptor. *IEEE Transactions on Image Processing (TIP)* 19(2): 533–544.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE conference on computer vision and pattern recognition (CVPR)*.