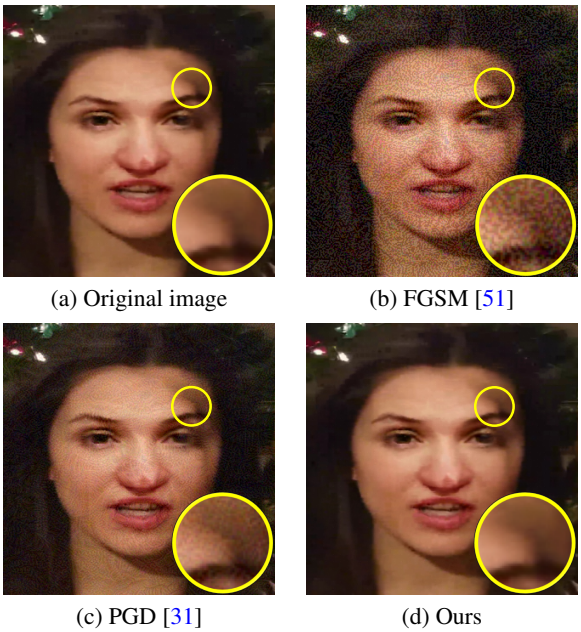
Exploring Frequency Adversarial Attacks for Face Forgery Detection

Shuai Jia[1] Chao Ma[1*] Taiping Yao[2] Bangjie Yin[2] Shouhong Ding[2] Xiaokang Yang[1]
[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2] Youtu Lab, Tencent
{jiashuai,chaoma,xkyang}@sjtu.edu.cn
{taipingyao,ericshding}@tencent.com, jamesyin10@gmail.com

Abstract

Various facial manipulation techniques have drawn serious public concerns in morality, security, and privacy. Although existing face forgery classifiers achieve promising performance on detecting fake images, these methods are vulnerable to adversarial examples with injected imperceptible perturbations on the pixels. Meanwhile, many face forgery detectors always utilize the frequency diversity between real and fake faces as a crucial clue. In this paper, instead of injecting adversarial perturbations into the spatial domain, we propose a frequency adversarial attack method against face forgery detectors. Concretely, we apply discrete cosine transform (DCT) on the input images and introduce a fusion module to capture the salient region of adversary in the frequency domain. Compared with existing adversarial attacks (e.g. FGSM, PGD) in the spatial domain, our method is more imperceptible to human observers and does not degrade the visual quality of the original images. Moreover, inspired by the idea of meta-learning, we also propose a hybrid adversarial attack that performs attacks in both the spatial and frequency domains. Extensive experiments indicate that the proposed method fools not only the spatial-based detectors but also the state-of-the-art frequency-based detectors effectively. In addition, the proposed frequency attack enhances the transferability across face forgery detectors as black-box attacks.



(a) Original image (b) FGSM [51] (c) PGD [31] (d) Ours

Figure 1. Illustration of adversarial examples generated by FGSM [15], PGD [31] and our method. The original image is classified as a fake face by the face forgery detector. After implementing these attacks, the adversarial examples are misclassified as real faces. Compared with FGSM [15] and PGD [31], our method associated with the frequency adversarial attack generates more natural perturbations, where the image quality of the adversarial example is much closer to the original image.

1. Introduction

With the rapid development of generative adversarial network (GAN), face forgery generation attracts increasing attention, such as Deepfake [46], FaceSwap [25], Face2Face [45], and NeuralTextures [44]. These techniques derive plenty of interesting applications, for instance, trying on makeup virtually and editing faces in the film industry. However, despite the positive aspect, face forgery generation may be maliciously abused, causing serious problems of security and privacy. Therefore, it is essential to design face forgery detection methods to distinguish the manipulated face from the real one.

Various face forgery detectors [1,7,8,38] are proposed to learn the decision boundary between real and fake faces and achieve significant performance on multiple datasets [9,37]. However, existing methods are vulnerable to the adversarial examples, which leaves a serious backdoor for the security of detectors. For instance, a forged face image that is classified correctly as fake by adding adversarial perturba-

* Corresponding author.

tions can fool the detector to make a wrong decision as real. Existing works [4, 13, 21, 26, 34] have explored the robustness of face forgery detection methods, but these methods add adversarial perturbations or patches on the original images, which are easily recognized by human eyes. In brief, the adversarial examples aim to fool a face forgery detector, while the objective of face forgery generation is to fool humans. An implicit attack that fools humans and detectors at the same time brings out a more serious problem of security. Meanwhile, more and more works [5, 36] consider the frequency diversity between real and fake faces as the essential clues for face forgery detection. It inspires us to conduct the adversarial attack in the frequency domain to boost the transferablility across various detectors.

To address the above issues, we propose a frequency adversarial attack method to add adversarial perturbations in the frequency domain. First, we apply discrete cosine transform (DCT) to transfer the input images into the frequency domain. Specifically, we utilize a fusion module to slightly modify the energy in different frequency bands via the adversarial loss. The indirect injection of adversary into frequency domain avoids the redundant noise of attacks in the spatial domain (e.g., FGSM [51], PGD [31]) and does not degrade the visual quality of original images. After that, we apply inverse DCT back to the spatial domain and obtain the final adversarial examples. For face forgery detectors, some existing methods [41, 53] only consider the noise pattern in the spatial domain to detect the fake faces, while others [28, 30, 36] utilize the frequency information as a clue. Moreover, some methods [5, 27, 32] combine the discriminative features from both domains to learn the boundary between real and fake faces. Therefore, in order to enhance the generalization of the proposed attack method, we propose a hybrid adversarial attack to integrate the spatial adversarial attack and frequency adversarial attack into a whole framework. Inspired by the idea of meta-learning [35], we alternately optimize the perturbations based on the adversarial gradients in different domains. The compatible ensemble of adversarial attacks can reserve the virtues of attacks in both domains. Adversarial examples with different attacks are illustrated in Figure 1.

Our main contributions can be summarized as follows.

- For the task of face forgery detection, we propose a novel adversarial attack method to generate perturbations in the frequency domain. Compared with the previous attacks, our method generates more imperceptible perturbations for human observers.
- To further boost the transferability of the attack, we propose a hybrid adversarial attack based on the strategy of meta-learning to simultaneously perform attacks on the spatial and frequency domain.
- We perform the proposed method both on the spatial-based face forgery detectors and the state-of-the-art frequency-based detectors. Extensive experiments on benchmarks demonstrate the effectiveness of our attack under both white-box and black-box settings.

2. Related Work

In this section, we briefly introduce the development of face forgery generation and detection. Besides, we review recent adversarial attack methods, especially for face forgery detection.

2.1. Face Forgery Generation

Face forgery generation [14, 23, 24] aims to craft the face image that is authentic in the eyes of human beings, which brings numerous productive applications, e.g., virtual shopping, online education, film production, etc. In summary, face forgery generation can be divided into four categories [33]: reenactment, replacement, editing and synthesis. One typical application of reenactment is to use one's expression or mouth to drive another one, e.g., RecycleGAN [3], STGAN [29]. FaceSwap [25] is the most common type of replacement. Averbuch-Elor et al. [2] animate the expressiveness of the subject through 2D warps and transfer it to the target automatically. Face2Face [45] considers the facial expressions as under-constrained problems to transfer the deformation between source and target. Editing and synthesis are used to add or remove ones' attributes consisting of hair, glasses, age, makeup, etc. In this paper, we choose the fake face images from the pubic datasets [9, 37] rather than generated by ourselves. In other words, we do not get access to the concrete approaches to manipulate the fake face.

2.2. Face Forgery Detection

Despite the creative applications of face forgery generation, this technology can be used abusively for malicious and unethical ways. Regarding its potential maleficence by the academic community, researchers attempt to detect if an image is manipulated or not to alleviate the danger, which is considered as a binary classification problem. Some works [1, 7, 38, 53] apply deep neural networks to extract discriminative features for face forgery detection. These methods only utilize the information from the spatial domain, which generally overfits the classification boundary. On the other hand, some works [11, 28, 30, 36, 40, 49] observe the diversity of real faces and fake faces in the frequency domain and propose the face forgery detection method with the frequency clues. F^3-Net [36] integrates the frequency-aware decomposition and local frequency statistics into a whole learning framework to classify the real and fake faces. Luo et al. [30] design several modules by taking full advantage of the high-frequency features at multiple scales to achieve higher accuracy. Furthermore, some methods [5, 16–18, 27, 32] integrate the spatial and frequency in-

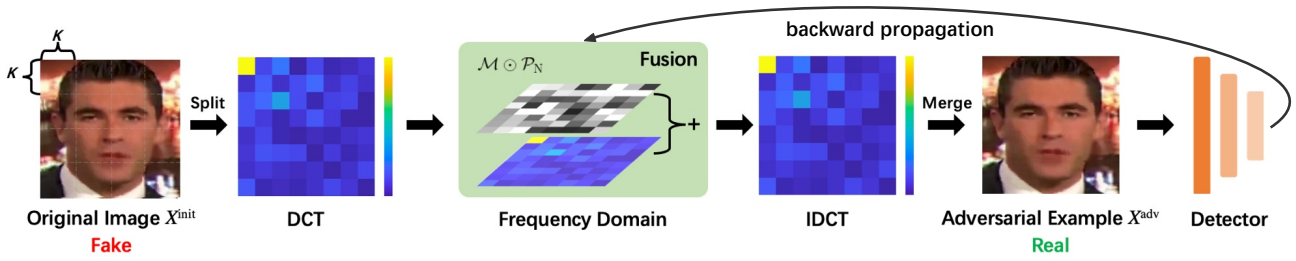


Figure 2. The pipeline of frequency adversarial attack. We first split the input image into $K \times K$ blocks and apply DCT on each block to transfer them into the frequency domain. We then introduce the frequency perturbation $\mathcal{P}_N$ and a predefined weight matrix $\mathcal{M}$ that controls the step sizes in different frequency bands. After that, we implement IDCT and merge them into the adversarial example. In each iteration, we calculate the adversarial loss and update the perturbation $\mathcal{P}_N$ based on it.

formation into a whole framework to detect the fake face accurately. LRL [5] adopts a multi-task learning strategy with two output branches, where one branch is to learn the surface label and the other one aims to focus on the edge of the modified region. Li et al. [27] combine the frequency clues with the spatial features to enlarge the difference between real faces and fake faces in the embedding space. Motivated by the diversity in the frequency domain, the proposed hybrid adversarial attack considers the effect on both domains to learn the robustness of existing face forgery detectors. For a complete comparison, we both select the spatial-based models and the frequency-based models to validate the effectiveness of the proposed attack method.

2.3. Adversarial Attack

Different from face forgery generation, the aim of adversarial attack is to fool a machine rather than human beings. Generally, given a well-trained network, the goal of adversarial attack is to generate the adversarial examples that make the network predict wrongly. The category for adversarial attack can be divided into white-box attack [15,31,42] and black-box attack [10,47,51] roughly, which is based on the attacker gets access to the concrete structures and parameters of victim models or not. While the majority of existing attack methods focus on the multi classification task, adversarial attack has been investigated in many fields, such as object detection [50], face recognition [52], visual tracking [22], etc.

For adversarial attacks in face forgery detection, some works [4, 13, 21, 26, 34] explore the robustness of models in different settings. Li et al. [26] manipulate the noise vectors and latent vectors of Style-GAN [48] with gradients to fool the face forgery models. Neekhara et al. [34] perform adversarial attacks in a black-box setting for face forgery detection. Carlini et al. [4] present the robustness of face forgery classifiers under various types of attack methods. The methods mentioned above generate the adversarial examples on the spatial domain, while some works [19,39] explore the frequency attack in other tasks. Since face forgery detection has a high relation with the frequency domain, we propose a novel attack method combined with the aspects of frequency domain to generate more imperceptible adversarial examples.

3. Method

Let X^{init} denote the original image, $f(X, \theta)$ denote the face forgery detector, and y^{gt} denote the corresponding ground-truth label. Our aim is to generate the adversarial example X^{adv} that makes the face forgery detector predict wrongly, i.e., $f(X^{\text{adv}}, \theta) \neq y^{\text{gt}}$. During adversarial attack, the objective is to maximize the loss function $\mathcal{L}(X^{\text{adv}}, y^{\text{gt}})$, where $\mathcal{L}$ is the binary cross entropy loss in face forgery detection. The concrete optimization is defined as:

$$\arg\max \mathcal{L}(X^{\text{adv}}, y^{\text{gt}}), \ \text{s.t.} \ ||X^{\text{adv}} - X^{\text{init}}||_{\text{p}} < \epsilon, \quad (1)$$

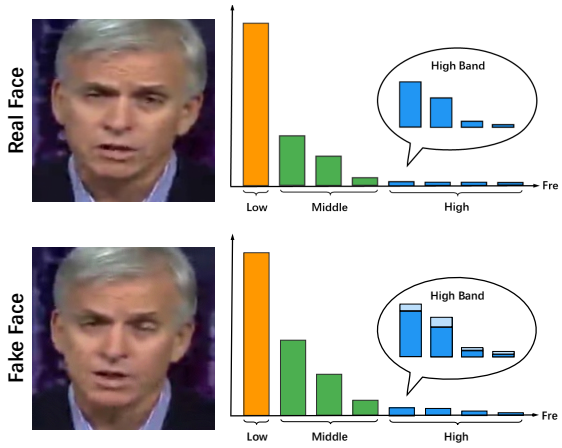
where p is l_{p}-norm to ensure the adversarial image close to the original image. We choose the untargeted attack to maximize the adversarial loss instead of the targeted attack due to the diversity of classification boundaries in different models. Although the targeted attack deteriorates the white-box model seriously, it has an extremely weak transferability to other models, which is prone to overfit on the specific network.

3.1. Spatial Adversarial Attack

Existing attack methods are mostly considered as spatial adversarial attacks that modify the adversarial examples on the pixels. Due to the limited page, we only introduce two spatial adversarial attack methods that are utilized for comparisons in the experiments. More variants of these methods can refer to [10].

Fast Gradient Sign Method (FGSM). FGSM [15] is a single-step attack method that calculates the perturbations based on the gradient of the adversarial loss. The optimization is defined as:

$$X^{\text{adv}} = X^{\text{init}} + \epsilon \cdot \text{sign}(\nabla_X \mathcal{L}(X^{\text{adv}}, y^{\text{gt}})). \quad (2)$$



(a) Image (b) Frequency distribution

Figure 3. The diversity between real faces and fake faces in the frequency domain. We select two examples with different labels from FaceForensics++ [37] and calculate the energy in different frequency bands. The energy of fake faces in the high frequency bands is richer than the one in real faces.

Projected Gradient Descent (PGD). PGD [31] is a multi-step variant of FGSM [51]. Meanwhile, it adopts a random initialization of perturbations at the first step. The update procedure is defined as:

$$\begin{aligned} X_0^{\text{adv}} &= X^{\text{init}}, \\ X_{\text{n+1}}^{\text{adv}} &= \text{Clip}\left\{X_{\text{n}}^{\text{adv}} + \alpha \cdot \text{sign}(\nabla_X \mathcal{L}(X_{\text{n}}^{\text{adv}}, y^{\text{gt}}))\right\}. \end{aligned} \tag{3}$$

3.2. Frequency Adversarial Attack

Previous studies [5, 36] have proven the difference between the real face and the fake face in the frequency domain. Figure 3 demonstrates the diversity of energy in different frequency bands between a real face and a fake face. The low frequency region is related to the content of images accounting for most of the energy, while the high frequency region is related to the edge and texture information of images. The fake face shows more energy in the high frequency regions compared to the real one. Inspired by this observation, we propose a frequency adversarial attack to directly modify the energy in the frequency domain. Compared to the spatial attacks, our attack method hides the adversary in the frequency bands and decreases the redundant noise in the pixel level, leading to a more invisible attack. The pipeline of frequency adversarial attack is illustrated in Figure 2. We summarize the optimization procedure as follows:

$$\begin{aligned} &\arg\ \max\ \mathcal{L}(\mathcal{D}'(\mathcal{F}(\mathcal{D}(X^{\text{adv}}))), \theta, y^{\text{gt}}), \\ &\text{s.t.}\ \ ||\mathcal{D}(\text{X}^{\text{adv}}) - \mathcal{D}(\text{X}^{\text{init}})||_{\text{p}} < \epsilon, \end{aligned} \tag{4}$$

where $\mathcal{D}(\cdot)$ denotes discrete cosine transform (DCT), $\mathcal{D}'(\cdot)$ denotes inverse discrete cosine transform (IDCT), $\mathcal{F}$ represents the fusion module to modify the energy in the frequency domain. Meanwhile, we utilize the l_{p}-norm to constrain the deviation of the original distribution of frequency.

Algorithm 1: Frequency Adversarial Attack

Input: Input image X^{init}, forensic detector $f(\cdot)$;
Output: Adversarial examples X^{adv};

Classify $f(X_0^{\text{adv}}, \theta)$ to get the true label y^{gt} ;
Generate the initial perturbations $\mathcal{P}_0 \sim \mathcal{U}(0, 1)$;
Initialize $X_0^{\text{adv}} = X^{\text{init}}$, $\hat{y} = y^{\text{gt}}$;
for $n = 0$ **to** N **do**
  Split $X_{\text{n}}^{\text{adv}}$ into $K \times K$ blocks;
  Apply the DCT on each block $\mathcal{D}(X_{\text{n}}^{\text{adv}})$;
  Calculate the adversarial loss $\mathcal{L}$ via Eq. 4;
  Update the perturbation $\mathcal{P}_{\text{n+1}}$ via Eq. 7;
  Fuse $\mathcal{P}_{\text{n+1}}$ and $\mathcal{M}$ into $\mathcal{D}(X_{\text{n}}^{\text{adv}})$ via Eq. 6;
  Apply the IDCT on each block $\mathcal{D}'(\mathcal{F}(X_n^{\text{adv}}))$;
  Merge $K \times K$ blocks into $X_{\text{n+1}}^{\text{adv}}$;
  Classify $f(X_{\text{n+1}}^{\text{adv}}, \theta)$ to get the predicted result $\hat{y}$;
end
return $X_{\text{n+1}}^{\text{adv}}$;

For details, we first implement DCT to transfer the image from spatial domain to frequency domain by following [36]. To balance the efficiency and quality of transformations, we split the original image into $K \times K$ blocks before DCT. For each block, we apply the DCT as follows:

$$\begin{aligned} \mathcal{D}(u, v) = \ & c(u) \cdot c(v) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} X(i, j) \\ & \cos[\frac{(2i+1)\pi}{2N} u] \cdot \cos[\frac{(2j+1)\pi}{2N} v], \end{aligned} \tag{5}$$

where $X(i, j)$ is the value on the coordinate (i, j) of image, $c(u)$ and $c(v)$ aim to make the DCT matrix orthogonal and N is the size of each block. Then, we generate the initial perturbations $\mathcal{P} \sim \mathcal{U}(0, 1)$ to inject on the frequency band. When the RGB image transfers into the frequency domain, the range of energy in different frequencies are lopsided as shown in Figure 3. Therefore, we propose a matrix $\mathcal{M}$ with adaptable step sizes, which is based on the proportion of each frequency band to balance the influence of lopsided energy. Moreover, the matrix $\mathcal{M}$ is dynamically reset for diverse inputs to maintain the visual quality. The complete fusion module is defined as:

$$\mathcal{F}(X_{\text{n}}^{\text{adv}}) = \mathcal{D}(X_{\text{n}}^{\text{adv}}) + \mathcal{M} \odot \mathcal{P}_{\text{n+1}}, \tag{6}$$

where $\odot$ is Hadamard product. During the optimization, $\mathcal{P}_{\text{n+1}}$ is updated as follows:

$$\mathcal{P}_{\text{n+1}} = \mathcal{P}_{\text{n}} + \lambda \cdot \text{sign}(\nabla_{\mathcal{P}} \mathcal{L}(\mathcal{D}'(\mathcal{F}(\mathcal{D}(X_{\text{n}}^{\text{adv}}))), \theta, y^{\text{gt}})), \tag{7}$$

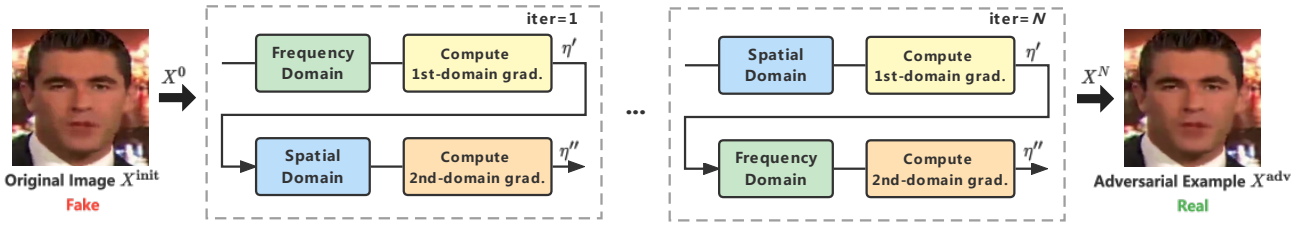


Figure 4. The procedure of hybrid adversarial attack. To combine the adversarial attack in different domains, we calculate gradients from both domains in order and update the perturbations. Then, we switch the order of domains in the next step. After iterations, the adversarial example gathers the gradients from both domains, leading to a stronger adversarial attack on both white-box and black-box settings.

where λ is the step size in each iteration. After that, we apply IDCT to transfer each block in the frequency domain back to the spatial domain. Note that the solo operation of DCT and IDCT is non-destructive, where no block artifacts are introduced in the process. When it reaches the maximum iteration or classifies the X^{adv} as a wrong label, we end up the loop and output the final adversarial example X^{adv}. The pseudo code is shown in Algorithm 1.

3.3. Hybrid Adversarial Attack

Meta-learning [12] aims to train a model that can quickly adapt to a new task with only a few training steps and training data, which is summarized as 'learn to learn'. Inspired by the idea of meta-learning [35], we propose a hybrid adversarial attack combined with the spatial domain and the frequency domain. Different from the vanilla meta-learning that updates a new model through the training data, we directly utilize the gradient from different domains to iteratively update the adversarial perturbations. Our hybrid adversarial attack can gather the virtues from both domains and integrate them in a compatible way. The hybrid perturbations improve the effectiveness of adversarial attacks in white-box settings and have a strong transferability on other models. The procedure of our hybrid adversarial attack is shown in Figure 4.

Let $\mathcal{A}_S$ and $\mathcal{A}_F$ denote the adversarial attack in spatial and frequency domain, respectively. At first, we compute the gradient based on the adversarial loss in the frequency domain. The optimization of $\mathcal{A}_F$ in the frequency domain is calculated by:

$$\eta' = \eta - \gamma_f \cdot \nabla_\eta \mathcal{L}_{\mathcal{A}_F}(\eta, \theta, y^{\text{gt}}), \tag{8}$$

where η' is frequency values and γ_f is the step size in the frequency domain. Then, we compute the gradient based on the adversarial loss in the spatial domain. The process of spatial attack $\mathcal{A}_S$ is formulated as:

$$\eta'' = \eta' - \gamma_s \cdot \nabla_{\eta'} \mathcal{L}_{\mathcal{A}_S}(\eta', \theta, y^{\text{gt}}), \tag{9}$$

where η'' is pixel values and γ_s is the step size in the spatial domain. The detailed procedure of adversarial attacks in each domain follows the above sections. We select PGD [31] as our spatial attack and the proposed frequency attack method as our frequency attack, respectively. Note that we remove the clip function in both attacks and add it at the end of each iteration. After each iteration, we switch the order of frequency attack $\mathcal{A}_F$ and spatial attack $\mathcal{A}_S$, and repeat the whole procedure. The complete algorithm of our hybrid adversarial attack is presented in Algorithm 2.

Algorithm 2: Hybrid Adversarial Attack

Input: Input image X^{init}, forensic detector $f(\cdot)$, spatial attack $\mathcal{A}_S$, frequency attack $\mathcal{A}_F$;
Output: Adversarial examples X^{adv};

Classify $f(X_0^{\text{adv}}, \theta)$ to get the true label y^{gt} ;
Initialize $X_0^{\text{adv}} = X^{\text{init}}$, $\hat{y} = y^{\text{gt}}$;
for $n = 0$ **to** N **do**
  Calculate the adversarial loss $\mathcal{L}_{\mathcal{A}_F}$ via Eq. 4;
  Update the perturbation η' via Eq. 8;
  Calculate the adversarial loss $\mathcal{L}_{\mathcal{A}_S}$ via Eq. 1;
  Update the perturbation η'' via Eq. 9;
  Clip the output images X_{n+1}^{adv} with the l_p-norm;
  Classify $f(X_{n+1}^{\text{adv}}, \theta)$ to get the predicted result $\hat{y}$;
  Switch the order of attacks $\mathcal{A}_S$ and $\mathcal{A}_F$;
end
return X_{n+1}^{adv};

4. Experiment

In this section, we first introduce the experimental setup. Then, we evaluate the performance of the proposed attack method with the single attacks and the ensemble attacks on the spatial-based models. We further validate our attack method on the frequency-based models. In addition, we conduct ablation studies on the variations of our method and different frequency bands. We finally evaluate the image quality of our method qualitatively and quantitatively.

4.1. Experimental Setup

Datasets. DFDC [9] is a challenging dataset with a variety of anonymous manipulations and perturbations. We randomly select 1000 fake face images from the DFDC dataset.

Table 1. The accuracy of spatial-based and frequency-based face forgery detectors on the DFDC [9] and FaceForensics++ [38] datasets.

Dataset	EfficientNet_b4 [43]	ResNet_50 [20]	XceptionNet [6]	F^3-Net [5]	LRL [36]
DFDC [9]	91.1%	78.7%	88.0%	69.8%	90.4%
FaceForensics++ [37]	94.3%	89.1%	92.7%	88.8%	98.2%

Table 2. The attack success rate of fake faces on spatial-based models on the DFDC [9] dataset.

Model	Attack	Eff_b4 [43]	Res50 [20]	Xcep [6]
Eff_b4 [43]	FGSM	33.2%	7.1%	2.3%
	PGD	77.7%	8.7%	1.8%
	Ours	**97.1%**	**20.1%**	**2.7%**
Res50 [20]	FGSM	0.0%	36.7%	0.9%
	PGD	0.0%	85.4%	0.0%
	Ours	**23.2%**	**87.8%**	**24.1%**
Xcep [6]	FGSM	0.0%	8.4%	45.6%
	PGD	0.0%	10.1%	72.3%
	Ours	**1.2%**	**14.3%**	**77.5%**

Table 3. The attack success rate of fake faces on spatial-based models on the FaceForensics++ [37] dataset.

Model	Attack	Eff_b4 [43]	Res50 [20]	Xcep [6]
Eff_b4 [43]	FGSM	38.7%	4.8%	0.9%
	PGD	71.6%	1.3%	0.3%
	Ours	**83.2%**	**22.7%**	**1.4%**
Res50 [20]	FGSM	3.2%	32.0%	2.1%
	PGD	3.9%	60.2%	2.3%
	Ours	**41.4%**	**65.4%**	**49.6%**
Xcep [6]	FGSM	1.1%	4.1%	18.9%
	PGD	1.1%	7.7%	61.6%
	Ours	**1.5%**	**8.5%**	**70.5%**

FaceForensics++ [37] is a popular dataset containing real videos from YouTube and corresponding fake videos, consisting of Deepfake [46], Face2Face [45], FaceSwap [25] and NeuralTextures [44]. We totally choose 560 (140×4) individual frames from each fake face video.

Models. For the spatial-based face forgery detectors, we choose three spatial-based classification networks, i.e., EfficientNet_b4 [43], ResNet_50 [20], and XceptionNet [6]. For the frequency-based models, we consider the state-of-the-art face forgery detectors, i.e., F^3-Net [36] and LRL [5]. All these models are trained by following the corresponding papers. The accuracy of these models on the selected images from different datasets are summarized in Table 1.

Evaluation metrics. For DFDC and FaceForensics++, we both choose the attack success rate as the evaluation metric. It is defined as the proportion of successfully attack images in all images that are classified as fake faces, i.e., $\frac{1}{N}\sum_{n=1}^{N} f(X^{\text{adv}}, \theta) \neq f(X^{\text{init}}, \theta)$. For the image quality assessment, we utilize MSE, PSNR and SSIM as the evaluation metrics to present the difference between the generated adversarial example and the original image.

Implementation details. The input size of images for three spatial-based models is 320×320×3. And the input sizes for F^3-Net [36] and LRL [5] are 299×299×3 and 320×320×3, respectively. We resize the adversarial examples to the corresponding size for the transfer attack. As for the parameters for attacks, we set the maximum perturbation of each pixel to be $\epsilon = 0.1$ for both FGSM and PGD. We also use PGD as the spatial attack for the hybrid attack.

4.2. Attack on Spatial-based Models

We compare the proposed method with FGSM [15] and PGD [31] on attacking spatial-based models. Table 2 and Table 3 report the attack success rates on the DFDC [9] and FaceForensics++ [37] datasets, respectively. We consider the basic classifiers in the first column to generate the adversarial examples and transfer them on the other networks to evaluate. The diagonal blocks indicate white-box attacks, while the off-diagonal blocks indicate their transferablility as black-box attacks. As Table 2 and Table 3 report, the proposed method outperforms FGSM and PGD for the white-box attack and gains higher attack success rates for the black-box attack. For instance, the adversarial examples generated by Res50 with our method get success rates of 41.4% on Eff_b4 and 49.6% on Xcep on the FaceForensics++ [37] dataset, which is 38.2% higher than FGSM and 37.5% higher than PGD on Eff_b4, and 47.5% greater than FGSM and 47.3% higher than PGD on Xcep. It suggests that the proposed method combined with the frequency attack enhances the transferability of adversarial examples. Due to the obvious diversity of structure between Eff_b4 and Xcep, the adversarial attacks between two networks have limited transferablility on each other.

4.3. Ensemble Attack on Spatial-based Models

As stated in [51], the adversarial examples with an ensemble of multiple networks achieve much stronger attack performance. We utilize an ensemble of two networks to attack the other one in three ways: ensemble in pixel, ensemble in loss, and ensemble in logits. The attack results on two datasets are summarized in Table 4 and Table 5, respectively. We consider all three networks and the sign '-' in the first column indicates the network not used during attacks. Thus, the diagonal blocks indicate transfer attacks (i.e., black-box setting) and the off-diagonal blocks indicate the white-box attacks. From both datasets, we observe that the ensemble in logits performs the strongest attack perfor-

Table 4. The attack success rate of fake faces with ensemble attacks on the DFDC [9] dataset.

Model	Ens.	Eff_b4 [43]	Res50 [20]	Xcep [6]
-Eff_b4 [43]	Pixel	**4.1**%	86.9%	44.4%
	Loss	2.3%	72.0%	57.3%
	Logit	2.4%	**88.5**%	**77.5**%
-Res50 [20]	Pixel	79.7%	20.4%	28.5%
	Loss	71.1%	18.1%	46.2%
	Logit	**93.1**%	**22.1**%	**65.3**%
-Xcep [6]	Pixel	72.4%	86.2%	13.1%
	Loss	76.4%	75.8%	**14.7**%
	Logit	**95.4**%	**95.6**%	12.7%

Table 5. The attack success rate of fake faces with ensemble attacks on the FaceForensics++ [37] dataset.

Model	Ens.	Eff_b4 [43]	Res50 [20]	Xcep [6]
-Eff_b4 [43]	Pixel	23.1%	64.0%	71.1%
	Loss	27.0%	52.9%	**79.1**%
	Logit	**27.5**%	**69.5**%	68.6%
-Res50 [20]	Pixel	77.3%	26.6%	49.1%
	Loss	55.2%	21.8%	57.9%
	Logit	**78.0**%	**29.8**%	**76.1**%
-Xcep [6]	Pixel	78.9%	**64.6**%	30.7%
	Loss	**83.4**%	57.3%	36.2%
	Logit	82.0%	64.4%	**38.2**%

mance in most cases. In the DFDC [9] dataset, when attacking on Res50 network, the ensemble in logits of Res50 and Eff_b4 obtains a 7.8% higher than the single network Res50 under the white-box setting. Besides, for the FaceForensics++ [37] dataset, the ensemble in logits of Res50 and Eff_b4 achieves a 38.2% success rate on Xcep, while the single network Res50 only gets an 8.5% success rate under the black-box setting. To sum up, an ensemble of different models can increase the diversity of structures, leading to a greater transferability to other models.

4.4. Attack on Frequency-based Models

The proposed hybrid attack is related to the frequency domain. To further illustrate its effectiveness, we also select two frequency-based face forgery detection methods, i.e., F^3-Net [5] and LRL [36]. Both methods collect the frequency information to distinguish the diversity between the real and fake faces to detect. Table 6 and Table 7 report the attack results on the DFDC [9] and FaceForensics++ [37] datasets, respectively. For better comparison, we also conduct the experiments of two detectors with FGSM [15] and PGD [31]. Moreover, we test the transferability of adversarial examples that our hybrid attack generates when attacking the spatial-based detectors. Briefly, our hybrid adversarial attack associated with the frequency domains achieves favorable white-box attacks in both datasets, where ∼90% of fake images are classified wrongly as real faces. For the

Table 6. The attack success rate of fake faces on frequency-based models on the DFDC [9] dataset.

Model	Attack	F^3-Net [5]	LRL [36]
F^3-Net [5]	FGSM	43.5%	9.6%
	PGD	97.6%	4.0%
	Ours	**98.7**%	**10.3**%
LRL [36]	FGSM	2.3%	71.3%
	PGD	3.0%	**100.0**%
	Ours	**5.5**%	**100.0**%
Eff_b4 [43]	Ours	7.4%	8.5%
Res50 [20]	Ours	**12.8**%	**43.6**%
Xcep [6]	Ours	7.6%	9.1%

Table 7. The attack success rate of fake faces on frequency-based models on the FaceForensics++ [37] dataset.

Model	Attack	F^3-Net [5]	LRL [36]
F^3-Net [5]	FGSM	24.8%	7.7%
	PGD	80.9%	28.7%
	Ours	**82.5**%	**36.2**%
LRL [36]	FGSM	0.2%	68.6%
	PGD	0.0%	98.7%
	Ours	**0.5**%	**99.3**%
Eff_b4 [43]	Ours	0.5%	11.8%
Res50 [20]	Ours	**7.1**%	**57.5**%
Xcep [6]	Ours	1.1%	19.5%

transfer attack, our method is marginally greater than the spatial attacks. When using the spatial-based detectors for transfer attacks, Res50 outperforms the other two networks with success attack rates of 12.8% for F^3-Net and 43.6% for LRL on DFDC [9], and 7.1% for F^3-Net and 57.5% for LRL on FaceForensics++ [37]. The proposed hybrid attack with the frequency domain strengthens the transferability of networks on the frequency-based models as well.

4.5. Ablation Study

We conduct a series of ablation studies on the proposed attack method. Due to the limited page, we only use Res50 [20] as the threat model to generate adversarial examples on the FaceForensics++ [37] dataset and show its transferability on Eff_b4 [43] and Xcep [6].

Variants of our method. We perform some variants of our method to analyze the effects on different domains. Concretely, we first only apply the spatial attack to craft adversarial examples. Then, we only implement the frequency attack. Besides, we simply sum the perturbations from the spatial and frequency domain. The concrete results are summarized in Table 8. While the sole spatial attack is prone to overfit the existing model and yields a limited transferability, the sole frequency attack always falls into local optimization and retains a fixed loss for binary classification, leading to a limited attack ability. For the combination of perturbations, the simple sum of perturbations

Table 8. The attack success rate of fake faces with variants of our method on the FaceForensics++ [37] dataset.

Method	Eff_b4 [43]	Res50 [20]	Xcep [6]
Spatial attack	3.9%	60.2%	2.3%
Frequency attack	14.5%	29.3%	15.9%
Sum attack	7.7%	61.2%	12.1%
Hybrid attack	**41.4%**	**65.4%**	**49.6%**

Table 9. The attack success rate of fake faces for attacking different frequency bands on the FaceForensics++ [37] dataset.

Frequency	Eff_b4 [43]	Res50 [20]	Xcep [6]
Low bands	39.5%	65.2%	49.1%
Middle bands	38.4%	62.7%	48.0%
High bands	36.2%	61.2%	47.1%
All bands	**41.4%**	**65.4%**	**49.6%**

Table 10. Quantitative evaluation of adversarial examples generated by FGSM [15], PGD [31] and our method on the FaceForensics++ [37] dataset.

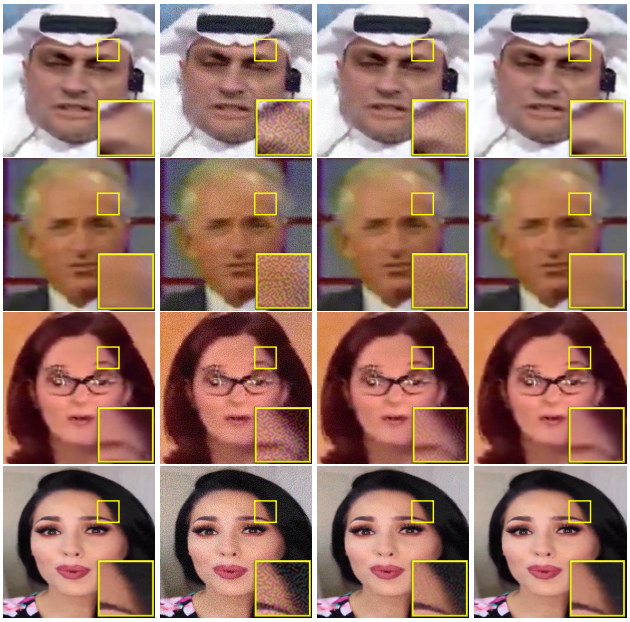
Attack method	MSE (↓)	PSNR (↑)	SSIM (↑)
FGSM	0.0279	23.3	0.0881
FGD	0.0238	30.4	0.1343
Ours	**0.0027**	**42.7**	**0.1763**

from two domains improves the performance of white-box attack but weakens its transferability. Our hybrid adversarial attack has more aggressive attack performance on both white-box and black-box attacks, which indicates that our method maintains both benefits from each domain and integrates them in a compatible way.

Frequency bands. To study the effect of different frequency bands, we divide the whole frequency band into three bands, i.e., low band, middle band, and high band, and only attack one of the bands with the hybrid adversarial attack. Table 9 reports the attack performance for different frequency bands. Compared with middle bands and high bands, the hybrid adversarial attack on low bands performs favorably under both white-box attacks (e.g., Res50 [20]) and black-box attacks (e.g., Eff_b4 [43] and Xcep [6]), since the low bands carry more content information of images and generate more perturbations in the spatial domain.

4.6. Image Quality Assessment

In order to illustrate the superior image quality by our method, we analyze the generated adversarial examples qualitatively and quantitatively. In Figure 5, we visualize adversarial examples crafted by FGSM [15], PGD [31] and our method. The adversarial examples by FGSM and PGD have obvious noise patterns when zooming in, while the ones generated by our method are more imperceptible to observers. In addition, we use the common metrics for image quality assessment to calculate the difference to the original images. Table 10 reports the quantitative results of MSE, PSNR and SSIM, where the image quality of our method outperforms other attacks by a large margin. It suggests that the proposed hybrid attack has a strong attack ability and maintains the image quality highly.



(a) Original (b) FGSM [15] (c) PGD [31] (d) Ours

Figure 5. Qualitative evaluation of adversarial examples generated by FGSM [15], PGD [31] and our method on the FaceForensics++ [37] dataset. These samples contain four types of face forgery generation, i.e., Deepfake, Face2Face, FaceSwap, and NeuralTextures. Although all adversarial examples fool the detectors as real faces successfully, the ones crafted by our hybrid adversarial attack obtain a superior image quality.

5. Conclusion

In this paper, we propose a frequency adversarial attack method for face forgery detection, which achieves a better image quality compared to the spatial attacks. To further improve its generalization, we propose a hybrid adversarial attack associated with the attacks both in the spatial domain and the frequency domain. The combination of multiple domains reserves their virtues and achieves favorable attack performance both on spatial-based and frequency-based face forgery detectors. Extensive experiments on two datasets indicate that the proposed method not only attacks the white-box models successfully but also enhances the transferability on other models under black-box settings. We hope that our work can draw more attention to the robustness of face forgery detectors.

Acknowledgements. This work was supported by NSFC (61906119, U19B2035), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and CCF-Tencent Open Research Fund.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE WIFS*, 2018.

[2] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM TOG*, 2017.

[3] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *ECCV*, 2018.

[4] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *CVPRW*, 2020.

[5] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI*, 2021.

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[7] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *ACM MM*, 2017.

[8] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, 2020.

[9] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[10] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019.

[11] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[13] Apurva Gandhi and Shomik Jain. Adversarial perturbations fool deepfake detectors. In *IJCNN*, 2020.

[14] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *CVPR*, 2014.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[16] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. *arXiv preprint arXiv:2112.13977*, 2021.

[17] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *ACM MM*, 2021.

[18] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *AAAI*, 2022.

[19] Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[21] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *WACV*, 2021.

[22] Shuai Jia, Chao Ma, Yibing Song, and Xiaokang Yang. Robust tracking against adversarial attacks. In *ECCV*, 2020.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[25] Marek Kowalski. Faceswap. https://github.com/marekkowalski/faceswap, 2018.

[26] Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. Exploring adversarial fake images on face manifold. In *CVPR*, 2021.

[27] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *CVPR*, 2021.

[28] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *CVPR*, 2021.

[29] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, 2019.

[30] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, 2021.

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[32] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, 2020.

[33] YISROEL MIRSKY and WENKE LEE. The creation and detection of deepfakes: A survey. *arXiv preprint arXiv:2004.11138*, 2020.

[34] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. Adversarial threats to deepfake detection: A practical perspective. In *CVPRW*, 2021.

[35] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[36] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020.

[37] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. In *ICCV*, 2019.

[38] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

[39] Yash Sharma, Gavin Weiguang Ding, and Marcus A Brubaker. On the effectiveness of low frequency perturbations. In *IJCAI*, 2019.

[40] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Rongrong Ji, et al. Dual contrastive learning for general face forgery detection. *arXiv preprint arXiv:2112.13522*, 2021.

[41] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *CVPR*, 2021.

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[43] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[44] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG*, 2019.

[45] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[46] Matt Tora. Deepfakes. https://github.com/deepfakes/faceswap/tree/v2.0.0, 2018.

[47] Wenxuan Wang, Bangjie Yin, Taiping Yao, Li Zhang, Yanwei Fu, Shouhong Ding, Jilin Li, Feiyue Huang, and Xiangyang Xue. Delving into data: Effectively substitute training for black-box attack. In *CVPR*, 2021.

[48] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.

[49] Xinyao Wang, Taiping Yao, Shouhong Ding, and Lizhuang Ma. Face manipulation detection via auxiliary supervision. In *ICONIP*, 2020.

[50] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *CVPR*, 2017.

[51] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019.

[52] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *IJCAI*, 2021.

[53] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, 2021.