Restricted Black-Box Adversarial Attack Against DeepFake Face Swapping

Junhao Dong, Yuan Wang, Jianhuang Lai, Senior Member, IEEE, and Xiaohua Xie, Member, IEEE

Abstract—DeepFake face swapping presents a significant threat to online security and social media, which can replace the source face in an arbitrary photo/video with the target face of an entirely different person. In order to prevent this fraud, some researchers have begun to study the adversarial methods against DeepFake or face manipulation. However, existing works mainly focus on the white-box setting or the black-box setting driven by abundant queries, which severely limits the practical application of these methods. To tackle this problem, we introduce a practical adversarial attack that does not require any queries to the facial image forgery model. Our method is built on a substitute model based on face reconstruction and then transfers adversarial examples from the substitute model directly to inaccessible black-box DeepFake models. Specially, we propose the Transferable Cycle Adversarial Generative Adversarial Network (TCA-GAN) to construct the adversarial perturbation for disrupting unknown DeepFake systems. We also present a novel post-regularization module for enhancing the transferability of generated adversarial examples. To comprehensively measure the effectiveness of our approaches, we construct a challenging baseline of DeepFake adversarial attacks for future development. Extensive experiments impressively show that the proposed adversarial attack method makes the visual quality of DeepFake face images plummet so that they are easier to be detected by humans and algorithms. Moreover, we demonstrate that the proposed algorithm can be generalized to offer face image protection against various face translation methods.

Index Terms—DeepFake, black-box, adversarial attack, substitute model.

I. INTRODUCTION

DEEPAKE, a portmanteau of “deep learning” and “fake”, has drawn broad attention in recent years. This burgeoning technique can replace the source face in an existing image or a video with the target face of an arbitrary identity. However, the abuse of this intriguing technology may result in an underlying hazard to social media and online security. For instance, it can be maliciously used to blackmail individuals or to bypass the authentication mechanism [1].

Although these subtle DeepFake artifacts confuse human vision, several forgery detection methods can precisely distinguish these manipulated images [2], [3], [4]. However, there exists a time delay between the publishing of forged images and their corresponding detection results, which may damage personal reputation by impersonating the victim on social media platforms. Another way to defend against this malicious face manipulation is to disrupt the generation stage of DeepFake face swapping, i.e., to distort generated face-swapped images. This straightforward method eliminates the complicated forgery detection, which can tackle these DeepFake issues from the root thoroughly.

Deep Neural Networks (DNNs) have made spectacular progress in various computer vision tasks [5], [6], [7]. Nonetheless, DNNs are especially vulnerable to some tailored examples synthesized by attaching imperceptible perturbations to original images. These visually unconscious examples with the added perturbation are also regarded as adversarial examples [8]. Adversarial examples can drastically mislead the inference of DNNs to output wrong (or even specific) results, which poses a significant challenge to current deep learning applications [9], [10], [11]. A plausible reason for the emergence of the adversarial example is the linear behavior in high-dimensional feature spaces of DNNs [12]. Thus, investigating adversarial examples can further help explain the learned feature of DNNs and shows a path to attack the improper usage of DNN mechanisms like DeepFake. Correspondingly, it is also significant to ensure the privacy protection of trusted samples during the inference stage of DNNs [13]. Generally, the attack scenarios are various according to the accessible information of the target model and its output results, as shown in Fig. 1. The white-box adversarial attack has the complete information of the target model. Nevertheless, the semi-white-box setting [14] focuses on training a DNN to generate adversarial examples, which do not need the backward gradients at the inference stage. However, these two types of attacks are impractical in defending against unknown tampering. The black-box adversarial attack can only gain the DNN output. In contrast, the restricted black-box adversarial attack [15] is built on an entirely black box with the inaccessible output of the target DNN model, which means that we can not conduct even a single query to the black-box model. Therefore, this attack is more applicable to real-world scenarios. To achieve general applications, we consider the restricted black-box adversarial attack, which can defend the immutability of our photos from DeepFake face swapping or other face editing systems.

Manuscript received 30 March 2022; revised 13 February 2023 and 6 April 2023; accepted 6 April 2023. Date of publication 12 April 2023; date of current version 27 April 2023. This work was supported by the National Natural Science Foundation of China under Grant 622A2095 and Grant 62072482. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dusit Niyato. (Corresponding author: Xiaohua Xie.)

The authors are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China, also with the Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510006, China, and also with the Guangdong Key Laboratory of Information Security Technology, Guangzhou 510006, China (e-mail: dongh8@mail2.sysu.edu.cn; wangy975@mail2.sysu.edu.cn; stuljh@mail.sysu.edu.cn; xiaoxiaoh6@mail.sysu.edu.cn). Digital Object Identifier 10.1109/TIFS.2023.3266702

1556-6021 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
We design TCA-GAN to generate powerful adversarial examples against DeepFake systems. Moreover, the post-regularization module is proposed for better transferability. Extensive experiments show that our method can be generalized to various face attribute editing models with nearly no cost.

• We construct a new baseline on disrupting DeepFake systems when simulating a real-world scenario of arbitrary face swapping. In addition to disrupting the visual quality of generated fake images, we demonstrate that the disruption can further enhance the performance of several image-level DeepFake detection methods.

II. RELATED WORKS

A. Automatic Face Swapping

Generally, face swapping consists of replacing the face of a certain person with another face of a different identity. This tricky technique was first proposed in [16], which fits 3D morphable models on both exchanging faces with the help of manual annotation. For efficiency, Bitouk et al. [17] presented a fully automatic face swapping by replacing the source face with the candidate face image from a face gallery. Nevertheless, the target identity of this face swapping is uncontrollable and time-consuming. Korshunova et al. [18] regarded the face swapping problem as a style transfer to accomplish designated face swapping, which is conducted by a deep convolutional neural network with efficient preprocessing and postprocessing. Especially, Bao et al. [19] disentangled the identities and attributes of face images and proposed an identity-preserving attack against DeepFake face swapping. Specifically, we build the face reconstruction autoencoder as the substitute model for producing adversarial examples that can be directly transferred to unknown face manipulation models.

• We design TCA-GAN to generate powerful adversarial examples against DeepFake systems. Moreover, the post-regularization module is proposed for better transferability. Extensive experiments show that our method can be generalized to various face attribute editing models with nearly no cost.

• We construct a new baseline on disrupting DeepFake systems when simulating a real-world scenario of arbitrary face swapping. In addition to disrupting the visual quality of generated fake images, we demonstrate that the disruption can further enhance the performance of several image-level DeepFake detection methods.
GAN-based network for open-set face synthesizing. Modern deep learning-based face swapping methods eliminate several fussy face editing steps and synthesize photo-realistic face images [20], [21], [22]. The most popular technique of these is DeepFake [1], which induces a plethora of fake public scandals. However, numerous types of research have been explored on DeepFake detection. We consider a straightforward approach to disturbing DeepFake face swapping.

B. Adversarial Attacks Against Classifiers

As demonstrated by [8], the reason for adversarial examples is the discontinuities of DNNs. Alternatively, Goodfellow et al. [12] argued that adversarial examples are results from the linearity in high dimensional space in DNNs and presented the Fast Gradient Sign Method (FGSM). The iterative strategy is also exerted into FGSM to build more powerful adversarial attacks [23], [24]. Moosavi-Dezfooli et al. [25] extend the adversarial attack to a universal (image-agnostic) scenario, i.e., making a wrong prediction for each input image with high confidence. Especially, Su et al. [26] considered an extremely confined scenario and proposed a one-pixel adversarial perturbation generative method based on differential evolution. In addition, Croce and Hein [27] introduced a Fast Adaptive Boundary (FAB) attack to find minimal adversarial perturbation in distinct norm constraints. Particularly, Xiao et al. [14] proposed AdvGAN to generate more effective adversarial examples against the classification with GANs. Our method also concentrates on the same GAN-based adversary generation. However, we extend it to attacking generative models in the restricted black-box setting via constructing a robust cycle-consistent GAN. Zhang et al. [28] studied the optimization of the manifold of the classification boundary and introduced boundary projection to obtain low adversarial perturbation. Zhong and Deng [29] proposed a novel surrogate model to enhance the transferability of adversarial attacks against face recognition models. Likewise, our goal is to enhance the transferability of adversarial examples, while we mainly focus on adversarial examples against DeepFake face swapping to protect in the restricted black-box scenario. While our focus is not limited to adversarial attacks against malicious face swapping to protect the security of personal photos, it yields supportive generalization performance on attacking several face manipulation models as well.

C. Adversarial Attacks Against Generative Models

Although adversarial attacks against classification models have achieved satisfactory performance, few studies have focused on adversarial attacks against generative models. Tabacof et al. [30] first investigated adversarial attacks for autoencoders with latent representation. They indicated that generative models do not have a definite decision boundary as classification models and thus are more difficult to disrupt. Kos et al. [11] attached an auxiliary classifier to the encoder of the target generative models to obtain classification-based adversarial attacks. Recently, applying adversarial examples to disturb image-translation models has received heavy investigation [31], [32], [33]. Particularly, Dong and Xie [34] demonstrate impressive results in disrupting DeepFake face swapping in white-box circumstances. In contrast, we explore the restricted black-box scenario. Sun et al. [35] described a white-box adversarial attack to disturb the facial landmark extraction, thus can induce the misalignment of DeepFake face-swapped images. Their method mainly focuses on the preprocessing of DeepFake face swapping, while we concentrate on a universal way to disturb the process of face swapping. Note that both DeepTag [36] and Disrupting Deepfakes [31] aim at developing proactive defense mechanisms against DeepFake facial image manipulation in the white-box setting. However, the above-mentioned two works [31], [36] adopt the term “Deepfake” in a broader context that is used to refer to any altered media of someone’s likeness. Specifically, they concentrate on protecting facial images against GAN-based image translation networks, e.g., StarGAN [37] and AttGAN [38]. In comparison, we mainly focus on defending facial images against original DeepFake face swapping. Note that DeepFake face swapping models are trained on facial images of two specific human identities, which is much harder to be disrupted by adversarial examples due to the high similarity of facial data. Ruiz et al. [32] first constructed a multi-query-based adversarial attack against image translation models in the black-box scenario. Nevertheless, their black-box adversary generation is mainly conducted through numerous queries, which is impractical to defend the malicious face swapping in real-world conditions. In comparison, our method considers a more practical restricted black-box setting, with no need for multiple queries to the target face swapping model. Furthermore, our method can also be generalized to protect facial images against other face manipulation tasks, e.g., GAN-based image translation.

III. Method

A. Preliminaries

1) Brief Review of DeepFake: The original DeepFake face swapping model consists of one shared encoder and two separate decoders to two face swapping identities, respectively. The shared encoder converts the input face image into a latent representation. Then the decoder transforms the latent representation into a face image, whose identity is according to the selected decoder. In the training stage, DeepFake conducts face reconstruction with the decoder that is consistent with the input identity. For face swapping, DeepFake decodes the latent variable with the other decoder to obtain face-swapped images and vice versa. Note that these input face images have already been trained for robust face swapping. Hence, it is also challenging to disrupt DeepFake face swapping. In this paper, we mainly focus on attacking this original type of DeepFake face swapping mechanism.

2) Notation: In this paper, we intend to build transferable adversarial examples against DeepFake face swapping in the black-box scenario. We denote the adversarial example as $x_{adv} = x + r$, where $x$ is the legitimate facial image, and $r$ is the appended adversarial perturbation. Let $DF : X \rightarrow Y$
Fig. 3. The flow chart of our proposed method. Due to the accessibility to the substitute model in the training stage, we train TCA-GAN to produce the adversarial perturbation against the substitute model according to the input face image. In the application stage, we append a post-processing module to regularize the generated adversarial example for better transferability. The application stage simulates a real-world scenario of attacking DeepFake, while we can not obtain the face-swapped output or any details of DeepFake.

Consequently, we consider a restricted black-box adversarial attack without access to the target DeepFake model. Based on the transferability of adversarial examples [12], we distinctly build a substitute model for generalizable adversary generation. More specifically, we obtain adversarial examples from the substitute model and expect to transfer them to unknown DeepFake face swapping models. Here, the adversarial example is composed of a legitimate image and its corresponding adversarial perturbation generated by TCA-GAN with the prior information of the legitimate image. Then we apply a post-regularization on the adversarial example to enhance the transferability. The flow chart of our adversarial attack is shown in Fig. 3. We first obtain the adversarial example against the substitute model through TCA-GAN. Afterward, we can transfer it to unforeseen DeepFake face swapping models.

Motivated by [39], we extend the substitute model to the generative model to simulate the black-box face manipulation model, which is much harder than the classification model that has an absolute classification boundary. Considering that DeepFake and other face manipulation models can be virtually regarded as rebuilding the objective face in a special style, we employ a pre-trained DNN autoencoder as the substitute model to conduct facial image reconstruction. Note that existing works have also demonstrated that facial images can be embedded into a unified representation space and further be interpreted [40], [41], [42]. For instance, DeepFake face swapping can also be viewed as reconstructing the input facial image with a given facial style of a different human identity. Therefore, it is reasonable to adopt the face reconstruction
Fig. 4. Illustration of the training mechanism of TCA-GAN, including two generators and two domain discriminators. These two generators are utilized to produce and remove the adversarial perturbation. The corresponding discriminators differentiate both legitimate examples and adversarial examples.

The forward adversary generation aims at producing an adversarial perturbation to be combined with the input facial image, disrupting the reconstruction of the substitute model. Note that the generated adversarial examples not only disrupt the image-level reconstruction of the substitute model but also perturb the latent variable extracted by the substitute model. This intuition is based on the observation that adversarial perturbations against DNNs are transferable [45]. The similarity between the components of DNNs incurs similar feature representations between different DNNs. Hence, the obtained adversarial example is prone to be transferred to attack the unaccessible DeepFake model. The loss function of disturbing the reconstruction can be written as below:

$$\mathcal{L}_{recons} = \|S(x) - x\|_1 + \|S(\hat{x}) - \hat{x}\|_1, \quad (2)$$

where $\hat{x}$ is a slightly wrapped version of the input face image $x$. This wrapping operation comprises random rotation, scaling, and shifting. The reason we take the wrapped face image into consideration is that the substitute model should possess a certain extent of robustness towards some subtle perturbations. Therefore, the adversarial example produced against the substitute model can present a better-transferring performance on the target model.

After obtaining the trained substitute model, we then construct TCA-GAN to generate the adversarial perturbation to fool this model. Our proposed TCA-GAN is a cyclic structure, which mainly consists of two generative modules and two corresponding domain discriminators. These two cycle-consistent adversary generators separately produce and remove the adversarial perturbation. The generated adversarial perturbation can be further added to the original image to obtain the adversarial example. Note that both the natural and adversarial domains have been demonstrated to be interchangeable by adding or eliminating adversarial perturbations [43], [44]. In our work, we take these two image domains into account by constructing the bidirectional adversary generation and elimination mechanism. Respectively, two domain discriminators are utilized to distinguish the legitimate example $x$ and its adversarial counterpart $x^{adv}$. The mechanism of TCA-GAN is presented in Fig. 4.

The forward adversary generation aims at producing an adversarial perturbation to be combined with the input facial image, disrupting the reconstruction of the substitute model.
as closer as possible. Hence, it can bring richer supervision on generating puissant adversarial perturbations. This cycle-consistent structure can further enhance the generalizability performance of the adversarial examples.

To balance generators relatively, we construct two domain discriminators to distinguish adversarial examples and legitimate examples, respectively. The main reason for constructing two domain discriminators is that we need them to assist two corresponding generators to simultaneously learn the reversible transformation between natural and adversarial image domains. In the meantime, two domain discriminators contribute to learning the discriminative features of natural and adversarial examples, reducing the risks of over-fitting for the adversary generation or its opposite purification. As the training of TCA-GAN continues, both generators and discriminators gradually become powerful to approximate any real-valued functions, resulting in the Nash equilibrium of this minimax optimization. Considering the adversarial domain is not the absolute complement of the legitimate domain, it is thus reasonable to construct two domain discriminators for the reversible adversary generation. The adversarial loss function of our GAN structure can be formalized as follows:

\[
L_{\text{adv}} = D_L(x_{\text{adv}}) - D_L(x) + D_A(x_{\text{adv}} - G_P(x_{\text{adv}})) - D_A(x_{\text{adv}}), \tag{5}
\]

where \(D_L(\cdot)\) denotes the discriminator of the domain on legitimate examples, while \(D_A(\cdot)\) represents the discriminator of the domain on adversarial examples. For brevity, \(x_{\text{adv}} = x + G_P(x)\) denotes the adversarial example created by the forward adversary generator. Consequently, the adversarial process will further enhance the performance of generators until Nash equilibrium. Overall, the total objective of TCA-GAN can be formalized as below:

\[
L = L_{\text{adv}} + \lambda_{\text{cyc}} \cdot L_{\text{cyc}} + \lambda_{\text{disr}} \cdot L_{\text{disr}}, \tag{6}
\]

where \(\lambda_{\text{cyc}}\) and \(\lambda_{\text{disr}}\) manage the relative significance of the objective function. We conduct the GAN mechanism by solving the min-max game, in which we minimize the objective function when training the generator and maximize it when training the discriminator.

### C. Post-Regularization

In order to further enhance the transferability of our adversarial examples, we append a post-regularization to weaken their specificity on the substitute model. Inspired by the observation from [34] and [46] that the more specific adversarial examples against the deep learning model, the less generalizable they become. We then slightly shift the attention of generated adversarial examples away from the substitute model, which means that we aim at obtaining a second-best adversarial example towards the substitute model for better generalization.

One effective way to regularize the off-the-shelf adversarial example is to make a distillation. As aforementioned, we disrupt the substitute model by increasing the discrepancy of latent variables between the natural and adversarial examples. Similarly, we also incorporate feature-level constraints to post-regularization to obtain more transferable adversarial examples. We first initialize the new adversarial example by appending random noise to escape the non-smooth vicinity of the adversarial example. Then we guide this new example to approximate the adversarial example generated by TCA-GAN. The corresponding optimization can be written as follows:

\[
\max_{x_{\text{adv}}} \left[ S_c(x_{\text{adv}}) - S_c(S(x)) \right] \circ [S_c(x_{\text{adv}}) - S_c(S(x))],
\]

s.t. \(\|x_{\text{adv}} - x\|_\infty < \epsilon, \tag{7}\)

where \(x_{\text{adv}}\) denotes the regularized adversarial example, and \(\circ\) is the Hadamard product. Note that the second term in Equation (7) is served as a weight on increasing latent discrepancy during the optimization. To solve the optimization problem, we adopt the iterative Projected Gradient Descent (PGD) method [24]. This approach involves conducting gradient ascent on the negative target function to increase its value while applying a projection operation to constrain the adversarial perturbation within an \(\epsilon\)-ball during each iteration. By implementing this approach, we effectively solve the constrained optimization problem to generate more transferable adversarial examples. As aforementioned, the post-regularization can be regarded as a distillation from the adversarial example generated from TCA-GAN. We also introduce the reconstruction of the original example as the reference to facilitate the optimization process. Specifically, we maximize the feature-level distance between regularized adversarial examples and reconstructed outputs instead of original examples. The main reason why we choose to involve reconstructed examples but not original ones is that reconstructed examples can be viewed as an augmentation of original counterparts. Thus we can obtain more generalizable adversarial examples by attacking both the original example and their augmented versions simultaneously. In other words, the main goal of our post-regularization module is to obtain a transferable adversarial example, and at the same time, its neighboring examples are also adversarial against the substitute model. This optimization can also induce the regularized adversarial example to be more robust to small transformations. Furthermore, the regularized adversarial example is optimized to keep the same high latent discrepancy. To simplify this optimization procedure, we adopt an iterative gradient ascent strategy on the regularized adversarial example. Afterward, we project the regularized adversarial example to a restriction to limit its magnitude. The overall procedure is shown in Algorithm 1, where \(\text{Clip}_e\) and \(\text{Clip}_{\text{image range}}\) represent the clipping in the \(\epsilon\)-ball and the image range, respectively. Note that the randomization in the initial stage of adversarial examples can further prevent the label leaking and gradient masking problem.

### IV. Experiments

In this section, we first introduce our proposed dataset on disrupting DeepFake face swapping and the measurement of our experimental results. Next, we make a comparison with other transferrable adversarial attack methods and conduct an ablation study on the component modules. To further explore the effectiveness of our method, we verify that the perturbed
Algorithm 1 Post-Regularization

Require:
- Adversary generator of TCA-GAN $G_p$;
- Original face image $x$;
- Substitute model $S$;
- Adversarial perturbation bound $\epsilon$;
- Iteration number $N$;
- Iterative step size $\alpha$.

Ensure:
- Regularized adversarial example $x_{adv}$.

1: $x_{adv} = x + G_p(x)$;
2: Randomly initialize $x_{adv}$ from the neighborhood of $x_{adv}$;
3: $x_{rec} = S(x)$;
4: $W = S_e(x_{adv}) - S_e(x_{rec})$;
5: for $k = 1, \ldots, N$ do
6: $\mathcal{L} = \left(\left(S_e(x_{adv}) - S_e(x_{rec})\right) \circ W\right) / \|x\|_F$;
7: $r = \text{Clip}_P(x_{adv} - x + \alpha \cdot \text{sign}(\nabla_{x_{adv}} \mathcal{L}))$;
8: $x_{adv} = \text{Clip}_{\text{image}_r}(x + r)$;
9: end for
10: return $x_{adv}$.

face-swapped images can facilitate DeepFake detection. Comprehensively, we also apply our method to several face manipulation models to validate the generalization performance.

A. Experimental Setup

1) Dataset: Following [34], we construct our DeepFake face-swapping dataset based on a high-quality facial image dataset: Face Scrub [52]. Our dataset consists of 6,274 images of 256 × 256 resolution, incorporating 40 male identities and 38 female identities. To support the future development of DeepFake adversarial attacks, we also train the corresponding DeepFake face-swapping models on the proposed dataset. Note that in order to simulate the real-world application scenario, the database to train both the substitute model and TCA-GAN is disjoint with the evaluation database on disrupting DeepFake face swapping. Specifically, we utilize 1,439 facial images (randomly selected 6 male and 6 female identities) to train the substitute model and also TCA-GAN. The rest 4,835 facial images of different identities are used for evaluating the generated transferable adversarial examples against black-box DeepFake face swapping. Apart from the evaluation of our constructed dataset for DeepFake face swapping, we also incorporate a third-part dataset, CelebFaces Attributes (CelebA) [53], to evaluate the generalizability of our method on other face manipulation models. CelebA contains 202,509 facial images that are annotated with 40 binary attributes. Following [31], [37], we utilize the resized and aligned facial images of 128 × 128 size and randomly select 5,000 images from CelebA as the test set. Likewise, these facial images have no intersection with the training set of TCA-GAN and the substitute model to simulate the real-world black-box scenario.

2) Metrics: To comprehensively evaluate the effectiveness of our proposed method, we conduct both referenced and non-referenced image quality assessments on the face-swapped images. Structure SIMilarity (SSIM) [54] index and Feature SIMilarity (FSIM) [55] are measured on both pairs of input face images and face swapped images in terms of legitimate examples and adversarial examples. Besides, Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [56] is also utilized to assess the image quality. A smaller BRISQUE score represents better visual quality.

3) Implementation Details: For the purpose of simulating the practical adversarial attack scenario, the generated adversarial examples will be further resized to the original size. Consequently, they will be resized and randomly transformed before entering the target DeepFake model. In this paper, the image value is normalized to [0, 1]. To obtain the visually undetectable adversarial perturbation, we restrict the bound $\epsilon$ to 0.03. The hyper-parameters for training TCA-GAN are set as $\lambda_{\text{CYC}} = 1.0$ and $\lambda_{\text{dist}} = 10.0$. We run the post-regularization for 10 iterations with the step size of 0.006.

B. Experimental Results

To begin with, we present several face-swapped examples in Fig. 5, depending on the input of legitimate images and adversarial images. In addition, we perform a sequence of experiments to provide empirical evidence of our method of attacking against DeepFake face swapping. Note that all experiments in this section are conducted in the restricted black-box scenario, i.e., the target DeepFake model is inaccessible until the completion of adversary generation. We generate

![Fig. 5. Examples of both original and adversarial (disrupted) face swapping in the restricted black-box scenario. Deepfake face swapping replaces the source face in the legitimate image with the fake face of the targeted identity.](image)

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
adversarial facial images against the substitute model and directly feed them into black-box DeepFake face swapping models. For a fair comparison, all the hyper-parameters of compared methods are set according to the suggestions in their original manuscripts. The comparison with other adversarial attack methods is reported in Table I.

The disruption against DeepFake face swapping is evaluated through the visual similarity between face-swapped images that are generated from legitimate examples and adversarial examples, respectively. In other words, the more extensive disruption on the face-swapped outputs represents the effectiveness of our adversarial examples. Due to the infinite norm restriction on the size of adversarial perturbation, there remains only a tiny gap in the visual distances between legitimate examples and adversarial examples during the comparison. According to the table, we can observe that our method presents an effective perturbation on DeepFake face swapping while preserving the adversarial perturbation visually imperceptible. Moreover, we evaluate the non-referenced image quality assessment BRISQUE as shown in Table II. Note that our goal is to obtain inferior face-swapped images that are induced by adversarial examples while keeping the image quality of input adversarial examples high. Both the referenced and non-referenced image quality assessments demonstrate that our methods can effectively give rise to facial distortion on DeepFake face-swapped images.

### C. Time Cost Comparison

In addition to the efficacy assessment of our method, we also evaluate the efficiency of our approach for adversary generation in the application stage. As presented in Table III, we provide the computational cost of producing adversarial facial images against the local substitute model related to several attack methods. We conduct all the experiments on a single NVIDIA Tesla A100 with different batch sizes. It can be seen that our method does not suffer from too heavy computational overheads at the application stage. Hence, generating adversarial examples by our method can be an efficient way to protect facial images against black-box face manipulation.

Note that we have to admit our method requires an extra computational cost to train TCA-GAN to generate adversarial examples compared with optimization-based adversary generation methods. However, utilizing a generative model to produce adversarial examples in the application stage can save more time in comparison with Multi-iterative adversary generation methods. The further acceleration of our adversary generation method will be investigated in our future work.

### D. Ablation Study

In this section, we analyze the effectiveness of each component module in our proposed method. To further explore the effectiveness of adopting the face reconstruction model as the substitute model, we provide comparison experiments related to different choices of the substitute model for adversary generation, as shown in Table IV. Specifically, we consider several (generative) image restoration models as substitute models, including denoising and deblurring models. We provide both referenced and non-referenced image quality assessments of the output face-swapped images with the input of adversarial examples against different substitute models. It is worth noting that the face reconstruction model tends to promote producing transferable adversarial examples against unforeseen black-box DeepFake models. On account of the high similarity of the generating process between face reconstruction and manipulation, adversarial examples against the face reconstruction model (substitute model) are easier to be transferred to other face manipulation models. Hence, the generated face-swapped results are much more easily to be visually disrupted.

On account of only possessing a single field of legitimate examples, our cycle-consistent loss considers the unidirectional adversary circulation. The unidirectional generation maintains the consistency between legitimate examples and reconstructed images that are generated through two generators in TCA-GAN, respectively. Furthermore, the bidirectional cycle-consistent version of TCA-GAN considers the cycle reconstruction of adversarial examples on the basis of the unidirectional one. Besides, these adversarial examples against the substitute model are regenerated by PGD [24] per iteration.
The comparison of the different versions of cycle-consistency is presented in Table V. It is worth mentioning that we mainly focus on perturbed face-swapped images. We can observe that TCA-GAN with the unidirectional cycle-consistency conducts much better than the one with bidirectional cycle-consistency. The main reason is that the obtained adversarial examples are excessively biased towards the substitute model, which is hard to transfer to other DNN models. On the contrary, our goal is to acquire a sub-optimal adversarial example and then generalize it to other face manipulation models like DeepFake. As a consequence, the guidance of an excessively powerful adversarial example may lead to the opposite effect of adversary transferring.

Concerning the main design of our proposed framework, we also explore the effect of component modules. Concretely, we mainly assess the disruption on face-swapped images in terms of the cycle-consistency, the latent variable disruption, and the post-regularization. The conducted ablation results are reported in Table VI. Note that the latent variable disruption denotes that we induce a disruption on the latent variable of the substitute model when training TCA-GAN. As the table presented, the latent variable disruption shows a beneficial impression on enhancing the transferred adversary disruption. Moreover, we can derive a 3.1% SSIM performance enhancement with the latent variable disruption. The disruption of latent variables of the substitute model can be regarded as disturbing the facial feature during the facial reconstruction. This also demonstrates that the perturbation towards the latent representation is more likely to generalize to other DNN models. Remarkably, the post-regularization presents a 3.9% SSIM performance-boosting result. This post-processing can further help to distill the generated adversarial examples from TCA-GAN, meanwhile keeping a long discrepancy on the latent variables in terms of legitimate examples and regularized adversarial examples. Likewise, the result supports that the sub-optimal adversarial example has a more robust generalization performance than the optimal adversarial example. The proposed modules and the integration of them are demonstrated to be valid for disrupting the DeepFake face swapping mechanism. In summary, there is a large room for performance improvement on our proposed DeepFake face swapping dataset.

Table III: Average time cost over 10 runs for transferable adversary generation against black-box DeepFake face swapping systems with different batch sizes

| Batch size | FGSM [12] s | PGD [24] s | DIM [47] s | TIM [48] s | ILA [46] s | APGD [49] s | Ours s |
|------------|-------------|------------|------------|------------|------------|------------|--------|
| 16         | 380         | 1278       | 1307       | 1196       | 1712       | 3544       | 716    |
| 32         | 221         | 746        | 742        | 703        | 1047       | 2138       | 471    |
| 64         | 198         | 662        | 679        | 636        | 920        | 1805       | 284    |

Table IV: Ablation study on the choice of the substitute model. The best performance is marked in bold

| Substitute model | SSIM↑ | FSIM↑ | BRISQUE↑ |
|------------------|-------|-------|----------|
| CBDNet [57]      | 0.784 | 0.880 | 41.29    |
| RIDNet [58]      | 0.770 | 0.882 | 43.07    |
| DeblurGAN [59]   | 0.802 | 0.906 | 39.65    |
| PBGAN [60]       | 0.762 | 0.889 | 44.20    |
| **Ours (Face reconstruction)** | **0.731** | **0.873** | **47.13** |

Table V: Ablation study on the cycle-consistency type in TCA-GAN. The best performance is marked in bold

| Cyclic type | SSIM↑ | FSIM↑ | BRISQUE↑ |
|-------------|-------|-------|----------|
| w/o cycle-consistency | 0.752 | 0.878 | 45.43    |
| w/ bidirectional cycle-consistency | 0.738 | 0.876 | 46.07    |
| w/ unidirectional cycle-consistency | **0.731** | **0.873** | **47.13** |

Table VI: Ablation results for component modules. The best result in each column is bold

| CC | LVD | PRM | SSIM↑ | FSIM↑ | BRISQUE↑ |
|----|-----|-----|-------|-------|----------|
| 1  | ✓   | ✓   | 0.79  | 0.903 | 40.95    |
| 2  | ✓   | ✓   | 0.783 | 0.896 | 42.11    |
| 3  | ✓   | ✓   | 0.764 | 0.887 | 43.27    |
| 4  | ✓   | ✓   | 0.761 | 0.884 | 43.94    |
| 5  | ✓   | ✓   | 0.752 | 0.878 | 45.43    |
| 6  | ✓   | ✓   | 0.755 | 0.882 | 45.18    |
| 7  | ✓   | ✓   | **0.731** | **0.873** | **47.13** |

CC means the cycle consistency mechanism.
LVD means the latent variable disruption.
PRM means the post-regularization module.

Fig. 6. Hyper-parameter sensitivity of our method on both reference (SSIM) and non-reference (BRISQUE) image quality assessments of the disrupted DeepFake face swapping images. We report the hyper-parameters adjustment of $\lambda_{cyc}$ in (a) and $\lambda_{disr}$ in (b).
Fig. 7. Exemplary results for StarGAN [37] with the input of legitimate face images and adversarial face images. The generated adversarial example is applicable to arbitrary style translation (Hair color, Gender, and Aged).

| TABLE VII |
|------------------|----------|---------|
| Method           | Original accuracy | Distorted accuracy |
| Xception [2]     | 67.1%    | 71.4%   |
| MesoNet [3]      | 71.9%    | 73.3%   |
| Capsule-forensics [4] | 60.5%    | 65.4%   |
| EfficientNetB4Att[61] | 69.8%    | 73.6%   |
| CNNDetection[62] | 78.3%    | 79.7%   |

We can easily observe that several DeepFake detection methods achieve a higher detection accuracy on the distorted face-swapped outputs compared with the original face-swapped images. In other words, our method contributes to the performance enhancement of DeepFake detection, which demonstrates the effectiveness of our DeepFake disruption method. The main reason behind the facilitation is that these DeepFake detection methods are focused on the imperceptible abnormal area of the input face images, while adversarial examples can enlarge the corresponding distortion of output face-swapped images.

Note that under the disturbance of cross-domain DeepFake face-swapped examples, these corresponding detection methods exhibit a slightly poor performance on our constructed DeepFake face swapping dataset. This also demonstrates the significance of our method in preventing malicious face swapping from the generation stage. Apart from decreasing the visual quality of face-swapped images, our adversary-based disruption method can also be served as assistance for enhancing the performance of DeepFake detection.

F. Generalization to Facial Style Translation

In order to investigate the generalizability ability of our method on other face manipulation models, we transfer our adversarial attacks against several face image translation models in the restricted black-box scenario. With respect to StarGAN [37] and AttGAN [38], the corresponding face image translation results of both legitimate examples and adversarial examples are shown in Fig. 7 and Fig. 8. It is clear that adversarial examples are nearly the same in human vision, yet they can strongly distort the face translation with respect to different attributes. Note that different perturbed image translation results are all induced by the same input of adversarial example. In other words, the generated adversarial perturbation...
Fig. 8. Exemplary results for AttGAN [38] with the input of legitimate face images and adversarial face images. The generated adversarial example is applicable to arbitrary style translation (Hair color, Gender, and Glasses).

| Method          | SSIM | FSIM | BRISQUE |
|-----------------|------|------|---------|
| StarGAN [37]    | 0.591| 0.806| 49.51   |
| GANImation [63] | 0.796| 0.898| 42.27   |
| SAGAN [64]      | 0.864| 0.919| 39.23   |
| AttGAN [38]     | 0.841| 0.908| 43.76   |

is image-agnostic against these facial style translation models and thus can be adapted to arbitrary face images.

To further quantify the visual distortion caused by our generated adversarial examples, we present both reference and non-reference image quality assessments on the disrupted face translation results, as shown in Table VIII. Note that the face style translation is mainly conducted on the CelebA [53] dataset. Moreover, we follow the same setting as the DeepFake disruption under restricted black-box attacks to simulate the real-world scenario. In consideration of the fixation of the adversarial perturbation upper bound, we mainly focus on the disturbance on face style-translated images. The generalization to face style translation models also represents that our proposed method can be served as an effective way to defend our facial images against unforeseen face manipulation.

V. Conclusion

In this paper, we systematically explore the practical DeepFake defense scenario with a restricted black-box adversarial attack. We design the TCA-GAN method to generate transferrable adversarial perturbation via attacking a substitute model. Afterward, we introduce a novel post-regularization on the synthesized adversarial example to further enhance the generalization ability, which can induce intense disruption on DeepFake face swapping. For ease of subsequent research on DeepFake disruption, we construct a new baseline to defend against DeepFake face swapping, including the dataset and models. To comprehensively evaluate the disruption induced by our adversarial examples, we conduct experiments on both referenced and non-referenced image quality assessments. Notably, we also show that our method can enhance the performance of DeepFake detection, which is more beneficial to defend against malicious face swapping. Subsequently, the extension to other face manipulation methods can further demonstrate the generalization performance of our method.

Diffusion models have recently achieved state-of-the-art performance on a series of generative tasks. Our method can also be applied to diffusion models by replacing the target loss function that is utilized for training the proposed adversary generation model. For instance, we can learn an adversary generator with an optimization target of minimizing the likelihood of the generated adversarial examples. Furthermore, constructing transferable adversaries among various types of generative models is also a significant aspect of defending personal photos from deep learning-based manipulation, which is left for our future work. We hope our work can shed light on protecting personal photos from unauthorized face manipulation in real-world scenarios.

REFERENCES

[1] P. Korshunov and S. Marcel, “DeepFakes: A new threat to face recognition? Assessment and detection,” 2018, arXiv:1812.08685.
[2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Dec. 2019, pp. 1–11.
[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a compact facial video forgery detection network,” in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), Mar. 2018, pp. 1–7.
[54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[55] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” IEEE Trans. Image Process., vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[56] A. Mittal, A. K. Moothy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” IEEE Trans. Image Process., vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[57] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, “Toward convolutional blind denoising of real photographs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 1712–1722.

[58] S. Anwar and N. Barnes, “Real image denoising with feature attention,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Dec. 2019, pp. 3155–3164.

[59] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, “DeblurGAN: Blind motion deblurring using conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8183–8192.

[60] J. Pan et al., “Physics-based generative adversarial models for image restoration and beyond,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 7, pp. 2449–2462, Jul. 2021.

[61] N. Bonetini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, “Video face manipulation detection through ensemble of CNNs,” in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5012–5019.

[62] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-generated images are surprisingly easy to spot · · · for now,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2020, pp. 8695–8704.

[63] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “GANimation: Anatomically-aware facial animation from a single image,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 818–833.

[64] G. Zhang, M. Kan, S. Shan, and X. Chen, “Generative adversarial network with spatial attention for face attribute editing,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 417–432.

Junhao Dong received the B.S. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2020. He is currently pursuing the M.S. degree in computer science and technology with Sun Yat-sen University, Guangzhou, China. His research interests include adversarial examples, few-shot learning, and face recognition.

Yuan Wang received the B.S. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2020, and the M.S. degree in computer technology from Sun Yat-sen University, Guangzhou, China, in 2022. His research interests include person re-identification, image generation, and face recognition.

Jianhuang Lai (Senior Member, IEEE) received the Ph.D. degree in mathematics from Sun Yat-sen University, China, in 1999. In 1989, he joined Sun Yat-sen University as an Assistant Professor, where he is currently a Professor with the School of Computer Science and Engineering. He has published more than 250 scientific papers in the international journals and conferences on image processing and pattern recognition. His current research interests include the areas of computer vision, pattern recognition, and its applications. He serves as the Deputy Director of the Image and Graphics Association of China.

Xiaohua Xie (Member, IEEE) received the Ph.D. degree in applied mathematics from Sun Yat-sen University, China, in 2010. He was an Associate Professor with the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences. He is currently an Associate Professor with Sun Yat-sen University. His current research interests include image processing, computer vision, pattern recognition, and machine learning.