AI-generated image detectors are surprisingly easy to mislead... for now

Zihang Lyu, Jun Xiao, Cong Zhang and Kin-Man Lam
The Hong Kong Polytechnic University, Kowloon, Hong Kong
E-mail: {zihang.lyu, jun.xiao, cong-clarence.zhang}@connect.polyu.hk, enkmlam@polyu.edu.hk

***Abstract*—AI-generated image detectors, also known as fake image detectors, have demonstrated remarkable performance across different datasets and generators, achieving superior detection accuracy and generalizability. Considering the structures of existing AI-generated detectors, which perform binary classification, we propose a simple yet effective adversarial attack method, namely Binary Fast Gradient Sign Method (BFGSM), in this paper. We demonstrate that existing AI-generated image detectors are sensitive to subtle and imperceptible distortions, which raises serious safety risk for these models and makes them inadequate for real-world applications. Experimental results show that our proposed attack successfully misleads current AI-generated image detectors, reducing the attack distortion level by 7.72% with negligible impact on the misleading success rate.**

I. Introduction

As large language models and deep generative models have made great progress in recent years, artificial intelligence-generated content (AIGC) has become increasingly prevalent in people's daily lives. Consequently, the associated security and privacy issues have garnered significant attention from researchers, with generated image detection being one of the most critical aspects.

The detection of generated images initially focused on generative adversarial networks (GANs). Some studies first applied co-occurrence matrices [1] and frequency domain analysis [2]. Wang *et al.* [3] found that, with proper data augmentation and pre-processing, CNN-generated images can be easily spotted, and a detector trained on a single type of generator can generalize to different types of generated images. Recent studies [4], [5] further extend this idea into diffusion models with enhanced feature encoding and artifact representation. However, despite the fact that current generative image detectors have demonstrated progressively higher detection accuracy and precision, these methods typically simplify the detection problem into a binary classification task and employ straightforward classifiers. This approach introduces a significant security vulnerability: these detectors lack the ability to resist adversarial attacks.

Adversarial attacks [6], [7] play an important role in evaluating and enhancing the robustness of classifiers. These attacks involve manipulating the input data to deceive machine learning models. For image classification, a general strategy is to generate a small and unnoticeable perturbation to input images, leading classifiers to produce incorrect labels. Fig. 1 shows some adversarial examples generated by the Binary Fast Gradient Sign Method (BFGSM) proposed in this paper.

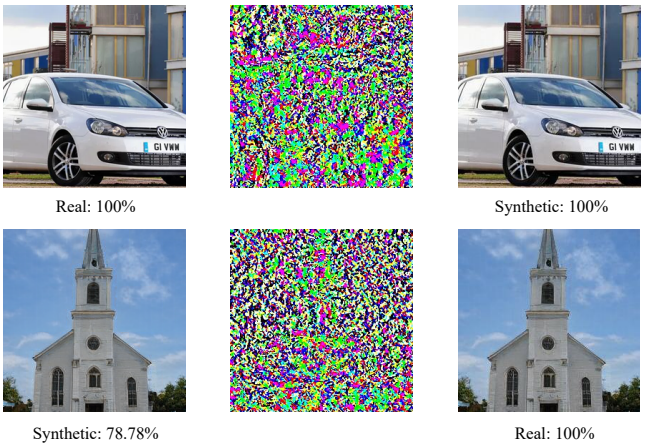

Fig. 1. Visualization of two adversarial examples generated by our proposed Binary Fast Gradient Sign Method (BFGSM): the **left column** represents the input image from real and fake datasets, respectively; the **middle column** shows the adversarial noise to be applied to the original image; the **right column** demonstrates the generated adversarial example. Common AI-generated image detectors are successfully misled with unnoticeable perturbations.

It clearly demonstrates that by adding a small noise that is unnoticeable to human eyes, the classifier can be easily misled. It is worth noting that for conventional classifiers, the security threat posed by adversarial attacks is relatively minor, as misleading these classifiers does not yield significant benefits. However, in the domain of generative image detection, evading the discriminator is a primary objective for malicious users, thereby imparting exceptional importance to adversarial attacks in this field.

In this paper, to address the research gap in adversarial attacks within the domain of generated image detection and to provide a paradigm for future studies, we first analyze the weaknesses of existing detectors when facing adversarial attacks and the underlying reasons for these vulnerabilities. Subsequently, we propose a simple yet effective adversarial attack method based on the traditional Fast Gradient Sign Method (FGSM), named BFGSM, to specifically mislead generated image detectors. This approach retains the simplicity of the traditional FGSM [9] structure while requiring only minimal modifications to the images. We aim for this method to provide a new baseline for misleading binary classification and to contribute to the development of more robust detectors in future research.

The main contributions of this paper are summarized as

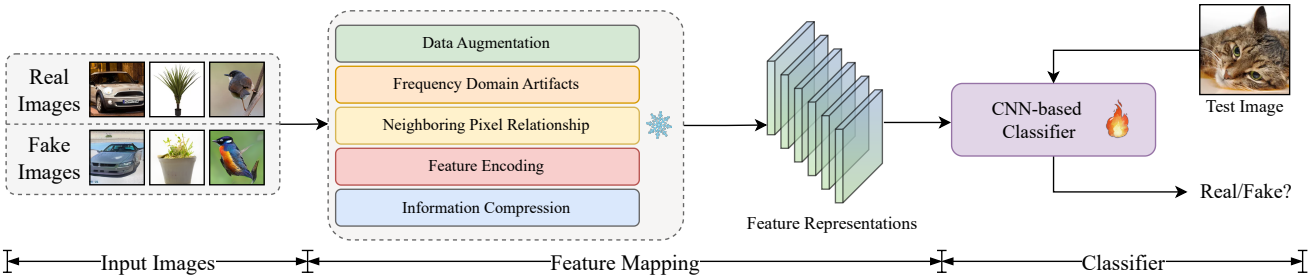

Fig. 2. Illustration of the overall framework of existing AI-generated image detectors: A variety of research studies have been conducted to generate feature representations that aim to distinguish generated images from real images, including data augmentation [3], frequency domain artifacts [2], neighboring pixel relationship [5], feature encoding [4], and information compression [8]. These feature mapping techniques are usually fixed in the training process, with only the CNN-based classifier being trainable. Such framework design makes it vulnerable to adversarial attack.

follows:

1) We introduce adversarial attacks to the field of AI-generated image detection and show that existing detectors can be easily misled with a simple adversarial attack method, which is an essential step for AI-generated image detectors in industrial-level practice.
2) We propose a novel Binary Fast Gradient Sigh Method that can mislead current AI-generated image detectors effectively.
3) Experimental results show that our proposed method can mislead existing detectors with smaller distortion compared to current attacking methods, demonstrating superior effectiveness.

II. Related Work

A. AI-generated Image Detection

While generated images can refer to a large number of images, the detection of AI-generated image mainly focuses on deep generative models, e.g. GANs and diffusion models, and some low-level vision tasks like image enhancement [10]–[12] and super-resolution [13]–[18]. This task initially focused on forged faces, known as deepfakes [19], and gradually tends to generalize fake image detection. Some methods are tailored for domain-specific generation, for example, frequency-level artifacts [2] for GAN images and DIRE [20] for Diffusion images. Other research aims for a universal fake image detector that trains the detectors with images from a single generator and is capable of detecting images generated from various sources. Wang *et al.* [3] initially proposed a simple yet effective detector that using data from ProGAN [21] and has the ability to generalize to detect images from other GAN models and some low-level vision tasks. Ojha *et al.* [4] further analyzed the detection of GAN and Diffusion images and proposed a feature representation for generated images, which extends the idea of universal fake image detectors to diffusion models.

B. Adversarial Attack

With the widespread application of deep neural networks [22]–[28], adversarial attacks have become a highly focused research area. The study of adversarial attacks began with Szegedy *et al.* [29], who discovered several intriguing properties of neural networks. They demonstrated, for the first time, that neural networks could be misled with slight distortions. The initial objective was to develop the equation of minimum distortion to mislead general classifier. However, due to high complexity, this objective was turned to find the minimum loss function addition that leads to misclassification. As the first approach, the proposed method constructs reliable adversarial examples, but with low efficiency and a complex algorithm structure. This idea was further enhanced through the Fast Gradient Sign Method (FGSM) [9], a gradient-based attack method that generates adversarial examples by computing the gradient of the loss function with respect to the input image and adding small perturbations in the direction of the gradient. This method is simple and fast, and some research also extends FGSM into iterative versions [30] and integrates a momentum term [31] to achieve a stronger attack effect. These attacking methods require full knowledge of the model structure and are known as white-box attack. Other research [32], [33] also focuses on black-box attacks where the attacker knows nothing about the model's internal information.

III. Methodology

A. Analysis of AI-generated image detectors

Revisiting the model structure of existing AI-generated image detectors, we find that they commonly focus on feature representation and data pre-processing. Fig. 2 illustrates the overall pipeline of these detectors as explored in various research studies. Researchers have experimented with various sophisticated methods to identify representations that differentiate between real and fake images. However, they have often overlooked the design of the classifier. Typically, a CNN-based network is chosen, sometimes incorporating a ResNet [34] block. This simplistic design leaves the detectors vulnerable to adversarial attacks. Additionally, the binary classification design represents another reason that existing detectors fail to resist adversarial attacks, which will be elaborated on in the following section.

B. Misleading AI-generated image detectors

We propose a simplified version of FGSM [9] targeted for binary classification, namely Binary FGSM (BFGSM), to

retain the original simplicity and success rate while reducing the distortion level applied to images. Suppose x and y are the original input image and label, θ represents the model parameters, the adversarial example $\bar{x}$ is generated by the traditional FGSM [9] as follows:

$$\bar{x} = x + \epsilon \cdot sign(\nabla_x(\mathcal{L}(\theta, x, y)), \tag{1}$$

where $\mathcal{L}$ represents the model loss function and $\nabla_x(\mathcal{L}(\theta, x, y))$ calculates the model gradient with respect to x. Here the perturbation level is controlled by the hyper-parameter ϵ. However, unlike multi-class classification problems, where multiple categories compete, misleading a binary classifier only requires the target class output to fall below a fixed threshold, typically set to 0.5 with the assumption that the input dataset is balanced. This means that when attacking binary classifiers, we can apply relatively smaller noise to samples where the model is less certain, without worrying about a decrease in success rate. In contrast to cases where the model is highly confident, for inputs near the threshold, even a very small perturbation can cause the model to make incorrect predictions. Specifically, it can be formulated as follows:

$$\min_{\alpha_x} \|x - \underbrace{(x + \alpha_x \cdot sign(\nabla_x(\mathcal{L}(\theta, x, y))))}_{\bar{x}} \|_2, \tag{2}$$

subject to $P(\bar{x}) = P(x)$ and $D(\bar{x}) \neq D(x)$, where $P(\cdot)$ denotes the data distribution, and $D(\cdot)$ denotes the AI-generated image detector. Intuitively, after adding the small perturbation, the data distribution is preserved but the AI-generated image detectors should produce inaccurate prediction labels. Therefore, in this paper, we consider the perturbation controlling term α_x to follow the distribution of classification confidence: for images that hold a high classification confidence, we apply a relatively larger distortion and vice versa, denoted as:

$$\alpha_x = \epsilon \cdot \frac{|\hat{y} - \gamma|}{z}, \tag{3}$$

where γ represents the classification threshold, $\hat{y}$ is the model output of x through a sigmoid activation function, and $z = max(|\hat{y} - \gamma|)$ is the normalizing term. The proposed BFGSM then generates the adversarial example as follows:

$$\bar{x} = x + \alpha_x \cdot sign(\nabla_x(\mathcal{L}(\theta, x, y)). \tag{4}$$

Let N represent the number of samples. We adopt the binary cross entropy as the loss function, represented as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}(y \log(\hat{y}) + ((1-y)\log(1-\hat{y}))). \tag{5}$$

Fig. 3 shows the framework of the proposed BFGSM. We can observe that the adversarial examples are produced dynamically with the model output.

IV. Experiments

A. Dataset and Implementation Details

Dataset Information. We evaluate our method on the UnivFD(UniversalFakeDetection) [4] dataset, which consists of various of generative models from GAN and diffusion models to some low-level vision tasks, including ProGAN [21], CycleGAN [35], BigGAN [36], StyleGAN [37], GauGAN [38], StarGAN [39], Deepfakes [19], SITD [10], SAN [13], CRN [40], IMLE [41], Guided Diffusion Model [42], LDM [43], Glide [44] and DALL-E [45]. Different variants of LDM on the number of steps and classifier-free diffusion guidance (CFG) are adopted. Glide has different settings for downsampling and upsampling.

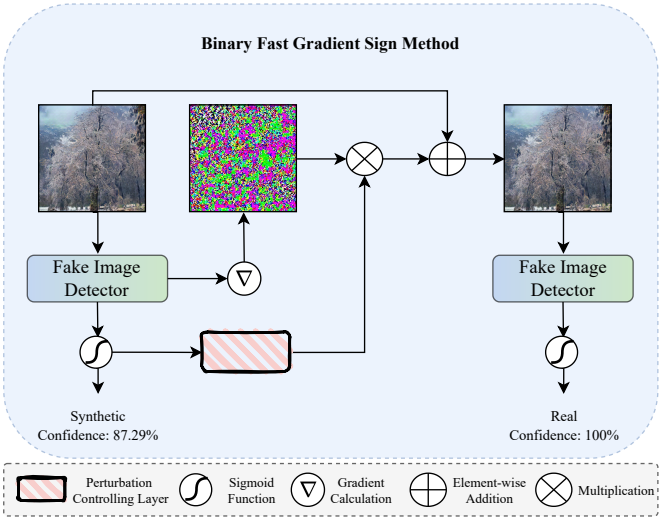

Fig. 3. Illustration of our proposed BFGSM. The perturbation controlling layer controls the magnitude of the adversarial noise, making it adaptive to the original image and corresponding target detector, thus reducing the perturbation level.

Evaluation Metrics. Following the target classifiers [3], [4], we select the classification accuracy and average precision as the evaluation metrics. To demonstrate the effectiveness of the proposed attacking method, we also adopt the L2 norm as a standard metric to evaluate the magnitude of the perturbations.

Implementation Details. We selected CNN Detection [3] and UnivFD [4] as targets for our attacks. The former method pioneered the field of AI-generated image detection through appropriate data augmentation, while the latter advanced the field with its feature extraction techniques. These two classifiers aptly represent the current state of the field. Both classifiers are trained on real and fake images from ProGAN [21] and have the ability to generalize to other generative models. All experiments were conducted on PyTorch on a single NVIDIA RTX 4090 GPU.

B. Effect of Binary Fast Gradient Sign Method

The accuracy and average precision (AP) of the target detectors before and after the attack are illustrated in Table I and Table II. It can be observed that with our proposed BFGSM, the performance of both target detectors has significantly declined, with an average decrease of 45.23% in accuracy and 49.28% in AP. The experimental results demonstrate the capability of our proposed attacking method to mislead classifiers, confirming that existing fake image detectors are indeed vulnerable to

TABLE I

THE ACCURACY OF THE TARGET DETECTORS FOR AI-GENERATED IMAGES ON THE UNIVFD DATASET BEFORE AND AFTER THE ATTACK. THE AVERAGE ACCURACY AFTER THE ATTACK IS HIGHLIGHTED IN BOLD.

Detection method	Adversarial attack	Generative Adversarial Networks						Deep fakes	Low level vision		Perceptual loss		Guided	LDM			Glide			DALL-E	Total
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		Avg. Acc.
CNN classifier [3]	✗	99.99	85.20	70.20	85.70	78.95	91.70	53.47	66.67	48.69	86.31	86.26	60.07	54.03	54.96	54.14	60.78	63.80	65.66	55.58	69.58
	✓	68.01	15.52	15.60	14.43	22.03	3.68	0.02	10.83	29.45	7.46	13.37	12.65	18.10	18.20	18.10	18.60	18.30	18.90	18.05	**17.96**
UnivFD [4]	✗	100.0	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.50	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
	✓	77.35	53.2	53.75	45.85	63.65	29.05	6.40	6.00	38.00	45.75	46.00	34.15	47.30	41.10	48.65	42.15	41.50	41.50	46.05	**42.49**

TABLE II

ILLUSTRATION OF THE ACCURACY OF DIFFERENT DETECTORS FOR AI-GENERATED IMAGES ON THE UNIVFD DATASET. THE BEST AND THE SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Detection method	Adversarial attack	Generative Adversarial Networks						Deep fakes	Low level vision		Perceptual loss		Guided	LDM			Glide			DALL-E	Total
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		mAP
CNN classifier [3]	✗	100.0	93.47	84.50	99.54	89.49	98.15	89.02	73.75	59.47	98.24	98.40	73.72	70.62	71.00	70.54	80.65	84.91	82.07	70.59	83.58
	✓	92.83	31.38	30.77	30.87	31.93	31.03	30.63	30.83	30.80	32.16	35.61	30.71	30.75	30.77	30.75	30.83	30.80	30.97	30.75	**34.48**
UnivFD [4]	✗	100.0	99.46	99.59	97.24	99.98	99.60	82.45	61.32	79.02	96.72	99.00	87.77	99.14	92.15	99.17	94.74	95.34	94.57	97.15	93.38
	✓	87.04	55.21	54.03	44.37	68.19	34.93	30.79	31.00	32.59	31.21	33.05	32.84	49.44	37.45	50.27	37.43	37.28	37.43	46.24	**43.73**

adversarial attacks. Moreover, unlike multi-class classifiers, adversarial attacks on binary classifiers only need to decrease their performance to chance levels rather than zero in order to significantly mislead them, as such reduced performance is practically meaningless. An interesting phenomenon is that generative models evaluated on the same dataset of real images always exhibit similar performance, as shown in Table I with examples, such as LDM, Glide, and DALL-E. This can be explained by the fact that images more resistant to adversarial attacks are concentrated in the real dataset. However, malicious users can easily mislead detectors by presenting generated images as real ones. This further illustrates the vulnerability of current detectors to adversarial attacks.

C. Effect of Dynamic Perturbation Controlling

Table III compares the misleading success rate of our proposed BFGSM with the traditional FGSM on the two target detectors, as well as the L2 distance between the generated adversarial examples and the original images. To mislead a binary classifier, it only needs to reduce the classification accuracy to chance performance. Therefore, we maintain a balance between the level of perturbation and the success rate to compare the effectiveness of different attack methods under the premise of successfully misleading the classifiers. It is clear that our proposed method significantly reduces the perturbations to the original images with almost no impact on the success rate. The noise level is reduced by 7.72% with only a negligible success rate decrease of 0.04%, demonstrating the superiority of our method. Such performance indicates that the proposed BFGSM is a more suitable adversarial attack method in the field of AI-generated image detection. It is sufficiently simple and efficient while significantly reducing the distortion level of the attack, making it more difficult to detect under the same level of perturbation and serving as a proper baseline for further adversarial attack and detector robustness evaluation.

TABLE III

ILLUSTRATION OF THE MISLEADING SUCCESS RATE (%) AND AVERAGE L2 DISTANCE BETWEEN ORIGINAL IMAGES AND THE ADVERSARIAL EXAMPLES GENERATED BY FGSM AND OUR PROPOSED BFGSM W.R.T THE TARGET DETECTORS ON THE UNIVFD DATASET.

Detection methods	CNN classifier [3]		UnivFD [4]	
	Success rate	L2 Distance	Success rate	L2 Distance
FGSM	74.20	2.29	47.86	1.94
BFGSM (Ours)	74.19	**2.22**	47.79	**1.70**

V. CONCLUSIONS

As AI-generated content continues to integrate into human society, the importance of AI-generated image detection in fields, such as security, privacy, and copyright, continues to increase. Despite the remarkable performance achieved by existing AI-generated image detectors, their vulnerability to adversarial attacks allows malicious users to easily bypass them, hindering their true industrial-level application. In this paper, we first analyze the structure of general detectors and point out that the lack of classifier design and the nature of binary classification are the reasons for existing shortcomings. Subsequently, we propose a simple yet effective adversarial attack method tailored for AI-generated image detection, namely the Binary Fast Gradient Sign Method (BFGSM). This method utilizes an adaptive controlling term that changes dynamically with the classification confidence of the input image with respect to the target detector, applying only the necessary noise level to mislead each specific image. Experimental results demonstrate that the proposed BFGSM can easily mislead existing classifiers with minimal perturbations, showcasing the effectiveness and superiority of our method. This study establishes a new research paradigm for adversarial attacks in the field of AI-generated image detection and provides a new adversarial attack baseline for misleading binary classification. We anticipate it will contribute to the development of more robust detectors in future studies.

REFERENCES

[1] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, *et al.*, "Detecting GAN generated fake images using co-occurrence matrices," *Electronic imaging*, 2019.

[2] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in *WIFS*, 2019.

[3] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *CVPR*, 2020.

[4] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *CVPR*, 2023.

[5] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection," in *CVPR*, 2024.

[6] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *ICPR*, 2021.

[7] T. Zheng, C. Chen, and K. Ren, "Distributionally adversarial attack," in *AAAI*, 2019.

[8] Z. Lyu, J. Xiao, C. Zhang, and K.-M. Lam, "AI-generated image detection with Wasseerstein distance compression and dynamic aggregation," in *ICIP*, 2024.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[10] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018.

[11] J. Xiao, Q. Ye, T. Liu, C. Zhang, and K.-M. Lam, "Deep progressive feature aggregation network for multi-frame high dynamic range imaging," *Neurocomputing*, vol. 594, p. 127 804, 2024.

[12] J. Xiao, Z. Lyu, C. Zhang, Y. Ju, C. Shui, and K.-M. Lam, "Towards progressive multi-frequency representation for image warping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2995–3004.

[13] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019.

[14] J. Xiao, Q. Ye, R. Zhao, K.-M. Lam, and K. Wan, "Deep multi-scale feature mixture model for image super-resolution with multiple-focal-length degradation," *Signal Processing: Image Communication*, vol. 127, p. 117 139, 2024.

[15] J. Xiao, Q. Ye, R. Zhao, K.-M. Lam, and K. Wan, "Self-feature learning: An efficient deep lightweight network for image super-resolution," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4408–4416.

[16] C. Yang, J. Xiao, Y. Ju, G. Qiu, and K.-M. Lam, "Improving robustness of single image super-resolution models with monte carlo method," in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 2135–2139.

[17] C. Yang, R. Dong, J. Xiao, *et al.*, "Geometric distortion guided transformer for omnidirectional image super-resolution," *arXiv preprint arXiv:2406.10869*, 2024.

[18] J. Xiao, W. Jia, and K.-M. Lam, "Feature redundancy mining: Deep light-weight image super-resolution model," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1620–1624.

[19] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, 2019.

[20] Z. Wang, J. Bao, W. Zhou, *et al.*, "Dire for diffusion-generated image detection," in *CVPR*, 2023.

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.

[22] C. Zhang, K.-M. Lam, and Q. Wang, "Cof-net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023.

[23] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023.

[24] C. Zhang, K.-M. Lam, T. Liu, Y.-L. Chan, and Q. Wang, "Structured adversarial self-supervised learning for robust object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2024.

[25] R. Dong and K.-M. Lam, "Bi-center loss for compound facial expression recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 641–645, 2024.

[26] Y. Ju, M. Jian, J. Dong, and K.-M. Lam, "Learning photometric stereo via manifold-based mapping," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, 2020, pp. 411–414.

[27] Y. Ju, B. Shi, Y. Chen, H. Zhou, J. Dong, and K.-M. Lam, "Gr-psn: Learning to estimate surface normal and reconstruct photometric stereo images," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[28] Y. Ju, K.-M. Lam, W. Xie, H. Zhou, J. Dong, and B. Shi, "Deep learning methods for calibrated photometric stereo and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[29] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *ICLR*, 2014.

[30] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, 2018.

[31] Y. Dong, F. Liao, T. Pang, *et al.*, "Boosting adversarial attacks with momentum," in *CVPR*, 2018.

[32] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017.

[33] Y. Dong, H. Su, B. Wu, *et al.*, "Efficient decision-based black-box adversarial attacks on face recognition," in *CVPR*, 2019.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[36] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *ICLR*, 2018.

[37] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[38] J. F. O'brien and H. Farid, "Exposing photo manipulation with inconsistent reflections," *ACM Transactions on Graphics*, 2012.

[39] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018.

[40] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *ICCV*, 2017.

[41] K. Li, T. Zhang, and J. Malik, "Diverse image synthesis from semantic layouts via conditional imle," in *ICCV*, 2019.

[42] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *NeurIPS*, 2021.

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[44] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *ICML*, 2022.

[45] A. Ramesh, M. Pavlov, G. Goh, *et al.*, "Zero-shot text-to-image generation," in *ICML*, 2021.