Exploring Adversarial Fake Images on Face Manifold

Dongze Li[1,2], Wei Wang[2], Hongxing Fan[1,2], Jing Dong[2]
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] Center for Research on Intelligent Perception and Computing, CASIA
lidongze2020@ia.ac.cn,hongxing.fan@cripac.ia.ac.cn
{wwang, jdong}@nlpr.ia.ac.cn

Abstract

Images synthesized by powerful generative adversarial network (GAN) based methods have drawn moral and privacy concerns. Although image forensic models have reached great performance in detecting fake images from real ones, these models can be easily fooled with a simple adversarial attack. But, the noise adding adversarial samples are also arousing suspicion. In this paper, instead of adding adversarial noise, we optimally search adversarial points on face manifold to generate anti-forensic fake face images. We iteratively do a gradient-descent with each small step in the latent space of a generative model, e.g. Style-GAN, to find an adversarial latent vector, which is similar to norm-based adversarial attack but in latent space. Then, the generated fake images driven by the adversarial latent vectors with the help of GANs can defeat main-stream forensic models. For examples, they make the accuracy of deepfake detection models based on Xception or EfficientNet drop from over 90% to nearly 0%, meanwhile maintaining high visual quality. In addition, we find manipulating style vector z or noise vectors n at different levels have impacts on attack success rate. The generated adversarial images mainly have facial texture or face attributes changing.

1. Introduction

Nowadays, it is increasingly hard for human eyes to tell a real image from a fake one with the rapid improvement of image generation techniques. Manipulated or generated fake images may draw social and privacy concern if being abused by malicious attackers. An attacker may register an account with photos belonging to a non-existent person or swap one person's face to another, thus causing privacy and security issues. Image forensic models are designed to clarify those images from real ones, and have gained considerable performance on several benchmarks and datasets [17, 30]. However, a smarter attacker may attempt to generate images which can bypass those detectors while keeping high visual quality. These images may escape the detection procedure and spread in social media. In order to combat the generation and spread of undetectable "deep fakes", it is necessary for image forensics researchers themselves to develop and study anti-forensic operations[32] to improve the robustness and generalization ability of the existed forensic models.

Figure 1. Adversarial images generated by different methods. Upper left is the original image. Upper right is the image generated by our method. Lower left and Lower right are adversarial images generated by FGSM[8] and PGD[21] L_{inf} norm-based attack respectively under the same perturbation level. Although all these images can bypass the target forensic model, images generated by our method are more invisible to human eyes.

In this paper, we propose to efficiently generate adversarial high visual quality fake images to fool forensic detectors. By adversarially exploring on the manifold of the recent powerful generative model Style-GAN[13], we can therefore generate the adversarial fake face images that fool forensic models. Though StyleGAN is capable of generat-

ing high-resolution images with various styles and stochastic details, the generated fake images are easily detected by models based on Xception or EfficientNet with accuracy of over 90%. But with intentionally iteratively searching these adversarial vectors in its latent space with a gradient descent manner, we successfully screen out fakes images that will be detected by forensic models as real ones.

Although one can exploit existing adversarial attack methods [8, 21] to deceive a forensic model, it may hold visible perturbations brought by the optimization process in image space, which make it detectable by human eyes or specially designed detectors [24]. Recently proposed unrestricted adversarial attack methods [31, 36, 11] could generate adversarial images with less suspicious visual artifacts by training GAN models, but they mainly focus on defeating classification or recognition tasks.

Our method has superiority in the following aspects: First, because we do modifications on the manifold, we don't have to care much about the image pixel constraint, which makes a higher updating strength possible. Second, unlike norm-based attack which leave visible artifacts onto the image, our method can generate the same image without obvious artifacts, see in Figure 1.

Our contribution are as follows:

1. We propose a novel method of generating adversarial anti-forensic images via exploring Style-GAN's manifold. Images generated by our method can successfully bypass two image forensic models, Xception [2] and EfficientNet [33]. indicating the demand for more robust forensic models.

2. We compare our method with nowadays widely-applied norm-based adversarial attacks and show that the proposed method can achieve the same attack success rate while introducing less visible perturbation, making it harder for our adversarial image to be detected by human eyes.

3. We conduct our attack in different ensemble ways and have shown our adversarial images can transfer between different forensic models, causing a threat even in the situation where the architecture of forensic models is unknown to the attacker.

4. We show some interesting effects between the adversarial strength, the level of input noise vector and the attributes of our generated images, which are worthy of investigation in future.

2. Related Work

2.1. Forged Image Generation and Detection

Forged Image Generation. Generating forged images manually with image-editing tools can be time-consuming and may be easily detected by both human eyes and other forensic methods. Recent deep-learning methods lower the threshold of synthesizing these fake images, making it possible to yield a large volume in a short period. Nowadays forged images have drawn more attention since detection on them are much harder then before, and have caused serious privacy and security issues, in which fake facial images account for a large proportion. There exists several methods to synthesize forge facial images[35]: Entire face synthesis [13, 22, 12], which our method belongs to, uses GANs or other generative models to generate a fake image; face identity swap[15, 34, 23], which swap one person's face to another one; face attributes editing [9, 40],which manipulates face attributes with image editing software or deep learning models and face expression manipulating [6, 34, 23], which transplant a target's facial expression to a source person.

GAN framework was first introduced by Goodfellow *et al*. [7] and has been widely applied to a series of unsupervised and semi-supervised image generation tasks . The main idea of a GAN is to fit the data distribution with an adversarial game procedure, where a generator tries to deceive a discriminator which aims to tell an image is from the real data distribution or not. GAN framework can also realize cross-domain translations [10, 43] with additional information and carefully designed loss functions. GANs have shown awesome results in face generation [22, 12, 13],and have drawn broad awareness for the vital role face images play nowadays. We choose StyleGAN [13] pretrained on FFHQ dataset for the diversity of its latent space and its ability to generate high resolution images with various styles and fine grained details. With a strong generator, we can generate more realistic images to bypass detectors as well as human eyes.

Image Forensics. Traditional Image forensic approaches are usually based on specific artifacts left by a certain forge method, which lack versatility and is fragile to the change of forge method and data distribution. Recent machine-learning and deep-learning methods are capable to handle more complex forge approaches. Zhang *et al*. [41] used SVMs and Random Forests to classify forged facial image, which is the first work to use machine-learning method on image forensics. A two-stream network was proposed to by Zhou*et al*. [42] for face manipulation detection. MesoNet proposed by Yamagishi*et al*. [1] uses a network with low layer numbers, focusing on the mesoscopic properties of images, to detect manipulated images. Rossler *et al*. [30] shows that a Xception model outperforms than other model on the forged image classification task. In practice, EfficientNet [33] also show good performance.

Recently, a Face X-Ray method [16] which focus on the detection of image blending artifact has been proposed and have shown good performance on identity-swapping.

For the simplicity, we focus on bypassing deep learning forensic models which are aimed to detect whether an image derives from a GAN or not. Images generated by our

method is able to escape the detection of two selected forensic models, Xception [2] and EfficientNet [33].

2.2. Adversarial Examples

Adversarial Examples are images crafted with small modifications to fool a target classifier. Given a classifier f and an clean image X as well as its ground truth label y which belong to a label set S,where $f(x) = y$ for an well-behaved classifier. The goal of an adversarial attack is to get a modified image X' and make the threat model predict $f(X') = y'$ while $y' \in S$ and $y' \neq y$. The type of the attack can be categorized as non-targeted attack and targeted attack based on whether its goal is to make the target model classify the adversarial image as any $y' \neq y$ or a specified y'. Based on how much information about the target model is known, attack method can be divided into white-box attack, in which model weights and architectures are accessible, and black-box attack, where attackers hold limited knowledge about the model. Besides, adversarial examples have shown transferability [25, 19], that is, adversarial examples crafted on one model may also fool other models although their architectures are different. The norm-based attack methods require the distance between the crafted adversarial image X' and the original image X should satisfy the p-norm constraint $\|X' - X\|_p < \epsilon$. Series of work have followed this protocol [4, 21, 29, 37] to improve the attack strength and transferability. Several ways to resist adversarial examples have also been proposed, such as adversarial training[8, 21], gradient distillation[26], high level guided denoise[18] and so on. Our work has taken advantage of several attack methods below but is free from the norm constraint.

FGSM. FGSM [8] is an one-step attack method which add perturbations onto the original images, hoping to maximize the loss function $J(X', y)$, while J is usually the cross-entropy loss, FGSM is formally defined by

$$X' = X + \epsilon \cdot sign(\nabla_X J(X, y)), \tag{1}$$

where ϵ denotes the max perturbation scale. We have deployed FGSM attack onto our generated images, trying to fool the forensic model as a controlled experiment.

PGD. Madry et.al [21] deploy a strong iterative attack method called Projected Gradient Descent. in each step the perturbation is projected to a $\epsilon-$ based ball. Defined by

$$X_{t+1} = X_t + \alpha \cdot sign(\nabla_{X_t} J(X_t, y)), \tag{2}$$

in which X_t denotes the attack image in t-th iteration.

In our work,we use a similar way to update our input vector to search on the generator's manifold, we also deployed L_{inf} and L_2 PGD attack onto the Style-GAN generated image to compare with on method on visual quaility and attack success rate.

Unrestricted Adversarial Examples. The traditional adversarial perturbations are constrained by norm-bounds, where unrestricted adversarial examples are free from. To generate an unrestricted adversarial image, one can apply various modifications to the original image,such as spatial transformations [38], rotating [5],attribute editing [28], translations[11], or even construct an image from scratch [31], as soon as the synthesized image still belongs to its previous class, which are usually judged by an auxiliary classifier[31].A recent work[27] also takes advantage of Style-GAN, generating unrestricted adversarial images via modifying Style-GAN's style vectors and noise inputs, attacking image classification models on ImageNet[3], CelebA[20] and Lsun[39], while we are aimed at deceiving forensic models and focus on face images generated by Style-GAN for the unique role face images plays in image forensic applications.

3. Method

Although highly realistic images can be generated by Style-GAN, they are easily to be detected by a forensic model. In this paper, we try to generate fake images which can fool forensic detectors without quality degradation. We do manipulations on Style-GAN's input latent vector z or noise vector n in the neighborhood of the given vector to find adversarial images on the face manifold. similar to work [27], while keeping the original architecture of Style-GAN unchanged.

3.1. Style-GAN

In our task, what we need is the original pretrained Style-GAN and its weight. Style-GAN architecture consists of a 8-layer linear mapping network f which maps a latent code z to an intermediate style space to get style vector w, and a generator g whose each layer takes style vectors and random noise inputs as input and generates image progressively. High-level patterns are determined by style vectors and stochastic details are controlled by noise inputs, respectively.

3.2. Anti-Forensics Fake Image Generation

In this work, we aim at generating our image adversarial once for all avoiding extra adding perturbations operation. As the iteration process goes further, the adversary of our generate images becomes stronger. Our search method focus on finding the right direction which makes the generated images have the ability to escape the forensic models. Motivated by the traditional norm-based iterative adversarial attack, we apply gradient descent on the noise and latent vector of Style-GAN, updating those vectors towards the

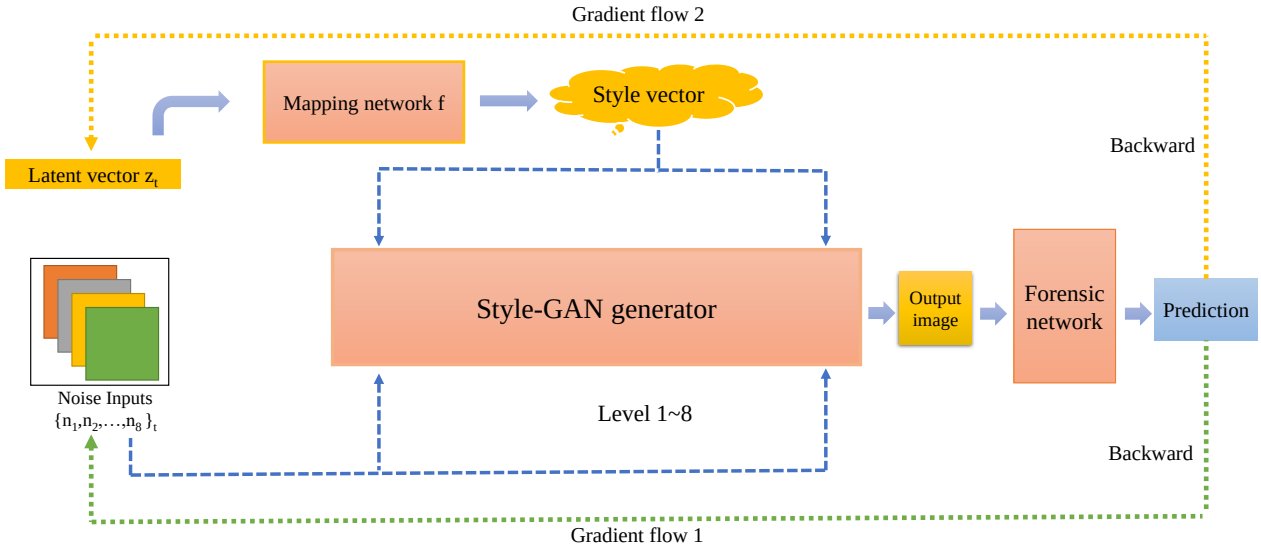

Figure 2. The overall pipeline of our method. We perform gradient descent on the latent vector and noise inputs of Style-GAN, respectively or together, maximizing loss function of the target forensic model(s).

direction which maximizes the loss of the forensic model's prediction. Each time after updating, we validate whether the forensic detector predicts the fake image as real. Because forged image detection task is a two classification problem, a non-target method will suffice. We bypass the target detection models successfully if they misclassify our generated images as real. The whole pipeline of our method is shown in Figure 2.

In the basic attack setting, we minimize the binary cross-entropy loss of the final prediction of a single forensic model, while in the ensemble setting, the predictions of the model are combined in different strategies, to achieve higher attack success rates on both models.

3.3. Adversarially Searching on Face Manifold

We perform gradient descent to the latent vector z before the mapping network f and the noise inputs $n_1, n_2 ... n_8$ inserted at each level of the generator's upsampling layer. Each time, we use the one step gradient sign of the loss function to update our latent vectors with a fixed step size hyperparameter ϵ_1. The modification formula can then be described as

$$z_{t+1} = z_t + \epsilon_1 \cdot sign(\nabla_{z_t}(J(g(z_t, n_t), y))), \quad (3)$$

where z denotes the input latent vector and n represents the noise injected into each layer, y is the ground truth label of the generated images and is always set to fake, g is our pre-trained Style-GAN generator. It is similar for each level's noise inputs. The iteration formula for noise inputs can be written as

$$n_{t+1} = n_t + \epsilon_2 \cdot sign(\nabla_{n_t}(J(g(z_t, n_t), y))). \quad (4)$$

The reason why we don't update the style vector after the mapping network f directly is that we want to keep the integrity of the whole Style-GAN architecture.

Updating on noise inputs $n_1, n_2, ..., n_8$ and latent vector z can be carried out separately or together. Each level of the generator network gets style vectors derived from the same z, at the first few steps of the iteration process, the images remain the same as they were. When iteration number increases, the texture of the resultant images become much deeper, see in Figure 3. Modification on $n_1, n_2, ...n_8$ will have more interesting effects on generated image for it will gradually change the appearance of the output image after more iteration steps and will bring different adversarial strength. Synthesized images have the ability to bypass the detector with ignorable changes after the first 1-2 steps, while after 3-5 steps, some attributes of the generated person's face have obviously changed. We also found that different update step size and iteration times have impact on the generate images visual quality. So we have to select feasible step sizes and iteration times if we want to keep the certain appearance of a person. The fact also means we might be able to deploy successful attacks with some of the image attributes modified, while others stay constant.

3.4. Gradients Ensemble

To increase the attack success rate of adversarial images generated by our method, we propose to conduct ensemble attack. We organize our ensemble attack method in three different manners: an alternatively attack manner, ensemble in loss and ensemble in network predicted logits. The success of the ensemble attack suggests our method can be used to bypass a set of forensic models as far as we know their architecture. To bypass a single model we need to search through a direction which maximizes the target model's predict error, while ensembling can constrain our searching towards the direction where we can find the images which make the prediction of both models deteriorate. Gradients ensemble can effectively avoid searching into local maximums.

4. Experiment

4.1. Basic Setting

Our method generates images totally from scratch. We use Style-GAN generator pretrained on FFHQ dataset as our generator. Random latent vectors and noise inputs are sampled from a standard Gaussian distribution. As for the forensic models, we use EfficientNet-B3 [33]and Xception [2] network for training, both of which are loaded with pretrained weights on the imagenet dataset. The training data set is composed of images from FFHQ Dataset and Style-GAN generated images, we use the first 50,000 images in both image sets as our training set, the 50,000-60,000th picture as the validation set, and the 60,000-70,000th picture as the test set. The size of the pictures in the dataset is 1024×1024. We resize those images to 299×299 and 300×300, and feed them into the EfficientNet-B3 and Xception networks respectively. We use Adam [14] as the optimizer and set the learning rate to 0.0002, weight decay to 0.001, and epoch to 3. In each epoch, we verified accuracy of our model for 5 times, and we choose the model with the best performance in the verification process as our final model and evaluate them on the testset.

To construct our attack dataset, for each strategy, we follow the same process: we first generate 5,000 images from scratch as our fake GAN-generated images, labelled as 1, and select another 5,000 real image from FFHQ dataset as real images, labelled as 0. Image size is 1024×1024 after generation process and is resized then normalized to $[0, 1]$ before being fed into forensic models. We test the accuracy of the detectors on those real images and fake images without attack as models accuracy on clean images. Average accuracy of each model has reached over 90%. Next, we apply our attack on those generated fake images and test the accuracy of the detector on the attacked images.

4.2. White-box Attack

In our basic white-box attack setting, we conduct three ways to attack. Noise method means we perform gradient decent on all noise inputs with the latent vector z fixed, while latent method means we only update the latent vector z with the noise inputs fixed. The third method, noise and latent means simultaneously updating noise inputs $n_1, n_2...n_8$ and latent vector z. As our threat model, Xception and EfficientNet are attacked separately. The hyperparameter ϵ_1 and ϵ_2 are set to 0.004 and 0.05, respectively. Although most of the attack images are able to bypass the target forensic models after 1-2 steps, we continue updating on these vectors till iteration times reached 10 for we want to see how the distortion the updating process bring to the output images will be.

As baseline experiments, we develop a FGSM attack under L_{inf} constraint and two PGD attacks under L_{inf} and L_2 norm constraint on the target forensic models. For L_{inf} and L_2 PGD attack, the iteration step size ϵ is set to 0.01, and will last 40 times. All pixel values are normalized to [0,1] in our experiment, and will be clipped after each iteration to avoid invalid outputs. and the allowed maximum perturbation size is 0.3 in all norm-based attacks.

Our basic attack success rate on two forensic models, comparing with the PGD attack, are shown in Table 1. In our attack setting, before attack, the accuracy on both real images and fake images is over 90%, while after our attack, the accuracy have dropped to less than 1%. Baseline methods also perform well in degrading the prediction of both forensic models, except for PGD L_2 attack, which only decrease the accuracy of the model by about 30%. We analyze that it is because of L_2 adversarial attack takes the whole image's information in consideration and thus requires much larger perturbation scale. While most methods shown in Table 1 can successfully deceive the model, our method is better in the visual quality of generated images. Images generated by our method, L_{inf} PGD attack and FGSM are shown in Figure 1 . PGD L_2 attack is not in our consideration due to its low attack success rate.

4.3. Black-Box Attack

In this subsection, we report the performance of our generated images in a black-box setting. We explore the black-box transferability of our generated adversarial image from one single forensic model to another. First we set our attack strategy to be the latent noise method as mentioned above, and fix the iteration step to 3. Under this situation, variations in images attributes is relatively small, meanwhile result images have enough adversarial strength. For the two forensic models, we generate 5000 adversarial images, denoted as Img_e and Img_x. We then test the accuracy Xception model reaches on adversarial images generated on EfficientNet and vice versa.

Method	Model	
	EfficientNet	Xception
Clean image	97%	93%
PGD $L_{inf}(\epsilon = 0.3)$	**0%**	5%
PGD $L_2(\epsilon = 0.3)$	60%	63%
FGSM L_{inf} $(\epsilon = 0.3)$	13%	5%
Noise(ours)	**0%**	**0%**
Latent(ours)	**0%**	**0%**
Noise and latent(ours)	**0%**	**0%**

Table 1. Accuracy different models perform on our method and other adversarial attack method, PGD L_2, PGD L_{inf} and FGSM attack. our method has the same ability to bypass the forensic detectors as norm-based adversarial attack, PGD L_{inf} and has better adversarial strength than FGSM and PGD L_2 attack. PGD L_2 attack shows poor performance on both models because of the limited perturbation scale.

Target Model	Method	Test Model	
		EfficientNet	Xception
EfficientNet	FGSM L_{inf}	**0%**	**0%**
	PGD L_{inf}	**0%**	**0%**
	PGD L_2	60%	81%
	Ours	**0%**	5%
Xception	FGSM L_{inf}	13%	5%
	PGD L_{inf}	86%	**0%**
	PGD L_2	95%	63%
	Ours	**9%**	**0%**

Table 2. Model accuracy under black-box setting. Adversarial examples generated from norm-based attacks (except L_2 PGD attack shows better transferability. while our method has also shown transferability to some extent.

We observe that images generated on EfficientNet by our method shows good black-box transferability, decreasing the accuracy of Xception from 93% to 50% by nearly a half. While images generated on Xception shows little transferability, the accuracy of EfficientNet is 97% on clean GAN-generated images before attack, while it is 90% on our adversarial images. Images generated by L_{inf} PGD attack, L_2 PGD attack and FGSM are also tested in the same way, and similar result was found. The model accuracy of black-box attacks can be found in Table 2. From the chart we can find transferability of norm-based attacks are better than our method. We speculate the reason for the fact is 1: Our method are generated from the specific Style-GAN's manifold and they still have artifacts(although invisible to human eyes) that can be caught by detectors and 2: norm based attack have less overfitting than our method on the target model.

Figure 3. Images generated though solely updating latent vector z at different steps. ϵ_1 is set to 0.004. While successfully deceiving the target model, our adversarial images are the same as the original image except for some nearly invisible textures under proper iteration times and step size.

4.4. Ensemble Attack

We have found that attack between different models which relies on transferability is kind of fragile and lacks robustness. To improve the performance, we conducted our experiment in an ensemble way. we have tried three ensemble methods: ensemble in loss, ensemble in logits, and a an alternative attack manner. Ensemble in loss means loss function of the two forensic models are added together after each calculation, while in ensembling in logits setting, we fuse output logits of image forensic models to get the final cross-entropy loss function, which is then used to get the gradient with a backward pass, and in this setup,the weight of the EfficientNet and Xception are both 0.5 for we observed their gradients are in the similar scale. In the alternative attack scenario, we alternatively carry out gradient descent updating process according to one model in the ensemble models in each step. Other settings are same as the base attack strategy. Result of model accuracy on different ensemble attack method are shown in Table 3. Result on Xception and EfficientNet model show all the ensemble attack method are able to deceive the target forensic model with a high confidence. The success of different ways of ensemble attack suggests we can improve anti-forensic ability of our generated images by expand our target model set.

Ensemble Method	Model	
	EfficientNet	Xception
Ensemble in loss	0.5%	**0%**
Ensemble in logits	**0.1%**	**0%**
Alternative attack	0.4%	**0%**

Table 3. Model accuracy under different ensemble attacks. Images generated from all the attack method can bypass the model successfully

4.5. Impact of Noise Inputs on Generated Images

Noise Level. Updating on the latent vector z and noise inputs $n_1, n_2, ...n_8$ can both achieve the goal of deceiving the target model. When iteration times are small enough,

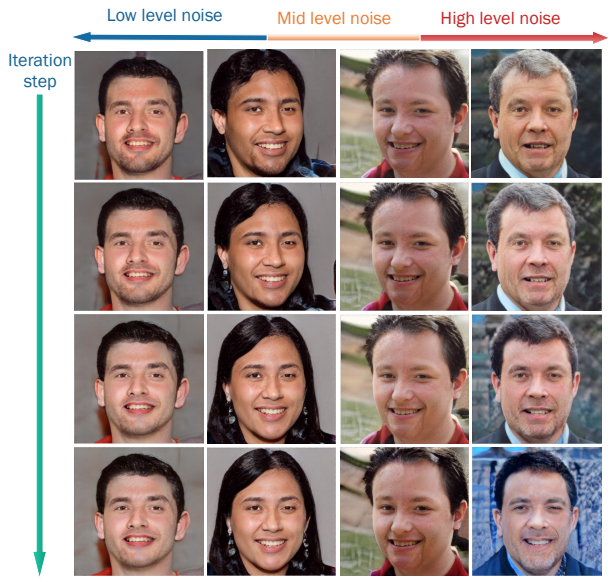

Figure 4. Images generated though solely updating noise vector at a specific level of the generator at different iteration steps. Updating low-level noise inputs results in more obvious attribute shifts.

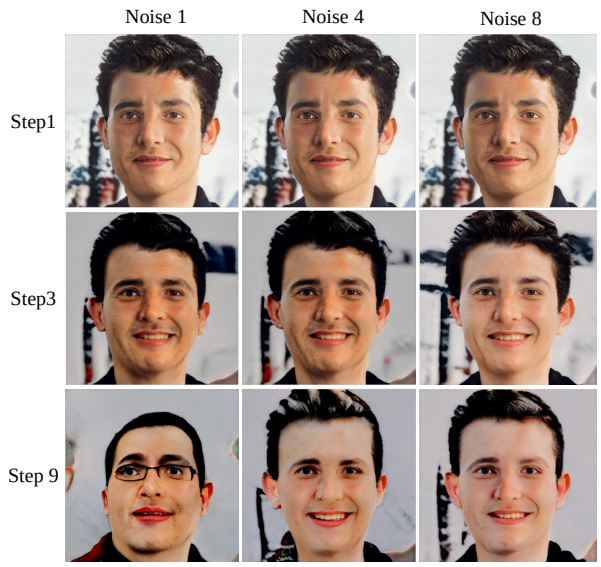

Figure 5. Images for a single person generated with a relatively large stride at different iteration step. After 3 steps, resultant images are still similar to their counterpart. While after 9 steps, distortion has become large enough to be noticed by human eyes.

the result image have little change comparing with the original one in both settings. However, When the iteration times goes larger, updating on z only brings irregular textures to the output image, making the attack more fragile facing with human eyes,see in Figure 3. while updating on noise vector shows more interesting results, controlling both the iteration step scale and updating noise injected in different level of the generator can affect both the attack success rate and the character's appearance of the generated image.

Concretely, changing the noise injected in lower level of the generator may affect the output character's appearance such like gender, beard, hair cuts and so on, while changing higher level noise may affect more general features of the images like color or textures. Updating on a certain level also leads to different attack strength.

We fix the latent vector z, and conduct attack by respectively updating noise vector from level 1 to level 8. We choose images generated in iteration step 3 and set a proper step length $\epsilon_2 = 0.05$ for it won't be too large to make the result images suffer from serious distortions. We have found that while the iteration steps and the latent vector z is fixed, the higher level noise vector impact the model prediction most, in other words, updating high level noise have strongest adversarial impact. Figure 4 shows our generate images under this setting and attack success rates of updating different level of noises are shown in Table 4.

Iterations & Step Size. We also found when updating step on our noise vector become larger, Changing in lower level noise inputs can also result in variations on some face

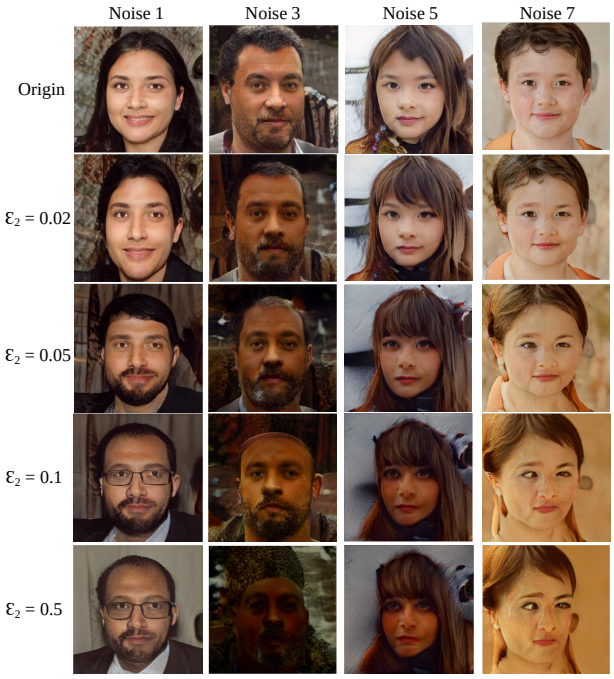

Figure 6. Images generated though with different stride at iteration step 3. Overlarge stride will lead to poor image quality.

attributes. Images generated through updating on different noise level with different iteration steps are shown in Figure 5 (with fixed $\epsilon_2 = 0.05$). More iteration steps bring

Attack level	Model	
	EfficientNet	Xception
0 (no attack)	97%	93%
1	99%	87%
2	98%	88%
3	98%	85%
4	75%	57%
5	35%	23%
6	**1%**	**0%**
7	**1%**	**0%**
8	**1%**	**0%**

Table 4. Model accuracy on adversarial images generate with only one level's noise vector taking part in the updating process. Updating higher level noise vectors results in stronger adversary.

larger image diversity. While proper modifications on noise inputs yield feasible results, excessively updating or overlarge step-size can cause serious distortions. We show images of a single person under different iteration steps with $\epsilon_2 = 0.1$ in figure and images with different iteration step size while step=3 in Figure 6.

While the clear relationship among adversarial strength, image attributes and noise level is still not clear. Further research may make us have the ability to make slight modifications on a certain feature of the image to bypass forensic models in future.

5. Conclusion & Future Work

In this paper, we proposed a novel method to generate GAN images which can bypass certain image forensic classifiers by searching on the manifold of Style-GAN. Images generated by our method can lower the forensic models' accuracy from over 90% to less than 1%, and have better visual quality than norm-based adversarial attacks. We also explored the transferability of our images over two forensic models and deployed our method in different ensemble manners and successfully generated unrestricted adversarial images which are able to bypass several forensic models. Our method suggests more robust image forensic models are needed to identify GAN-generated fake images from real ones. In our future work, we are about to find out ways to improve transferability of our adversarial images over different model architectures. Relationship among noise level, face attributes and adversarial strength is also an attractive realm to explore.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 2

[2] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2, 3, 5

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 3

[5] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 1(2):3, 2017. 3

[6] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4217–4224, 2014. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2, 3

[9] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[11] Kazuya Kakizaki and Kosuke Yoshida. Adversarial image translation: Unrestricted adversarial examples in face recognition systems. *arXiv preprint arXiv:1905.03421*, 2019. 2, 3

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[15] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017. 2

[16] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 2

[17] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 2019. 1

[18] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 3

[19] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 3

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 3

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 3

[22] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[23] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE international conference on computer vision*, pages 7184–7193, 2019. 2

[24] Tianyu Pang, Chao Du, Yinpeng Dong, and Jun Zhu. Towards robust detection of adversarial examples. 2017. 2

[25] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 3

[26] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016. 3

[27] Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge Belongie, and Ser-Nam Lim. Fine-grained synthesis of unrestricted adversarial examples. *arXiv preprint arXiv:1911.09058*, 2019. 3

[28] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *arXiv preprint arXiv:1906.07927*, 2019. 3

[29] Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4322–4330, 2019. 3

[30] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 1, 2

[31] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, pages 8312–8323, 2018. 2, 3

[32] M. C. Stamm and K. J. R. Liu. Anti-forensics of digital image compression. *IEEE Transactions on Information Forensics and Security*, 6(3):1050–1065, 2011. 1

[33] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 2, 3, 5

[34] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2

[35] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020. 2

[36] Xiaosen Wang, Kun He, Chuanbiao Song, Liwei Wang, and John E Hopcroft. At-gan: An adversarial generator model for non-constrained adversarial examples. *arXiv preprint arXiv:1904.07793*, 2019. 2

[37] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1161–1170, 2020. 3

[38] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018. 3

[39] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 3

[40] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–432, 2018. 2

[41] Ying Zhang, Lilei Zheng, and Vrizlynn LL Thing. Automated face swapping and its detection. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, pages 15–19. IEEE, 2017. 2

[42] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017. 2

[43] 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. 2