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

Nataniel Ruiz Sarah Adel Bargal Stan Sclaroff
Boston University
{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.

1. Problem Definition

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 [15, 23]. Other face manipulation systems modify the age, hair color, gender or other attributes of the person [4, 5].

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 and companies such as Facebook have banned uploading modified pictures and video of people [1].

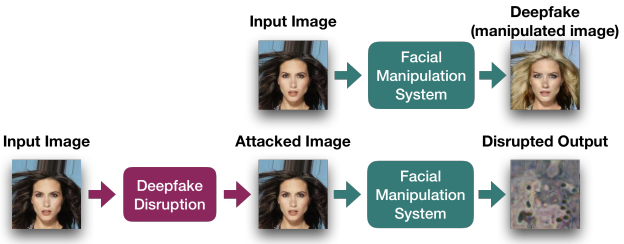

Figure 1. **Deepfake disruption:** a real example. After applying an imperceptible filter on the image using our disruption, the output of the face manipulation system (StarGAN [4]) is successfully disrupted.

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 [19, 20].

There is work showing that classifier deep neural networks are vulnerable to adversarial attacks [16, 7, 12, 14, 3, 13, 11], where an attacker applies imperceptible perturbations to an image causing it to be incorrectly classified. Generative neural networks are also susceptible to attacks [17, 8, 6, 2]. 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). In this work, 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

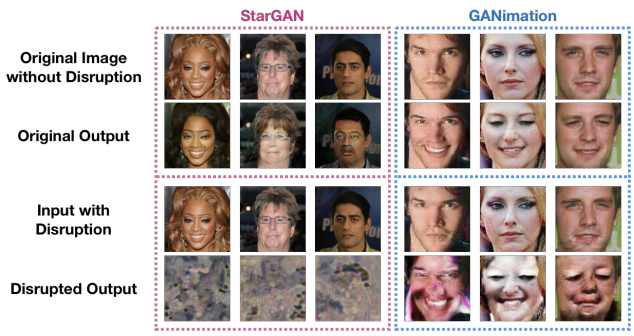

Figure 2. An example of our deepfake disruptions on StarGAN [4] and GANimation [15]. Some image translation networks are more prone to disruption.

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 2. An illustration of our method lies in Figure 1 and we present example disruptions in Figure 2.

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.

Blurring is a broken defense in the white-box scenario, where a disruptor knows the type and magnitude of preprocessing 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. Previous and concurrent work [22, 18] do not tackle the following problems.
- 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. Method

We describe methods for image translation disruption (Section 2.1), our proposed conditional image translation disruption techniques (Section 2.2), our proposed adversarial training techniques for GANs (Section 2.3) and our proposed spread-spectrum disruption (Section 2.4).

2.1. 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.

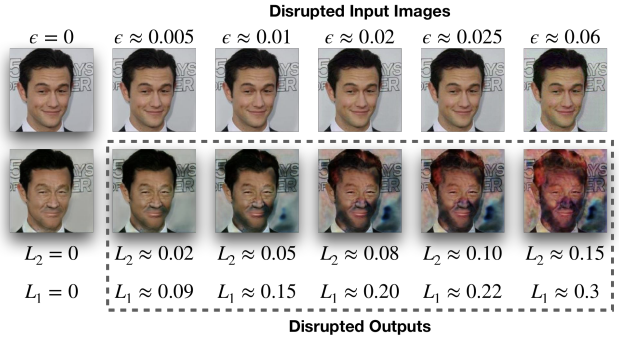

Figure 3. Equivalence scale between L_2 and L_1 distances and qualitative distortion on disrupted StarGAN images. We also show the original image and output with no disruption. Images with $L_2 \geq 0.05$ have very noticeable distortions.

We can generate a targeted disruption by adapting well-established adversarial attacks: FGSM, I-FGSM, and PGD. Fast Gradient Sign Method (FGSM) [7] 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) [9] 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) [10], 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.

2.2. 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:

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

2.3. GAN Adversarial Training

Adversarial training for classifier deep neural networks was proposed by Madry *et al.* [10]. 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:

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

Architecture (Dataset)	PGD L^1	PGD L^2	PGD % dis.
StarGAN (CelebA)	1.119	1.479	100%
GANimation (CelebA)	0.139	0.044	30.4%
GANimation (CelebA, $\epsilon = 0.1$)	0.190	0.077	83.7%
pix2pixHD (Cityscapes)	0.922	1.084	100%
CycleGAN (Horse)	0.402	0.253	100%
CycleGAN (Monet)	0.881	0.898	100%

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 $\epsilon = 0.05$ unless noted. We notice that strong disruptions are successful on all tested architectures.

2.4. 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

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

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 $k = 1$.

3. 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 [15], StarGAN [4], pix2pixHD [21] and CycleGAN[24]. We show that we can attack several different architectures in Table 1. For reference, in Figure 3 we show qualitative examples of differing L^2 distortion levels. In Table 2, we show that our image-level disruption is superior than adapted related work. In Table 3 and Figure 4, we demonstrate that both our *class transferable disruptions* are able to successfully transfer to different conditioning classes. In Table 4, we show that our proposed *G+D adversarial training* is most effective at alleviating disruptions, although strong disruptions are able to overcome this defense. *G+D adversarial training* is a first step towards robust image translation architectures. Finally, in Figure 5 we show that our *spread-spectrum adversarial disruption* effectively evades blur defenses in a gray-box scenario.

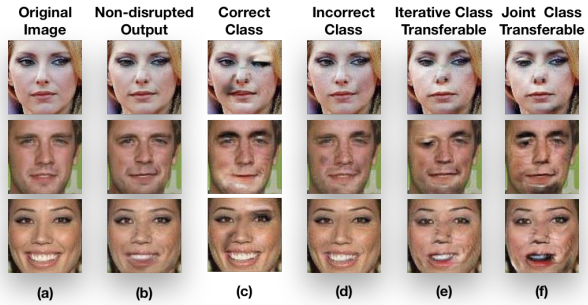

Figure 4. Examples of our class transferable disruptions. (a) Input image. (b) Ground truth GANimation output without disruption. (c) Disruption using the correct Action Unit (AU) correctly is successful. (d) Disruption with an incorrect target AU is not successful. (e) Our iterative class transferable disruption and (f) joint class transferable disruption are able to transfer across different AUs and successfully disrupt the deepfake generation.

Layer	Kos *et al.* [8] 4	5	6	7	Ours
L^1	0.671	0.661	0.622	0.573	**1.066**
L^2	0.656	0.621	0.558	0.478	**1.365**

Table 2. Comparison of our image-level PGD disruption with an adapted feature-level disruption from Kos *et al.* [8] on StarGAN.

	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 3. 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.

Defense	PGD L^1	PGD L^2	PGD % dis.
No Defense	0.863	0.981	100
Blur	0.279	0.133	89.2
Adv. G. Training	0.319	0.186	95.2
Adv. G+D Training	0.281	0.136	87.6
Adv. G. Train. + Blur	0.224	0.099	61.2
Adv. G+D Train. + Blur	**0.184**	**0.062**	**37.2**

Table 4. Disruptions on StarGAN with different defenses.

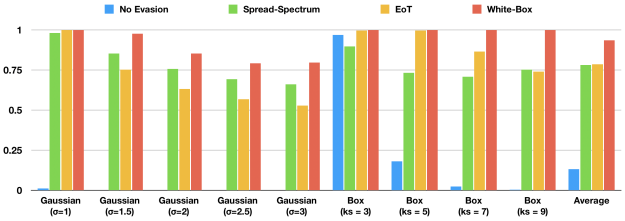

Figure 5. Proportion of disrupted images ($L^2 \geq 0.05$) for different blur evasions under different blur defenses.

References

[1] Facebook to ban 'deepfakes'. https://www.bbc.com/news/technology-51018758. Accessed: 2020-1-10. 1

[2] Dina Bashkirova, Ben Usman, and Kate Saenko. Adversarial self-defense for cycle-consistent gans. In *Advances in Neural Information Processing Systems*, pages 635–645, 2019. 1

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 1

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 1, 2, 4

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. *arXiv preprint arXiv:1912.01865*, 2019. 1

[6] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017. 1

[7] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 3

[8] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 36–42. IEEE, 2018. 1, 4

[9] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 3

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 3

[11] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 1

[12] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 1

[13] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 1

[14] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 1

[15] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. 1, 2, 4

[16] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[17] Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*, 2016. 1

[18] Lin Wang, Wonjune Cho, and Kuk-Jin Yoon. Deceiving image-to-image translation networks for autonomous driving with adversarial perturbations. *IEEE Robotics and Automation Letters*, PP:1–1, 01 2020. 2

[19] Run Wang, Lei Ma, Felix Juefei-Xu, Xiaofei Xie, Jian Wang, and Yang Liu. Fakespotter: A simple baseline for spotting AI-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019. 1

[20] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 1

[21] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 4

[22] Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *The IEEE Winter Conference on Applications of Computer Vision (WACV) Workshops*, March 2020. 2

[23] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9459–9468, 2019. 1

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 4