Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems

Nataniel Ruiz, Sarah Adel Bargal, Stan Sclaroff

Boston University, Boston, MA, USA
{nruiz9,sbargal,sclaroff}@bu.edu

Abstract. Face modification systems using deep learning have become increasingly powerful and accessible. Given images of a person's face, such systems can generate new images of that same person under different expressions and poses. Some systems can also modify targeted attributes such as hair color or age. This type of manipulated images and video have been coined Deepfakes. In order to prevent a malicious user from generating modified images of a person without their consent we tackle the new problem of generating adversarial attacks against such image translation systems, which disrupt the resulting output image. We call this problem *disrupting deepfakes*. Most image translation architectures are generative models conditioned on an attribute (e.g. put a smile on this person's face). We are first to propose and successfully apply (1) class transferable adversarial attacks that generalize to different classes, which means that the attacker does not need to have knowledge about the conditioning class, and (2) adversarial training for generative adversarial networks (GANs) as a first step towards robust image translation networks. Finally, in gray-box scenarios, blurring can mount a successful defense against disruption. We present a spread-spectrum adversarial attack, which evades blur defenses. Our open-source code can be found at `https://github.com/natanielruiz/disrupting-deepfakes`.

Keywords: adversarial attacks, image translation, face modification, deepfake, generative models, GAN, privacy

1 Introduction

Advances in image translation using generative adversarial networks (GANs) have allowed the rise of face manipulation systems that achieve impressive realism. Some face manipulation systems can create new images of a person's face under different expressions and poses [21,30]. Other face manipulation systems modify the age, hair color, gender or other attributes of the person [6,7].

Given the widespread availability of these systems, malicious actors can modify images of a person without their consent. There have been occasions where faces of celebrities have been transferred to videos with explicit content without their consent [1] and companies such as Facebook have banned uploading modified pictures and video of people [2].

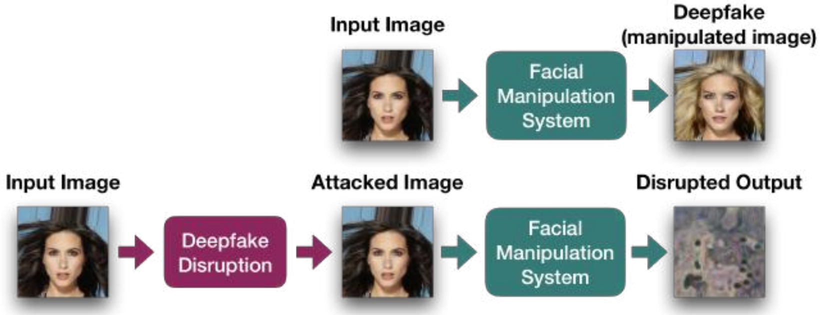
Input Image
Facial Manipulation System
Deepfake (manipulated image)
Input Image
Deepfake Disruption
Attacked Image
Facial Manipulation System
Disrupted Output

Fig. 1: Illustration of deepfake disruption with a real example. After applying an imperceptible filter on the image using our I-FGSM disruption the output of the face manipulation system (StarGAN [6]) is successfully disrupted.

One way of mitigating this risk is to develop systems that can detect whether an image or video has been modified using one of these systems. There have been recent efforts in this direction, with varying levels of success [26, 27].

There is work showing that deep neural networks are vulnerable to adversarial attacks [5, 12, 20, 23], where an attacker applies imperceptible perturbations to an image causing it to be incorrectly classified. We distinguish different attack scenarios. In a *white-box* scenario the attacker has perfect knowledge of the architecture, model parameters and defenses in place. In a *black-box* scenario, the attacker is only able to query the target model for output labels for chosen inputs. There are several different definitions of *gray-box* scenarios. In this work, a gray-box scenario denotes perfect knowledge of the model and parameters, but ignorance of the pre-processing defense mechanisms in place (such as blurring). We focus on white-box and gray-box settings.

Another way of combating malicious actors is by *disrupting the deepfaker's ability to generate a deepfake*. In this work we propose a solution by adapting traditional adversarial attacks that are imperceptible to the human eye in the source image, but interfere with translation of this image using image translation networks. A successful disruption corresponds to the generated image being sufficiently deteriorated such that it has to be discarded or such that the modification is perceptually evident. We present a formal and quantifiable definition of disruption success in Section 3.

Most facial manipulation architectures are conditioned both on the input image and on a target conditioning class. One example, is to define the target expression of the generated face using this attribute class (e.g. put a smile on the person's face). In this example, if we want to prevent a malicious actor from putting a smile on the person's face in the image, we need to know that the malicious actor has selected the smile attribute instead of, for instance, eye closing. In this work, we are first to formalize the problem of disrupting class conditional image translation, and present two variants of class transferable disruptions that improve generalization to different conditioning attributes.

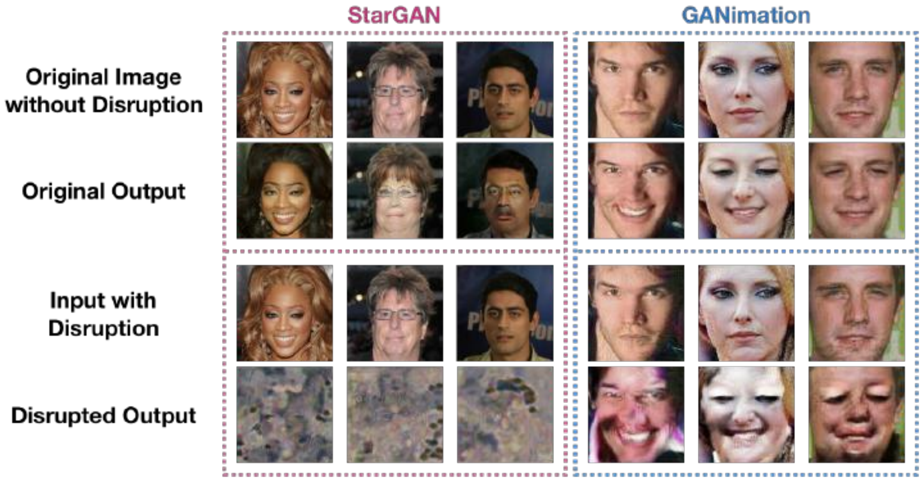

Fig. 2: An example of our deepfake disruptions on StarGAN [6] and GANimation [21]. Some image translation networks are more prone to disruption.

Blurring is a broken defense in the white-box scenario, where a disruptor knows the type and magnitude of pre-processing blur being used. Nevertheless, in a real situation, a disruptor might know the architecture being used yet ignore the type and magnitude of blur being used. In this scenario the efficacy of a naive disruption drops dramatically. We present a novel spread-spectrum disruption that evades a variety of blur defenses in this gray-box setting.

In summary:

- We present baseline methods for disrupting deepfakes by adapting adversarial attack methods to image translation networks.
- We are the first to address disruptions on conditional image translation networks. We propose and evaluate novel disruption methods that transfer from one conditioning class to another.
- We are the first to propose and evaluate adversarial training for generative adversarial networks. Our novel *G+D adversarial training* alleviates disruptions in a white-box setting.
- We propose a novel spread-spectrum disruption that evades blur defenses in a gray-box scenario.

2 Related Work

There are several works exploring image translation using deep neural networks [6, 7, 13, 21, 28, 30, 31]. Some of these works apply image translation to face images in order to generate new images of individuals with modified expression or attributes [6, 7, 21, 30].

There is a large amount of work that explores adversarial attacks on deep neural networks for classification [5, 12, 17–20, 23]. Fast Gradient Sign Method (FGSM), a one-step gradient attack was proposed by Goodfellow *et al.* [12]. Stronger iterative attacks such as iterative FGSM (I-FGSM) [15] and Projected Gradient Descent (PGD) [16] have been proposed. Sabour *et al.* [22] explore feature-space attacks on deep neural network classifiers using L-BFGS.

Tabacof *et al.* [24] and Kos *et al.* [14] explore adversarial attacks against Variational Autoencoders (VAE) and VAE-GANs, where an adversarial image is compressed into a latent space and instead of being reconstructed into the original image is reconstructed into an image of a different semantic class. In contrast, our work focuses on attacks against image translation systems. Additionally, our objective is to disrupt deepfake generation as opposed to changing the output image to a different semantic class.

Chu et al. [8] show that information hiding occurs during CycleGAN training and is similar in spirit to an adversarial attack [23]. Bashkirova et al. [4] explore self-adversarial attacks in cycle-consistent image translation networks. [4] proposes two methods for defending against such attacks leading to more honest translations by attenuating the self-adversarial hidden embedding. While [4] addresses self-adversarial attacks [8], which lead to higher translation quality, our work addresses adversarial attacks [23], which seek to disrupt the image translation process.

Wang *et al.* [25] adapt adversarial attacks to the image translation scenario for traffic scenes on the pix2pixHD and CycleGAN networks. Yeh *et al.* [29], is concurrent work to ours, and proposes adapting PGD to attack pix2pixHD and CycleGAN networks in the face domain. Most face manipulation networks are conditional image translation networks, [25, 29] do not address this scenario and do not explore defenses for such attacks. We are the first to explore attacks against conditional image translation GANs as well as attacks that transfer to different conditioning classes. We are also the first to propose adversarial training [16] for image translation GANs. Madry *et al.* [16] propose adversarial training using strong adversaries to alleviate adversarial attacks against deep neural network classifiers. In this work, we propose two adaptations of this technique for GANs, as a first step towards robust image translation networks.

A version of spread-spectrum watermarking for images was proposed by Cox *et al.* [10]. Athalye *et al.* [3] proposes the expectation over transformation (EoT) method for synthesizing adversarial examples robust to pre-processing transformations. However, Athalye *et al.* [3] demonstrate their method on affine transformations, noise and others, but do not consider blur. In this work, we propose a faster heuristic iterative spread-spectrum disruption for evading blur defenses.

3 Method

We describe methods for image translation disruption (Section 3.1), our proposed conditional image translation disruption techniques (Section 3.2), our proposed adversarial training techniques for GANs (Section 3.3) and our proposed spread-spectrum disruption (Section 3.4).

3.1 Image Translation Disruption

Similar to an adversarial example, we want to generate a disruption by adding a human-imperceptible perturbation $\boldsymbol{\eta}$ to the input image:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\eta}, \tag{1}$$

where $\tilde{\boldsymbol{x}}$ is the generated disrupted input image and $\boldsymbol{x}$ is the input image. By feeding the original image or the disrupted input image to a generator we have the mappings $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{y}$ and $\boldsymbol{G}(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{y}}$, respectively, where $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$ are the translated output images and $\boldsymbol{G}$ is the generator of the image translation GAN.

We consider a disruption successful when it introduces perceptible corruptions or modifications onto the output $\tilde{\boldsymbol{y}}$ of the network leading a human observer to notice that the image has been altered and therefore distrust its source.

We operationalize this phenomenon. Adversarial attack research has focused on attacks showing low distortions using the L^0, L^2 and L^∞ distance metrics. The logic behind using attacks with low distortion is that the larger the distance, the more apparent the alteration of the image, such that an observer could detect it. In contrast, we seek to *maximize* the distortion of our output, with respect to a well-chosen reference $\boldsymbol{r}$.

$$\max_{\boldsymbol{\eta}} L(\boldsymbol{G}(\boldsymbol{x} + \boldsymbol{\eta}), \boldsymbol{r}), \quad \text{subject to } ||\boldsymbol{\eta}||_\infty \leq \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon}$ is the maximum magnitude of the perturbation and L is a distance function. If we pick r to be the ground-truth output, $\boldsymbol{r} = \boldsymbol{G}(\boldsymbol{x})$, we get the *ideal* disruption which aims to maximize the distortion of the output.

We can also formulate a *targeted* disruption, which pushes the output $\tilde{\boldsymbol{y}}$ to be close to $\boldsymbol{r}$:

$$\boldsymbol{\eta} = \arg\min_{\boldsymbol{\eta}} L(\boldsymbol{G}(\boldsymbol{x} + \boldsymbol{\eta}), \boldsymbol{r}), \quad \text{subject to } ||\boldsymbol{\eta}||_\infty \leq \boldsymbol{\epsilon}. \tag{3}$$

Note that the ideal disruption is a special case of the targeted disruption where we minimize the negative distortion instead and select $\boldsymbol{r} = \boldsymbol{G}(\boldsymbol{x})$. We can thus disrupt an image *towards* a target or *away from* a target.

We can generate a targeted disruption by adapting well-established adversarial attacks: FGSM, I-FGSM, and PGD. Fast Gradient Sign Method (FGSM) [12] generates an attack in one forward-backward step, and is adapted as follows:

$$\boldsymbol{\eta} = \boldsymbol{\epsilon} \ \text{sign}[\nabla_{\boldsymbol{x}} L(\boldsymbol{G}(\boldsymbol{x}), \boldsymbol{r})], \tag{4}$$

where $\boldsymbol{\epsilon}$ is the size of the FGSM step. Iterative Fast Gradient Sign Method (I-FGSM) [15] generates a stronger adversarial attack in multiple forward-backward steps. We adapt this method for the targeted disruption scenario as follows:

$$\tilde{\boldsymbol{x}}_t = \text{clip}(\tilde{\boldsymbol{x}}_{t-1} - \boldsymbol{a} \ \text{sign}[\nabla_{\tilde{\boldsymbol{x}}} L(\boldsymbol{G}(\tilde{\boldsymbol{x}}_{t-1}), \boldsymbol{r})]), \tag{5}$$

where $\boldsymbol{a}$ is the step size and the constraint $||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \boldsymbol{\epsilon}$ is enforced by the clip function. For disruptions *away from* the target $\boldsymbol{r}$ instead of *towards* $\boldsymbol{r}$, using the negative gradient of the loss in the equations above is sufficient. For an adapted Projected Gradient Descent (PGD) [16], we initialize the disrupted image $\tilde{\boldsymbol{x}}_0$ randomly inside the $\boldsymbol{\epsilon}$-ball around $\boldsymbol{x}$ and use the I-FGSM update function.

3.2 Conditional Image Translation Disruption

Many image translation systems are conditioned not only on the input image, but on a target class as well:

$$\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{c}), \tag{6}$$

where $\boldsymbol{x}$ is the input image, $\boldsymbol{c}$ is the target class and $\boldsymbol{y}$ is the output image. A target class can be an attribute of a dataset, for example blond or brown-haired.

A disruption for the data/class pair $(\boldsymbol{x}, \boldsymbol{c}_i)$ is not guaranteed to transfer to the data/class pair $(\boldsymbol{x}, \boldsymbol{c}_j)$ when $i \neq j$. We can define the problem of looking for a class transferable disruption as follows:

$$\boldsymbol{\eta} = \arg\min_{\boldsymbol{\eta}} \mathbb{E}_{\boldsymbol{c}}[L(\boldsymbol{G}(\boldsymbol{x} + \boldsymbol{\eta}, \boldsymbol{c}), \boldsymbol{r})], \quad \text{subject to } ||\boldsymbol{\eta}||_\infty \leq \boldsymbol{\epsilon}. \tag{7}$$

We can write this empirically as an optimization problem:

$$\boldsymbol{\eta} = \arg\min_{\boldsymbol{\eta}} \sum_{\boldsymbol{c}} [L(\boldsymbol{G}(\boldsymbol{x} + \boldsymbol{\eta}, \boldsymbol{c}), \boldsymbol{r})], \quad \text{subject to } ||\boldsymbol{\eta}||_\infty \leq \boldsymbol{\epsilon}. \tag{8}$$

Iterative Class Transferable Disruption In order to solve this problem, we present a novel disruption on class conditional image translation systems that increases the transferability of our disruption to different classes. We perform a modified I-FGSM disruption:

$$\tilde{\boldsymbol{x}}_t = \text{clip}(\tilde{\boldsymbol{x}}_{t-1} - \boldsymbol{a}\ \text{sign}[\nabla_{\tilde{\boldsymbol{x}}} L(\boldsymbol{G}(\tilde{\boldsymbol{x}}_{t-1}, \boldsymbol{c}_k), \boldsymbol{r})]). \tag{9}$$

We initialize $k = 1$ and increment k at every iteration, until we reach $k = K$ where K is the number of classes. We then reset $k = 1$.

Joint Class Transferable Disruption We propose a disruption which seeks to minimize the expected value of the distance to the target $\boldsymbol{r}$ at every step t. For this, we compute this loss term at every step of an I-FGSM disruption and use it to inform our update step:

$$\tilde{\boldsymbol{x}}_t = \text{clip}(\tilde{\boldsymbol{x}}_{t-1} - \boldsymbol{a}\ \text{sign}[\nabla_{\tilde{\boldsymbol{x}}} \sum_{\boldsymbol{c}} L(\boldsymbol{G}(\tilde{\boldsymbol{x}}_{t-1}, \boldsymbol{c}), \boldsymbol{r})]). \tag{10}$$

3.3 GAN Adversarial Training

Adversarial training for classifier deep neural networks was proposed by Madry *et al.* [16]. It incorporates strong PGD attacks on the training data for the classifier. We propose the first adaptations of adversarial training for generative adversarial networks. Our methods, described below, are a first step in attempting to defend against image translation disruption.

Generator Adversarial Training A conditional image translation GAN uses the following adversarial loss:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}}\left[\log \boldsymbol{D}(\boldsymbol{x})\right] \; + \mathbb{E}_{\boldsymbol{x},\boldsymbol{c}}[\log\left(1 - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{x},\boldsymbol{c}))\right)], \tag{11}$$

where $\boldsymbol{D}$ is the discriminator. In order to make the generator resistant to adversarial examples, we train the GAN using the modified loss:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}}\left[\log \boldsymbol{D}(\boldsymbol{x})\right] \; + \mathbb{E}_{\boldsymbol{x},\boldsymbol{c},\boldsymbol{\eta}}[\log\left(1 - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta},\boldsymbol{c}))\right)]. \tag{12}$$

Generator+Discriminator (G+D) Adversarial Training Instead of only training the generator to be indifferent to adversarial examples, we also train the discriminator on adversarial examples:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\eta}_1}\left[\log \boldsymbol{D}(\boldsymbol{x}+\boldsymbol{\eta}_1)\right] \; + \mathbb{E}_{\boldsymbol{x},\boldsymbol{c},\boldsymbol{\eta}_2,\boldsymbol{\eta}_3}[\log\left(1 - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}_2,\boldsymbol{c})+\boldsymbol{\eta}_3)\right)]. \tag{13}$$

3.4 Spread-Spectrum Evasion of Blur Defenses

Blurring can be an effective test-time defense against disruptions in a gray-box scenario, where the disruptor ignores the type or magnitude of blur being used. In order to successfully disrupt a network in this scenario, we propose a spread-spectrum evasion of blur defenses that transfers to different types of blur. We perform a modified I-FGSM update

$$\tilde{\boldsymbol{x}}_t = \text{clip}(\tilde{\boldsymbol{x}}_{t-1} - \boldsymbol{\epsilon}\ \text{sign}[\nabla_{\tilde{\boldsymbol{x}}} L(\boldsymbol{f}_k(\boldsymbol{G}(\tilde{\boldsymbol{x}}_{t-1})), \boldsymbol{r})]), \tag{14}$$

where $\boldsymbol{f}_k$ is a blurring convolution operation, and we have K different blurring methods with different magnitudes and types. We initialize $k = 1$ and increment k at every iteration of the algorithm, until we reach $k = K$ where K is the total number of blur types and magnitudes. We then reset $k = 1$.

4 Experiments

In this section we demonstrate that our proposed image-level FGSM, I-FGSM and PGD-based disruptions are able to disrupt different recent image translation architectures such as GANimation [21], StarGAN [6], pix2pixHD [28] and CycleGAN [31]. In Section 4.1, we show that the ideal formulation of an image-level disruption presented in Section 3.1, is the most effective at producing large distortions in the output. In Section 4.2, we demonstrate that both our *iterative class transferable disruption* and *joint class transferable disruption* are able to transfer to different conditioning classes. In Section 4.3, we test our disruptions against two defenses in a white-box setting. We show that our proposed *G+D adversarial training* is most effective at alleviating disruptions, although strong disruptions are able to overcome this defense. Finally, in Section 4.4 we show that blurring is an effective defense against disruptions in a gray-box setting, in which the disruptor does not know the type or magnitude of the pre-processing blur. We then demonstrate that our proposed *spread-spectrum adversarial disruption* evades different blur defenses in this scenario. All disruptions in our experiments use $L = L^2$.

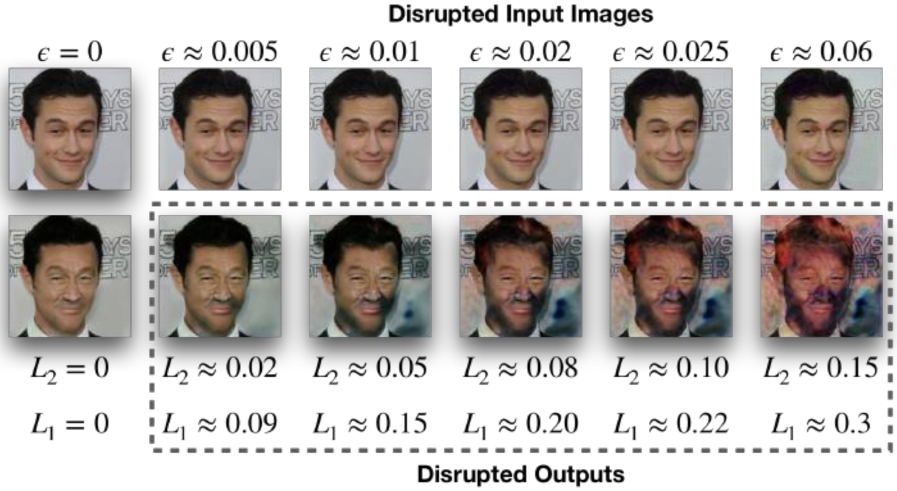

Fig. 3: Equivalence scale between L_2 and L_1 distances and qualitative distortion on disrupted StarGAN images. We also show the original image and output with no disruption. Images with $L_2 \geq 0.05$ have very noticeable distortions.

Architectures and Datasets We use the GANimation [21], StarGAN [7], pix2pixHD [28] and CycleGAN [31] image translation architectures. We use an open-source implementation of GANimation trained for 37 epochs on the CelebA dataset for 80 action units (AU) from the Facial Action Unit Coding System (FACS) [11]. We test GANimation on 50 random images from the CelebA dataset (4,000 disruptions). We use the official open-source implementation of StarGAN, trained on the CelebA dataset for the five attributes black hair, blond hair, brown hair, gender and aged. We test StarGAN on 50 random images from the CelebA dataset (250 disruptions). For pix2pixHD we use the official open-source implementation, which was trained for label-to-street view translation on the Cityscapes dataset [9]. We test pix2pixHD on 50 random images from the Cityscapes test set. For CycleGAN we use the official open-source implementation for both the zebra-to-horses and photograph-to-Monet painting translations. We disrupt 100 images from both datasets. We use the pre-trained models provided in the open-source implementations, unless specifically noted.

4.1 Image Translation Disruption

Success Scenario In order to develop intuition on the relationship between our main L^2 and L^1 distortion metrics and the qualitative distortion caused on image translations, we display in Fig. 3 a scale that shows qualitative examples of disrupted outputs and their respective distortion metrics. We can see that when the L^2 and L^1 metric becomes larger than 0.05 we have very noticeable distortions in the output images. Throughout the experiments section, we report

Table 1: Comparison of L^1 and L^2 pixel-wise errors, as well as the percentage of disrupted images (% dis.) for different disruption methods on different facial manipulation architectures and datasets. All disruptions use $\boldsymbol{\epsilon} = 0.05$ unless noted. We notice that strong disruptions are successful on all tested architectures.

	FGSM			I-FGSM			PGD		
Architecture (Dataset)	L^1	L^2	% dis.	L^1	L^2	% dis.	L^1	L^2	% dis.
StarGAN (CelebA)	0.462	0.332	100%	1.134	1.525	100%	1.119	1.479	100%
GANimation (CelebA)	0.090	0.017	0%	0.142	0.046	34.9%	0.139	0.044	30.4%
GANimation (CelebA, $\boldsymbol{\epsilon} = 0.1$)	0.121	0.024	1.5%	0.212	0.098	93.9%	0.190	0.077	83.7%
pix2pixHD (Cityscapes)	0.240	0.118	96%	0.935	1.110	100%	0.922	1.084	100%
CycleGAN (Horse)	0.133	0.040	21%	0.385	0.242	100%	0.402	0.253	100%
CycleGAN (Monet)	0.155	0.039	22%	0.817	0.802	100%	0.881	0.898	100%

the percentage of successfully disrupted images (% dis.), which correspond to the percentage of outputs presenting a distortion $L^2 \geq 0.05$.

Vulnerable Image Translation Architectures We show that we are able to disrupt the StarGAN, pix2pixHD and CycleGAN architectures with very successful results using either I-FGSM or PGD in Table 1. Our white-box disruptions are effective on several recent image translation architectures and several different translation domains. GANimation reveals itself to be more robust to disruptions of magnitude $\boldsymbol{\epsilon} = 0.05$ than StarGAN, although it can be successfully disrupted with stronger disruptions ($\boldsymbol{\epsilon} = 0.1$). The metrics reported in Table 1 are the average of the L^1 and L^2 errors on all dataset samples, where we compute the error for each sample by comparing the ground-truth output $\boldsymbol{G}(\boldsymbol{x})$ with the disrupted output $\boldsymbol{G}(\tilde{\boldsymbol{x}})$, using the following formulas $L^1 = ||\boldsymbol{G}(\tilde{\boldsymbol{x}}) - \boldsymbol{G}(\boldsymbol{x})||_1$ and $L^2 = ||\boldsymbol{G}(\tilde{\boldsymbol{x}}) - \boldsymbol{G}(\boldsymbol{x})||_2$. For I-FGSM and PGD we use 20 steps with step size of 0.01. We use our ideal formulation for all disruptions.

We show examples of successfully disrupted image translations on GANimation and StarGAN in Fig. 2 using I-FGSM. We observe different qualitative behaviors for disruptions on different architectures. Nevertheless, all of our disruptions successfully make the modifications in the image obvious for any observer, thus avoiding any type of undetected manipulation of an image.

Ideal Disruption In Section 3.1, we derived an ideal disruption for our success metric. In order to execute this disruption we first need to obtain the ground-truth output of the image translation network $\boldsymbol{G}(\boldsymbol{x})$ for the image $\boldsymbol{x}$ being disrupted. We push the disrupted output $\boldsymbol{G}(\tilde{\boldsymbol{x}})$ to be maximally different from $\boldsymbol{G}(\boldsymbol{x})$. We compare this ideal disruption (designated as *Away From* Output in Table 2) to targeted disruptions with different targets such as a black image, a white image and random noise. We also compare it to a less computationally intensive disruption called *Away From* Input, which seeks to maximize the distortion between our disrupted output $\boldsymbol{G}(\tilde{\boldsymbol{x}})$ and our original input $\boldsymbol{x}$.

Table 2: Comparison of efficacy of FGSM, I-FGSM and PGD methods with different disruption targets for the StarGAN generator and the CelebA dataset.

	FGSM		I-FGSM		PGD	
Target	L^1	L^2	L^1	L^2	L^1	L^2
Towards Black	0.494	0.336	0.494	0.335	0.465	0.304
Towards White	0.471	0.362	0.711	0.694	0.699	0.666
Towards Random Noise	**0.509**	**0.409**	0.607	0.532	0.594	0.511
Away From Input	0.449	0.319	1.086	1.444	1.054	1.354
Away From Output	0.465	0.335	**1.156**	**1.574**	**1.119**	**1.480**

Table 3: Comparison of our image-level PGD disruption to an adapted feature-level disruption from Kos *et al.* [14] on the StarGAN architecture.

	Kos *et al.* [14]											
Layer	1	2	3	4	5	6	7	8	9	10	11	Ours
L^1	0.367	0.406	0.583	0.671	0.661	0.622	0.573	0.554	0.512	0.489	0.778	**1.066**
L^2	0.218	0.269	0.503	0.656	0.621	0.558	0.478	0.443	0.384	0.331	0.817	**1.365**

We display the results for the StarGAN architecture on the CelebA dataset in Table 2. As expected, the *Away From* Output disruption is the most effective using I-FGSM and PGD. All disruptions show similar effectiveness when using one-step FGSM. *Away From* Input seems similarly effective to the *Away From* Output for I-FGSM and PGD, yet it does not have to compute $\boldsymbol{G}(\boldsymbol{x})$, thus saving one forward pass of the generator.

Finally, we show in Table 3 comparisons of our image-level *Away From* Output disruption to the feature-level attack for Variational Autoencoders (VAE) presented in Kos *et al.* [14]. Although in Kos *et al.* [14] attacks are only targeted on the latent vector of a VAE, here we attack every possible intermediate feature map of the image translation network using this attack. The other two attacks presented in Kos *et al.* [14] cannot be applied to the image-translation scenario. We disrupt the StarGAN architecture on the CelebA dataset. Both disruptions use the 10-step PGD optimization formulation with $\boldsymbol{\epsilon} = 0.05$. We notice that while both disruptions are successful, our image-level formulation obtains stronger distortions on average.

4.2 Class Transferable Adversarial Disruption

Class Conditional Image Translation Systems such as GANimation and StarGAN are conditional GANs. Both are conditioned on an input image. Additionally, GANimation is conditioned on the target AU intensities and StarGAN is conditioned on a target attribute. As the disruptor we *do know* which image the malicious actor wants to modify (our image), and in some scenarios we *might know* the architecture and weights that they are using (white-box disruption),

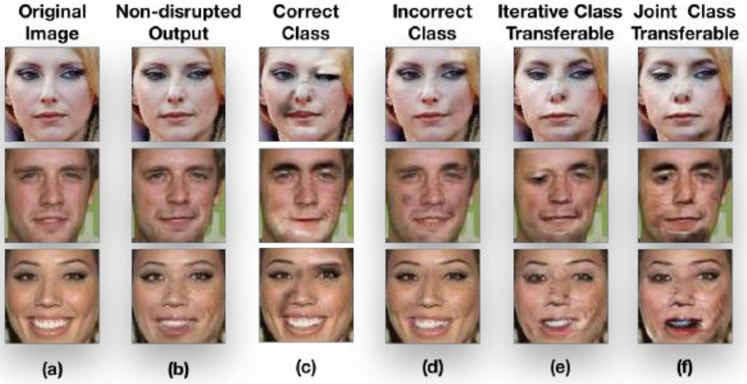

Fig. 4: Examples of our class transferable disruptions. (a) Input image. (b) The ground truth GANimation output without disruption. (c) A disruption using the correct action unit correctly is successful. (d) A disruption with a incorrect target AU is not successful. (e) Our iterative class transferable disruption and (f) joint class transferable disruption are able to transfer across different action units and successfully disrupt the deepfake generation.

yet in almost all cases we *do not know* whether they want to put a smile on the person's face or close their eyes, for example. Since this non-perfect knowledge scenario is probable, we want a disruption that transfers to all of the classes in a class conditional image translation network.

In our experiments we have noticed that attention-driven face manipulation systems such as GANimation present an issue with class transfer. GANimation generates a color mask as well as an attention mask designating the parts of the image that should be replaced with the color mask.

In Fig. 4, we present qualitative examples of our proposed iterative class transferable disruption and joint class transferable disruption. The goal of these disruptions is to transfer to all action unit inputs for GANimation. We compare this to the unsuccessful disruption transfer case where the disruption is targeted to the incorrect AU. Columns (e) and (f) of Fig. 4 show our iterative class transferable disruption and our joint class transferable disruption successfully disrupting the deepfakes, whereas attempting to disrupt the system using the incorrect AU is not effective (column (c)).

Quantitative results demonstrating the superiority of our proposed methods can be found in Table 4. For our disruptions, we use 80 iterations of PGD, magnitude $\boldsymbol{\epsilon} = 0.05$ and a step of 0.01.

For our second experiment, presented in Table 5, instead of disrupting the input image such that the output is visibly distorted, we disrupt the input image such that the output is the identity. In other words, we want the input image to be untouched by the image translation network. We use 80 iterations of I-FGSM, magnitude $\boldsymbol{\epsilon} = 0.05$ and a step of 0.005.

Table 4: Class transferability results for our proposed disruptions. This disruption seeks maximal disruption in the output image. We present the distance between the ground-truth non-disrupted output and the disrupted output images, *higher distance* is better.

	L^1	L^2	% dis.
Incorrect Class	0.144	0.053	45.7%
Iterative Class Transferable	**0.171**	**0.075**	**75.6**%
Joint Class Transferable	0.157	0.062	53.8%
Correct Class	0.166	0.071	68.7%

Table 5: Class transferability results for our proposed disruptions. This disruption seeks minimal change in the input image. We present the distance between the input and output images, *lower distance* is better.

	L^1	L^2
Incorrect Class	1.69×10^{-3}	3.09×10^{-4}
Iterative Class Transferable	6.07×10^{-4}	8.02×10^{-5}
Joint Class Transferable	3.86×10^{-4}	$\mathbf{1.67 \times 10^{-5}}$
Correct Class	$\mathbf{9.88 \times 10^{-5}}$	4.73×10^{-5}
No Disruption	9.10×10^{-2}	2.15×10^{-2}

4.3 GAN Adversarial Training and Other Defenses

We present results for our *generator adversarial training* and *G+D adversarial training* proposed in Section 3.3. In Table 6, we can see that *generator adversarial training* is somewhat effective at alleviating a strong 10-step PGD disruption. *G+D adversarial training* proves to be even more effective than *generator adversarial training.*

Additionally, in the same Table 6, we present results for a Gaussian blur test-time defense ($\sigma = 1.5$). We disrupt this blur defense in a white-box manner. With perfect knowledge of the pre-processing, we can simply backpropagate through that step and obtain a disruption. We achieve the biggest resistance to disruption by combining blurring and *G+D adversarial training*, although strong PGD disruptions are still relatively successful. Nevertheless, this is a first step towards robust image translation networks.

We use a 10-step PGD ($\boldsymbol{\epsilon} = 0.1$) for both *generator adversarial training* and *G+D adversarial training.* We trained StarGAN for $50,000$ iterations using a batch size of 14. We use an FGSM disruption $\boldsymbol{\epsilon} = 0.05$, a 10-step I-FGSM disruption $\boldsymbol{\epsilon} = 0.05$ with step size 0.01 and a 10-step PGD disruption $\boldsymbol{\epsilon} = 0.05$ with step size 0.01.

4.4 Spread-Spectrum Evasion of Blur Defenses

Blurring can be an effective defense against our adversarial disruptions in a gray-box setting where the disruptor does not know the type and magnitude of

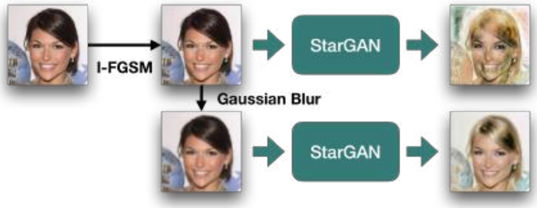

Fig. 5: An example of a successful Gaussian blur defense on a disruption.

Table 6: Image translation disruptions on StarGAN with different defenses.

	FGSM			I-FGSM			PGD		
Defense	L^1	L^2	% dis.	L^1	L^2	% dis.	L^1	L^2	% dis.
No Defense	0.489	0.377	100	0.877	1.011	100	0.863	0.981	100
Blur	0.160	0.048	37.6	0.285	0.138	89.6	0.279	0.133	89.2
Adv. G. Training	0.125	0.032	15.6	0.317	0.183	96	0.319	0.186	95.2
Adv. G+D Training	0.141	0.036	17.2	0.283	0.138	87.6	0.281	0.136	87.6
Adv. G. Train. + Blur	0.138	0.039	21.6	0.225	0.100	63.2	0.224	0.099	61.2
Adv. G+D Train. + Blur	**0.116**	**0.026**	**10.4**	**0.184**	**0.062**	**36.8**	**0.184**	**0.062**	**37.2**

blurring being used for pre-processing. In particular, low magnitude blurring can render a disruption useless while preserving the quality of the image translation output. We show an example on the StarGAN architecture in Fig. 5.

If the image manipulator is using blur to deter adversarial disruptions, the adversary might not know what type and magnitude of blur are being used. In this Section, we evaluate our proposed *spread-spectrum adversarial disruption* which seeks to evade blur defenses in a gray-box scenario, with high transferability between types and magnitudes of blur. In Fig. 6 we present the proportion of test images successfully disrupted ($L^2 \geq 0.05$) for our spread-spectrum method, a white-box perfect knowledge disruption, an adaptation of EoT [3] to the blur scenario and a disruption which does not use any evasion method. We notice that both our method and EoT defeat diverse magnitudes and types of blur and achieve relatively similar performance. Our method achieves better performance on the Gaussian blur scenarios with high magnitude of blur, whereas EoT outperforms our method on the box blur cases, on average. Our iterative spread-spectrum method is roughly K times faster than EoT since it only has to perform one forward-backward pass per iteration of I-FGSM instead of K to compute the loss. Additionally, in Fig. 7, we present random qualitative samples, which show the effectiveness of our method over a naive disruption.

5 Conclusion

In this paper we presented a novel approach to defend against image translation-based deepfake generation. Instead of trying to detect whether an image has

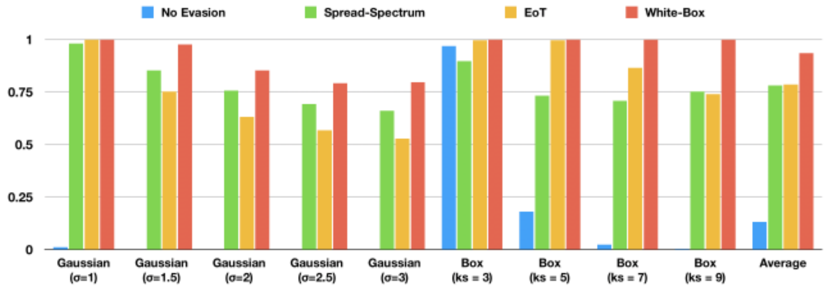

Fig. 6: Proportion of disrupted images ($L^2 \geq 0.05$) for different blur evasions under different blur defenses.

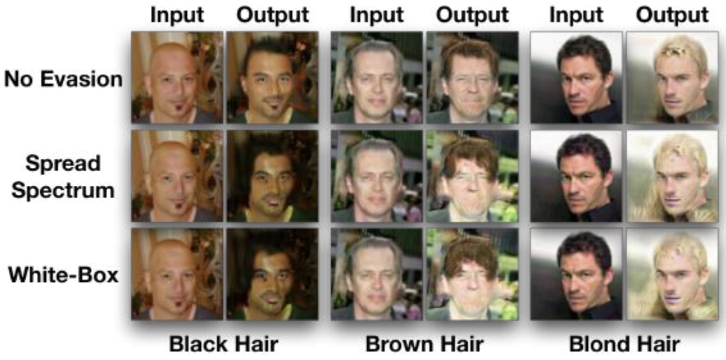

Fig. 7: An example of our spread-spectrum evasion of blur defenses for a Gaussian blur ($\sigma = 1.5$). The first row shows a naive disruption, the second row shows our spread-spectrum evasion and the last row shows a white-box disruption.

been modified after the fact, we defend against the non-authorized manipulation by disrupting conditional image translation facial manipulation networks using adapted adversarial attacks.

We operationalized our definition of a successful disruption, which allowed us to formulate an ideal disruption that can be undertaken using traditional adversarial attack methods such as FGSM, I-FGSM and PGD. We demonstrated that this disruption is superior to other alternatives. Since many face modification networks are conditioned on a target attribute, we proposed two disruptions which transfer from one attribute to another and showed their effectiveness over naive disruptions. In addition, we proposed adversarial training for GANs, which is a first step towards image translation networks that are resistant to disruption. Finally, blurring is an effective defense against naive disruptions in a gray-box scenario and can allow a malicious actor to bypass the disruption and modify the image. We presented a spread-spectrum disruption which evades a wide range of blur defenses.

References

1. Deepfakes, revenge porn, and the impact on women. https://www.forbes.com/sites/chenxiwang/2019/11/01/deepfakes-revenge-porn-and-the-impact-on-women/#7dfb95bf1f53, accessed: 2019-12-10
2. Facebook to ban 'deepfakes'. https://www.bbc.com/news/technology-51018758, accessed: 2020-1-10
3. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: International Conference on Machine Learning. pp. 284–293 (2018)
4. Bashkirova, D., Usman, B., Saenko, K.: Adversarial self-defense for cycle-consistent gans. In: Advances in Neural Information Processing Systems. pp. 635–645 (2019)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57. IEEE (2017)
6. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
7. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. arXiv preprint arXiv:1912.01865 (2019)
8. Chu, C., Zhmoginov, A., Sandler, M.: Cyclegan, a master of steganography. arXiv preprint arXiv:1712.02950 (2017)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
10. Cox, I.J., Kilian, J., Leighton, F.T., Shamoon, T.: Secure spread spectrum watermarking for multimedia. IEEE transactions on image processing **6**(12), 1673–1687 (1997)
11. Ekman, R.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA (1997)
12. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015), http://arxiv.org/abs/1412.6572
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
14. Kos, J., Fischer, I., Song, D.: Adversarial examples for generative models. In: 2018 IEEE Security and Privacy Workshops (SPW). pp. 36–42. IEEE (2018)
15. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
16. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rJzIBfZAb
17. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
18. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)

19. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
20. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P). pp. 372–387. IEEE (2016)
21. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 818–833 (2018)
22. Sabour, S., Cao, Y., Faghri, F., Fleet, D.J.: Adversarial manipulation of deep representations. arXiv preprint arXiv:1511.05122 (2015)
23. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
24. Tabacof, P., Tavares, J., Valle, E.: Adversarial images for variational autoencoders. arXiv preprint arXiv:1612.00155 (2016)
25. Wang, L., Cho, W., Yoon, K.J.: Deceiving image-to-image translation networks for autonomous driving with adversarial perturbations. IEEE Robotics and Automation Letters **PP**, 1–1 (01 2020). https://doi.org/10.1109/LRA.2020.2967289
26. Wang, R., Ma, L., Juefei-Xu, F., Xie, X., Wang, J., Liu, Y.: Fakespotter: A simple baseline for spotting AI-synthesized fake faces. arXiv preprint arXiv:1909.06122 (2019)
27. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot...for now. In: CVPR (2020)
28. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
29. Yeh, C.Y., Chen, H.W., Tsai, S.L., Wang, S.D.: Disrupting image-translation-based deepfake algorithms with adversarial attacks. In: The IEEE Winter Conference on Applications of Computer Vision (WACV) Workshops (March 2020)
30. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9459–9468 (2019)
31. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)