Generalizable AI-Generated Image Detection Based on Fractal Self-Similarity in the Spectrum

Shengpeng Xiao[1] spxiao@buaa.edu.cn
Yuanfang Guo[1*] andyguo@buaa.edu.cn
Heqi Peng[1] penghq@buaa.edu.cn
Zeming Liu[1] zmliu@buaa.edu.cn
Liang Yang[2] yangliang@vip.qq.com
Yunhong Wang[1] yhwang@buaa.edu.cn
Beihang University[1]
Hebei University of Technology[2]

Abstract

The generalization performance of AI-generated image detection remains a critical challenge. Although most existing methods perform well in detecting images from generative models included in the training set, their accuracy drops significantly when faced with images from unseen generators. To address this limitation, we propose a novel detection method based on the fractal self-similarity of the spectrum, a common feature among images generated by different models. Specifically, we demonstrate that AI-generated images exhibit fractal-like spectral growth through periodic extension and low-pass filtering. This observation motivates us to exploit the similarity among different fractal branches of the spectrum. Instead of directly analyzing the spectrum, our method mitigates the impact of varying spectral characteristics across different generators, improving detection performance for images from unseen models. Experiments on a public benchmark demonstrated the generalized detection performance across both GANs and diffusion models.

1. Introduction

In recent years, AI image generation technology has made significant strides, with the emergence of a variety of diverse image generation models. Generative Adversarial Networks (GANs) [9], diffusion models (DMs) [30] and their variants [3, 4, 6, 10, 14–16, 24, 37] are the most effective motheds. These generative models are capable of producing highly realistic images from text description or other reference image(s). While the AI generation technology significantly enhances the effectiveness of image content creation, it may also be maliciously utilized to produce deep image forgeries, such as Deepfake [28]. These deep image forgeries may induce significant security and ethical issues for the government, society, and individuals.

To address this issue, extensive research on AI-generated image detection has been conducted [13, 17–19, 23, 32, 33, 36]. Some methods identify the generated images/videos with the help of certain prior knowledge, such as detecting deep facial forgeries (a.k.a. deepfakes) based on anomalous facial features [11, 20]. Unfortunately, these methods are only applicable to specific categories of images.

To effectively identify the generated images containing various content, certain approaches focus on the low-level features of generated images [2, 5, 7, 8, 19, 35]. Before diffusion model emerges, earlier methods usually focused on studying the texture and spectral artifacts in the GANs-generated images [7, 8, 19, 35]. As the continuously development of diffusion model, forensic researchers also paid attentions to the detection of diffusion-generated images. They observed that noise residuals in diffusion-generated images still exhibited low-level features similar to GANs [2, 5]. By leveraging these low-level features, existing AI-generated image detection techniques can effectively identify GAN-based and diffusion based generated images across various semantic categories.

However, while the existing detection methods can detect images generated by various generators, their performances tend to less satisfactory when the fake images are produced by unseen image generation techniques. As shown in a recently proposed benchmark AIGCDetect [36], most of the state-of-the-art detection methods [8, 13, 18, 23, 31–33] suffer obvious performance drops when processing the images generated by models not included in the training set. Some detectors [17, 36] demonstrate decent generalization performances. Unfortunately, they either directly extract spectral features from the image noises or adopt the inconsistent inter-pixels relationships bewteen the rich and poor textural image patches. Under such circumstance, the extracted features tend to be highly correlated to the param-

eters of the generators, which limits the generalization performance.

In this work, we propose a novel generalizable AI-generated image detection method, by leveraging the spectral fractal self-similarity feature, to further improve the generalization performance against the unseen generators. Although existing studies have observed that the spectral artifacts universally exist in generated images from diverse generators [8, 32], the generator-related variations of spectral artifacts limit their generalization ability against different generator architectures. To address this limitation, we further analyze the formation of spectral artifacts and propose to model the fractal self-similarity structure of the artifacts. Specifically, the spectrum of feature map is repeatedly duplicated by upsample operations during the generation process, which results in the fractal structure of the spectrum and artifacts. While features of the artifacts vary with the parameters of the generators, the self-similarity of the artifact remains consistent, which is only related to the ratio of upsampling. Therefore, we proposed Fractal CNN, an fractal-structured network to capture the self-similarity feature from AI-generated image detection. Rather than directly analyzing the spectral artifacts, our method could mitigate the impact of the variations of the generators, and thus increase the generalization performance.

Our main contributions can be summarized as follows:

- We discover the fractal structure in the spectrum of AI-generated images, which provides a more detailed explanation to the spectral artifacts in generated images.
- We propose a spectral fractal self-similarity model to explicitly describe the spectral artifacts, which are uncorrelated to the parameters of the generator and thus avoid the generalization issue caused by variations in the spectral artifact of different generative models.
- We construct a fractal-structured convolution neural network to better capture the self-similarity feature for AI-generated image detection.
- Experiments demonstrate that we significantly outperform the existing state-of-the-art methods in detecting images generated by unseen models.

2. Related Work

Image Generation Deep learning-based image generation learns the probability distribution of real images and generates highly realistic images. Here we introduce some mainstream image generation techniques.

Generative Adversarial Network (GANs) [9] adversarially train convolution neural networks(CNNs) to generate images. ProGAN [14] utilizes cascaded units with nearest-neighbor upsampling, convolution, and pixel-wise normalization. StyleGAN [15] enhances ProGAN by introducing Gaussian noise for randomness and adaptive instance normalization (AdaIN) to control image style, while StyleGAN2 [16] further refines the architecture. BigGAN [3] uses residual blocks (ResBlocks) as basic units, and CycleGAN [37] and StarGAN [4] adopt U-Net [27] with instance normalization.

Diffusion Models(DMs) differ from GANs by incorporating Gaussian noise into backbone outputs, effectively adding a residual structure. Despite this, their architectures are fundamentally similar, both relying on visual deep learning models. For example, DDPM [12] and Stable Diffusion [6] use U-Net [27] as denoising networks, where Stable Diffusion integrates cross-attention for conditional image generation. DALL-E [25] and Glide [22] employ Transformer-based ViT [29] as noise prediction models. These backbones share common operations like upsampling.

AI-generated Image Detection AI-generated image detection researches aim at identifying images synthesized by generative models. Some methods adopt prior knowledge and the inconsistent semantic information in the image for generated image detection. For example, in the filed of facial forgeries detection, [11] leverages irregular pupil shapes to detect fake faces, and [20] exploits visual artifacts like lighting inconsistencies. These methods are only suitable for detecting generated images of specific categories, which limits their application.

To detect generated images of more semantic categories, approaches based on low-level artifacts are widely researched. [13] combined global and local image information for detection. [19] observed the generator-related noise patterns in GANs-generated images. [35] adopts a simulation generator to introduce GANs-generated artifacts, and trained detection models in a self-supervised manner. [8] discovered that different upsampling operations introduce spectral artifacts and proposed to detect generated images in frequency domain. [18] enhanced generalizability by adding global texture feature extraction units to ResNet. [32] used data augmentation improve the generalization, and succeed to detect images generated by different GANs. After diffusion models outperform GANs, related detectors are also proposed. [33] uses reconstruction errors as detection criteria. [2, 5] revealed that GANs and diffusion models have regular autocorrelation patterns in the spectrum of their noise residual.

These detection methods have achieved high performance on detecting images generated by specific generator, but their performance decreases significantly faced with unseen generators. For better real-world application, researches for detection images from unseen generator have been proposed. LNP [17] adopts the noise extracted by a denoising network as the fingerprint to improve the generalizability. [36] utilized the differences in pixel-correlations between rich-textured and poor-texture regions, improving the generalization performance. But the these methods ex-

tract features related to the parameter of generators. As generator become more and more diverse, their generalization performance is still limited.

3. Methodology

In this section, we will introduce the fractal self-similarity in the spectrum of AI-generated images, and demonstrate how this feature is applied to increase the generalization performance on detecting images from unseen generators.

3.1. Fractal Spectrum of AI-generated Images

To increase the generalization performance for detecting images from unseen generators, it is of significant importance to capture a common feature across different generator architectures. Previous works have widely discovered spectral artifacts in images generated by different GANs [8, 35]. With the advent of diffusion models, similar periodic spectral artifacts are observed in the noise residuals of DM-generated images [2, 5]. It is suggested that the generation of spectral artifacts is associated with upsampling [8], a common operation in the backbone of various existing generative models. Therefore, spectral artifacts could to be used as a generalizable feature for AI-generated image detection.

However, while spectral artifacts exist widely in the spectrum of images generated by different generators, the specific feature of the artifacts varies. As shown in Fig. 1, the average spectrum of images of different generators and their corresponding real images. This fact indicates the reason why directly adopting the spectral artifacts feature has limitation in generalizable AI-generated image detection.

To address this limitation, we performed a further analysis of the fine-grained structure of the spectral artifacts. Before presenting our analysis results, we first give the definition of the image spectrum. Given input image I with resolution of $M \times N$, the spectrum of I is represented as

$$\mathcal{F}(I)[u,v] = \left| \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I[x,y] e^{-j2\pi(\frac{ux}{M}+\frac{vy}{N})} \right|, \quad (1)$$

where $x, u \in \{0, \cdots, M-1\}, y, v \in \{0, \cdots, N-1\}$, and only the magnitude of the spectrum is preserved. We then analyze the primary cause of spectral artifacts of AI-generated images—the upsample operations—in the perspective of spectrum. We view the upsampled image as the Direct Sequence Spread Spectrum(DSSS) signal, i.e., the signal is multiplied by a noise code, spreading the data signal over a larger frequency range. For AI-generated images, the noise code is the parameters of the upsample module such as transposed convolution.

In this perspective, the impact of upsample operations on the spectrum of images, or feature maps, is equivalent to copying the original spectrum to a wider band, i.e., the spectrum is periodically extended and subjected to certain post-processing. To better approximate the real image spectrum, this post-processing usually takes the form of low-pass filtering. According to the convolution theorem, convolution in the spatial domain equals multiplying in the spectral domain. Therefore, this low-pass filtering multiplies the filter kernel $\mathcal{K}_n$ and the spectrum of feature map element-wise in the spectral domain. This specific filter kernel is related to upsample method. For linear upsmaple methods, such as interpolated zeros, nearest neighbors, the filter kernel $\mathcal{K}_n(I_n)$ is fixed, i.e., each element of its spectrum is constant. For AI generated images, the filtering is performed using nonlinear transposed convolution and convolution layers. In this case, the corresponding $\mathcal{K}_n(I)$ is aa function of the input image or feature map I. Therefore, the spectrum of the AI generated image can be expressed as follows,

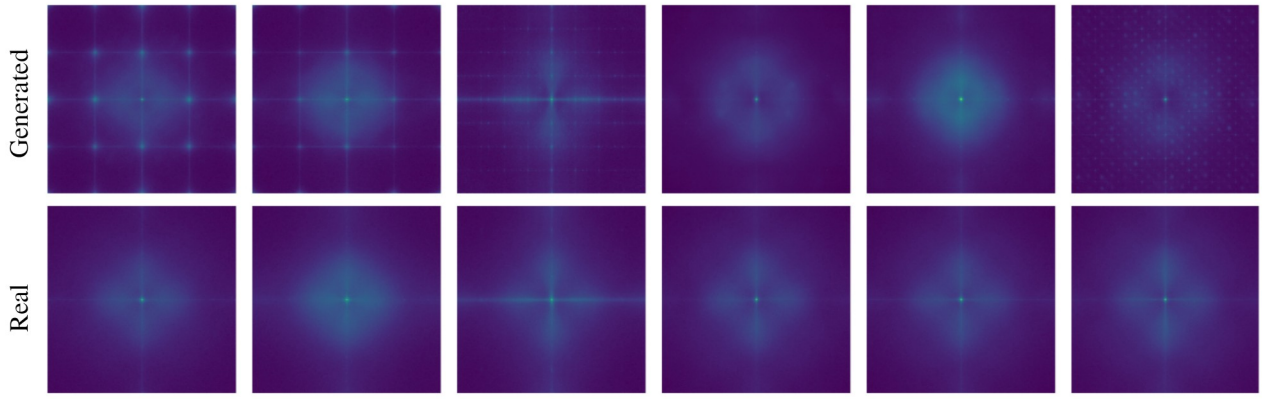
$$\mathcal{F}(I_{n+1}) = \mathcal{K}_n(I_n) \odot \begin{bmatrix} \mathcal{F}(I_n) & \mathcal{F}(I_n) \\ \mathcal{F}(I_n) & \mathcal{F}(I_n) \end{bmatrix}, \quad (2)$$

where $\mathcal{F}(I_{n+1})$ denoted the upsampled image or feature map, $\mathcal{K}_n(I_n)$ denoted the low-pass filter in spectral domain. Considering the limited size of the kernels in convolution neural networks, the inconsistent spectral traces in the high-frequency cannot be completely removed. As the energy in the low-frequency region becomes more and more concentrated with the effectiveness of the low-pass filters, traces in the high-frequency areas become more and more difficult to remove. These traces finally become the spectral artifacts of AI-generated images.

Fig. 2 visualizes the formation process of the fractal-structured spectrum of AI-generated images. $\mathcal{F}(I_0)$ is the spectrum of a real image with a watermark in the shape of letter 'A' added to $\mathcal{F}(I_0)$ for better demonstrating the process of spectrum replication. The origin spectrum $\mathcal{F}(I_0)$ is upsampled with 3 different methods for 3 times. The different up-sampling methods adopted are zero-interpolation, nearest-neighbor and transposed convolution, respectively. Two convolution layers are added after transposed convolution to introduce nonlinearity, with their kernel weights initialized randomly. It could be observed that the spectrum of the origin image gradually forms a self-similar fractal structure through continuous self-replication and low-pass filtering in multiple up-sampling.

3.2. Generalizable Fractal Self-Similarity Feature

Having revealed the formation progress of the spectral artifacts and discovered its self-similarity fractal structure, we propose using the spectral fractal self-similarity feature to improve the generalizability of AI-generated image detection. Although the spectral artifacts of AI-generated images vary with the generator's architecture, parameters, their self-similarity structure is commonly shared. Hence,



(a) CycelGAN [37] (b) StarGAN [4] (c) StyleGAN [15] (d) SDv1.4 [25] (e) Midjurney [21] (f) VQDM [10]

Figure 1. **The average spectrums of images generated by different models and corresponding real images.** While spectral artifacts widely existing in the spectrum of images generated by different generators, the specific feature of the artifacts varies.

by assuming the spectrum of AI-generated images has commonly higher self-similarity feature, we propose a detection method for images from unseen generators.

We use a simple method to extract the spectral fractal self-similarity features-fusing the feature of different fractal branches by element-wise multiplication, and recursively analyzing each fractal in the same manner. Specifically, we first extract the noise residual from the origin image. Then the spectrum of the noise residuals calculated with FFT, is split into 4 fractal branches,

$$\mathcal{F}(I) = \begin{bmatrix} \mathcal{F}(I)_{00} & \mathcal{F}(I)_{01} \\ \mathcal{F}(I)_{10} & \mathcal{F}(I)_{11} \end{bmatrix}, \tag{3}$$

where $\mathcal{F}(I)_{00}$, $\mathcal{F}(I)_{01}$, $\mathcal{F}(I)_{10}$ and $\mathcal{F}(I)_{11}$ are the different fractal branches of the image spectrum. From the analysis in the previous section, we have conclude that these 4 branches originate from the self-replication of the same original spectrum in the process of image generation. Therefore, we believe that there is some positive or negative correlation between these 4 branches of AI generated images. We measure the relevance of these 4 fractal branches $S[\mathcal{F}(I)]$ by element-wise multiplication and a measure function d,

$$S[\mathcal{F}(I)] = d(\mathcal{F}(I)_{00} \odot \mathcal{F}(I)_{01} \odot \mathcal{F}(I)_{10} \odot \mathcal{F}(I)_{11}). \tag{4}$$

According to Eq. (2), $\mathcal{F}(I)_{00}$, $\mathcal{F}(I)_{01}$, $\mathcal{F}(I)_{10}$ and $\mathcal{F}(I)_{11}$ of AI-generated images are generated from the same low-resolution spectrum. Therefore, these 4 sub-spectrum of AI-generated images should be similar to each other, and $S(\mathcal{F}(I)$ will be different from real images. Recursively, $S[\mathcal{F}(I)_{00}]$, $S[\mathcal{F}(I)_{01}]$, $S[\mathcal{F}(I)_{10}]$ and $S[\mathcal{F}(I)_{1}]$ could be calculated in the same way as each of them has the fractal self-similarity feature due to multiple upsample operations. With multi-level self-similarity feature, the detection performance could be improved.

3.3. Fractal Convolution Neural Network

To capture the proposed spectral fractal self-similarity feature more effectively, we designed Fractal-CNN, a fractal-structured convolution neural network. The framework of Fractal-CNN is shown in Fig. 3. The most important parts of our model are the Fractal Units, which capture multi-level self-similarity features $S = [S^{(0)}, ..., S^{(N)}]$ from the spectrum. The multi-level self-similarity features S will be adopted as the learned representation of AI-generated images, and be put into the full-connected layers for classification.

In addition, before the spectrum of images are put in to Fractal Units, a high-pass filter is exploited to strengthen the spectral artifacts for a more accurate analysis of the self-similarity . This filter first capture the noise residual of the image I_{res} by subtracting the blurred version from the origin image,

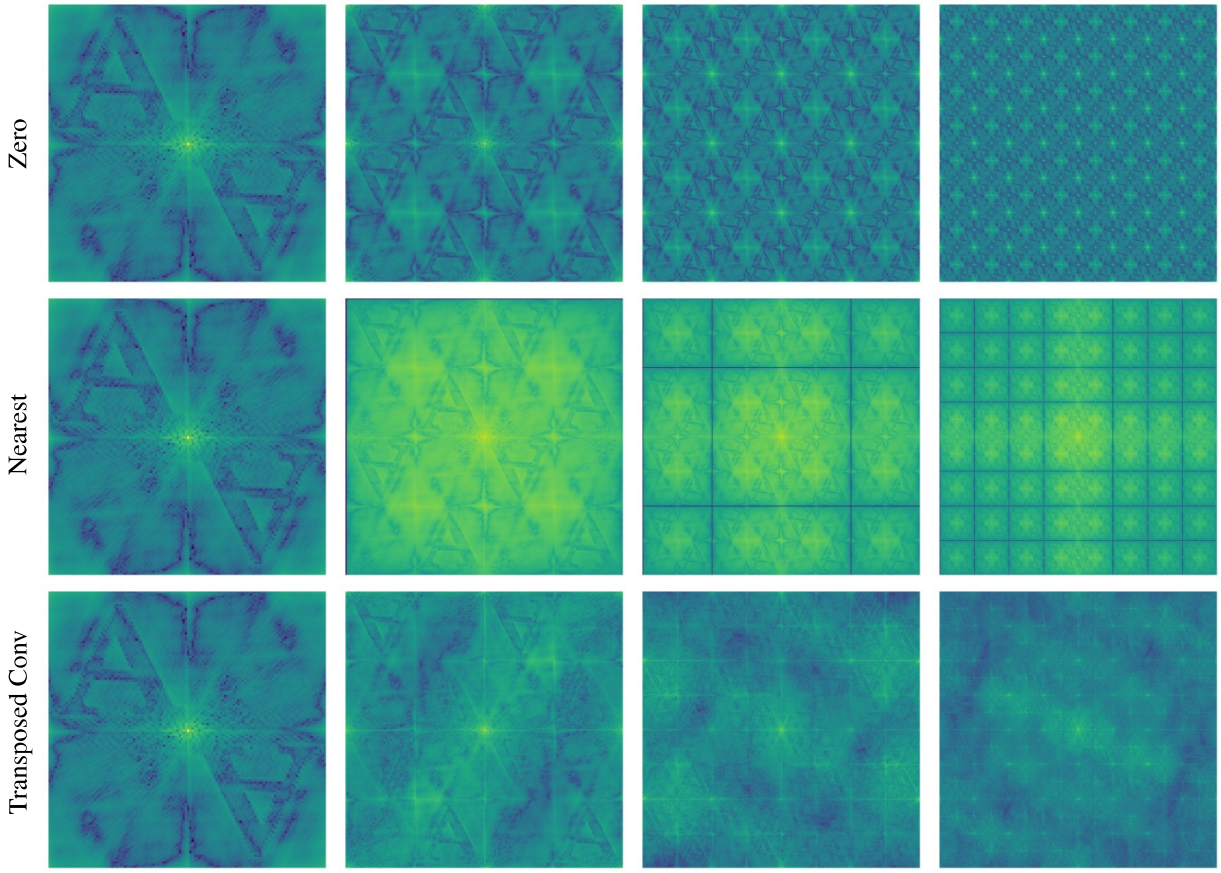
$$I_{res} = I - Blur(I), \tag{5}$$

where the blur filter is implemented with median blur following previous works [19, 32]. To strengthen the spectral artifacts in the spectrum of the noise residual, convolution layers are adopted in both spatial and spectral domain. Ultimately, the feature maps with enhanced spectral artifacts are put into Fractal Units for self-similarity analysis.

The Fractal Units are the core of our design, employing a recursive architecture to achieve efficient extraction of self-similar features. We refer to the spectral feature map extracted by the high-pass filtering as the zero-level spectrum H^0. And all levels of spectrums $H^{(n)}$ will be split into different fractal branch according to the method in Eq. (3),

$$H^{(n)} = \begin{bmatrix} H_{00}^{(n)} & H_{02}^{(n)} \\ H_{10}^{(n)} & H_{11}^{(n)} \end{bmatrix}. \tag{6}$$

As previously analyzed, these four fractal branches are



(a) Origin spectrum $\mathcal{F}(I_0)$. (b) Upsampled once $\mathcal{F}(I_1)$. (c) Upsampled twice $\mathcal{F}(I_2)$. (d) Upsampled three times $\mathcal{F}(I_3)$.

Figure 2. **The formation process of the fractal structure in the spectrum of AI-generated images.** Images in column (a) are the spectrum of the original image, embedded with a watermark in the shape of the letter 'A' for better visualization. Column (b), (c) and (d) represent the spectrum of the upsampled images. The upsampling method used in the first row is interpolation of zeros, the second row is nearest-neighbour upsampling, and the third row is non-lineaner transposed convolution and convolution which is widely used in image generation. It can be observed that the watermark 'A' replicates itself along with the spectrum and forms fractal-structured spectral artifacts.

generated by the same origin low-resolution spectrum in the spectrum of AI-generated image. They are simply replicated from the low-resolution spectrum and subsequent low-pass filtering. For each branch, we first use a convolution layer to further extract its features $\hat{H}_{ij}^{(n)}$. We then fuse the feature maps from the four branches into a single feature map by element-wise multiplication. Next, different from Eq. (4) which directly calculates the self-similarity value, we adopt convolution layers to transform the fused feature map into a vector $S^{(0)}$. This is to avoid mistaking plain real images like solid color images, which also have trivial self-similarity feature.

$$S^{(n)} = Conv\left(\hat{H}_{00}^{(n)} \odot \hat{H}_{01}^{(n)} \odot \hat{H}_{10}^{(n)} \odot \hat{H}_{11}^{(n)}\right). \quad (7)$$

To make full use of the multi-level fractal self-similarity feature, the fractal branches $\hat{H}_{ij}^{(n)}$ are recursively analyzed in the same manner. To avoid the number of the fractal branches growing exponentially, we compute the average spectrum of $H_{ij}^{(n)}$ to obtain the next-level spectrum,

$$H^{(n+1)} = \frac{1}{4}\left(H_{00}^{(n)} + H_{01}^{(n)} + H_{10}^{(n)} + H_{11}^{(n)}\right). \quad (8)$$

Using the same recursive process, we extract the self-similarity of each level of the fractal structure and derive the next level of the fractal. Finally, we concatenate the fractal self-similarities from different levels of the spectrum as $S^{(0)}, ..., S^{(N)}$ and pass them through a fully connected module for classification.

4. Experiment

In this section, we evaluate the detection performance on a public benchmark and analyze the effectiveness of the self-

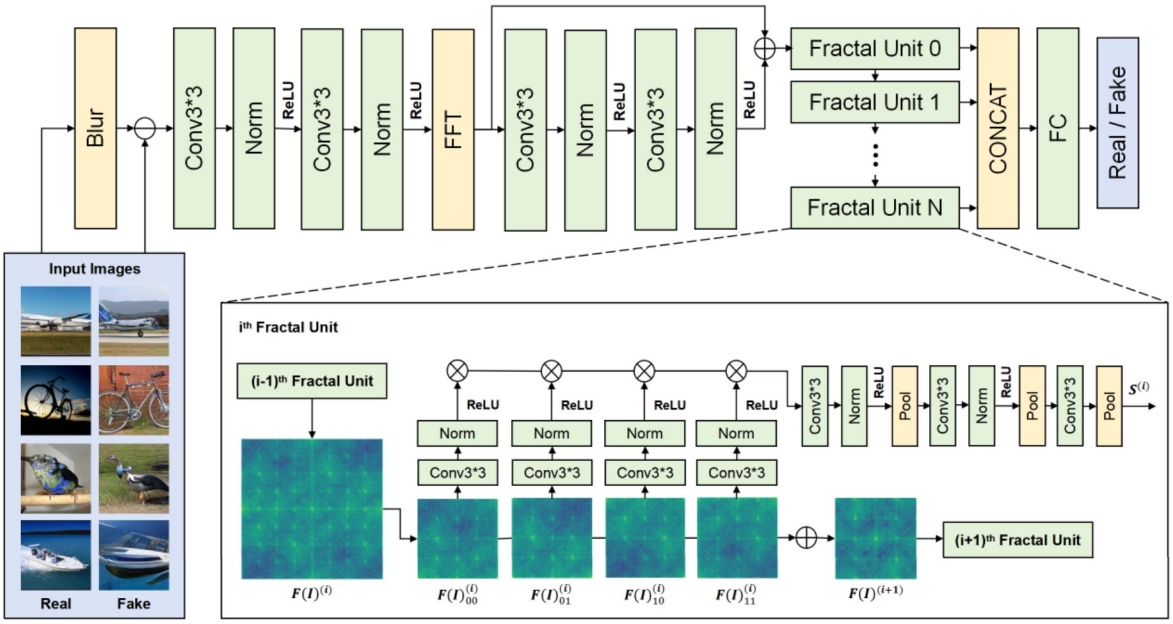


Figure 3. Fractal Convolution Neural Network

similarity feature and Fractal CNN through ablation studies and visualization.

4.1. Experiment Setup

Benchmark We adopt a public benchmark AIGCDetect [36] for generalizable AI-generated image detection to evaluate our model. This benchmark contains images generated by 16 different models, including GANs such as ProGAN [14], StyleGAN [15], BigGAN [3], CycleGAN [37], StarGAN [4], GauGAN [24], and StyleGAN2 [16], and diffusion models such as ADM [6], SDv1.4, SDv1.5 [6], and VQDM [10], as well as images generated via commercial platforms such as Midjourney [21], DALLE2 [25], Glide [22], WFIR[34], and Wukong [1]. The training set consisted of 360,000 images generated by ProGAN and 360,000 corresponding real images. The performance of various state-of-the-art detection methods on the benchmark has been reported [8, 13, 17, 18, 23, 31–33, 36]. LNP [17] and PatchCraft [36] are the methods with the best generalization performance. LNP analyzes the spectrum of the noise residual extracted with a well-trained denoising model. PatchCraft adopts the inconsistent between rich and poor texture areas of the AI-generated image.

Data augmentation We used the same data augmentation methods of the benchmark to ensure a fair comparison, i.e. randomly applied the following data augmentations, each with 10% probability: (1) JPEG using Python Image Library(PIL), with quality $\sim$ Uniform{70, ..., 100}, (2) Gaussian blur with σ $\sim$Uniform[0,1]. (3) Downsampling with a ratio of 0.5. After performing the above operations, each image is cropped to a size of 224×224 from the center. If the original size of the image is not sufficient for cropping, it is padded to the target size using a reflect extension.

Implementation details In the experiment, we first extract the noise residual with median blur with kernel size of 7. The number of feature maps in the hidden layers is 32. Each channel undergoes a Fast Fourier Transform (FFT), and the normalized magnitude spectrum is used as the next feature map. All convolution layers have a kernel size of 3, with padding set to 1. Leaky ReLU is used as the activation function and instance normalization is used for normalization. During training, we randomly split 10% of the training set as the validation set. And early stopping with patience of 2 is adopted. The epoch with the lowest validation loss is selected as the final model.

4.2. Detection Performance

During the experiment, each detection model is trained only over fake images generated by ProGAN and corresponding real images, and is then evaluated by images generated by different GANs and diffusion models that are unseen during the training process (except ProGAN). As shown in Tab. 1, our methods outperforms SoTA methods in terms of the average detection performance over images generated by unseen models. Although images generated by the models are unseen in the training dataset, our methods still learn the ability to recognize them. Even the images generated by

Generator	No Distortion			JPEG Compression			Downsample			Blur		
	LNP	PCFT	**Ours**	LNP	PCft	**Ours**	LNP	PCFT	**Ours**	LNP	PCFT	**Ours**
ProGAN [14]	99.67	**100.00**	99.96	71.16	**97.84**	94.86	71.50	**99.92**	98.43	84.67	**99.01**	95.63
StyleGAN [15]	91.75	92.77	**97.89**	56.02	82.49	**82.62**	69.19	**90.37**	85.52	76.85	**90.38**	84.32
BigGAN [3]	77.75	95.80	**97.95**	51.20	65.25	**93.80**	61.60	72.35	**88.68**	57.30	63.00	**79.39**
CycleGAN [37]	84.10	70.17	**95.84**	57.27	71.09	**92.15**	67.71	83.76	**92.34**	54.39	75.47	**85.14**
StarGAN [4]	99.92	99.97	**100.00**	50.75	60.21	**90.25**	56.08	**99.90**	99.45	78.06	78.71	**93.42**
GauGAN [24]	75.39	71.58	**87.77**	50.06	**73.71**	63.59	49.73	62.07	**74.66**	52.00	60.65	**69.68**
StyleGAN2 [16]	94.64	89.55	**97.55**	58.81	**82.71**	74.21	74.46	**89.00**	85.39	86.67	**91.99**	78.45
WFIR [34]	70.85	85.80	**95.72**	50.12	**79.40**	79.33	55.05	**79.55**	77.62	47.75	62.30	**78.63**
ADM [6]	**84.73**	82.17	57.05	51.28	**62.64**	50.95	53.97	71.12	**73.91**	**77.05**	69.58	58.31
Glide [22]	80.52	**83.79**	81.07	50.97	**68.01**	66.87	48.52	58.37	**73.32**	**82.86**	72.52	76.34
Midjourney [21]	65.55	90.12	**91.69**	51.60	57.87	**63.69**	54.00	57.87	**74.24**	54.21	**76.28**	63.96
SDv1.4 [26]	85.55	**95.38**	94.22	52.66	**75.00**	68.03	55.25	**81.39**	72.48	63.47	**78.85**	68.14
SDv1.5 [26]	85.67	**95.30**	93.94	52.31	**74.87**	**67.90**	55.15	**81.10**	72.57	64.03	**78.61**	67.52
VQDM [10]	74.46	88.91	**92.49**	50.98	64.94	**73.17**	46.14	75.30	**84.55**	64.82	70.53	**74.99**
Wukong [1]	82.06	**91.07**	91.03	51.55	**67.91**	65.98	57.58	78.74	**68.11**	62.64	**74.23**	70.66
DALLE2 [25]	88.75	**96.60**	84.47	50.50	**70.35**	53.67	45.30	73.40	**82.20**	**79.75**	72.00	65.65
Average	83.84	89.31	**91.17**	53.58	72.48	**73.82**	57.58	78.36	**81.47**	67.91	75.09	**75.90**

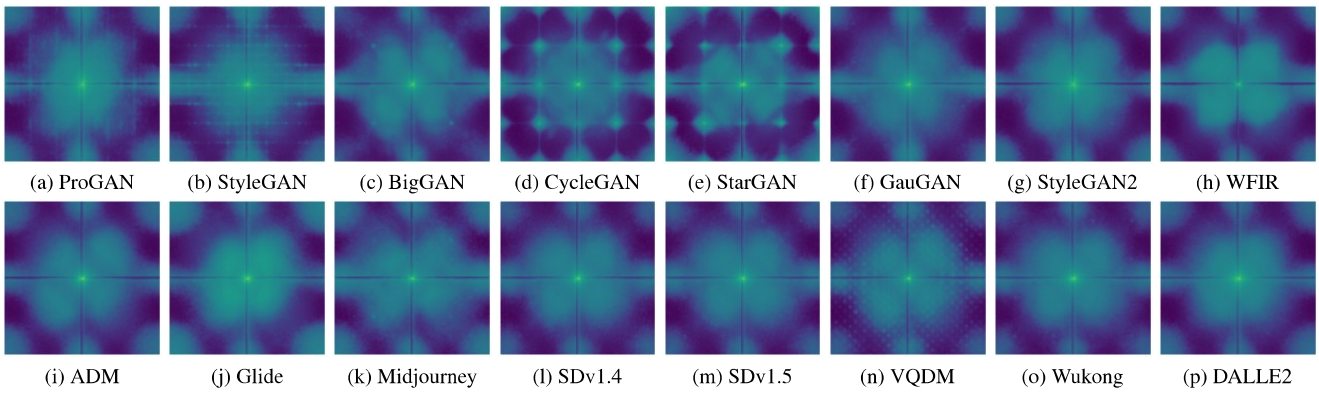
Table 1. The detection accuracy of SoTA methods LNP [17] and PCFT (PatchCraft) [36] and ours on images from unseen generators, including GANs and diffusion models. **Only images generated by ProGAN are used for training the detection models, images generated by other generative models are totally unseen in the training dataset**.

diffusion models, which are quietly different from images generated by GANs, are also detected sensitively. The performance of our method only decreased when detecting the fake image generated by ADM, it could be observe in Fig. 4 and Fig. 5 that the spectral artifact in the noise residual of ADM is dense and the self-similarity representation are in close proximity to the real images, and better noise residual extraction methods should be proposed. In summary, the results demonstrated that our method has superior generalizability. The experimental results of the baselines are referred to from [36], LNP [17], PCFT [36] are the best methods. Other detection methods whose performance reported by this benchmark include [8, 13, 18, 23, 31–33].

Besides, our method also outperforms SoTA methods under distortions in real-world applications. Following prior work, we adopted three real-world image distortions to evaluate the robustness of our model, including JPEG compression with quality of 95, down-sampling with a ratio of 0.5 and Gaussian blur with sigma of 1 [36]. The detection accuracy under distortion is also shown in Tab. 1, which demonstrates that our method outperforms SoTA methods in terms of average detection accuracy under each distortion condition. This result indicates leveraging the similarity among the multilevel sub-bands of the spectrum of AI-generated images could capture more AI-generated traces, which benefits the robustness.

Generator	$N = 0^*$	$N = 1$	$N = 2$	$N = 3$	$N = 4$
ProGAN	99.99	99.92	99.94	99.99	99.96
StyleGAN	95.55	90.14	94.25	93.86	97.89
BigGAN	65.52	98.51	98.89	98.89	97.65
CycleGAN	85.44	93.33	96.84	96.00	95.84
StarGAN	97.33	100.00	100.00	99.97	100.00
GauGAN	76.11	86.60	89.76	83.19	87.77
StyleGAN2	88.78	93.20	97.11	96.71	97.55
WFIR	68.75	98.89	92.54	94.71	95.72
ADM	72.76	55.76	56.90	56.61	57.05
Glide	80.72	65.25	80.01	77.86	81.07
Midjourney	56.79	91.17	92.42	92.57	91.69
SDv1.4	75.58	94.88	93.66	95.00	94.22
SDv1.5	75.08	95.08	93.48	94.90	93.94
VQDM	73.17	92.87	92.39	95.66	92.49
Wukong	74.11	93.32	92.00	93.62	91.03
DALLE2	86.39	75.50	76.41	82.46	84.47
Average	79.51	89.03	90.41	90.75	91.17

Table 2. **The Effectiveness of Fractal Units.** The number of the Fractal Units used is denoted as N. For $N = 0^*$, no fractal unit is adopted and the spectrum is directly used for detection.



(a) ProGAN (b) StyleGAN (c) BigGAN (d) CycleGAN (e) StarGAN (f) GauGAN (g) StyleGAN2 (h) WFIR

(i) ADM (j) Glide (k) Midjourney (l) SDv1.4 (m) SDv1.5 (n) VQDM (o) Wukong (p) DALLE2

Figure 4. **The average feature map of generated images before Fractal Units.** It could be observed that our model has autonomously learned to enhance fractal-structured spectral artifacts under without additional loss applied.

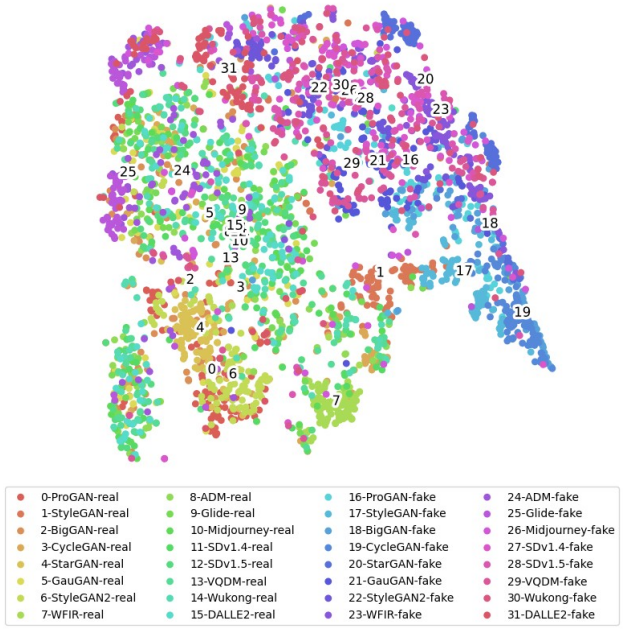


Figure 5. T-SNE visualization of the fractal self-similarity feature of images from different generators.

4.3. Ablation Studies

To analyze the effectiveness of the most crucial part of our method, the Fractal Units, we analyze the impact of the number of Fractal Units on the detection performance. We evaluated the performance under different number of Fractal Units as well as without Fractal Units. Considering that the times of upsample operations in the AI image generation process is finite, we focused only on the use of up to 4 Fractal Units. As shown in Tab. 2, the performance decreased significantly if all Fractal Units are removed from the model, especially for the detection of images generated by diffusion models. This fact proves that the fractal self-similarity feature is more generalizable than the origin spectral feature. In addition, the detection performance gradually improves when more Fractal Units are used, indicating that multilevel spectral fractal self-similarity feature could make full use of the AI-generated traces in the spectrum.

4.4. Visualization

To make a comprehensive evaluation of the effectiveness of Fractal CNN, we conduct visualization analysis on the feature map to be put into Fractal Units and the learn fractal self-similarity features. Fig. 4 demonstrates the average feature map before the Fractal Units. It could be observed that the convolution layers for high-pass filtering automatically learned to strengthen the fractal-structured spectral artifacts. This observation proves that fractal self-similarity features play a crucial role in generated image detection. Fig. 5 shows the hidden fractal self-similarity representation of images from different generators and real images. It could be observed that the representations of generated images from most generators are distinguished from the real images. The abnormal situation of ADM [6], Glide [22] could result from the noise residual extraction methods according to Fig. 4. The median blur might be too simple for complicated samples and more adaptive solutions should be suggested.

5. Conclusions

In this paper, we proposed a generalizable AI-generated image detection method based on the fractal self-similarity in the spectrum. The proposed method has the ability to detect images generated by both GANs and diffusion models, even these models are entirely unseen in the training progress. The most critical contribution of our approach is the discovery of the mechanism for the formation of periodic spectral domain artifact, i.e. the fractal structure of the spectrum of AI-generated images. This discovery motivated us to adopt

the self-similarity of spectral artifact as the share feature of the image generated by different models. Although the feature of spectral artifacts varies with the architecture and parameters of the generator, their self-similarity feature is general. Based on the fractal self-similarity feature, we proposed a fractal-structure neural network-FractalCNN-to capture multi-level self-similarity feature. As multi-level fractal self-similarity features are adopted, our model made better use of the spectral information, which brought better performance and robustness. The results of the experiments indicate that our method has superior generalizable detection performance and has superior robustness under real-world distortions compared to SoTA methods.

References

[1] Wukong. Accessed 26 Janaury 2025. https://xihe.mindspore.cn/modelzoo/wukong, 2025. 6, 7

[2] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023. 1, 2, 3

[3] Andrew Brock. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 2, 6, 7

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 1, 2, 4, 6, 7

[5] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 973–982, 2023. 1, 2, 3

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2, 6, 7, 8

[7] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020. 1

[8] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 1, 2, 3, 6, 7

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 2

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 1, 4, 6, 7

[11] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2904–2908. IEEE, 2022. 1, 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[13] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3465–3469. IEEE, 2022. 1, 2, 6, 7

[14] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1, 2, 6, 7

[15] Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019. 2, 4, 6, 7

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1, 2, 6, 7

[17] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022. 1, 2, 6, 7

[18] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020. 1, 2, 6, 7

[19] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019. 1, 2, 4

[20] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019. 1, 2

[21] Midjourney. Midjourney. Accessed 26 Janaury 2025. https://www.midjourney.com/home/, 2025. 4, 6, 7

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 6, 7, 8

[23] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 1, 6, 7

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 1, 6, 7

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 4, 6, 7

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[28] 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/CVF international conference on computer vision*, pages 1–11, 2019. 1

[29] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*, 2021. 2

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[31] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 1, 6, 7

[32] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 1, 2, 4

[33] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 1, 2, 6, 7

[34] Jevin West and Carl Bergstrom. Which face is real? Accessed 26 Janaury 2025. https://www.whichfaceisreal.com/, 2025. 6, 7

[35] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019. 1, 2, 3

[36] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. 1, 2, 6, 7

[37] 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. 1, 2, 4, 6, 7