Explainable Synthetic Image Detection Through Diffusion Timestep Ensembling

Yixin Wu[*1,2], **Feiran Zhang**[*1,2], **Tianyuan Shi**[*1,2], **Ruicheng Yin**[1,2], **Zhenghua Wang**[1,2], **Zhenliang Gan**[1,2], **Xiaohua Wang**[1,2], **Changze Lv**[1,2], **Xiaoqing Zheng**[†1,2], **Xuanjing Huang**[1,2,3]

[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]Shanghai Key Laboratory of Intelligent Information Processing
[3]IEIT System Co., Ltd.
yixinwu23@m.fudan.edu.cn, zhengxq@fudan.edu.cn

Abstract

Recent advances in diffusion models have enabled the creation of deceptively real images, posing significant security risks when misused. In this study, we empirically show that different timesteps of DDIM inversion reveal varying subtle distinctions between synthetic and real images that are extractable for detection, taking the forms of such as Fourier power spectrum high-frequency discrepancies and inter-pixel variance distributions. Based on these observations, we propose a novel detection method named ESIDE that directly utilizes features of intermediately noised images by training an ensemble on multiple noised timesteps, circumventing the overtime of conventional reconstruction-based strategies. To enhance human comprehension, we introduce a metric-grounded explanation refinement module to identify and explain AI-generated flaws. Additionally, we present the benchmarks GenHard and GenExplain, offering detection samples of greater difficulty and high-quality rationales for fake images. Extensive experiments show that ESIDE achieves state-of-the-art performance with 98.91% and 95.89% detection accuracy on regular and challenging samples respectively, and demonstrates generalizability and robustness.

Code, Datasets — https://github.com/Shadowlized/ESIDE
Extended version — https://arxiv.org/abs/2503.06201

1 Introduction

With the booming development of diffusion models such as Stable Diffusion (Rombach et al. 2021), DALL-E 3 (Betker et al. 2023), Midjourney and Flux, the proliferation of artificially generated images has reached unprecedented levels. While users marvel at the stunning quality of these synthetic visuals, a growing conundrum has also risen: distinguishing these creations from genuine photographs has become increasingly difficult, and the risks of malicious use have also skyrocketed. Can prevailing detection methods keep pace with the sophistication of knockoffs? Moreover, can current detectors provide robust explanations to satisfy skeptics with more than just a feeble *yes* or *no*? Existing methods fail to cover challenging images, and human comprehension of results is yet to be fully explored. We seek to address these gaps, improving performance on harder detection samples and integrating high-quality explanations into our pipeline.

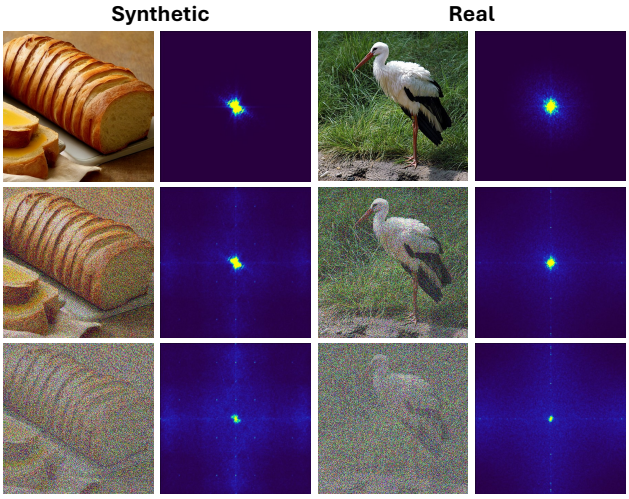

Figure 1: Fourier power spectra of synthetic and real images at various DDIM inversion timesteps. Artifacts of synthetics manifest as high-frequency peaks in spectral backgrounds, becoming more pronounced with increasing timesteps.

Previous studies on synthetic image detection have employed deep neural networks (Wang et al. 2019; Tan et al. 2023; Sha et al. 2023; Cozzolino et al. 2024), or exploited distinguishable fingerprints within frequency and spatial domains (Dzanic, Shah, and Witherden 2020; Liu et al. 2022; Corvi et al. 2023; Zhong et al. 2023). Methods utilizing diffusion-based characteristics such as DIRE, SeDID, LaRE and DRCT (Wang et al. 2023; Ma et al. 2023; Luo et al. 2024; Chen et al. 2024) focus on detecting discrepancies through noising and denoising reconstruction errors. These strategies inherently require ***both*** forward and reverse processes, as they rely on diffusion-based reconstruction.

However, features usable for detection in partially-noised images exist and are often overlooked. As shown in Figure 1, synthetic and real images manifest differing high-frequency

[*]These authors contributed equally.
[†]Corresponding Author.

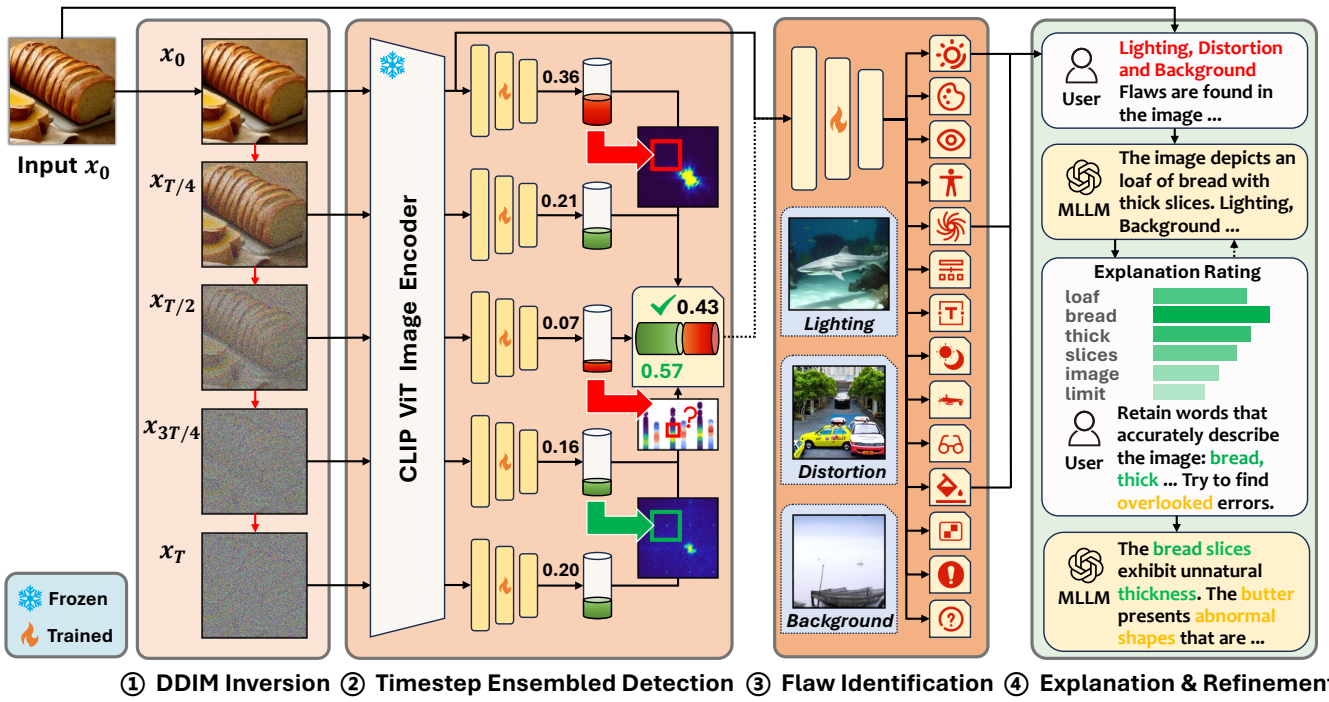

Figure 2: Illustration of the **ESIDE** pipeline. **Stage 1:** DDIM inversion progressively adds noise to input images, creating intermediately noised images. **Stage 2:** Synthetic image detection based on ensembling noised timesteps, discriminators are trained on noised images corresponding to distinct diffusion-induced data distributions, capturing various intermediate features. **Stage 3:** Multi-label flaw identification for synthetic images. **Stage 4:** Explanation generation with MLLMs and rated refinement.

Fourier spectral structures throughout Gaussian diffusion-based noising. Since Fourier transform inherently captures transformation-invariant cues useful for image analysis (Reddy and Chatterji 1996) and is mathematically linked to convolution via the convolution theorem, these frequency-domain features can be directly extracted. Inspired by (Li et al. 2023), we also observe that inconsistent inter-pixel variance distributions across timesteps in terms of spread and peak intensity are exhibited, depicted in Figure 3. Synthetic and real images demonstrate disparate characteristics at each timestep, providing additional clues for detection.

Building upon these findings, we propose a pipeline requiring only a ***single*** noising pass, aiming to directly utilize features within these intermediately-noised steps, named **ESIDE**: **E**xplainable **S**ynthetic **I**mage Detection through **D**iffusion Timestep **E**nsembling, as illustrated in Figure 2. Our framework bypasses conventional reconstruction measures and is designed for detecting challenging generated images while synthesizing high-quality explanations. By removing denoising, we halve the time consumed by DDIM (Song, Meng, and Ermon 2021) pre-processing of methods like DIRE and DRCT. Underexplored research gaps regarding human comprehension of synthetic images are also addressed, grounding content generated by multimodal large language models (MLLMs) on computable evaluation metrics. Our main contributions are four-fold:

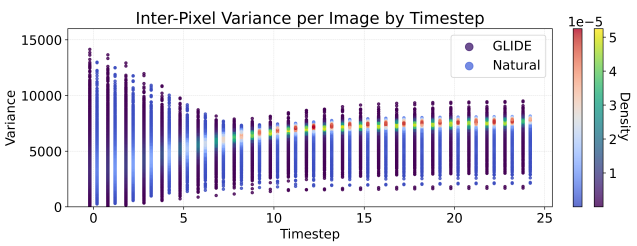

Figure 3: Inter-pixel variances of GLIDE-generated images and natural images noised through DDIM inversion.

- We propose a performance-oriented synthetic image detection method based on ensemble learning of diffusion-noised images, achieving state-of-the-art (SOTA) performance, while further pushing the limits of detectors to challenging samples that truly demand reliable detection.
- We are the first to circumvent image reconstruction measures and provide the insight that varying DDIM inversion intermediate timesteps reveal different discriminative features directly extractable. Manual feature engineering is avoided, consistent with modern deep-learning practice.
- We extend MLLM-based ungrounded explainers and propose a 3-step paradigm of flaw detection, explanation generation, and refinement, applicable for future research.
- We construct two datasets: GenHard and GenExplain, providing researchers access to images of greater detection difficulty, with synthetic flaws and explanations.

2 Related Work

2.1 Synthetic Image Detection

Methods that analyze standalone image characteristics without supplementary captions or additional contextual information for synthetic detection could be broadly categorized into three main approaches (Laurier et al. 2024): deep learning detectors, frequency analysis and spatial analysis.

Deep learning-based detection methods are the most commonly adopted. Early detectors primarily targeted images generated by traditional convolutional neural networks (Wang et al. 2019). More recently, methods leveraging vision transformers (ViTs) gained prominence, utilizing CLIP-ViT (Radford et al. 2021) image encoders for feature extraction, combined with additional classifiers networks or similarity metrics (Ojha, Li, and Lee 2023; Sha et al. 2023; Cozzolino et al. 2024; Lin et al. 2024; Xu et al. 2024a). LGrad used pretrained StyleGAN to convert images into gradients (Tan et al. 2023). Other studies that leverage diffusion methods (Wang et al. 2023; Ma et al. 2023; Luo et al. 2024; Chen et al. 2024) mainly focus on identifying synthetical discrepancies by comparing original images with their reconstructions through diffusion noising and denoising.

On the other hand, frequency analysis generally classify synthetic images based on their high-frequency features. Generated images share systematic shortcomings in replicating attributes of high-frequency Fourier modes (Dzanic, Shah, and Witherden 2020; Corvi et al. 2023). Frequency inconsistencies and patterns among images generated with different models could also be effectively utilized for detection (Liu et al. 2024; Song, Ye, and Zhang 2024; Tan et al. 2024a). However, utilizing the amplification of distinctions between synthetic and authentic images through intermediate steps of DDIM inversion is yet to be investigated.

Meanwhile, spatial analysis methods detect fake images by computing pixel-level relations and noise patterns. Inter-pixel relationships and contrasts could be captured and used to train detectors (Zhong et al. 2023; Tan et al. 2024b), while noise patterns of real and synthetic images extracted through spatial models exhibit distinct characteristics usable for classification (Liu et al. 2022; Chen, Yao, and Niu 2024).

2.2 Detection Explainability

The explanation of synthetic images is an underexplored field of research. Previous studies have delved into image forgery explanation using MLLMs (Huang et al. 2024; Xu et al. 2024b; Kang et al. 2025), focusing on identifying manipulations and modifications instead of explaining images synthesized from scratch, lying outside our scope of discussion. Existing synthetic explainers introduce benchmarks rather limited in size, and rely entirely on MLLMs for detection without integrating specialized models and metrics (Li et al. 2024; Ye et al. 2024). Latest work (Wen et al. 2025) utilizes LLaVA (Liu et al. 2023) for detection, but requires substantial training resources for moderate performance, while categorizing based on image content rather than the actual reason an image is identified as synthetic.

Instead, our work proposes a light-weighted unified framework combining detection, explanation and automated refinement. We introduce an explanation benchmark for synthetic images named GENEXPLAIN, larger in scale than current benchmarks and categorizing images by their actual synthetical appearance. Manual data pruning revealed a maximal of 67.4% of identified flaws to be incorrect, demonstrating that exclusive reliance on MLLMs proves untrustworthy. By anchoring explanations in classified synthetical errors, and incorporating a refinement process guided by quantitative scoring metrics, we effectively mitigate the limitations of MLLMs in standalone detection tasks.

3 Method

3.1 Preliminaries

Diffusion models (Rombach et al. 2021; Betker et al. 2023) generate images through a two-stage process of forward noise addition and reverse denoising. The forward process gradually transforms data into Gaussian noise, while the reverse trains a neural network to iteratively denoise for distribution restoration. Diffusion models build a symmetric Markov chain connecting the processes, aiming to minimize the KL divergence between data and noise distributions.

Denoising Diffusion Probabilistic Models (DDPM) (Sohl-Dickstein et al. 2015) parameterize the reverse process as a Markov chain, where the mean μ_θ is derived from a neural network that predicts the noise $\epsilon_\theta(x_t, t)$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}\right), \tag{1}$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right). \tag{2}$$

Denoising Diffusion Implicit Models (DDIM) (Song, Meng, and Ermon 2021) accelerate sampling by defining a non-Markovian forward process while maintaining the same marginal distribution $q(x_t|x_0)$. The reverse combines deterministic generation with stochastic noise injection:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\epsilon_\theta + \sigma_t\epsilon_t, \tag{3}$$

3.2 Synthetic Image Detection

We propose a novel method for detecting synthetics through ensemble learning of intermediate noise. AdaBoost (Freund and Schapire 1997) is a boosting algorithm that aggregates the predictions of multiple weak models through a weighted sum, constructing a stronger learner with enhanced accuracy for discrimination. Specifically, we follow previous ensembling measures, but train classifiers on distinct data distributions, each corresponding to a different diffusion timestep.

Given an input image x_0, we apply a T-timestep DDIM inversion process to yield intermediate samples, generating a sequence of stepwise noised images: $\{x_0, x_1, x_2, \ldots, x_T\}$. For each timestep with an interval of a stride s, a base classifier m_k is trained exclusively on corresponding timestep noised images, resulting in a collection of models M:

$$M = \{m_0, m_s, m_{2s}, \ldots, m_T\} \tag{4}$$

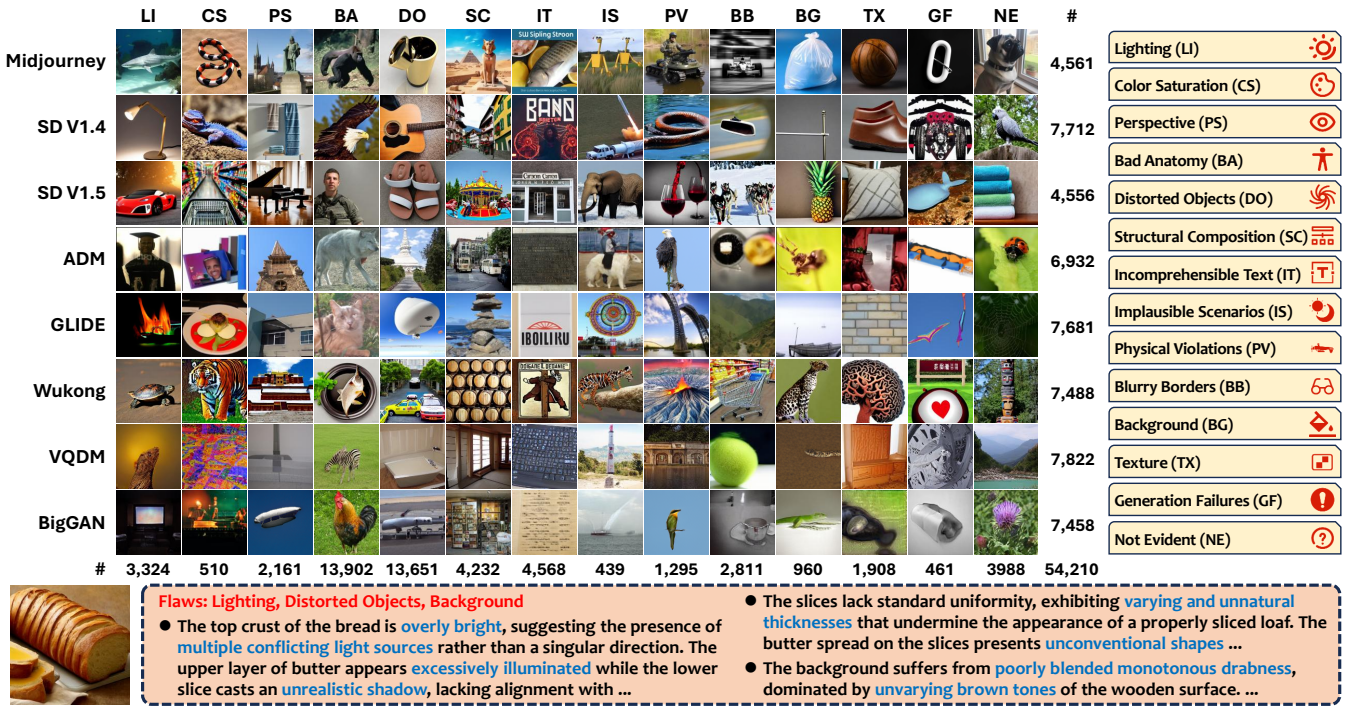

Figure 4: Visualization of the **GenExplain** benchmark. Synthetic images from GenImage are divided into 14 categories of flaws. Each image is matched with one or multiple categories, with a corresponding explanation for each flaw type.

A **sample weight** $w_{k,i}$ is assigned to each image of the training set to emphasize samples previously misclassified, while reducing the significance of correctly predicted cases. As each model operates on a noised dataset corresponding to a different diffusion timestep, the sample weights are separately initialized for each model, where N represents the total number of images in the training set.

$$w_{k,1} = w_{k,2} = \cdots = w_{k,n} = \frac{1}{N} \tag{5}$$

By incorporating $w_{k,i}$ into binary cross-entropy (BCE), a weighted loss function $\mathcal{L}_{\text{WB}}$ is obtained, where y denotes the true label and $\hat{y}$ is the predicted probability:

$$\mathcal{L}_{\text{WB}}(k, y, \hat{y}) = -\sum_{i=1}^{N} w_{k,i} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] \tag{6}$$

The weighted error e_k of m_k can then be calculated, where $h_k(x_i) \in \{-1, 1\}$ is the prediction of m_k for x_i:

$$e_k = \frac{\sum_{i=1, h_k(x_i) \neq y_i}^{N} w_{k,i}}{\sum_{i=1}^{N} w_{k,i}} \tag{7}$$

To form the final prediction, a **model weight** α_k is assigned to each classifier and updated throughout training:

$$\alpha_k = \frac{1}{2} \ln\left(\frac{1 - e_k}{e_k}\right) \tag{8}$$

Sample weights are then adjusted according to the prediction results and normalized across timestep samples:

$$\tilde{w}_{k,i} = w_{k,i} \cdot \mathrm{e}^{-\eta \alpha_k \cdot h_k(x_i) y_i} \tag{9}$$

$$w'_{k,i} = \frac{\tilde{w}_{k,i}}{\sum_{i=1}^{N} \tilde{w}_{k,i}} \tag{10}$$

Here, η is a learning rate factor applied to limit change rate. After training the base classifiers, the model weights α_k are used to calculate a weighted sum $H(x_i)$ of the predictions from each m_k for the final prediction of image x_i:

$$H(x_i) = \text{sign}\left(\sum_{k=0}^{T} \alpha_k \cdot h_k(x_i)\right) \tag{11}$$

In practice, a threshold is applied to the weighted error to normalize model weights and prevent degradation.

3.3 Multimodal Explanation Refinement

When an image is identified as synthetic, a multi-label classifier is employed to identify potential flaws present. MLLMs are then utilized to generate a rough explanation for each image based on the identified types. An refinement process is iteratively conducted to enhance explanation quality, segmenting the initial explanation into multiple phrases and assigning each a rating based on its semantic similarity with the image. Text-image cross attention (Lee et al. 2018) mechanisms are leveraged to compute these ratings. Faster R-CNN (Ren et al. 2015) is first applied for object detection, identifying n sub-image regions, which are then encoded into visual embeddings $\{v_1, v_2, \ldots, v_n\}$ using a CLIP ViT,

Method	Training & Test Subsets								Avg. Acc (%)
	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	
ResNet-50	83.90/86.75	80.63/94.17	76.17/90.19	75.81/91.00	**95.82**/99.00	73.72/89.50	92.94/98.67	**99.10**/96.08	84.76/93.17
DeiT-S	58.29/76.08	73.92/81.08	74.32/78.75	63.37/68.42	72.84/93.83	71.01/83.83	66.87/77.50	46.23/76.00	65.86/79.44
Swin-T	70.81/91.00	74.63/88.08	78.02/90.00	70.30/79.50	85.19/97.83	74.92/85.92	86.96/90.17	88.24/99.67	78.63/90.27
CNNSpot	70.91/83.92	78.02/89.25	80.15/88.06	68.32/72.58	78.57/87.08	78.95/87.83	89.97/97.42	70.44/96.08	76.92/87.78
CBSID	67.10/93.25	73.03/90.92	72.22/93.63	94.20/99.83	88.53/99.17	75.67/92.75	84.10/98.17	98.27/**99.91**	81.64/95.95
DIRE	88.86/92.83	95.72/97.42	96.02/96.25	90.52/94.67	81.40/**99.83**	84.17/92.67	94.90/97.83	93.29/99.67	90.91/96.40
LGrad	72.25/87.33	72.28/83.92	76.45/84.37	70.86/79.17	82.19/96.00	65.50/79.75	74.80/81.08	83.74/94.58	74.76/85.78
UnivFD	40.80/87.75	41.92/89.33	35.40/88.25	65.28/85.58	79.13/94.25	65.35/89.42	67.84/91.92	71.07/94.75	58.35/90.16
FreqNet	87.13/94.33	93.47/94.58	**96.73**/95.12	95.69/91.50	82.13/98.25	90.88/90.17	**98.60**/95.42	85.12/99.17	91.22/94.82
NPR	85.10/87.25	89.90/95.25	95.23/97.12	99.36/99.75	86.29/92.08	**96.94/98.42**	93.74/97.33	92.87/99.58	92.43/95.85
DRCT	**92.89**/97.50	**97.51/100.00**	96.10/**99.38**	88.03/96.67	85.59/98.33	91.29/95.83	96.17/**100.00**	97.93/99.17	93.19/98.36
ESIDE	92.38/**98.42**	96.65/99.17	**96.73**/98.63	**99.43/100.00**	94.98/99.00	91.50/97.25	97.90/99.33	97.58/99.50	**95.89/98.91**

Table 1: Synthetic image detection accuracy on GenImage subsets. Models are trained individually on each GenImage subset, with the **original** samples training set only, while tested on both the **original** samples test set, and a previously unseen **hard** samples test set from GenHard. The prior number for each cell marks the test accuracy on the **hard** samples, while the posterior marks the test accuracy on the **original** samples. The best scores are highlighted in **bold**, and the second best are underlined.

while also concatenating the embedding of the full image v_0. Phrases p are then encoded into the same vector space, and their similarities with each region are calculated and normalized, pairing these phrases with visual regions. The weighted combination of the region embeddings, denoted as a, is then derived based on the similarities:

$$\tilde{s}_i = \frac{\cos\langle p, v_i\rangle}{\sum_{i=0}^{n} \cos\langle p, v_i\rangle} \tag{12}$$

$$a = \sum_{i=0}^{n} \frac{v_i \cdot \mathrm{e}^{\lambda \tilde{s}_i}}{\sum_{i=0}^{n} \mathrm{e}^{\lambda \tilde{s}_i}} \tag{13}$$

where λ denotes the inverse temperature of the softmax function. The **rating** r of the phrase p is then calculated as its cosine similarity with a.

For refinement, Top-K sampling is performed according to phrase relevance, and the model is instructed to retain these phrases while identifying additional overlooked flawed regions. This process is iteratively repeated with the revised explanations, resulting in a final explanation that describes the flaws present in the image with greater accuracy.

3.4 Construction of GenHard and GenExplain

We construct two datasets, GenHard and GenExplain, based on GenImage (Zhu et al. 2023). The former comprises synthetic and natural images more challenging to detect, while the latter aims to categorize and provide explanations for flaws commonly found in artificial images.

GenHard To extract samples of greater difficulty, we employed CBSID (Cozzolino et al. 2024) with a minimalist linear network classifier under-fittingly trained for a single epoch on the validation subsets of GenImage, and subsequently tested on training subsets. Across the 8 subsets tested, the 108,704 synthetic images and 112,682 natural images misclassified were identified as hard samples, which were then partitioned into training and validation sets.

GenExplain Extending prior taxonomies (Mathys, Willi, and Meier 2024; Li et al. 2024; Xu et al. 2024b), we identified 14 common categories of flaws associated with realistic synthetic images, and constructed a dataset comprising 54,210 groups of images, flaws and explanations, illustrated in Figure 4. Images from GenImage validation subsets were fed into `gpt-4o`, prompted with flaw definitions and corresponding instructions, yielding a preliminary categorization of approximately 11,000 to 14,000 image-flaw pairs per subset. Manual data pruning removed 30.1% to 67.4% of images incorrectly categorized from each subset, and the final explanations are obtained through iterative refinement.

4 Experiments

4.1 Synthetic Image Detection

We train an ensemble on noised images derived from DDIM inversion intermediate timesteps, following the noise sampling implementation of (Dhariwal and Nichol 2021; Wang et al. 2023) and using pre-trained diffusion models of sizes 256×256 and 512×512. `CLIP ViT-L/14` (Radford et al. 2021) is employed to directly extract image features, which are passed through multi-layer perceptrons for classification.

We evaluate our results on GenImage (Zhu et al. 2023), a million-scale dataset covering 8 generator subsets: Midjourney, Stable Diffusion V1.4 (Rombach et al. 2021), Stable Diffusion V1.5 (Rombach et al. 2021), ADM (Dhariwal and Nichol 2021), GLIDE (Nichol et al. 2021), Wukong (Wukong), VQDM (Gu et al. 2021), and BigGAN (Brock, Donahue, and Simonyan 2019). We partition the validation subsets of GenImage by a 9:1 ratio for training and evaluation due to computational resource constraints.

The baselines ResNet-50 (He et al. 2015), DeiT-S (Touvron et al. 2020), Swin-T (Liu et al. 2021), CNNSpot (Wang et al. 2019), CBSID (Cozzolino et al. 2024), DIRE (Wang et al. 2023), LGrad (Tan et al. 2023), UnivFD (Ojha, Li, and Lee 2023), FreqNet (Tan et al. 2024a), NPR (Tan et al.

Method	Test Subsets								Avg. Acc (%)
	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	
ResNet-50	45.27/71.17	80.63/94.17	74.06/90.33	**75.36**/48.17	24.67/57.25	76.91/90.33	49.76/59.75	26.58/41.75	56.66/69.12
DeiT-S	47.88/49.33	73.92/81.08	75.21/79.50	65.21/48.17	25.67/48.92	74.55/80.75	47.89/46.08	31.70/43.58	55.25/59.68
Swin-T	47.14/56.58	74.63/88.08	76.13/86.94	57.71/49.83	21.45/50.33	76.54/84.08	43.24/49.25	26.92/45.08	52.97/63.77
CNNSpot	49.28/56.58	78.02/89.25	78.34/86.19	72.77/48.50	26.36/51.50	74.77/83.42	49.39/50.25	27.13/47.83	57.01/64.19
CBSID	51.16/74.33	73.03/90.92	74.07/91.69	73.69/54.42	49.36/**74.83**	73.41/78.33	52.32/**65.58**	33.36/59.08	60.05/73.65
DIRE	67.38/62.08	95.72/97.42	95.59/96.75	49.01/30.50	31.38/17.17	62.18/56.50	33.80/29.25	37.58/19.50	59.08/51.15
LGrad	50.28/56.17	72.28/83.92	73.57/81.92	74.19/44.50	25.05/49.50	55.56/53.83	53.21/52.92	33.46/45.33	54.70/58.51
UnivFD	26.75/**82.42**	41.92/89.33	38.66/79.17	61.83/47.08	25.12/71.58	58.17/72.17	54.36/51.58	29.75/49.58	42.07/67.87
FreqNet	**68.96**/70.50	93.47/94.58	88.52/91.58	67.51/**63.92**	48.81/72.58	87.23/78.83	53.61/56.25	70.44/**71.83**	72.32/**75.01**
NPR	47.35/53.75	89.90/95.25	96.01/93.50	70.95/55.42	50.80/65.58	**95.53/96.08**	58.27/49.33	41.61/46.08	68.80/69.37
DRCT	51.52/46.17	**97.51/100.00**	**97.51**/86.50	69.36/53.50	**55.68**/67.42	86.42/89.50	**62.91**/56.25	44.39/52.25	70.66/68.95
ESIDE	67.40/74.33	96.65/99.17	97.15/**98.75**	63.79/37.08	50.21/39.00	82.60/67.33	59.73/50.42	**90.10**/53.67	**75.95**/64.97

Table 2: Cross-validation accuracy on GenImage subsets. Models are trained on GenImage/SD V1.4 with the **original** samples training set only, while tested on both the **original** and **hard** samples test set from another generator subset from GenHard.

2024b), and DRCT (Chen et al. 2024) are compared, on both the original GenImage dataset and the more challenging samples of GenHard. Two distinct scenarios are investigated: (1) train-test subsets from the same generator, and (2) train-test subsets sourced from different generators, assessing both optimal performance and generalizability.

Implementation Details Timesteps $T = 24$ are taken in total, with 9 classifiers resulting from a stride $s = 3$. Sample weights learning rate $\eta = 0.25$, while error threshold $\epsilon_k \in [0.001, 0.5]$ prevents ensemble degradation. Only a few simple 5-layer multi-layer perceptron networks are required, easily scalable to larger architectures. Batch normalization, LeakyReLU and dropout are sequentially applied after each layer. The AdamW optimizer alongside our modified $\mathcal{L}_{\text{WB}}$ is used. Low GPU memory consumption is exhibited, and model parallelism is easily enabled as models trained on different noised datasets could be distributively allocated.

Result Analysis As shown in Table 1, ESIDE achieves SOTA performance on both harder and original images, with an average absolute accuracy increase of 2.70% and 0.55% respectively, and 95% confidence Wilson Score Intervals of $[95.63, 96.13]$ and $[98.68, 99.10]$. We perform 15.10% better than baseline averages on harder samples and lead by 7.28% on originals. Meanwhile, DDIM preprocessing time for each image is reduced to 2.66 s on a L20 GPU, as compared to 5.71 s of DRCT and 5.24 s of DIRE. Training time on GLIDE is 44.2 s per epoch, remaining on par with baseline averages of 32.6 ± 19.1 s.

Cross-validation results in Table 2 show generalizability across images synthesized by other models. Notably, some methods perform worse than random guessing on hard samples as only incorrect identifications are included in GenHard, further underscoring its difficulty. Similar cross-validation results reveal that effective indicators for one generator may perform opposite on another. The anomaly of higher cross-validation accuracy on samples of greater difficulty is observed, as previously-simple original scenarios are under-trained in comparison to more controversial images.

Metric	MJ	SD1.4	SD1.5	ADM	GLI	WK	VQ	BG	Avg.
EM	41.60	52.03	58.08	50.41	49.72	33.81	54.61	49.07	48.67
mAP	34.70	35.71	33.01	40.04	34.57	33.80	54.05	34.50	37.55

Table 3: Performance of 14-type flaw classification on GenExplain subsets. Subset names are abbreviated.

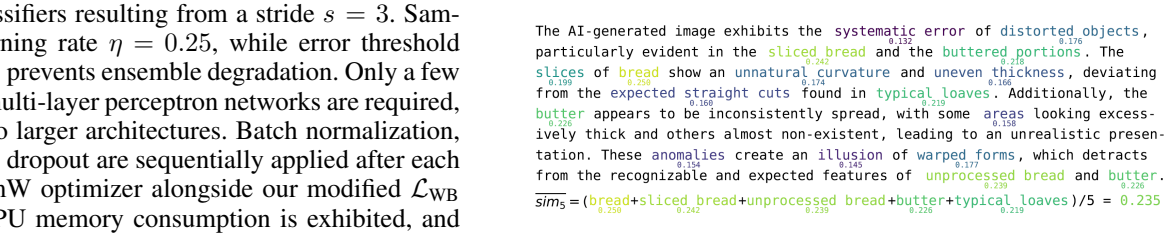

Figure 5: Illustration of phrase-image similarities in an explanation snippet and calculation of Top-5 average.

4.2 Error Explanation and Refinement

Flaw Classification We propose synthetic flaw detection as a new task, and introduce a simple baseline experimented on GenExplain. The same classifier architecture as in Section 4.1 is adopted, with output layer dimension adjusted to 14 to match the categories of flaw types. BCE Loss is used, and each label is predicted independently. Besides Mean Average Precision (mAP), Exact Match (EM) accuracy is used to measure the proportion of predictions where all 14 labels are correctly matched. Results are presented in Table 3.

Explanation Refinement We instruct `gpt-4o` to generate explanations, use `spacy` for phrase segmentation, and refine for 3 iterations. Top-5, Top-10, and Overall Similarity between text phrases and image regions are evaluated to guide refinement, and the top 10 phrases are retained, exemplified in Figure 5. Additionally, Type-Token-Ratio (TTR↑), normalized Shannon Entropy (SE↑) and Perplexity (PPL↓, tokenized using `gpt-2`) are employed to evaluate lexical di-

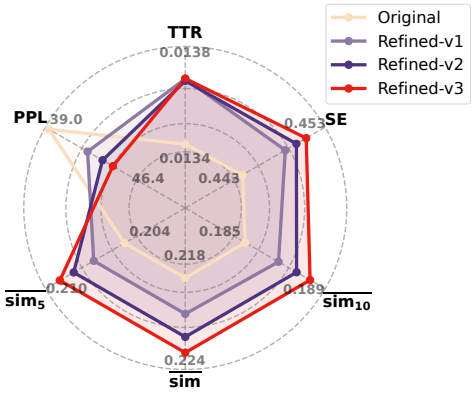

Figure 6: Metrics average across all 8 subsets of GenExplain regarding the initial explanations and their refined versions.

Method	MJ	SD1.4	SD1.5	ADM	GLI	WK	VQ	BG	Avg.
FreqNet	**85.67**	84.50	84.94	**70.25**	77.50	74.25	73.25	76.08	78.31
NPR	65.58	49.92	50.00	57.50	68.75	50.83	68.25	**78.17**	61.13
DRCT	**85.67**	81.58	**91.08**	70.00	78.58	70.92	68.25	73.75	77.48
ESIDE	84.08	**86.58**	86.06	70.00	**85.08**	**82.08**	**74.58**	76.58	**80.63**

Table 4: Robustness performance evaluated on perturbated test subsets of GenImage. Subset names are abbreviated.

versity, information density and fluency, with results shown in Figure 6. Throughout refinement iterations, all similarity metrics, TTR and SE improved. However, PPL also rose due to the retention of specialized or domain-specific terms uncommon in general language usage. Meanwhile, our *Original* setting reflects the effect of baselines directly utilizing MLLMs for explanation, falling short on most metrics.

4.3 Robustness Experiments

In real-world scenarios, images awaiting detection often exhibit degradation. Following established procedures (Wang et al. 2019, 2023; Lorenz, Durall, and Keuper 2023), and additionally incorporating post-processing operations to better mimic actual cases, we assess the resilience of our method to distribution shifts by introducing three types of perturbations to test images: Gaussian blur ($\sigma \in [0.5, 2.5]$ pixels), arbitrary rotation ($\theta \in [-45°, 45°]$), and illumination variation (brightness factor $\alpha \in [0.3, 1.8]$). All models are trained on the unmodified images, and the best-performing baselines are compared. As our method is exceptionally trained on those samples more challenging, superior robustness is also demonstrated, as observed in Table 4.

4.4 Ablation Studies

Noised Images and Ensembling Would simply using unnoised or fully-noised images perform better, and is performance increase merely due to ensembling? To test this hypothesis, we evaluated four different settings correspondingly leaving one unnoised and replacing intermediately-noised images with **fully-noised images**, ensembling on **unnoised images**, **removing ensembling**, and **eliminating misclassification-centric training**. Table 5 shows that our method achieves an accuracy increase of 5.71% on hard samples compared to ensembling entirely on unnoised images, while deactivating model weights and sample weights further decreases performance. Ensembling a model trained on unnoised images with multiple models trained on fully noised images degrades performance on both distributions, implying that features from varying timesteps are utilized.

Setting	DDIM	α_k	$w_{k,i}$	Acc (%)	Δ
ESIDE	✓	✓	✓	**94.98**/99.00	0.00/0.00
– DDIM-Interm.	✓	✓	✓	94.37/98.00	-0.61/-1.00
– DDIM	✗	✓	✓	89.27/99.00	-5.71/0.00
– α_k	✗	✗	✓	88.75/98.92	-6.23/-0.08
– $\alpha_k, w_{k,i}$	✗	✗	✗	88.53/**99.17**	-6.45/0.17

Table 5: Ablation studies of architectural components on detection results. Models are trained on GenImage/GLIDE.

Bandwidth	Suppression	Percentile	Acc (%)	Δ
0.06	0.1	0.15	92.52/98.83	-2.46/-0.17
0.08	0.2	0.10	92.74/98.92	-2.24/-0.08

Table 6: Effects of suppressing high-frequency peaks of Fourier power spectra quadrants. Models are trained and evaluated on reconstructions of GenImage/GLIDE.

High-Frequency Peaks Utilization For both synthetic and natural images, we suppressed the highest percentile of their Fourier frequencies to a fixed ratio, while masking the commonly-shared components located within a specific bandwidth along the axes, and then used images reconstructed based on these modified spectra for training and evaluation. Table 6 supports our insight that these frequency peaks could be captured by intermediate-step detectors to enhance detection capability regarding more questionable instances, as their suppression halves ensemble effect.

Additional illustrations, details and analyses, demonstrations of state-of-the-art performance on the recent Flux generator, and ablation studies on architecture, stride and timestep performance are presented in our extended version.

5 Conclusion

We present ESIDE, a novel pipeline for detecting and explaining synthetic images. We train an ensemble on noised images to directly utilize intermediate features introduced through DDIM inversion, circumventing conventional reconstruction measures. To improve human perception of fake images, we introduce an explanation generation and refinement module. Additionally, we construct two datasets, GenHard and GenExplain, comprising more challenging samples and providing categorized flaw types with explanations for AI-generated images. Extensive experiments show SOTA performance on both regular and harder images, with significant improvements on tougher samples. Our method also generalizes effectively, demonstrates robustness, and enables hybrid parallelism easily.

Acknowledgments

We sincerely thank the anonymous reviewers for their valuable feedback, which helped us increase the quality and rigor of this work. We are also grateful to our lab members for their arduous efforts in data curation during the construction of the GenExplain dataset. This research was supported partially by the National Natural Science Foundation of China (NSFC, No. 62076068).

References

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Chen, B.; Zeng, J.; Yang, J.; and Yang, R. 2024. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*.

Chen, J.; Yao, J.; and Niu, L. 2024. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*.

Corvi, R.; Cozzolino, D.; Poggi, G.; Nagano, K.; and Verdoliva, L. 2023. 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*, 973–982.

Cozzolino, D.; Poggi, G.; Corvi, R.; Nießner, M.; and Verdoliva, L. 2024. Raising the Bar of AI-generated Image Detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4356–4366.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dzanic, T.; Shah, K.; and Witherden, F. 2020. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33: 3022–3032.

Freund, Y.; and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119–139.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2021. Vector Quantized Diffusion Model for Text-to-Image Synthesis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10686–10696.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, Z.; Xia, B.; Lin, Z.; Mou, Z.; Yang, W.; and Jia, J. 2024. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*.

Kang, H.; Wen, S.; Wen, Z.; Ye, J.; Li, W.; Feng, P.; Zhou, B.; Wang, B.; Lin, D.; Zhang, L.; et al. 2025. Legion: Learning to ground and explain for synthetic image detection. *arXiv preprint arXiv:2503.15264*.

Laurier, L.; Giulietta, A.; Octavia, A.; and Cleti, M. 2024. The Cat and Mouse Game: The Ongoing Arms Race Between Diffusion Models and Detection Methods. *arXiv preprint arXiv:2410.18866*.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 201–216.

Li, M.; Qu, T.; Yao, R.; Sun, W.; and Moens, M.-F. 2023. Alleviating exposure bias in diffusion models through sampling with shifted time steps. *arXiv preprint arXiv:2305.15583*.

Li, Y.; Liu, X.; Wang, X.; Lee, B. S.; Wang, S.; Rocha, A.; and Lin, W. 2024. FakeBench: Probing Explainable Fake Image Detection via Large Multimodal Models. *arXiv preprint arXiv:2404.13306*.

Lin, L.; Amerini, I.; Wang, X.; Hu, S.; et al. 2024. Robust CLIP-based detector for exposing diffusion model-generated images. In *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–7. IEEE.

Liu, B.; Yang, F.; Bi, X.; Xiao, B.; Li, W.; and Gao, X. 2022. Detecting generated images by real images. In *European Conference on Computer Vision*, 95–110. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Y.; Li, X.; Zhang, J.; Hu, S.; and Lei, J. 2024. Dahfnet: Progressive fine-grained forgery image detection and localization based on dual attention. In *2024 3rd International Conference on Image Processing and Media Computing (ICIPMC)*, 51–58. IEEE.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.

Lorenz, P.; Durall, R. L.; and Keuper, J. 2023. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 448–459.

Luo, Y.; Du, J.; Yan, K.; and Ding, S. 2024. 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*, 17006–17015.

Ma, R.; Duan, J.; Kong, F.; Shi, X.; and Xu, K. 2023. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272*.

Mathys, M.; Willi, M.; and Meier, R. 2024. Synthetic Photography Detection: A Visual Guidance for Identifying Synthetic Images Created by AI. *arXiv preprint arXiv:2408.06398*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Reddy, B. S.; and Chatterji, B. N. 1996. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE transactions on image processing*, 5(8): 1266–1271.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Sha, Z.; Li, Z.; Yu, N.; and Zhang, Y. 2023. 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*, 3418–3432.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, J.; Ye, D.; and Zhang, Y. 2024. Trinity Detector: text-assisted and attention mechanisms based spectral fusion for diffusion generation image detection. *IEEE Signal Processing Letters*.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024a. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5052–5060.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024b. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28130–28139.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12105–12114.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and J'egou, H. 2020. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2019. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8692–8701.

Wang, Z.; Bao, J.; gang Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023. DIRE for Diffusion-Generated Image Detection. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 22388–22398.

Wen, S.; Ye, J.; Feng, P.; Kang, H.; Wen, Z.; Chen, Y.; Wu, J.; Wu, W.; He, C.; and Li, W. 2025. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*.

Wukong. 2022.

Xu, J.; Yang, Y.; Fang, H.; Liu, H.; and Zhang, W. 2024a. FAMSeC: A Few-shot-sample-based General AI-generated Image Detection Method. *IEEE Signal Processing Letters*.

Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; and Zhang, J. 2024b. FakeShield: Explainable Image Forgery Detection and Localization via Multi-modal Large Language Models. *ArXiv*, abs/2410.02761.

Ye, J.; Zhou, B.; Huang, Z.; Zhang, J.; Bai, T.; Kang, H.; He, J.; Lin, H.; Wang, Z.; Wu, T.; et al. 2024. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732*.

Zhong, N.; Xu, Y.; Qian, Z.; and Zhang, X. 2023. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *CoRR*.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2023. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36: 77771–77782.