TRINITY DETECTOR:TEXT-ASSISTED AND ATTENTION MECHANISMS BASED SPECTRAL FUSION FOR DIFFUSION GENERATION IMAGE DETECTION

Jiawei Song, Dengpan Ye, Yunming Zhang

ABSTRACT

Artificial Intelligence Generated Content (AIGC) techniques, represented by text-to-image generation, have led to a malicious use of deep forgeries, raising concerns about the trustworthiness of multimedia content. Adapting traditional forgery detection methods to diffusion models proves challenging. Thus, this paper proposes a forgery detection method explicitly designed for diffusion models called **Trinity Detector**. Trinity Detector incorporates coarse-grained text features through a CLIP encoder, coherently integrating them with fine-grained artifacts in the pixel domain for comprehensive multimodal detection. To heighten sensitivity to diffusion-generated image features, a Multi-spectral Channel Attention Fusion Unit **(MCAF)** is designed, extracting spectral inconsistencies through adaptive fusion of diverse frequency bands and further integrating spatial co-occurrence of the two modalities. Extensive experimentation validates that our Trinity Detector method outperforms several state-of-the-art methods, our performance is competitive across all datasets and up to 17.6% improvement in transferability in the diffusion datasets.

***Index Terms*—** Diffusion, forgery detection, deepfake

1. INTRODUCTION

Recently, diffusion models have rapidly advanced the field of image generation. AI generation technologies, exemplified by text-image generation, have significantly reduced the barriers to synthetic image creation. Unfortunately, this capability has the potential to be abused for malicious purposes. For instance, text-image generation can be utilized in zero-shot scenarios to craft deepfake attacks targeting prominent political figures worldwide [1]. This misuse can potentially engender severe trust issues within our societal fabric. The diffusion generation mechanism differs from previous approaches, and existing detection methods exhibit poorly in its transferability. Thus, it is of high significance to develop a forgery detection method for diffusion models.

There have been some recent research on image detection for diffusion model generation, some research [2] [3] employs convolutional neural networks trained on different datasets for extracting features and performing classification. Zhong et al. [4] proposed a detection model that extracts the difference features by analyzing the pixel correlation between rich texture and weak texture regions in the image, but only considering the single feature of texture leads to poor data generalization, etc. Huang et al. [5] proposed to incorporate image-text bimodality into the detection of false content by stacking cross-aligned bimodal features, but this did not achieve alignment in semantic space.

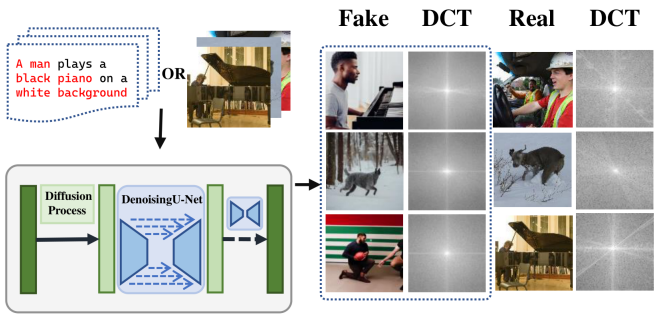

Fig. 1. Diffusion generation process and comparison of spectrogram of real and fake images after DCT transformation.

To address this challenge, our work analyzes the differential representation of diffusion models and real images. It has been found that the prompt input has a significant impact on whether the generated image is more realistic or not [6], the semantic coarseness of textual prompts and the structural integrity of its semantic space have a direct impact on the quality of diffusion model generation, and it can be expected that the relevant characteristics of the text also have a feature-level mapping impact on the detection of images generated by the diffusion model. Given that U-Net has been widely adopted in various diffusion models, the up-sampling layer included in U-Net has also become a vital component of the diffusion model, and the frequency-domain content changes brought about by the up-sampling layer also have a great potential to be used as a generalized diffusion forgery detection. As shown in Fig 1, through the frequency domain transformation, we find that the frequency domain of the real image is more balanced in all directions, while the diffusion-generated image is concentrated in specific directions.

We propose a new method called Trinity Detector. Trinity Detector provides a reliable way to distinguish between

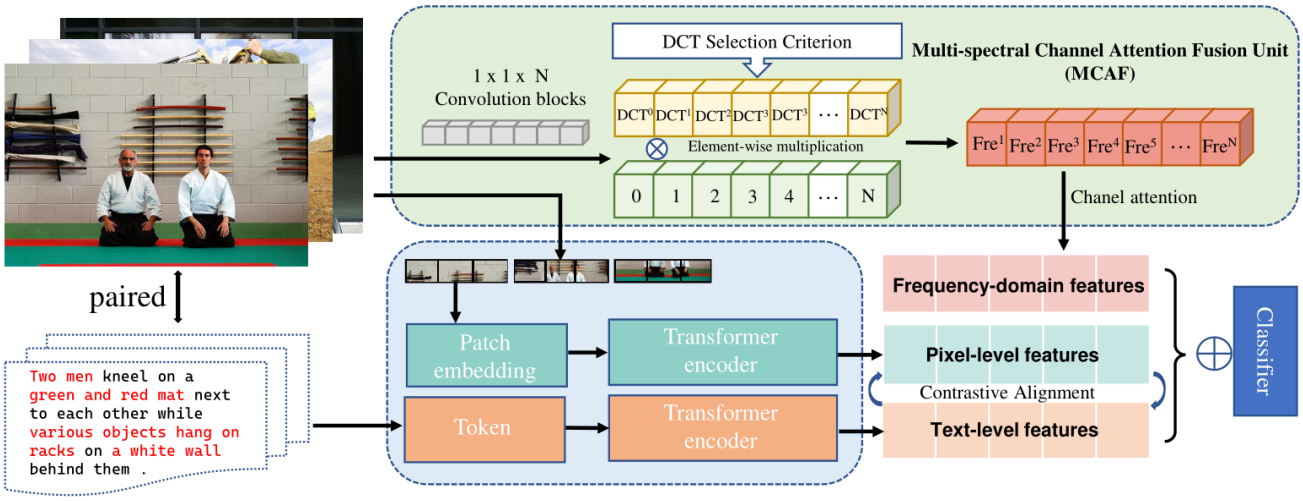

Fig. 2. **An illustration of diffusion—generated images detection.** The Text-Image Alignment and Extraction Module processes textual and visual information pairs, extracting aligned content information. Within the SpectraFuse Unit, DCT vectors extracted through the DCT Selection Criterion are applied to perform DCT transformation on individual channels of the image. Subsequently, a channel attention mechanism is employed to fuse frequency domain information.

real images and diffusion-generated images. Precisely, the Trinity Detector method consists of two parts: (1) We specially design a multispectral channel attention fusion unit to extract the spectral inconsistencies between real images and diffusion model-generated images through adaptive fusion of different frequency bands to further incorporate the spatial co-occurrence of these two modalities. (2) Semantic spatial alignment fusion by introducing coarse-grained features of textual information with fine-grained artifacts in the pixel domain through the CLIP encoder.

Additionally, to train and evaluate the diffusion-generated image detector, we created a comprehensive diffusion-generated dataset, which includes images generated by training on Stable Diffusion and GLIDE.

In summary, our major contributions of this work are three-fold as follows:

- We propose a new forgery detection method called Trinity Detectorthat, it uses an attention mechanism to fuse the frequency domain and text-assisted visual content for diffusion image forgery detection.
- We propose a Multi-spectral Channel Attention Fusion Unit (MCAF), that extracts the spectral inconsistencies between real images and images generated by the diffusion model through adaptive fusion across different frequency bands.
- We produced a diffusion model-based image-text pair dataset for benchmarking the diffusion-generated image detectors.
- Extensive experiments show that our method has excellent performance on diffusion-generated images, strong generalization ability, and excellent robustness.

2. RELATED WORK

2.1. Diffusion Model Image Generation

Inspired by nonequilibrium thermodynamics, Ho et al. [7] proposed a new generation of paradigms, namely Denoised Diffusion Probabilistic Models (DDPMs), which achieved competitive performance compared with PGGAN [8] right out of the gate. Since then, more and more researchers have turned their attention to diffusion models. Song et al. [9] generalized DDPM to Denoising Diffusion Implicit Models (DDIMs) through a class of non-Markovian diffusion processes, which resulted in more high-quality samples with fewer sampling steps. Later work ADM [10] found a more efficient architecture that further achieves state-of-the-art performance compared to other generative models with classifier bootstrapping.LDM [11] modulates the input diffusion model through a cross-attention mechanism and proposes introducing latent diffusion models with latent space. The recently popular Stable Diffusion v1 and v2 are based on the LDM and have been further improved to achieve surprising performance in text-to-image generation. DALLE and DALLE2, released in the same period, utilize a priori models to generate image embedding of the input text based on CLIP [12] and then generate images from this embedding by using a diffusion-based decoder.

2.2. Fake Image Detection Techniques

Generative image detection has been extensively studied in the last few years. Early researchers focused on the subtle changes in the tampering boundaries of the forged image to determine the authenticity of the image based on boundary

artifacts and changes in statistical features [13] [14], wang et al. [15] used monitoring neurons to forensically examine GAN generated images. However, they neglected the ability to generalize for invisible generative models, and there are frequency-based methods in addition to spatial artifact detection. Subsequently, Frank et al. [16] proposed that in the frequency domain, GAN has a special texture due to the presence of up-sampling operations in the generator when generating images. However, with the rapid development of diffusion modeling, a general and robust detector for detecting diffusion model-generated images has yet to be developed. We note that the problem of diffusion-generated image detection has also been noted in several recent works, e.g., as it was pointed out that the lack of explicit 3D modeling of objects and surfaces leads to asymmetry in shadows and reflection images [17]. There are also some works based on pre-trained models of text images for detection, but they do not propose a new paradigm for the development of forged image detection techniques, and they are simply a simple exploitation of several existing works [6]. There are also papers that perform detection by measuring the error between the input image and the image reconstructed by a pre-trained Diffusion model [18], but this performance may be affected by the Diffusion model because different Diffusion models may produce different reconstruction error sizes, and it is necessary to propose some more generalized methods considering the current proliferation of Diffusion models. Unlike them, our work focuses on exploring a generalizable detector that can be applied to large-scale diffusion models.

3. METHOD

In this paper, we present a novel method named Trinity Detector for diffusion-generated image detection. The rest of this section is organized as follows. First, we briefly review DDPM and DCT. Then, we present details of Trinity Detector for diffusion-generated image detection. Finally, we introduce a new dataset, i.e., the diffusion-generated graphic pair dataset, for training and evaluating diffusion-generated image detectors.

3.1. Preliminaries

3.1.1. Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs generate images by simulating a random diffusion process involving two Markov chains: a forward chain that perturbs data into noise and a reverse chain that transforms noise back into data. The forward chain is typically manually designed to convert any data distribution into a simple prior distribution, while the Markov chain of the reverse chain is learned by parameterizing the transition kernel with deep neural networks to invert the former. By initially sampling a random vector from a prior distribution and subsequently performing ancestral sampling through a reverse Markov chain, new samples are generated.

In the forward chain process, the data distribution $x_0 \sim q(x_0)$ is transformed using the transition kernel $q(x_t \mid x_{t-1})$ through a Markov process that defined as:

$$q(x_1, \ldots, x_T \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \tag{1}$$

In DDPMs, this transformation kernel is classically designed by adding a Gaussian perturbation by sampling the Gaussian vector $\epsilon \sim \mathcal{N}(0, I)$ and applying the transformation. Finally, x_t is approximated as a Gaussian distribution, i.e., $x_t \sim q(x_T) \quad : \int q(x_T \mid x_0) q(x_0)\, dx_0 \approx \mathcal{N}(x_T; 0, I)$

The reverse process is also characterized as a Markov chain:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{2}$$

Diffusion models leverage a network, $\Phi(\cdot)$, to model the true distribution $p_\theta(x_{t-1} \mid x_t)$, where θ represents model parameters. The corresponding approximate distribution $q(x_{t-1} \mid x_t)$ is typically parameterized by a neural network. The overarching simplified optimization objective entails a sampling and denoising process, articulated as follows:

$$L_{\text{sim}}(\theta) = \mathbb{E}_{t, x_0, \epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\alpha t} x_0 + \sqrt{1-\alpha t}\epsilon, t)\|^2\right], \tag{3}$$

3.1.2. Discrete Cosine Transform (DCT)

Typically, the basis functions for two-dimensional (2D) Discrete Cosine Transform (DCT) is:

$$B_{i,j}^{h,w} = \cos\left(\frac{\pi h}{H}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(j + \frac{1}{2}\right)\right), \tag{4}$$

Thus, the two-dimensional DCT can be expressed as:

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} B_{i,j}^{h,w}, \tag{5}$$

$s.t. h \in \{0, 1, \ldots, H-1\}, w \in \{0, 1, \ldots, W-1\}$

Here, $\mathbf{f}^{2d} \in \mathbb{R}^{H \times W}$ represents the 2D DCT spectrum, $\mathbf{x}^{2d} \in \mathbb{R}^{H \times W}$ is the input, H is the height of $\mathbf{x}^{2d}$, and W is the width of $\mathbf{x}^{2d}$. Correspondingly, the inverse of the 2D DCT can be expressed as:

$$x_{i,j}^{2d} = \frac{1}{H} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} f_{h,w}^{2d} B_{i,j}^{h,w}, \tag{6}$$

$s.t. i \in \{0, 1, \ldots, H-1\}, j \in \{0, 1, \ldots, W-1\}.$

3.2. Trinity Detector

We find that the images generated by the diffusion model are significantly different from the real images in terms of frequency domain information distribution compared to the real images. To address this, Trinity Detector leverages a Multi-spectral Channel Attention Fusion Unit (MCAF) that utilizes channel attention to model and process channels in the frequency domain. This enhances the representational capacity for frequency domain features, which are then fused with the image-text features extracted by a pre-trained encoder. Building upon these enhancements, Trinity Detector imparts discriminative properties for distinguishing diffusion-generated images from real images. Overall, our detection process is shown in Eq 7, where $\widetilde{\phi}_{\text{Text}}$ and $\widetilde{\phi}_{\text{Image}}$ denote text image extraction and alignment in semantic space, and $\phi_{\text{Frequency}}$ is adaptive fusion based on the channel attention mechanism to extract multi-band frequency domain information

$$\begin{aligned}\text{Feature} =& \widetilde{\phi}_{\text{Text}}(\text{Text}) \oplus \widetilde{\phi}_{\text{Image}}(\text{Image}) \\ & \oplus \text{Attention}_{\text{Channel}}(\phi_{\text{Frequency}}(\text{Image})),\end{aligned} \tag{7}$$

MCAF To enhance channel compression and introduce more information, we propose the Multi-spectral Channel Attention Fusion Unit (MCAF), which extends frequency extraction to more frequency components of the 2D DCT, compressing additional information from multiple frequency components of the 2D DCT.

Initially, the input image is divided into multiple parts along the channel dimension through convolution, denoted as $[X_0, X_1, \ldots, X_{n-1}]$, $X_i \in \mathbb{R}^{C' \times H \times W}$, $\quad i \in \{0, 1, \ldots, n-1\}$ For each part, a 2D DCT transformation is applied to the selected frequency components determined by the DCT Selection Criterion. This serves as the frequency domain channel attention, represented as follows:

$$\begin{gathered}Freq_i = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X^i_{:,h,w} B^{u_i,v_i}_{h,w} \\ \text{s.t. } i \in \{0, 1, \ldots, n-1\}\end{gathered}, \tag{8}$$

The overall SpectraFuse process can be expressed as:

$$\begin{aligned}Freq &= FreqAttention\{cat([Freq_0, \ldots, Freq_n])\} \\ &= sigmoid(fc(cat([Freq_0, \ldots, Freq_n])))\end{aligned}, \tag{9}$$

Where the entire process can be seen as the extraction and compression of multi-channel frequency domain features, our approach extends the original methods focused on specific frequency domains to a framework that encompasses multiple frequency components.

DCT Selection Criterion To implement multi-spectrum channel attention, we adopted the criteria proposed by Qin et al. [19], which include FcaNet-LF (Low-Frequency), FcaNet-TS (Two-Step Selection), and FcaNet-NAS (Neural Architecture Search). The optimal frequency components for channel attention are searched through neural network exploration. The frequency components for this part can be expressed as:

$$Freq^i_{nas} = \sum_{(u,v) \in O} \frac{\exp(\alpha(u,v))}{\sum_{(u',v') \in O} \exp(\alpha(u',v'))} DCT^{u,v}_{2D}(X_i), \tag{10}$$

Image-Text Content Extraction Module Combining the formidable text-image feature extraction and alignment capabilities demonstrated by CLIP, our approach strategically employs its pretrained text-image encoder. Leveraging this pretrained encoder facilitates the deep extraction and fusion of image features and semantic information. Specifically, we use Vit-32 as the image encoder and a transformer encoder for text, creating a cohesive multimodal representation.

This combined strategy not only enhances feature extraction but also promotes a deeper integration of semantic information, contributing to the improved effectiveness and versatility of our proposed method.

3.3. TxtDiffusionForensics: Dataset for Evaluating Diffusion-Generated Image Detectors

Due to the lack of public datasets for generating fake images from diffusion models, this paper produces the TxtDiffusionForensics dataset, which consists of a diffusion model and the corresponding text pairs. We select text cues from the Flickr30K and MSCOCO public datasets and use the stable diffusion model and the GLIDE model to generate synthetic images.

4. EXPERIMENT

In this section, we first describe the experimental setup and then provide extensive experimental results to demonstrate the superiority of our approach.

4.1. Experimental Setup

Data Pre-processing All experiments were conducted on our TxtDiffusionForensics dataset. We randomly selected 5000 real images and 5000 synthesized images, each paired with corresponding textual prompts, for the training dataset. The evaluation test set for assessing the performance of the hybrid detection model consists of three parts. One dataset is generated by StyleGAN, while the other two are generated by stable diffusion and GLIDE models, respectively. In assessing model performance, we employ classification accuracy (ACC).

Baselines To validate the exceptional performance of our proposed method, we have chosen several state-of-the-art

Table 1. Comparison with other detectors. We performed white-box testing on the StableDiffusion dataset and transferability black-box experiments on the other two datasets and investigate the robustness tests with different JPEG compression rates (80%, 50%), different Gaussian blurring rates (α = 1, 2), * denotes re-training on the Stable Diffusion dataset of TxtDiffusion-Foresnsic, Ori denotes the original image.

	StableDiffusion					GLIDE					StyleGAN				
	Ori	JEPG		Gaussian		Ori	JEPG		Gaussian		Ori	JEPG		Gaussian	
Method		80%	50%	1	2		80%	50%	1	2		80%	50%	1	2
CNN[2]	0.493	0.534	0.387	0.456	0.389	0.645	0.674	0.612	0.681	0.616	0.827	0.707	0.611	0.714	0.598
ProGAN[20]	0.486	0.412	0.382	0.543	0.473	0.503	0.436	0.427	0.459	0.429	0.968	0.768	0.746	0.783	0.719
StyleGAN[20]	0.501	0.489	0.413	0.528	0.474	0.510	0.542	0.436	0.536	0.438	**1.000**	**0.820**	0.715	**0.881**	0.746
CNN*[2]	0.931	0.781	0.616	0.733	0.625	0.891	0.629	0.574	0.635	0.543	0.764	0.636	0.539	0.693	0.510
GAN*[20]	0.943	0.757	0.599	0.749	0.578	0.837	0.644	0.519	0.609	0.529	0.913	0.811	0.730	0.837	0.694
DE-FAKE[6]	0.853	0.748	0.682	0.767	0.739	0.703	0.651	0.617	0.737	0.675	0.657	0.607	0.544	0.611	0.537
Dire[18]	0.977	0.809	0.718	0.902	0.856	0.929	0.855	0.702	0.862	0.821	0.835	0.720	0.689	0.785	0.697
OURS	**0.993**	**0.931**	**0.894**	**0.957**	**0.895**	**0.945**	**0.868**	**0.790**	**0.889**	**0.845**	0.841	0.815	**0.779**	0.806	**0.755**

methods that have achieved outstanding results on diverse datasets for comparison:

1. CNNDetection [2]: This method introduces an image detection model generated by a convolutional neural network (CNN). The model undergoes training on a specific CNN dataset and demonstrates the ability to generalize to other images generated by CNNs.

2. GANDetection [20]: By training on ProGAN and StyleGAN, this method has achieved notable success in terms of generalization.

3. DiffusionDetection: DIRE [18] employs a pre-trained diffusion model to measure the error between input images and reconstructed images. DE-FAKE [6] combines graphic and text content through the CLIP pre-trained model, and then classifies it on the classifier.

4.2. Comparison to Existing Detectors

In this study, we employed pre-trained weights obtained from official repositories to assess the performance of CNNDetection, GANDetection, and DiffusionDetection on our curated dataset. Since the DE-FAKE method is not yet open source, we follow the method in the original article to reproduce it on our dataset.

Quantitative results are detailed in Table 1. Notably, existing detectors exhibited a significant decline in performance when tasked with handling diffusion-generated images, with an accuracy (ACC) falling below 60%. To address this limitation, we utilized diffusion-generated images as additional training data and conducted a retraining of CNNDetection and GANDetection. The resulting models demonstrated substantial improvements in detecting images generated by the same diffusion model used during training. However, their performance remained suboptimal when confronted with diffusion models not encountered during training.In contrast, our proposed method, Trinity Detector, showcases outstanding generalization capabilities.

Table 2. The ablation experiment data for each module.

	Stable Diffusion	GLIDE	StyleGAN
Trinity Detector	**0.973**	**0.945**	**0.841**
Fre_{ab}	0.676	0.559	0.513
$\text{Caption}_{\text{ab}}$	0.783	0.751	0.647
$\text{Caption}_{\text{BLIP}}$	0.968	0.937	0.903

4.3. Robustness Assessment

In addition to generalization to unknown generative models, robustness to unknown perturbations is also a general concern since, in practice, images are usually perturbed by various kinds of degradations. Here, we evaluate the robustness of the detector in two types of perturbations (i.e., Gaussian blur and JPEG compression). Perturbations are added in Gaussian blur ($\sigma = 1, 2$) and JPEG compression (quality = 80%, 50%). We explored the robustness of the Baseline and our Trinity Detector. The results are shown in Table 1. We observe that our DIRE achieves better performance at each level of Gaussian blurring and JPEG compression.

4.4. Ablation Study

Based on the description in Section 3.2, we introduce the Multi-spectral Channel Attention Fusion Unit (MCAF). In this section, we first perform a detailed ablation analysis of the module, evaluating and comparing its performance by training detectors that consider only text and image content. Considering that not all forged images contain textual descriptions, we performed ablation experiments with the text extraction unit and the corresponding experiments in the case of text generation using BLIP. The specific results are shown in Table 2, and the Trinity Detector performs more superiorly compared to all the ablation detectors. Meanwhile, regardless of whether the text is a natural language text or a BLIP-generated text, it presents better results than the detector without text.

This indicates that combining frequency domain informa-

tion with coarse-grained text and fine-grained visual content can effectively amplify the difference between the forged image and the real image, thus improving the detection of the forged image generated by the diffusion model.

5. CONCLUSION

In this paper, our goal is to develop a generalized detector to distinguish images generated by diffusion models. To address this challenge, we introduce the Trinity detector method. This approach is based on the observation that images generated by the diffusion model exhibit significant defects in the frequency domain. Therefore, we introduce a multispectral channel attention fusion unit, which adaptively fuses different frequency bands of real and diffusion model-generated images through an adaptive channel attention mechanism to extract their spectral inconsistencies in order to distinguish between real and faked images. Extensive experiments have verified that the proposed Trinity Detector method has better detection performance and robustness as well as generalization compared to other methods in detecting images generated from diffusion models..

6. REFERENCES

[1] Derui Zhu, Dingfan Chen, Jens Grossklags, and Mario Fritz, "Data forensics in diffusion models: A systematic analysis of membership privacy," .

[2] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros, "CNN-generated images are surprisingly easy to spot... for now," pp. 8695–8704.

[3] Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei, "Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection," 2023.

[4] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang, "Rich and poor texture contrast: A simple yet effective approach for AI-generated image detection," .

[5] Zhongqiang Huang, Yuxue Hu, Zhi Zeng, Xiang Li, and Ying Sha, "Multimodal stacked cross attention network for fine-grained fake news detection," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, July 2023, pp. 2837–2842.

[6] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 3418–3432.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*. vol. 33, pp. 6840–6851, Curran Associates, Inc.

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018.

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," .

[10] Prafulla Dhariwal and Alexander Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*. vol. 34, pp. 8780–8794, Curran Associates, Inc.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," pp. 10684–10695.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[13] Bo Liu and Chi-Man Pun, "Deep fusion network for splicing forgery localization," pp. 0–0.

[14] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," pp. 101–117.

[15] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu, "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," .

[16] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proceedings of the 37th International Conference on Machine Learning*. pp. 3247–3258, PMLR, ISSN: 2640-3498.

[17] Hany Farid, "Lighting (in)consistency of paint by text," .

[18] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li, "DIRE for diffusion-generated image detection," .

[19] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li, "FcaNet: Frequency channel attention networks," pp. 783–792.

[20] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? a critical analysis of the state-of-the-art," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, ISSN: 1945-788X.