FIRE: Robust Detection of Diffusion-Generated Images via Frequency-Guided Reconstruction Error

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You*, Linna Zhou*
School of CyberSpace Security, Beijing University of Posts and Telecommunications
beilin.chu@bupt.edu.cn

Abstract

*The rapid advancement of diffusion models has significantly improved high-quality image generation, making generated content increasingly challenging to distinguish from real images and raising concerns about potential misuse. In this paper, we observe that diffusion models struggle to accurately reconstruct mid-band frequency information in real images, suggesting the limitation could serve as a cue for detecting diffusion model generated images. Motivated by this observation, we propose a novel method called **F**requency-gu**I**ded **R**econstruction **E**rror (FIRE), which, to the best of our knowledge, is the first to investigate the influence of frequency decomposition on reconstruction error. FIRE assesses the variation in reconstruction error before and after the frequency decomposition, offering a robust method for identifying diffusion model generated images. Extensive experiments show that FIRE generalizes effectively to unseen diffusion models and maintains robustness against diverse perturbations.*

1. Introduction

The advent of diffusion models (DMs) [18] has made high-quality, controllable image generation significantly more accessible. Subsequent works leverage large-scale multi-modal learning [33, 34, 38], particularly involving text and image pairs, to further diversify the scenarios in which DMs can generate images. The realism of these generated images has reached a level where they can deceive the human visual system, raising concerns about the potential misuse of DMs to create misleading content [21]. Therefore, the development of effective methods to detect such generated images is critical to mitigate the potential societal risks posed by this technology.

Several works [5, 17, 48, 49] have been proposed specifically for detecting images generated by DMs. Among them, a novel class of detection methods based on diffusion re-

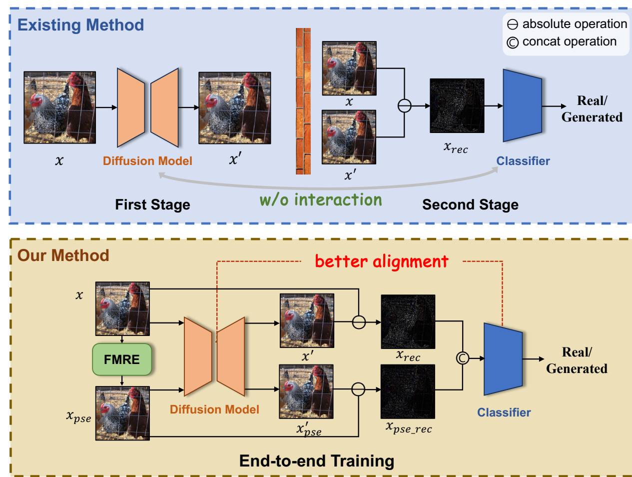

Figure 1. Comparison between existing reconstruction-based methods and FIRE. Existing approaches [32, 43] proceed in two steps: first, compute the reconstruction error of the image using a pre-trained diffusion model, and then train a backend classifier on the reconstruction error. FIRE integrates the classifier with the diffusion model, allowing **end-to-end** learning and better alignment of the latent space for artifact generation and detection. Additionally, FMRE can leverage **frequency-guided reconstruction** to identify the information that the diffusion model struggles to reconstruct.

construction are proposed [7, 26, 27, 32, 43], highlighting a promising research direction. These methods are founded on the assumption that generated images, having undergone the diffusion process, are closer to the latent space in DM. Consequently, they are less affected by reconstruction through a second denoising process compared to real images. By utilizing the reconstruction error, these methods can achieve satisfactory performance in detecting images generated by various models, often with lightweight models or even in a training-free manner. However, these methods still face challenges in generalizing to unseen datasets [7].

When rethinking the significance of reconstruction error representation in the diffusion process, we argue it indeed

*Equal corresponding authors.

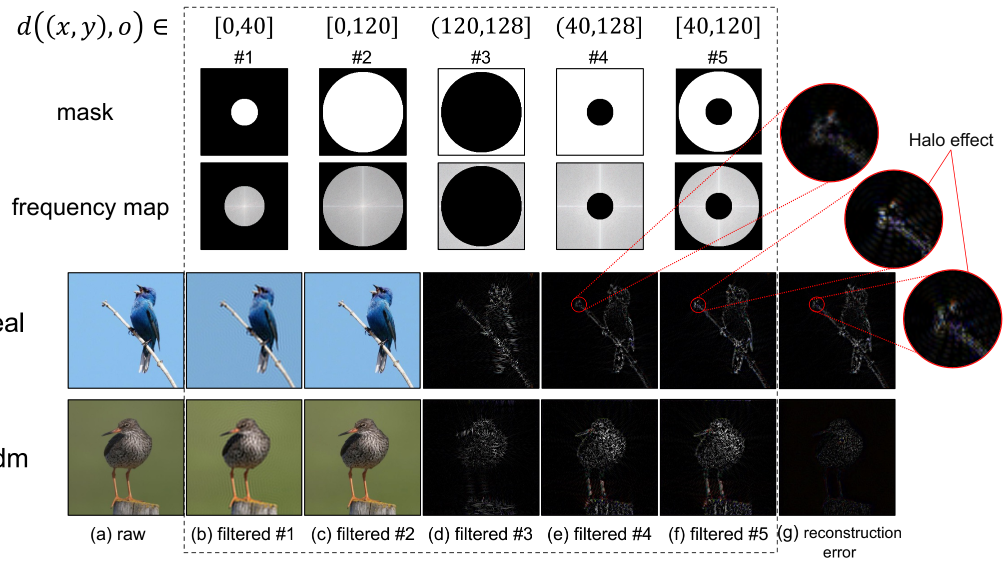

Figure 2. **Analysis between results of images filtered by different masks and their reconstruction errors.** (a) shows a real image from ImageNet and a generated counterpart produced by a pre-trained ADM [12] model. (b)-(f) are the results of applying different frequency masks to the image. Frequency maps are obtained by applying FFT to the images, with the low-frequency regions shifted to the center. $d((x, y), o) \in$ represents points within a specific distance range from the center of the mask, which are set to 1 (preserved), while the rest are filtered out (set to 0). (g) shows the reconstruction error using Stable Diffusion v1.5 [1]. The red circles highlight the halo effect observed in **filtered #5** and **the reconstruction error**. The reconstruction error of the real image visually resembles the band-pass filtered #5 image. Notably, as explained in prior work [43], the reconstruction error for generated images is much lower than for real ones. (To enhance print visibility, we apply 100% sharpening to the residual images in all figures presented in the main paper.)

reflects the parts of the image that the DM struggles to accurately reconstruct. Therefore, existing reconstruction-based methods can be viewed as a process of identifying which parts of the image that are challenging to reconstruct. However, current approaches input the entire reconstruction error into a backend classifier, which also contains both potential content biases and noise from the diffusion process. This might be the reason why existing methods struggle to generalize on unseen datasets [7]. Luo et al. [26] observe that the reconstruction error predominantly manifests in the high-frequency components of the image, yet their approach still lacks the refinement of the hard-to-reconstruct information. Therefore, a critical research questions arise: which specific parts of the image are particularly challenging for the DM to reconstruct?

To address the question, we conduct a frequency analysis of real and generated images. As shown in Figure 2, low-frequency information captures more of the color and general content, while noise and edge details are embedded in the high-frequency components. The mid-frequency band (filtered #5 in Figure 2), which contains both high- and low-frequency information, visually aligns well with the patterns seen in the reconstruction error maps. Based on this observation, **we hypothesize that the reconstruction error—i.e., the hard -to-reconstruct information, primarily resides in the mid-frequency region.** Visualization experiments in Section 4.5 corroborate this hypothesis. Moreover, the mid-frequency component and reconstruction error of real images exhibit stronger signals than generated images. This suggests that real images contain unique mid-frequency information that is particularly difficult to reconstruct. Intuitively, if we can isolate this mid-frequency signal from the image, comparing the reconstruction errors before and after this isolation could serve as a potential cue for detecting generated images.

Building on the above observation and hypothese, we propose a novel **F**requency-gu**I**ded **R**econstruction **E**rror (**FIRE**) method for detecting DM-generated images. FIRE consists of a **F**requency **M**ask **RE**finement module (FMRE) and a backend classifier. FMRE refines the localization of the mid-frequency regions that are difficult for the DM to reconstruct, and filters out such frequency component. The backend classifier then evaluates the change in reconstruction error before and after isolating the mid-frequency components, using this as a cue for classification. Notably, to better align the frequency guidance with its subsequent impact on the reconstruction result, we innovatively integrate the reconstruction pipeline into the detection framework by using only the autoencoder (AE) of a latent DM (LDM), thus bypassing the complex denoising process. This end-to-end learning approach better aligns the DM with the detection task, enhancing the consistency of the mask refinement process.

In summary, the contributions of our work are threefold:

- We are the first to integrate frequency decomposition into reconstruction-based detection methods, identifying the image components that contribute to higher reconstruction errors.
- We propose an innovative end-to-end learning approach that better aligns the classifier with the task of detecting images generated by DM.
- Extensive visualizations validate our hypothesis regarding the frequency distribution of reconstruction errors, and comprehensive experiments demonstrate the effectiveness of our proposed method.

2. Related Works

2.1. Generated Image Detection

With advancements in AI-based image generation, numerous detection methods have emerged to tackle the challenges of increasingly realistic synthetic images. The primary approach utilizes neural networks to capture artifacts within images [4, 19, 24]. Zhong et al. [49] detect generated images by segmenting them into patches and examining inter-pixel correlations, while Tan et al. [40] introduce Neighboring Pixel Relationships, which identify generated content by analyzing local pixel distribution patterns during upsampling.

Additionally, multimodal methods leverage semantic information to enhance detection [45, 50]. Chang et al. [8] and Jia et al. [20] reformulate detection as a visual question answering task, combining vision-language models such as InstructBLIP [10] and ChatGPT [2] to improve performance on unseen data. Shao et al. [37] propose HAMMER for detecting and attributing manipulated content by examining subtle interactions between image and text.

Frequency domain analysis has also proven effective [28, 41, 44], complementing spatial methods that struggle with certain artifacts. Dolhansky et al. [13] use frequency-based masking to extract common features, while Li et al. [23] classify frequency components into semantic, structural, and noise levels to locate regions with distinctive frequency distributions.

2.2. Detection Based on Reconstruction Error

The advent of DMs [18] has spurred the development of specialized detectors to identify DM-generated images. Wang et al. [43] propose Diffusion Reconstruction Error (DIRE), which differentiates real from DM-generated images by measuring reconstruction error. Ma et al. [27] enhance detection accuracy using multi-step error computation in their SeDID method. Ricker et al. [32] utilize autoencoder reconstruction error from latent DMs for a simple, training-free approach. Cazenavette et al. [7] develop FakeInversion to detect images generated by unseen text-to-image DMs using text-conditioned inversion. Chen et al. [9] introduce Reconstruction Contrastive Learning to improve generalization by generating hard samples. Additionally, Luo et al. [26] propose LaRE2, leveraging Latent Reconstruction Error (LaRE) with an Error-Guided Feature Refinement module for more distinct error feature extraction.

Compared to existing methods, our approach is the first to explore the impact of frequency decomposition on reconstruction error. By exploring the hard-to-reconstruct frequency bands hidden within the image, we build a robust detector specifically targeting DM-generated images.

3. Method

In this section, we first provide foundational knowledge on DDPM and the principles behind reconstruction error-based detection methods. Then, we introduce our method named **F**requency-gu**I**ded **R**econstruction **E**rror (**FIRE**), which leverages the observation that DMs struggle to reconstruct mid-frequency information in real images. By comparing the differences between real and fake images after frequency-guided reconstruction, FIRE determines whether an image is generated by a DM. The overall framework of FIRE is illustrated in Figure 3.

3.1. Diffusion Reconstruction Pipeline

Existing DMs can be summarized into two stages: the forward process and the reverse process. In the forward process, noise of increasing intensity is progressively added to the input image:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where $\epsilon_t \sim \mathcal{N}(0, I)$ for $t = 0, \ldots, T$. Here, x_t represents the noisy image at time step t, and x_0 represents the initial image without noise. $\sqrt{\bar{\alpha}t}$ denotes the noise scaling factor at step t, and $\bar{\alpha}t = \prod_{s=1}^{t} \alpha_s$.

In the reverse process, the image is gradually denoised back to its clean state, primarily through the use of a parameterized neural network that predicts the denoised result at each time step t, as follows:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\epsilon_t, \tag{2}$$

where $\alpha_{t-1} = \frac{\bar{\alpha}t-1}{\bar{\alpha}t}$ for $t = T, \ldots, 1$, and $\epsilon_\theta(x_t, t)$ is the predicted noise by the denoising neural network parameterized by θ, typically implemented as a U-Net.

3.2. FIRE

As illustrated in Section 1. We hypothesize that the mid-frequency components of real images are particularly difficult for DMs to reconstruct. Thus, if a real image has

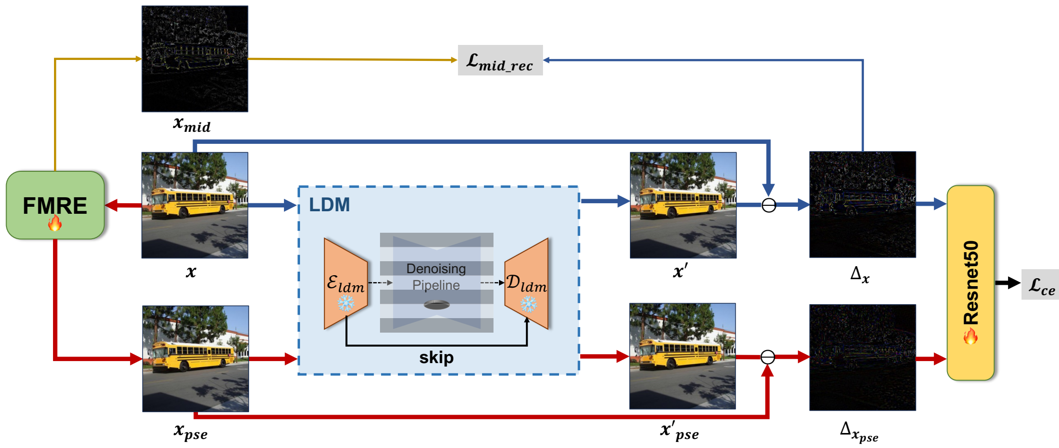

Figure 3. **The overview of FIRE.** We aim to **extract the frequency bands from the image that the diffusion model struggles to reconstruct**, i.e. information that is abundant in real images but lacking in generated ones, and then **compare the reconstruction errors before and after the extraction to determine whether the image is real or generated.** The original image first undergoes reconstruction error computation using an LDM, where we substitute the AE of LDM for the reconstruction process to avoid introducing the denoising pipeline. To effectively extract the frequency band information that is difficult for the diffusion model to reconstruct, we propose the **F**requency **M**ask **RE**finement Module (FMRE). The reconstruction error is then computed for the pseudo-generated image with such information removed. Finally, the two reconstruction error maps are concatenated along the channel dimension and fed into the classifier.

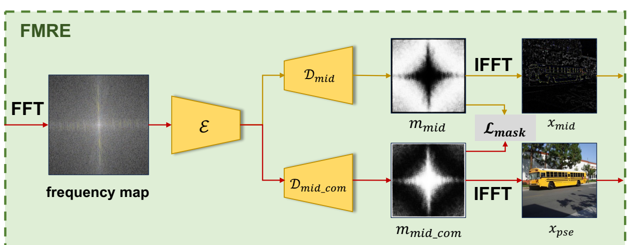

Figure 4. The architecture of our proposed FMRE, which consists of a shared encoder and two independent decoders.

these mid-frequency components removed, its reconstruction error will be smaller than that of the unaltered image. In other words, by stripping away some of the inherent information that differentiates real images from generated ones, the modified real image becomes more similar to a generated image—we term this a **pseudo-generated image**. Intuitively, the difference between a real image and its pseudo-generated version will be greater than the difference between a generated image and its pseudo-generated version. Hence, this frequency-guided difference can serve as a cue for distinguishing real from generated images.

To this end, we propose FIRE, which distinguishes real images from DM-generated ones by comparing reconstruction error differences between an image and its pseudo-generated version. First, we apply FFT to image x and center the low-frequency components. Then, a Frequency Mask Refinement module (FMRE) identifies mid-frequency regions that are difficult for the DM to reconstruct. The frequency mask is applied to obtain a band-pass filtered pseudo-generated image x_{pse}. Both the original and pseudo-generated images are then input to the DM for reconstruction.

$$x'_{pse} = \mathrm{R}(x_{pse}), \tag{3}$$

$$x' = \mathrm{R}(x), \tag{4}$$

where $\mathrm{R}(\cdot)$ represents the reverse and reconstruction process of a DM. Notably, recent work [32] has shown that the denoising process is not critical for the image generation detection task, and satisfactory performance can be achieved by using only the AE of LDM to compute the reconstruction error. Following this insight, we use only the encoder and decoder of a LDM to replace the entire diffusion reconstruction pipeline:

$$\mathrm{R} = \mathcal{D}_{\mathrm{ldm}}(\mathcal{E}_{\mathrm{ldm}}(\cdot)), \tag{5}$$

where $\mathcal{E}_{\mathrm{ldm}}$ and $\mathcal{D}_{\mathrm{ldm}}$ denote the encoder and decoder of the LDM, respectively. This design reduces training costs, which has driven existing methods to use a two-step process: reconstructing the image with a DM and then separately training on the reconstruction. Our approach directly links the detection module with the AE, enabling end-to-end training that aligns the learned representations of the detector with the latent space of the DM.

Once we obtain the reconstructed images x'_{pse} and x', we compute their reconstruction error as follows:

$$\Delta_x = |x' - x|, \tag{6}$$

$$\Delta_{x_{pse}} = |x'_{pse} - x_{pse}|, \tag{7}$$

where $|\cdot|$ denotes the absolute difference. Finally, we concatenate the two reconstruction error maps and feed them into a binary classifier for final prediction:

$$y' = \theta_{cls}(\text{cat}(\Delta_x, \Delta_{x_{pse}})), \tag{8}$$

where θ_{cls} represents the binary classifier, $\text{cat}(\cdot,\cdot)$ denotes concatenation along the channel dimension, and y' is the predicted label.

3.3. Frequency Mask Refinement Module

Our frequency mask refinement module uses an encoder-decoder architecture. The encoder has four convolutional layers (3 × 3 kernel, stride 2, padding 1), and the decoder includes one convolutional layer (3 × 3 kernel, stride 2, padding 1) followed by a PixelShuffle operation [39]. The encoder $\mathcal{E}$ encodes the frequency spectrum after FFT, and two decoders, $\mathcal{D}_{\text{mid}}$ and $\mathcal{D}_{\text{mid\_c}}$, generate two distinct masks: m_{mid} for mid-frequency regions and $m_{\text{mid\_c}}$ for complementary high- and low-frequency regions.

$$m_{\text{mid}} = \mathcal{D}_{\text{mid}}(\mathcal{E}(\mathcal{F}(x))), \tag{9}$$

$$m_{\text{mid\_c}} = \mathcal{D}_{\text{mid\_c}}(\mathcal{E}(\mathcal{F}(x))), \tag{10}$$

where $\mathcal{F}(\cdot)$ denotes the Fast Fourier Transform. The mask m_{mid} is used to isolate the mid-frequency information that is challenging for the DM to reconstruct, and $m_{\text{mid\_c}}$ is used to produce the pseudo-generated image:

$$x_{mid} = \mathcal{F}^{-1}(\mathcal{F}(x) \otimes m_{\text{mid}}), \tag{11}$$

$$x_{pse} = \mathcal{F}^{-1}(\mathcal{F}(x) \otimes m_{\text{mid\_c}}), \tag{12}$$

where $\mathcal{F}^{-1}(\cdot)$ denotes the Inverse Fast Fourier Transform, $\otimes$ denotes element-wise multiplication. The overall architecture of the module is shown in Figure 4.

3.4. Loss Function

Given an image, we aim for the frequency mask refinement module to precisely locate the mid-frequency band that aligns with the reconstruction error map. Intuitively, we transform the extracted mid-frequency components back to the spatial domain using IFFT and need to align them with the reconstruction error of raw image using the following **mid-frequency reconstruction alignment loss**:

$$\mathcal{L}_{\text{mid\_rec}} = \|x_{mid} - \Delta_x\|_2^2. \tag{13}$$

Additionally, based on our prior spectral analysis, we have a preliminary understanding of the difficult-to-reconstruct frequency distribution in real images. We want the mid-band and complementary masks, predicted by the frequency mask refinement module, to capture the desired frequency bands and remain complementary to some extent. To accelerate model convergence and ensure that the predicted masks align with our expectations, we predefine two ideal masks. Given a frequency spectrum $f \in \mathbb{R}^{1\times256\times256}$ with a center point at $o = (128, 128)$, the predefined mid-frequency mask is:

$$M_{mid}(u,v) = \begin{cases} 1 & \text{if } \mathrm{d}((u,v),o) \in [40,120], \\ 0 & \text{otherwise}, \end{cases} \tag{14}$$

$$M_{mid\_c} = \mathbf{1} - M_{mid}, \tag{15}$$

where $d(\cdot,\cdot)$ represents the Euclidean distance between coordinates, and $\mathbf{1} \in 1^{1\times256\times256}$. Therefore, we apply the following **mask refinement loss** to the learned masks:

$$\begin{aligned}\mathcal{L}_{\text{mask}} = &\|m_{mid} - M_{mid}\|_2^2 + \|m_{mid\_c} - M_{mid\_c}\|_2^2 \\ &+ \|\mathbf{1} - m_{mid} - m_{mid\_c}\|_2^2,\end{aligned} \tag{16}$$

Finally, the **overall loss objective** of our method minimizes the combination of the mid-frequency reconstruction alignment error, the mask refinement loss, and the cross-entropy loss:

$$\mathcal{L} = \lambda_0\mathcal{L}_{\text{mid\_rec}} + \lambda_1\mathcal{L}_{\text{mask}} + \lambda_2\mathcal{L}_{\text{ce}}, \tag{17}$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss for training the binary classifier, and we use ResNet-50 as the backbone network:

$$\mathcal{L}_{ce} = -[(y\log(y') + (1-y)\log(1-y'))], \tag{18}$$

where y represents the true label, y' is the predicted label, and $\lambda_0, \lambda_1, \lambda_2$ are balancing factors for the three loss terms. In our experiments, λ_0 and λ_1 are set to $\frac{1}{5}$, and λ_2 is set to $\frac{3}{5}$.

4. Experiment

In this section, we first describe the experimental details and then provide comprehensive experimental results to validate the effectiveness of FIRE.

4.1. Setup

Baselines. We compare our method against a set of state-of-the-art detectors. All methods use the official open-source code for training and inference on our experimental datasets. CNNDet [42] employs a CNN to detect images at the RGB level, revealing that generated images can be effectively identified by convolutional classifiers. AEROBLADE [32] adopts a training-free detection approach by calculating the reconstruction error after it passes through the AE of LDM. DIRE [43] explores image-level DDIM reconstruction error as detection cues. FakeInversion [7] utilizes the latent noise map from the inversion process of Stable Diffusion, guided by prompts, along with the reconstructed image as additional input signals for detection.
Datasets. We conduct experiments on two datasets, DiffusionForensics [43] and a self-collected dataset.

Eval set	Train set w. Gen. method	LSUN-B. w ADM					Imagenet w ADM				
	Gen. method / Model	CNNDet[42]	AEROBLADE[32]	DIRE[43]	FakeInversion[7]	FIRE(ours)	CNNDet	AEROBLADE	DIRE	FakeInversion	FIRE(ours)
Imagenet[11]	ADM[12]	73.4/57.6	99.1/98.3	99.4/96.4	**100.0**/99.8	**100.0/100.0**	99.6/99.3	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**
	SD-v1[33]	67.3/53.4	98.7/97.4	98.3/95.2	99.7/97.6	**100.0/100.0**	87.2/84.7	99.7/98.3	**100.0**/99.6	**100.0/100.0**	**100.0/100.0**
LSUN-B.[46]	ADM	96.8/93.7	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	76.0/54.2	98.4/97.5	**100.0**/99.7	**100.0/100.0**	**100.0/100.0**
	DDPM[18]	75.3/55.1	98.9/97.8	**100.0/100.0**	**100.0/100.0**	**100.0**/99.8	68.4/45.1	99.1/98.2	**100.0**/97.7	**100.0/100.0**	**100.0/100.0**
	IDDPM[29]	76.2/49.5	99.7/98.2	**100.0/100.0**	99.8/98.4	**100.0/100.0**	77.6/53.8	99.5/98.7	**100.0/100.0**	98.6/97.9	**100.0/100.0**
	PNDM[25]	81.1/49.0	99.2/97.9	99.7/88.6	**100.0**/99.7	**100.0/100.0**	73.2/38.7	97.9/97.4	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**
	SD-v1[33]	69.7/49.7	99.6/98.4	**100.0**/99.8	**100.0/100.0**	**100.0/100.0**	66.8/54.3	99.3/98.1	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**
	SD-v2[33]	78.7/50.4	98.3/97.5	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	82.4/63.1	99.0/97.8	**100.0/100.0**	**100.0**/99.9	**100.0/100.0**
	LDM[33]	59.4/49.2	99.8/98.5	**100.0/100.0**	**100.0**/99.8	**100.0/100.0**	60.8/47.5	98.6/97.7	**100.0**/98.6	**100.0/100.0**	**100.0/100.0**
	VQD[15]	71.5/51.9	98.5/97.2	100.0/99.8	**100.0/100.0**	**100.0/100.0**	72.1/50.2	99.8/98.4	**100.0/100.0**	**100.0**/99.7	**100.0/100.0**
	IF[35]	78.0/50.3	99.3/97.6	**100.0/100.0**	**100.0**/99.9	**100.0/100.0**	75.6/49.8	98.9/97.3	**100.0/100.0**	**100.0**/99.9	**100.0/100.0**
	DALLE·2[31]	78.3/67.1	99.0/98.1	**100.0**/98.2	**100.0/100.0**	**100.0/100.0**	77.1/57.3	99.2/98.5	**100.0**/99.9	**100.0/100.0**	**100.0/100.0**
	Mid.	86.1/74.6	97.9/97.3	**100.0/100.0**	**100.0**/99.8	**100.0/100.0**	87.9/73.2	98.7/97.9	**100.0/100.0**	**100.0**/99.7	**100.0/100.0**
Average		76.3/57.8	99.08/98.0	99.8/98.3	**100.0**/99.6	**100.0/100.0**	77.3/59.3	99.1/98.1	**100.0**/99.7	99.9/99.7	**100.0/100.0**

Table 1. **Comparisons of our FIRE and other competitive state-of-the-art detectors.** We evaluate models on the offical Diffusion-Forensics dataset [43]. All the modles are re-trained with the official codes. We report AUC (%) and ACC (%) (ACC/AUC in the table).

Train set	LAION[36] + DALLE·3					LAION + Kan.3				
Eval set / Model	CNNDet	AEROBLADE	DIRE	FakeInversion	FIRE(ours)	CNNDet	AEROBLADE	DIRE	FakeInversion	FIRE(ours)
DALLE·3[6]	52.1/49.8	55.6/52.7	56.2/53.9	74.1/65.9	**78.3/72.6**	44.8/48.9	51.2/49.7	45.4/47.7	64.5/61.2	**76.6/72.0**
Kan.3[3]	46.3/50.0	53.5/51.0	52.4/49.6	68.3/62.7	**74.5/68.7**	51.5/47.6	56.6/53.8	55.2/52.1	78.7/73.8	**86.5/79.9**
Mid.	50.6/50.0	47.6/50.2	53.1/50.2	59.6/54.5	**66.2/64.7**	45.6/50.0	48.6/49.9	52.1/49.8	65.0/58.2	**71.4/66.9**
SDXL[30]	48.4/50.1	53.3/49.6	53.2/52.6	**72.1/69.4**	70.2/67.2	47.8/50.0	53.2/49.5	51.2/49.6	70.7/66.4	**79.8/73.7**
Vega[16]	50.0/50.0	52.8/49.5	47.2/47.0	69.7/65.1	**70.6/67.2**	47.3/50.0	51.3/50.0	48.3/49.9	73.5/68.6	**76.2/71.9**
Average	49.5/50.0	52.6/50.6	52.4/50.7	68.8/63.5	**72.0/68.1**	47.7/49.3	52.2/50.6	50.4/49.8	70.5/65.6	**78.1/72.9**

Table 2. **Performance comparison of FIRE and baselines on updated generation methods.** Each model is loaded with weights pre-trained on the ImageNet subset of DiffusionForensics, then fine-tuned and tested on specific datasets. We report AUC (%) and ACC (%) (ACC/AUC in the table).

DiffusionForensics is a relatively simple open-source benchmark. We chose to use the LSUN-bedroom and ImageNet subsets for experiments. LSUN-bedroom subset collects bedroom images from LSUN-Bedroom [46] and generates fake images using various DMs, including: ADM [12], DDPM [18], IDDPM [29], PNDM [25], SD-v1 [33], SD-v2 [33], LDM [33], VQ-Diffusion [15], IF [35], DALLE·2 [31], and Midjourney [1]. The training set consists of 40,000 real images and 40,000 ADM-generated images. For each subset, we select 1,000 real and 1,000 generated images for testing. Additionally, the ImageNet subset is crafted to evaluate models in general scenarios. Specifically, they use ADM [12] and SD-v1 [33] to generate images. We also use 40,000 real and 40,000 ADM-generated images for training, with 5,000 real and 5,000 generated images selected for testing across both ADM and SD-v1 subsets.

Self-collected dataset. To assess the performance of our model in more realistic scenarios, we follow [7] and form a new dataset using several novel DMs. For real images, we randomly sample 10,000 images from LAION [36] for training, with 1,000 images for testing. For generated images, we use prompts from Midjourney and generate 10,000 images using several open-source text-to-image models for training, with 1,000 for testing. These models include: DALL·E 3 [6], Kandinsky 3 [3], Midjourney[1], SDXL [30], and Segmind Vega [16].

Implementation details. For the data preprocessing, we apply a series of random augmentations, including flip, crop, color jitter, grayscale, cutout, noise, blur, jpeg, and rotate. Each image is resized to 256 $\times$ 256 along the shortest side. During training, the batch size is set to 16, and we use the Adam optimizer with an initial learning rate of 1e-4. We train for 100 epochs, and all experiments are conducted on a single NVIDIA A100 GPU. We adopt two widely used metrics for image generation detection: Area Under ROC (AUC) and accuracy (ACC), to evaluate the effectiveness of models. It is worth mentioning that we use the AE from SD-v1.5 [1] to compute reconstruction errors. Stable Diffusion uses a variational autoencoder (VAE) [22] with Kullback-Leibler regularization.

4.2. Comparison to Baselines

4.2.1. Performance on DiffusionForensics

Table 1 shows the performance of different detection models under various generation methods on the LSUN-Bedroom and ImageNet datasets. We observe that existing detectors, such as CNNDetection, exhibit significant performance drops when dealing with unseen datasets and un-

[1] https://www.midjourney.com

seen diffusion methods. While DIRE and FakeInversion reach satisfactory performance, they still show slight drops in a few unseen domains. In comparison, our FIRE method achieves an average 100% AUC and ACC across all tests.

4.2.2. Performance on Newer Generation Models

The above results demonstrate that the DiffusionForensics dataset is not a challenging task for our method. We then evaluate models on the five newer generation methods, as shown in Table 2. The results show that CNNDet, AEROBLADE, and DIRE exhibit obvious deficiencies in cross-model generalization. The result suggests a fundamental limitation of these methods when handling unseen generation models. In contrast, our proposed FIRE method achieved significant performance improvements in all test scenarios. Specifically, FIRE outperform the next-best method FakeInversion by 3.2% AUC and 4.6% ACC when trained on LAION + DALL·E 3, and by 7.6% AUC and 7.3% ACC on LAION + Kandinsky 3. These experiments reveal that FIRE strikes a better balance between performance and generalization.

4.3. Ablation Study

In this section, we conduct several ablation studies to evaluate the contribution of each component in the model and explore the impact of different predefined frequency masks.

Mask	DALLE·3	Kan.3	Mid.	SDXL	Vega
#1	58.5	57.3	55.8	62.1	55.2
#2	73.4	68.4	61.6	68.5	62.7
#3	62.7	56.9	54.4	66.8	54.3
#4	76.5	72.2	65.8	69.4	68.9
#5 (Ours)	**78.3**	**74.5**	**66.2**	**70.2**	**70.6**

Table 3. Detection performance using different preset frequency masks, measured in AUC.

$\mathcal{L}_{\text{mask}}$	$\mathcal{L}_{\text{mid\_rec}}$	DALLE·3	Kan.3	Mid.	SDXL	Vega
	✓	68.3	63.1	62.6	67.7	65.4
✓		54.3	55.6	53.4	55.3	51.3
✓	✓	**78.3**	**74.5**	**66.2**	**70.2**	**70.6**

Table 4. Ablation study on objective terms, measured in AUC.

4.3.1. Influence of Predefined Frequency Mask

To verify that our predefined frequency mask M_{mid} indeed helps the model focus on the mid-band frequency that DMs struggle to reconstruct, we designed an ablation study. We use different preset frequency masks as shown in Figure 2. We then applied these different mask settings and train on LAION + DALL·E 3 for performance comparison, as shown in Table 3. As indicated in the table, the experimental results confirm that the key information for effectively distinguishing real images is indeed hidden in the mid-frequency components of the image.

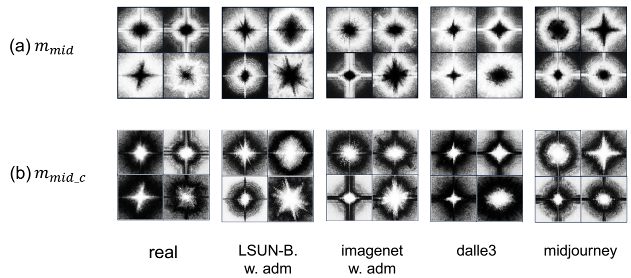

Figure 5. **Visualization of Filtered Frequency Maps in FMRE.** (a) highlights frequency bands the model focuses on, likely containing hard-to-reconstruct information. (b) shows the residual frequency map post-filtering, used for pseudo-generated images. The model primarily focuses on mid-band frequencies, supporting the hypothesis in Section 3.1. Additionally, (a) and (b) are complementary, indicating that (a) can be fully decoupled from (b).

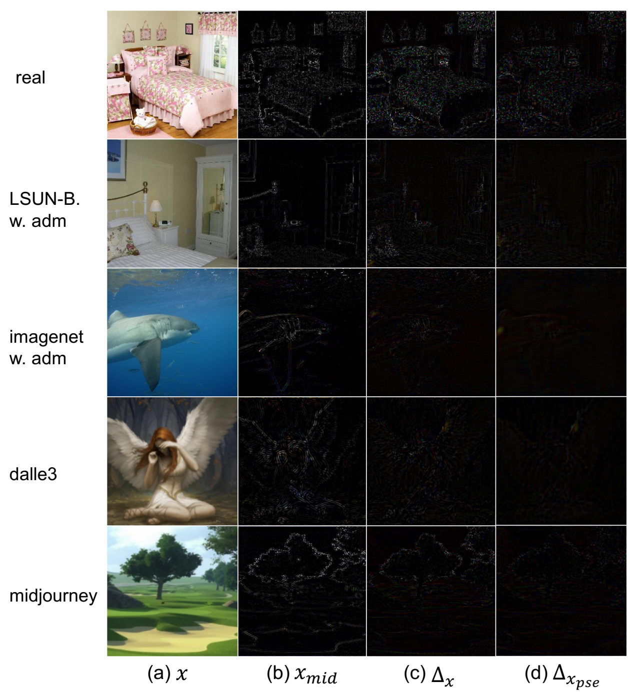

Figure 6. **Visualization of Mid-Band Frequency Information and Reconstruction Errors.** (a) shows original images. (b) displays images after applying IFFT to mid-band frequencies extracted by FMRE. (c) represents reconstruction errors of (a), while (d) shows those of pseudo-generated images. The reconstruction error for frequency-filtered pseudo-generated images is significantly lower than for real images, with real images showing a more pronounced change in error before and after filtering.

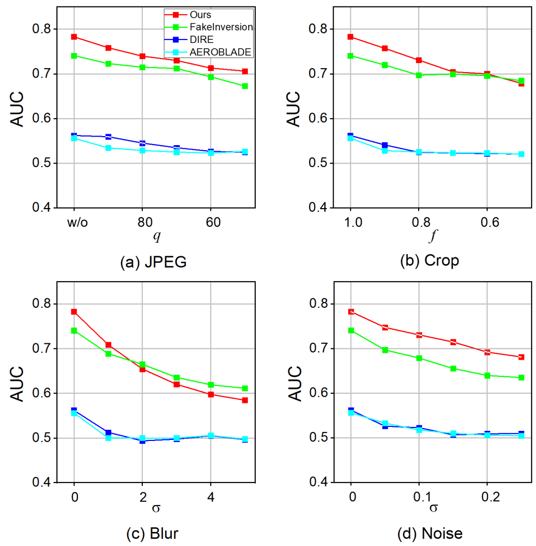

Figure 7. Detection performance of FIRE and baselines when handling perturbed images, measured in AUC.

4.3.2. Influence of Loss Terms

In this section, we conduct ablation studies on each objective term of the model, with the experiment also conducted on LAION + DALL·E 3. The results are shown in Table 4. We first ablate the $\mathcal{L}_{mask}$, meaning the model lacks the guidance of the predefined frequency mask and must spontaneously discover the frequency regions corresponding to the reconstruction error map. The results show that the model's performance drops under this setting, indicating that the predefined frequency mask helps the model better locate the frequency regions corresponding to the reconstruction error. Next, we ablate the $\mathcal{L}_{mid\_rec}$, resulting in a significant performance degradation. Since our method relies on the alignment between the reconstruction error and the extracted mid-frequency information, $\mathcal{L}_{mid\_rec}$ is crucial to the overall approach. Removing it reduces the method to a simpler approach similar to DIRE [43]. These analyses demonstrate that each objective term contributes to the overall performance.

4.4. Robustness to Unseen Perturbations

In real-world scenarios, the images to be detected are often post-processed, such as through quality compression or cropping. In this section, we evaluate the robustness of our method to various perturbations. Following previous works [14, 47], we use the LAION + DALL·E 3 dataset and apply JPEG compression (quality factor q), center cropping (crop factor f and subsequent resizing), Gaussian blur, and Gaussian noise (both with standard deviation σ). The results are shown in Figure 7. In the figure, we report the performance of several SOTA methods and our approach under varying levels of image perturbation. We find that our method outperforms other detectors across most metrics. Only under severe blur conditions does our model perform worse than FakeInversion, likely because blurring destroys much of the high-frequency information in the image, reducing the gap between real and generated images in terms of reconstruction error, making it harder for the model to distinguish. In future work, we plan to design a more robust frequency refinement scheme to improve the robustness of FIRE.

4.5. Visual Analysis

To better understand our model's behavior, we visualize the frequency maps decoded by two decoders of FMRE in Figure 5. The results show that the model mainly focuses on mid-band frequencies, supporting our hypothesis that these frequencies are crucial for distinguishing real from generated images.

In Figure 6, we further visualize mid-band frequency information and the corresponding reconstruction errors. Results indicate that reconstruction errors for frequency-filtered pseudo-generated images are significantly lower, as filtering reduces complex textures in the mid-band. This change is more pronounced in real images, suggesting that our model effectively captures frequency information that DMs struggle to reconstruct. Additional visualizations are in Section 7.

5. Conclusion

In this paper, we first analyze that real images contain inherent information that current DMs struggle to reconstruct, particularly concentrated in the mid-band frequency regions. Based on this observation, we propose a frequency-guided reconstruction error detection method, FIRE. FIRE captures such mid-band frequency information and removes it, detecting generated images by comparing the changes in reconstruction error before and after the removal. Unlike existing reconstruction-based approaches that rely on separate steps for calculating reconstruction errors and training classifiers, FIRE enables end-to-end learning, aligning the feature space for artifacts generation and detection. Extensive experiments demonstrate that FIRE outperforms state-of-the-art baselines, achieving superior performance both on standard datasets and under challenging conditions with perturbed images. We believe that FIRE offers a promising direction for improving the robustness and generalization for DM-generated image detection.

6. Acknowledgment

This work was supported in part by the National Key Research and Development Program of China (Grant 2023YFC3305401 and Grant 2022YFC3303301), the National Natural Science Foundation of China (Grant 62172053 and Grant 62302059), and Shanghai Key Laboratory of Forensic Medicine and Key Laboratory of Forensic Science, Ministry of Justice (KF202420).

References

[1] stable-diffusion-v1-5/stable-diffusion-v1-5 · models at hugging face, 2022. 2, 6

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[3] Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report. *arXiv preprint arXiv:2312.03511*, 2023. 6, 1, 3

[4] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 719–728, 2024. 3

[5] Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Alessandro Nicolosi, and Rita Cucchiara. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. *arXiv preprint arXiv:2407.20337*, 2024. 1

[6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 6, 1, 3

[7] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10759–10769, 2024. 1, 2, 3, 5, 6

[8] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023. 3

[9] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024. 3

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

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

[13] Chandler Timm Doloriel and Ngai-Man Cheung. Frequency masking for universal deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13466–13470. IEEE, 2024. 3

[14] 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. 8

[15] 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. 6

[16] Yatharth Gupta, Vishnu V Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss. *arXiv preprint arXiv:2401.02677*, 2024. 6, 1, 4

[17] Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112*, 2024. 1

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 6

[19] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023. 3

[20] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333, 2024. 3

[21] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *International journal of computer vision*, 130(7):1678–1734, 2022. 1

[22] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6

[23] Hanzhe Li, Yuezun Li, Jiaran Zhou, Bin Li, and Junyu Dong. Freqblender: Enhancing deepfake detection by blending frequency knowledge. *arXiv preprint arXiv:2404.13872*, 2024. 3

[24] Honggu Liu, Xiaodan Li, Wenbo Zhou, Han Fang, Paolo Bestagini, Weiming Zhang, Yuefeng Chen, Stefano Tubaro, Nenghai Yu, Yuan He, et al. Bifpro: A bidirectional facial-data protection framework against deepfake. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7075–7084, 2023. 3

[25] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 6

[26] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024. 1, 2, 3

[27] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272*, 2023. 1, 3

[28] Changtao Miao, Zichang Tan, Qi Chu, Huan Liu, Honggang Hu, and Nenghai Yu. F 2 trans: High-frequency fine-grained transformer for face forgery detection. *IEEE Transactions on Information Forensics and Security*, 18:1039–1051, 2023. 3

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 6

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6, 1, 4

[31] 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. 6

[32] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9130–9140, 2024. 1, 3, 4, 5, 6

[33] 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. 1, 6

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 6

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 6

[37] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2023. 3

[38] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 1

[39] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5

[40] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Ping Liu. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, pages 28130–28139, 2024. 3

[41] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5052–5060, 2024. 3

[42] 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. 5, 6

[43] 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, 3, 5, 6, 8

[44] Simon Woo et al. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 122–130, 2022. 3

[45] Zixin Yin, Jiakai Wang, Yisong Xiao, Hanqing Zhao, Tianlin Li, Wenbo Zhou, Aishan Liu, and Xianglong Liu. Improving deepfake detection generalization by invariant risk minimization. *IEEE Transactions on Multimedia*, 2024. 3

[46] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

[47] Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. *arXiv preprint arXiv:2012.08726*, 2020. 8

[48] Yichi Zhang and Xiaogang Xu. Diffusion noise feature: Accurate and fast generated image detection. *arXiv preprint arXiv:2312.02625*, 2023. 1

[49] 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*, pages 1–18, 2024. 1, 3

[50] Heqing Zou, Meng Shen, Yuchen Hu, Chen Chen, Eng Siong Chng, and Deepu Rajan. Cross-modality and within-modality regularization for audio-visual deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4900–4904. IEEE, 2024. 3