Practical Disruption of Image Translation Deepfake Networks

Nataniel Ruiz[1], **Sarah Adel Bargal**[2], **Cihang Xie**[3], **Stan Sclaroff**[1]

[1] Boston University
[2] Georgetown University
[3] University of California, Santa Cruz
nruiz9@bu.edu, sarah.bargal@georgetown.edu, cixie@ucsc.edu, sclaroff@bu.edu

Abstract

By harnessing the latest advances in deep learning, image-to-image translation architectures have recently achieved impressive capabilities. Unfortunately, the growing representational power of these architectures has prominent unethical uses. Among these, the threats of (1) face manipulation ("DeepFakes") used for misinformation or pornographic use (2) "DeepNude" manipulations of body images to remove clothes from individuals, etc. Several works tackle the task of disrupting such image translation networks by inserting imperceptible adversarial attacks into the input image. Nevertheless, these works have limitations that may result in disruptions that are not practical in the real world. Specifically, most works generate disruptions in a white-box scenario, assuming perfect knowledge about the image translation network. The few remaining works that assume a black-box scenario require a large number of queries to successfully disrupt the adversary's image translation network. In this work we propose *Leaking Transferable Perturbations (LTP)*, an algorithm that significantly reduces the number of queries needed to disrupt an image translation network by dynamically repurposing previous disruptions into new query efficient disruptions.

Introduction

Since its inception as an encoder-decoder based face-swapping technique (Güera and Delp 2018) the term "deepfake" has adopted a broader meaning and can be used to refer to any altered media of someone's likeness. Recently there have been remarkable advances in face modification algorithms and controllable face synthesis (Thies et al. 2016, 2018; Wiles, Sophia K., and Zisserman 2018; Kim et al. 2018; Ranjan et al. 2018; Usman et al. 2019; Geng, Cao, and Tulyakov 2019; Nguyen-Phuoc et al. 2019; Ghosh et al. 2020). Some algorithms only need a single image and can create modified versions of that person under different poses, expressions, lighting and other attribute changes (Choi et al. 2018; Pumarola et al. 2018; Choi et al. 2019). The most advanced algorithms can create puppeteering videos using as few as one image (Zakharov et al. 2019; Tewari et al. 2020). This few-shot deepfake technology based on image translation networks has gained popularity in the mainstream with apps such as FaceApp that allow for transformation of images such as making someone smile and making them appear older or younger, among other interventions. In parallel, image translation networks have been re-purposed to generate nude pictures of clothed bodies in what is called a "DeepNude" transformation.

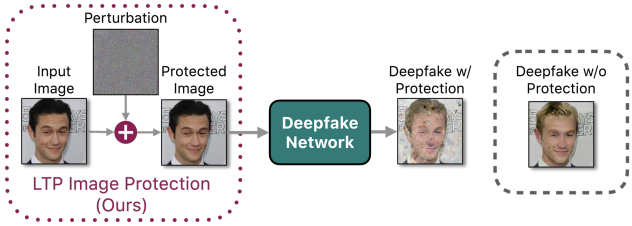

Figure 1: Our approach (LTP) protects face images from being manipulated by an image translation deepfake network. After applying an imperceptible filter on the input image, the deepfake system is forced to generate a corrupted output.

These technologies can be used in malicious ways to produce undesirable content of someone without their consent. This concern has already materialized in several ways, including creating non-consensual pornographic footage and producing videos with fake political speeches. Attempts to detect manipulated media are underway and there is an "arms race" between detecting deepfakes (Rossler et al. 2019; Yang, Li, and Lyu 2019; Li and Lyu 2019; Wang et al.) and evasion of deepfake detection (Neekhara et al. 2020; Gandhi and Jain 2020). Instead of detecting deepfakes after the fact, there is work (Ruiz, Bargal, and Sclaroff 2020; Yeh et al. 2020; Aneja, Markhasin, and Niessner 2021) that proposes using *white-box adversarial attacks* to protect an image from modification by disrupting the functioning of image translation networks. While these works assume that one has access to the model's structure, weights and gradients, in a real scenario, there is a high probability that these might not be accessible. In contrast, in this work, we focus on the *black-box scenario* where we solely have limited query access to the deepfake model instead of unlimited access to the model and its internals.

An image translation-based online deepfake generation service usually allows for API queries where a user sends an image and receives the translated output (e.g. FaceApp, DeepNude). This is an instance of the image translation

black-box threat model, where the model internals are unknown, but the user can query the model using selected input images and study the output of the model. This is similar in nature to the classification black-box threat model (Papernot et al. 2017; Chen et al. 2017; Ilyas et al. 2018; Ilyas, Engstrom, and Madry 2019). In this work we demonstrate attacks on image translation models under this threat model and show the vulnerability of facial attribute editing and expression editing image translation networks. Specifically, we are the first to explore black-box adversarial attacks on image translation systems with the application of disrupting deepfake generation, along with other concurrent work (Yeh et al. 2021; Huang et al. 2021). In contrast to this work, we pay special attention to the query-efficiency of our disruption generation. This is in order to (1) minimize the probability of a deepfake provider detecting our disruption attempt (2) provide practical disruptions that can be used in the real world for large amounts of images.

In our work, we reformulate classic black-box attacks for this new image translation scenario and demonstrate their effectiveness in preventing deepfake generation. However, the number of queries of such black-box attacks is prohibitive in a real-world scenario where an adversary might detect an attempted attack or the query budget might run out. We present highly effective algorithm called *Leaking Transferable Perturbations (LTP)* that sharply decreases the average number of queries required to generate successful image translation disruptions. We show an illustration of LTP in Figure 1. LTP is composed of two phases, a short *leaking phase* during which the network is attacked using a classic black-box attack on a small set of images and an *exploitation phase*, where the algorithm leverages the information obtained during the *leaking phase* to subsequently attack the network with high efficiency.

During the LTP *leaking phase* we attack a set of images using a classic black-box attack. Once these perturbations have been generated, PCA components are extracted from them. During the *exploitation phase* these PCA components are used as attack vectors in a query-based attack. Compared to state-of-the-art methods, we are able to reduce the number of necessary queries to protect an image by more than half on multiple image translation networks. Our code will be made publicly available upon acceptance with a permissive open-source license in order to promote research of protection of individuals from non-consensual deepfakes.

We summarize our contributions as follows:

- We present a framework to protect images from being modified by deepfake networks by using black-box adversarial attacks. In contrast to prior work, these disruptions can be used in real world scenarios where the deepfake network can only be accessed using queries and the model internals cannot be inspected.
- We present a novel method called *Leaking Transferable Perturbations (LTP)*, that significantly improves the efficiency of black-box deepfake disruptions by repurposing information gathered during initial attacks. This allows the attack to scale vastly more efficiently compared to other state-of-the-art methods.

Related Work

White-box Attacks on Classifiers Different threat models for adversarial attacks have been defined for the image classification scenario. They are defined by the amount of information that the adversary has regarding the target model. Under a white-box threat model in the classification scenario, the structure and weights of the classifier h are available to the adversary. This means that the classifier can be run locally on the adversaries' infrastructure, and gradients can be computed. Under this threat model (Szegedy et al. 2014) demonstrated the existence of adversarial examples for deep neural network classifiers. Since then, there has been a large amount of work on attacking models under this setting by performing gradient descent on the defined classification loss l or optimization methods using the gradient information (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Papernot et al. 2016; Carlini and Wagner 2017; Nguyen, Yosinski, and Clune 2015; Moosavi-Dezfooli et al. 2017; Kurakin, Goodfellow, and Bengio 2017; Madry et al. 2018).

Black-box Attacks on Classifiers In a real-world scenario the adversary might not have access to either the structure or the weights of the classifier h. Instead, she might have access to an API which allows queries to the model. The adversary might then have either access to the probability outputs, or uniquely to the classification decisions of the model. The goal is to attack the model while minimizing the number of queries as well as the magnitude of the attack under a suitable norm.

There is extensive work on black-box attacks on classification deep networks. One approach is to train a surrogate network and transfer white-box attacks generated using the surrogate network to the target network (Papernot et al. 2017; Liu et al. 2016). Another effective approach is to estimate the gradients using finite-differences, Monte Carlo sampling methods or other techniques and subsequently perform gradient descent (Chen et al. 2017; Ilyas et al. 2018; Ilyas, Engstrom, and Madry 2019; Cheng et al. 2019; Tu et al. 2019). Another class of approaches are local-search approaches that attack the network by probing the black-box without any gradient estimation (Narodytska and Kasiviswanathan 2016; Guo et al. 2019; Andriushchenko et al. 2020). This prior work generates adversarial attacks from scratch for each individual image. In contrast, our proposed LTP attack learns to perform more efficient attacks by re-purposing information from initial black-box attacks. (Bhagoji et al. 2018) use a PCA-based (principal component analysis) query reduction technique where the gradient for a sample is computed along the principal components of a representative data sample. Our work involves the use of PCA, albeit in a completely different manner. LTP computes a PCA decomposition of the *generated perturbations* and re-uses this information for future attacks. In essence, LTP learns transferable attack components that can be used to efficiently query the model in a local-search manner in subsequent attacks.

Image Translation Adversarial Attacks Image translation networks have recently achieved impressive results in

deepfake generation and face modification using few images (or one image) of an individual (Choi et al. 2018; Pumarola et al. 2018; Choi et al. 2019; Zakharov et al. 2019; Tewari et al. 2020). Some models allow for generation of video of a person saying things that they did not say, using a single image (Zakharov et al. 2019; Tewari et al. 2020). In general, most image translation models are trained using a GAN setup. Some are trained in a supervised manner (Isola et al. 2017; Wang et al. 2018; Zakharov et al. 2019), while others are trained in an unsupervised manner (Zhu et al. 2017; Pumarola et al. 2018; Choi et al. 2018, 2019).

There is previous work that demonstrates attacks on generative models, specifically autoencoders (Tabacof, Tavares, and Valle 2016; Kos, Fischer, and Song 2018). Recently, there has been work that proposes white-box attacks on image translation networks (Ruiz, Bargal, and Sclaroff 2020; Yeh et al. 2020) in order to disrupt output generation by either neutralizing the image transformation or corrupting/distorting the output image. Disrupting deepfakes reveals itself to be an interesting application for these types of attacks. (Ruiz, Bargal, and Sclaroff 2020) explore white-box attacks on image translation networks such as pix2pixHD (Wang et al. 2018) and CycleGAN (Zhu et al. 2017). They also show that their white-box attack allows for disruption of deepfake generation using StarGAN (Choi et al. 2018) and GANimation (Pumarola et al. 2018). (Yeh et al. 2020) present white-box attacks on CycleGAN (Zhu et al. 2017). In contrast, we explore black-box attacks on image translation models to disrupt deepfake generation.

There is also recent work, that explores black-box disruption attacks on image translation networks (Yeh et al. 2021; Huang et al. 2021). Yeh et al. (Yeh et al. 2021) present a black-box attack that seeks to neutralize or nullify the image translation process. Their method is tested on the CycleGAN architecture, trained for different manipulations such as putting glasses on faces, or changing hair color. Their method succeeds in neutralizing these manipulations with a high success rate. Nevertheless, the order of magnitude of queries needed to protect each individual image is in the tens of thousands. Given current security capabilities, websites hosting this type of deepfake service would be able to detect the attempt and either throttle or restrict network queries. Another concern is the time needed to generate any such attack, which can be in the order of magnitude of days for a set of one-hundred images.

Huang et al. (Huang et al. 2021) adopt the technique introduced by Papernot et al. (Papernot et al. 2017) of training a surrogate network that resembles the target image translation network, and then train a GAN to generate adversarial attacks, similar in spirit to Xiao et al. (Xiao et al. 2018). This type of approach has two main weaknesses: (1) the attack is situational, since the type of network and manipulation have to be roughly known before deciding on a surrogate architecture and surrogate training task (i.e. some surrogate tasks do not transfer to the target network (Li, Guo, and Chen 2020)) (2) the distribution of the training data of the surrogate and target model have to be similar (Papernot et al. 2017) and the attacker has to have a large amount of labeled training data for the surrogate (tens of thousands of images) - which is seldom the case (e.g. in a DeepNude scenario). In this work, we assume that we do not have knowledge over the target network and that we do not have large amounts of labeled data. Thus, we approach this problem by the optimization route. We take special care in addressing the query efficiency issues and produce a method that is significantly more query efficient than prior work.

Method

In this section we first provide a general formulation for image translation disruptions. Next, we present modifications of classic black-box attacks for the image translation scenario as baseline methods. Finally, we present our proposed method of Leaking Transferable Perturbations (LTP), that obtains more efficient attacks than baseline methods and state-of-the-art black-box image translation attacks.

Disrupting Image Translation Models Via Adversarial Attacks

An adversarial example is an image with small additive changes, that can be imperceptible to a human being, and affect the output label of the image classification model. In general an adversarial attack, which creates adversarial examples, on an image classification model h is defined by:

$$\min_{\boldsymbol{\eta}} l_{\boldsymbol{y}}(h(\boldsymbol{x}+\boldsymbol{\eta})), \quad \text{subject to } p(\boldsymbol{\eta}) \leq \boldsymbol{\epsilon}. \tag{1}$$

Different distance norms p have been proposed, and attacks usually use the L_2 or L_∞ norms. $l_{\boldsymbol{y}}$ is a surrogate loss that measures the degree of certainty that the model will classify the input as class $\boldsymbol{y}$. This surrogate loss can be defined in different ways, depending on the output of the model.

To delve into black-box attacks on image translation models, it is helpful to first present formulations for the white-box scenario. (Ruiz, Bargal, and Sclaroff 2020) formulates a targeted attack on an image translation generator $\boldsymbol{G}$, with target $\boldsymbol{r}$:

$$\min_{\boldsymbol{\eta}} L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}), \boldsymbol{r}), \quad \text{subject to } p(\boldsymbol{\eta}) \leq \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{x}$ is the input image, $\boldsymbol{\eta}$ is the generated perturbation, p is a chosen norm, $\boldsymbol{\epsilon}$ is the maximum attack magnitude and L is the chosen image-level regression loss. If $\boldsymbol{r} = \boldsymbol{x}$, the goal is to drive the output of the model towards the original input. We call this a *neutralizing attack*, since it neutralizes the image transformation brought by the generator $\boldsymbol{G}$.

They also define an untargeted attack seeking to maximize the distortion of the output image with respect to the non-attacked output (i.e. output without protection). We call this a *distortion attack*.

$$\max_{\boldsymbol{\eta}} L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}), \boldsymbol{G}(\boldsymbol{x})), \quad \text{subject to } p(\boldsymbol{\eta}) \leq \boldsymbol{\epsilon}. \tag{3}$$

Image Translation Black-Box Attacks

Here we propose formulations of baseline image translation black-box attacks. In essence, we modify attacks that were initially proposed for image classification for this scenario. We reformulate two gradient estimation-based approaches,

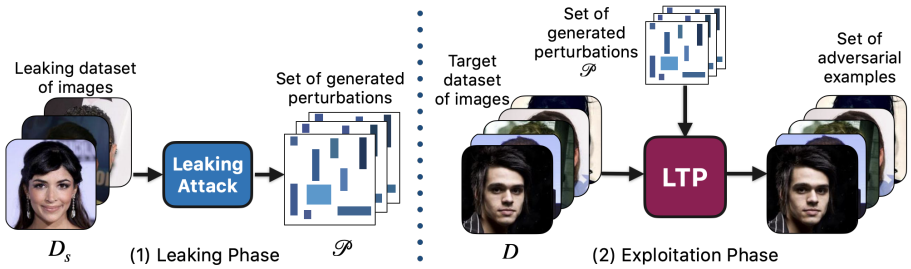

Figure 2: LTP method: During the leaking phase, image translation attacks are performed on the leaking dataset and a set of perturbations $\mathcal{P}$ are collected. The algorithm finds strong attacks efficiently during the exploitation phase by exploring the perturbation directions given by the principal components of $\mathcal{P}$.

Natural Evolution Strategies (Ilyas et al. 2018) and Bandits-TD (Ilyas, Engstrom, and Madry 2019) and two local search-based attacks SimBA (Guo et al. 2019) and Square Attack (Andriushchenko et al. 2020). We do so by replacing the classification loss l by an image-level regression loss L, that measures the distance between the target image $\boldsymbol{r}$ and the translated adversarial example $\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta})$.

Gradient Estimation-based Attacks We reformulate both Natural Evolution Strategies (NES) (Ilyas et al. 2018) and Bandits-TD (Ilyas, Engstrom, and Madry 2019) attacks for the image translation scenario. We show that they are able to produce effective attacks on image translation networks and are able to disrupt deepfake generation.

Our formulation of image translation NES (IT-NES) gradient estimate for the image-level regression loss L is $\nabla\mathbb{E}[L(\boldsymbol{G}(\boldsymbol{x}),\boldsymbol{r})] \approx \frac{1}{\sigma n}\sum_{i=1}^{n}\delta_i L(\boldsymbol{G}(\boldsymbol{x}+\sigma\delta_i),\boldsymbol{r})$, where $\boldsymbol{G}$ is the generator, n are the number of queries, $\boldsymbol{r}$ is the target image, σ is the variance of the Gaussian search distribution and using antithetic sampling we have $\delta_i \sim \mathcal{N}(0,I)$ for $i \in \{1,\ldots,\frac{n}{2}\}$ and set $\delta_j = -\delta_{n-j+1}$ for $j \in \{(\frac{n}{2}+1),\ldots,n\}$. The adversarial example is then updated using the estimated gradient $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \boldsymbol{\epsilon}\nabla\mathbb{E}[L(\boldsymbol{G}(\boldsymbol{x}),\boldsymbol{r})]$.

Bandits-TD introduces a time dependent prior and a data dependent prior. The method uses the antithetic NES gradient estimation method with $n = 2$. In similar fashion, we reformulate this attack for image translation. We call this formulation IT-Bandits-TD. We replace the classification criterion $l_{\boldsymbol{y}}(\boldsymbol{x})$ for image $\boldsymbol{x}$ and label $\boldsymbol{y}$ by the image-level regression criterion $L_{\boldsymbol{r}} = L(\boldsymbol{x},\boldsymbol{r})$, where $\boldsymbol{r}$ is the target image.

Local Search-based Attack We reformulate the SimBA attack (Guo et al. 2019) for the image translation scenario. Similar to (Narodytska and Kasiviswanathan 2016), SimBA iteratively changes single pixel values (in positive and negative directions) to find better adversarial attack candidates. In the classification scenario, SimBA iterates over all pixels and determines whether the change increases the loss $l_{\boldsymbol{y}}(h(\boldsymbol{x}))$. If yes, the pixel is modified. If both directions do not increase the loss then the pixel is skipped. The algorithm halts whenever the classifier misclassifies the perturbed image $h(\tilde{\boldsymbol{x}}) \neq \boldsymbol{y}$. We reformulate SimBA by replacing the classification loss $l_{\boldsymbol{y}}$ by the regression loss $L_{\boldsymbol{r}}$, and call this method IT-SimBA. We also reformulate the state-of-the-art Square Attack (Andriushchenko et al. 2020) for the image translation scenario in the same manner as IT-SimBA.

Leaking Transferable Perturbations (LTP)

In the black-box adversarial attack setting, we are given a budget of black-box queries for each image we would like to attack. In this setting, we have the same number of maximum allowed queries for all images in the dataset. That is, for each image $\boldsymbol{x}$ we want to solve the optimization problem

$$\min_{\boldsymbol{\eta}} L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}),\boldsymbol{r}), \quad \text{subject to } p(\boldsymbol{\eta}) \leq \boldsymbol{\epsilon}, \mathbf{Q} \leq \boldsymbol{B}, \quad (4)$$

where $\boldsymbol{G}$ is the generator of the image translation system, $\boldsymbol{\eta}$ is the perturbation, $\mathbf{Q}$ is the number of queries used and $\boldsymbol{B}$ is the maximum number of queries allowed for a single image.

An adversary would benefit from reducing the *total number of queries* required to attack a given dataset $\boldsymbol{B}_0$. Our proposed algorithm seeks to reduce $\boldsymbol{B}_0$ by, first, leaking elements of transferable perturbations from a small auxiliary dataset and then exploiting these transferable components on the images in the larger test dataset.

Intuitively, LTP works in two phases (1) a *leaking phase*, where the model is attacked using a classic attack on a small auxiliary dataset and information is gathered on successful attacks (2) an *exploitation phase*, where the model is attacked using this leaked information on the larger test set. This allows for a sharp reduction of amortized queries needed. Both phases are shown in Figure 2.

Leaking Phase During the *leaking phase* a small leaking dataset $\mathcal{D}_s$ is attacked using classic black-box attack. All images $\boldsymbol{x} \in \mathcal{D}_s$ are attacked until either success is achieved ($L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}),\boldsymbol{r}) < \tau$, where τ is the success threshold) or until a maximum number of queries $\mathbf{Q}$ are used. Our framework is general and any attack or combination of attacks can be used for the leaking phase. At this point, a set $\mathcal{P}$ of generated perturbations $\boldsymbol{\eta}$ is created. By applying principal component analysis (PCA) on perturbations $\boldsymbol{\eta} \in \mathcal{P}$, LTP extracts principal components $\boldsymbol{q} \in \mathcal{Q}$. These will serve as candidate vectors during the exploitation phase. Note that the leaking dataset can be a subset of the target dataset. This means that in general the distribution of the leaking dataset is very close to that of the target dataset such that no distribution shift issues should arise. Nevertheless, we study the performance variance when sampling different leaking datasets in the Experiments Section finding that it is small and does not substantially change results - showing the robustness of LTP with respect to variations in leaking dataset sampling.

Attack	Avg. Queries ↓	Avg. Norm ↓	SR ↑
IT-NES	598	1.82	98.8%
IT-Bandits-TD	855	4.38	96.3%
IT-SimBA	551	4.87	97.9%
IT-Square	531	5.00	98.8%
LaS-GSA	3,767	1.65	82.1%
LTP (Ours)	**231**	2.42	98.8%

Table 1: Comparison of black-box attacks with our proposed LTP method on an *expression editing* task. We observe that LTP achieves more than a x2 reduction in number of queries compared to the next best attack (IT-Square) with a much lower average norm and equal success rate.

Attack	Avg. Queries ↓	Avg. Norm ↓	SR ↑
IT-NES	1,001	2.90	99.8%
IT-Bandits-TD	4,901	4.99	52.2%
IT-SimBA	444	5.93	100%
IT-Square	3,856	5.00	98.7%
LTP (Ours)	**136**	4.88	100%

Table 2: Comparison of baseline image translation black-box attacks with our proposed LTP method on a *facial attribute editing* task. LTP achieves a x3 reduction in queries compared to the next best attack with lower average norm and equal success rate.

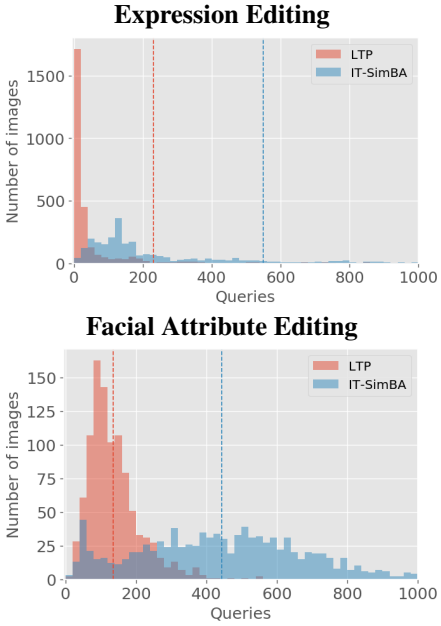

Figure 3: Histogram of queries for successful attacks for *expression editing* (top) and *facial attribute editing* (bottom). Vertical lines show mean queries. LTP achieves much more efficient attacks for both tasks.

Exploitation Phase This phase consists of using a local search-based querying attack that re-purposes the information gained during the leaking phase. Specifically, the image is iteratively queried using the leaked PCA components $\boldsymbol{q} \in \mathcal{Q}$ in the positive $(\boldsymbol{x}+\boldsymbol{q})$ and negative directions $(\boldsymbol{x}-\boldsymbol{q})$. If any of these directions increases the loss $L_{\boldsymbol{r}}(h(.))$, then the image is modified using that component. Given that $\mathcal{Q}$ does not necessarily span the image space (since $N_s < d^2$, where d is the image width/height), LTP switches to a full basis in image space after a number of iterations n_{sat} of saturating loss. The resulting attacks achieve strong results using substantially fewer queries $\mathbf{Q}$.

Discussion Whereas prior work uses statistical properties of data to improve query efficiency of attacks, to the best of our knowledge, we are the first to propose re-purposing previous attacks using dimensionality reduction techniques in order to perform more efficient local search-based attacks. Our method is inspired by the intuition that attacks on different images for a specific architecture should be correlated to a certain degree, given that models usually have specific vulnerabilities. The conclusions are surprising: *attacks on image translation models are decomposable in such a way that this decomposition is transferable to other images*. In essence, we postulate and verify the hypothesis that successful attacks can be mounted from a linear combination of components $\boldsymbol{\eta} = \alpha_1\boldsymbol{q}_1 + \alpha_2\boldsymbol{q}_2 + \alpha_3\boldsymbol{q}_3 + ...$ from previous attacks.

Experiments

In this section we introduce the datasets and models used. We then introduce our experimental setup. Finally, we present results on two image translation deepfake tasks.

Experimental Setup

Architectures and Datasets We apply our attack on two different tasks with vastly different network architectures: *expression editing* using GANimation (Pumarola et al. 2018) and *facial attribute editing* using StarGAN (Choi et al. 2018). For expression editing we attack three different expression changes and present averaged results. The expressions correspond to "closed eyes smile", "open eyes smile" and "surprised eyebrow raise". These expressions were selected because they enact salient changes in the image, as opposed to other more subtle expressions. For facial attribute editing we present averaged results over 5 different attribute classes. The classes are "black hair", "blond hair", "brown hair", "female" and "old". The dataset used for both architectures is the CelebA dataset (Liu et al. 2015). For expression editing we attack 1,000 images using each expression, yielding 3,000 individual attacks. For attribute editing we attack 200 images using 5 different classes, yielding 1,000 individual attacks. Our evaluation takes into account a larger diversity of attribute changes than related work (Ruiz, Bargal, and Sclaroff 2020; Yeh et al. 2021; Huang et al. 2021), which customarily evaluates attacks on a handful of attribute changes. We also evaluate on a larger number of images than prior work, which generally evaluates attacks on image counts in the low hundreds.

Implementation Details We adapt versions of the official NES, Bandits-TD, SimBA, Square and LaS-GSA code. For IT-NES, IT-Bandits-TD, and IT-Square we follow the parameter settings in the corresponding papers. For IT-SimBA

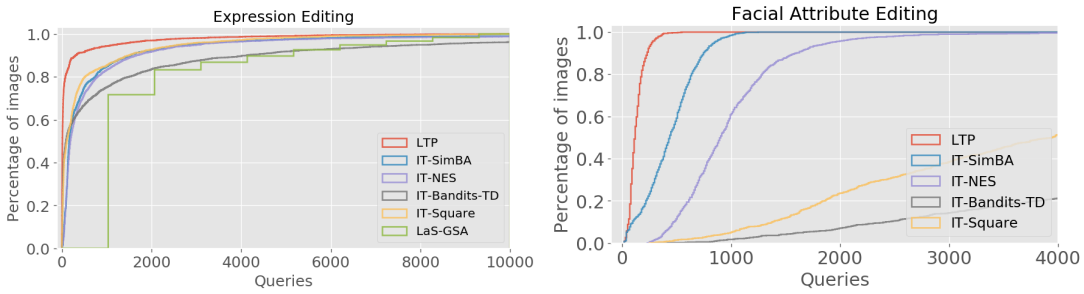

Figure 4: Success rate by number of queries for *expression editing* (left) and *facial attribute editing* (right). We observe that LTP converges to the top success rate much faster than other attacks.

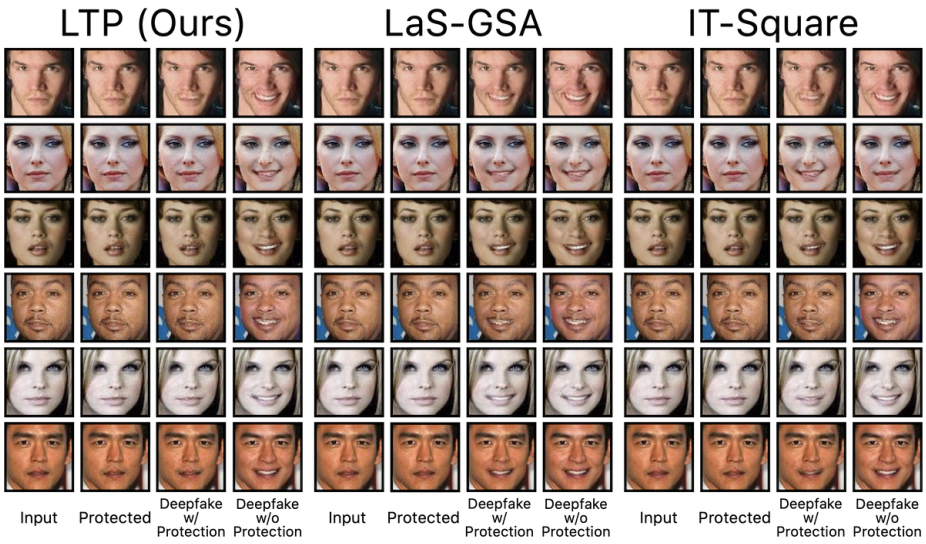

Figure 5: Qualitative examples of our LTP attack on expression editing (smile insertion). We show a *neutralizing attack* where we neutralize the deepfake expression transformation and seek to make the output of the network the same as the input.

and LTP we use the same parameters used in the SimBA paper, and step size of 0.4. We use a maximum number of saturating loss steps $n_{\text{sat}} = 20$ for LTP. For LaS-GSA we use the parameters in the paper with unbounded magnitude.

For the expression editing task we build our leaked PCA components using 100 random images, for each of the three expressions. We attack them using IT-NES with a 0.005 success threshold and 1,000 max iterations. We perform 351.9 queries on average per image. For the attribute editing task we build our PCA components using 10 random images and 5 classes. We attack them using IT-NES with a 0.05 success threshold and 1,000 max iterations. We perform 928.4 queries on average per image.

Experimental Results

In this section we disrupt an expression editing model (GANimation) and an attribute editing model (StarGAN) using IT-NES, IT-Bandits-TD, IT-SimBA, IT-Square, LaS-GSA and LTP. We compare the average number of queries required to successfully attack an image. We also present success rates, FID scores and average perturbation magnitudes. We evaluate over the type of attack (neutralizing or distortion) that is most effective for each architecture.

Expression Editing (GANimation) We attack GANimation using a *neutralizing attack*, where we select the target image $\boldsymbol{r}$ to be the input image $\boldsymbol{x}$, such that the network output is pushed to be the same as the input. We select a success threshold of $\tau = 0.005$, meaning that we halt the attack when $L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}), \boldsymbol{x}) \leq \tau$. At this threshold a successful attack renders the transformations by GANimation unnoticeable. We use a maximum number of queries $\boldsymbol{B} = 10,000$ for all methods. In Table 1 we show comparisons between IT-NES, IT-Bandits-TD, IT-SimBA, IT-Square, LaS-GSA and LTP. LaS-GSA is bounded using the L_∞ norm instead of the L_2 norm that we study in this threat model. In order to realize the comparison, we make the attack unbounded $\boldsymbol{\epsilon} = \infty$, instead of using $\boldsymbol{\epsilon} = 0.1$, which is used in the original formulation (Yeh et al. 2021). We report the L_2 norm of the attack, in order to compare with other attacks.

We observe that LTP is vastly more efficient than competing methods achieving a 56% reduction in average queries (231 *vs.* 531) compared to the next best method (IT-Square).

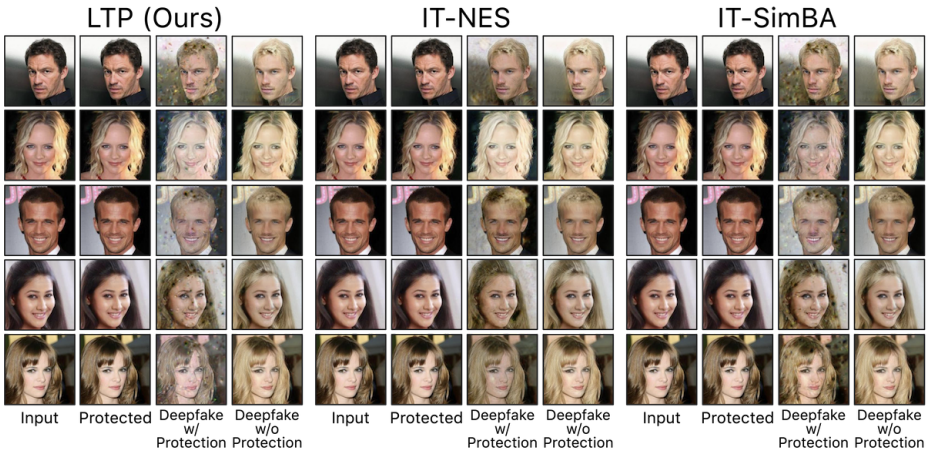

Figure 6: Examples of our LTP attack on an attribute editing task (blond hair transformation). We show a *distortion attack* where we successfully distort the output of the network and make it unusable. We compare to the next two most efficient methods.

Our method also achieves a lower average perturbation norm than the comparable IT-SimBA attack as well as an improved success rate. Additionally, we compute FID scores (Heusel et al. 2017) for the disrupted network outputs, comparing the feature distributions of attacked outputs with that of original images. Thus, FID measures how similar the attacked output images are to the intact inputs. LTP achieves the lowest FID score, reflecting that the images have been preserved to a greater extent under the *neutralizing attack*. We present results for other thresholds in the supp. material.

In Figure 3 (top) we present histograms of the number of queries needed for successful attacks. LTP is heavily skewed to the left and greatly outperforms IT-SimBA and IT-Square. Our method achieves successful attacks using fewer than 20 queries for 50% of the dataset, confirming our intuition that the information collected during the *leaking phase* allows us to build efficient transferable attacks. Figure 4 (left) shows the cumulative histogram of images successfully attacked for the number of queries represented by the x-axis. We observe that LTP achieves superior results than the two next best performing methods.

Finally, we show comparative qualitative results for the "open eye smile" expression transformation in Figure 5. We compare to the two next-best attacks in IT-SimBA and IT-Square, as well as LaS-GSA, the recent black-box image translation attack proposed by Yeh et al. (Yeh et al. 2021). We observe that our attack successfully neutralizes the smile transformation, whereas other attacks have less success when the transformation is very salient.

Facial Attribute Editing (StarGAN) We attack 200 images on StarGAN using 5 different attribute classes. We use a *distortion attack*, where the target image $\boldsymbol{r}$ is the non-attacked output image $\boldsymbol{G}(\boldsymbol{x})$ and we maximize the loss to achieve the maximum amount of distortion in the output image. We present results for a threshold $\tau = 0.05$, where the output image is visibly distorted. We use a maximum number of queries $\boldsymbol{B} = 10,000$ for all methods. In Table 2 we show comparisons between LTP and competing methods. In this case we cannot compare to LaS-GSA directly, since it is formulated uniquely as a neutralizing attack. We can see that LTP is much more efficient than other methods achieving a reduction in mean queries of 70% compared to the next best attack and achieving a 100% success rate. We also compute FID scores for the network outputs and LTP achieves a high FID score, reflecting that the images have highly corrupted using the *distortion* attack. The average norm is also slightly lower than the next-best method (IT-SimBA) and remains imperceptible as seen in Figure 6.

In Figure 3 (bottom) we show histograms of the number of queries required to attack dataset images. We see again that LTP is heavily skewed to the left compared to competing methods. Figure 4 (right) shows the cumulative histogram of images successfully attacked for a specific number of queries, demonstrating the superior efficiency of LTP.

Conclusion

We present successful black-box attacks on image translation models, with an application to disrupting the generation of deepfake images. This is a first step to combating real-world modern deepfake systems that puppeteer faces using few images and no longer rely on face swapping or large sets of face images. Our work also tackles the ever-growing threat of DeepNude applications.

A key limitation of existing attacks is the high number of queries needed. We show that Leaking Transferable Perturbations (LTP) reduces the number of queries necessary to attack models. This is a consequence of the transferability of the leaked PCA components that are subsequently used as candidate vectors during the exploitation phase. We find that image translation architectures have specific vulnerabilities and that there exist correlations between attacks constructed for different images. This is the surprising nugget of intuition that motivates our approach.

Acknowledgments

Cihang Xie is supported by a gift from Open Philanthropy.

References

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, 484–501. Springer.

Aneja, S.; Markhasin, L.; and Niessner, M. 2021. TAFIM: Targeted Adversarial Attacks against Facial Image Manipulations. *arXiv preprint arXiv:2112.09151*.

Bhagoji, A. N.; He, W.; Li, B.; and Song, D. 2018. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 154–169.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE.

Chen, P.-Y.; Zhang, H.; Sharma, Y.; Yi, J.; and Hsieh, C.-J. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 15–26.

Cheng, M.; Le, T.; Chen, P.; Zhang, H.; Yi, J.; and Hsieh, C. 2019. Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2019. StarGAN v2: Diverse Image Synthesis for Multiple Domains. *arXiv preprint arXiv:1912.01865*.

Gandhi, A.; and Jain, S. 2020. Adversarial perturbations fool deepfake detectors. *arXiv preprint arXiv:2003.10596*.

Geng, Z.; Cao, C.; and Tulyakov, S. 2019. 3D guided fine-grained face manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9821–9830.

Ghosh, P.; Gupta, P. S.; Uziel, R.; Ranjan, A.; Black, M.; and Bolkart, T. 2020. Gif: Generative interpretable faces. *arXiv preprint arXiv:2009.00149*.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *Proc. ICLR*.

Güera, D.; and Delp, E. J. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. IEEE.

Guo, C.; Gardner, J. R.; You, Y.; Wilson, A. G.; and Weinberger, K. Q. 2019. Simple Black-box Adversarial Attacks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 2484–2493. PMLR.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 6626–6637.

Huang, Q.; Zhang, J.; Zhou, W.; Zhang, W.; and Yu, N. 2021. Initiative Defense against Facial Manipulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2): 1619–1627.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box Adversarial Attacks with Limited Queries and Information. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2142–2151.

Ilyas, A.; Engstrom, L.; and Madry, A. 2019. Prior Convictions: Black-box Adversarial Attacks with Bandits and Priors. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

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.

Kos, J.; Fischer, I.; and Song, D. 2018. Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, 36–42. IEEE.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Li, Q.; Guo, Y.; and Chen, H. 2020. Practical no-box adversarial attacks against dnns. *Advances in Neural Information Processing Systems*, 33: 12849–12860.

Li, Y.; and Lyu, S. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 46–52.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2016. Delving into Transferable Adversarial Examples and Black-box Attacks. *CoRR*, abs/1611.02770.

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

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Narodytska, N.; and Kasiviswanathan, S. P. 2016. Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299*.

Neekhara, P.; Hussain, S.; Jere, M.; Koushanfar, F.; and McAuley, J. 2020. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. *arXiv preprint arXiv:2002.12749*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436.

Nguyen-Phuoc, T.; Li, C.; Theis, L.; Richardt, C.; and Yang, Y.-L. 2019. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, 7588–7597.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 506–519. ACM.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 372–387. IEEE.

Pumarola, A.; Agudo, A.; Martinez, A. M.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 818–833.

Ranjan, A.; Bolkart, T.; Sanyal, S.; and Black, M. 2018. Generating 3D faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 704–720.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Niessner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *The IEEE International Conference on Computer Vision (ICCV)*.

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

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *In Proc. ICLR*.

Tabacof, P.; Tavares, J.; and Valle, E. 2016. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*.

Tewari, A.; Elgharib, M.; Bharaj, G.; Bernard, F.; Seidel, H.-P.; Pérez, P.; Zollhöfer, M.; and Theobalt, C. 2020. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images. *arXiv preprint arXiv:2004.00121*.

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

Thies, J.; Zollhöfer, M.; Theobalt, C.; Stamminger, M.; and Nießner, M. 2018. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics (TOG)*, 37(4): 1–13.

Tu, C.-C.; Ting, P.; Chen, P.-Y.; Liu, S.; Zhang, H.; Yi, J.; Hsieh, C.-J.; and Cheng, S.-M. 2019. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 742–749.

Usman, B.; Dufour, N.; Saenko, K.; and Bregler, C. 2019. PuppetGAN: Cross-Domain Image Manipulation by Demonstration. In *The IEEE International Conference on Computer Vision (ICCV)*.

Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; and Liu, Y. ???? FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Wiles, O.; Sophia K., A.; and Zisserman, A. 2018. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 670–686.

Xiao, C.; Li, B.; Zhu, J.-Y.; He, W.; Liu, M.; and Song, D. 2018. Generating adversarial examples with adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3905–3911.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8261–8265.

Yeh, C.-Y.; Chen, H.-W.; Shuai, H.-H.; Yang, D.-N.; and Chen, M.-S. 2021. Attack As the Best Defense: Nullifying Image-to-Image Translation GANs via Limit-Aware Adversarial Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16188–16197.

Yeh, C.-Y.; Chen, H.-W.; Tsai, S.-L.; and Wang, S.-D. 2020. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 53–62.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, 9459–9468.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.