Fast and Generalizable AI-Generated Image Detection via Model-Agnostic Feature Reconstruction

Qinghui He[1,2], **Haifeng Zhang**[1,2], **Bo Liu** *[,1,3] and **Yang Wei** *[,1,3]
[1]Chongqing Key Laboratory of Image Cognition,
Chongqing University of Posts and Telecommunications, Chongqing, China
[2]School of Computer Science and Technology
[3]School of Artificial Intelligence
{D250201011, D240201065}@stu.cqupt.edu.cn, {boliu, weiyang}@cqupt.edu.cn

Abstract

Generative models continue to advance rapidly in both fidelity and diversity, posing increasing challenges for reliably distinguishing generated images from real ones. Existing reconstruction-based detection methods often rely on assumptions tied to specific generative models, which leads to limited cross-model generalization and substantial computational overhead. In this paper, we propose **G**eneral **F**eature **R**econstruction **E**rror (GFRE), a fast and generalizable detection paradigm that leverages reconstruction behavior in a general-purpose representation space. Our key insight is that real and generated images exhibit consistently different reconstruction stability when projected into universal visual representations. Instead of tracing generator-specific artifacts, GFRE employs a lightweight autoencoder to model the reconstructability of image representations, producing a reconstruction signal that is inherently generator-agnostic and transferable across diverse generative processes. Extensive experiments on images synthesized by 18 different generative models demonstrate that GFRE consistently outperforms existing state-of-the-art methods, achieving a 4.70% improvement in detection accuracy and an 8.20% gain in cross-model generalization. Moreover, by avoiding costly generator inversion and diffusion-based reconstruction, GFRE reduces reconstruction error extraction time by up to 150×, enabling efficient and scalable deployment.

1 Introduction

In recent years, the rapid advancement of generative adversarial networks (GANs) and diffusion models (DMs) has significantly enhanced the fidelity and accessibility of image synthesis. While such advances enable a wide range of creative and practical applications, they also raise growing concerns about the misuse of synthetic images, including the large-scale dissemination of misinformation and manipulated visual content. These risks threaten the reliability of visual media and undermine public trust in digital information. As a result, developing efficient, reliable, and generalizable techniques for detecting AI-generated images has become an increasingly important and urgent challenge.

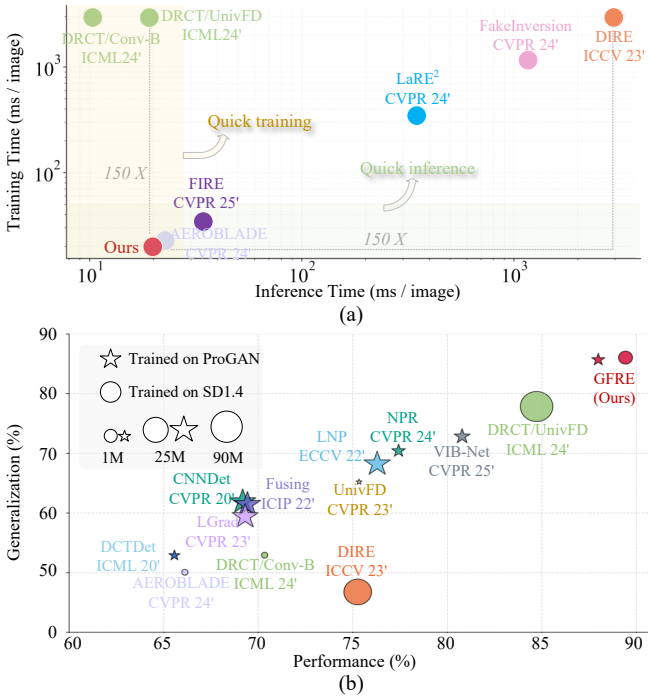


Figure 1: Efficiency and generalization comparison of AI-generated image detection methods. (a) Training and inference time per image. (b) Detection performance versus cross-model generalization under different training sources.

Early detection methods [Wang *et al.*, 2020; Frank *et al.*, 2020; Chandrasegaran *et al.*, 2021] primarily focus on identifying low-level artifacts introduced by convolutional generators. While these approaches are effective for images synthesized by GANs and early diffusion models, their performance degrades substantially when confronted with high-fidelity images produced by modern generative models. Even with retraining on higher-quality data, their generalization ca-

*Corresponding Authors.

pability across different generator families remains limited. More recent methods [Ojha *et al.*, 2023; Cozzolino *et al.*, 2024; Khan and Dang-Nguyen, 2024; Wang *et al.*, 2023; Chen *et al.*, 2024; Cazenavette *et al.*, 2024; Luo *et al.*, 2024; Ricker *et al.*, 2024] attempt to improve generalization by leveraging pre-trained representations, such as CLIP embeddings, or by adopting reconstruction-based strategies through model inversion. Despite these advances, robust detection under challenging and diverse generative settings remains difficult. In particular, diffusion-based reconstruction methods typically rely on multi-step DDIM sampling during feature extraction, which not only introduces cumulative reconstruction errors but also incurs substantial computational overhead. As a result, such approaches face significant limitations in real-time applicability and large-scale deployment.

To address the aforementioned challenges, we revisit source-invariant representations for AI-generated image detection and reconsider how reconstruction should be performed to support robust cross-model generalization. Through a series of exploratory analyses, we observe that existing reconstruction-based detectors, which operate primarily in the image space or generator-specific feature spaces, are often insensitive to subtle generative discrepancies and are easily influenced by redundant spatial patterns. Motivated by this observation, we reformulate reconstruction as a representation-centered process that focuses on intrinsic discrepancy patterns in general-purpose feature spaces—patterns that are frequently obscured by image-level reconstruction. This reformulation provides a more stable foundation for detection by emphasizing representation-level regularities rather than generator-specific artifacts. Building on these insights, we propose General Feature Reconstruction Error (GFRE), a detection framework that characterizes images based on their reconstruction behavior in a general-purpose representation space. GFRE does not rely on generator-specific inversion procedures and avoids the complex optimization processes commonly adopted in diffusion-based reconstruction methods. Instead, it employs a lightweight autoencoder to model the reconstructability of image representations, yielding a reconstruction signal that is both generator-agnostic and highly transferable across diverse generative processes. GFRE offers two key advantages. First, performing reconstruction entirely in a general representation space suppresses interference from redundant spatial signals and irrelevant semantics, resulting in stronger cross-model generalization—particularly when detecting GAN-generated images that differ structurally from diffusion models. Second, the lightweight autoencoder design enables fast and stable reconstruction, reducing reconstruction feature extraction time from seconds in prior approaches (e.g., DIRE [Wang *et al.*, 2023] and DRCT [Chen *et al.*, 2024]) to approximately 20 milliseconds. As a result, GFRE significantly improves detection efficiency while maintaining strong overall accuracy and cross-model generalization, as illustrated in Fig. 1. In summary, our contributions can be outlined as follows:

- We revisit source-invariant representations for AI-generated image detection and identify fundamental limitations of image-level reconstruction, showing that redundant spatial structures and semantic content often obscure subtle generative discrepancies.
- We propose GFRE, a reconstruction-based framework that reformulates reconstruction in a representation-centered manner, enabling generator-agnostic detection without relying on generator-specific inversion procedures.
- We design a lightweight feature autoencoder to model the reconstructability of general-purpose representations, leading to more stable reconstruction behavior and reducing time from seconds to milliseconds.
- Experiments across 18 generative models demonstrate the effectiveness of GFRE, achieving a 4.70% improvement in detection accuracy and an 8.20% gain in cross-model generalization over state-of-the-art methods.

2 Related Work

2.1 Types of Generated Images

The rapid development of image generation models has significantly improved the diversity and fidelity of synthetic content. Early GANs [Brock, 2018; Zhu *et al.*, 2017; Karras, 2017; Karras, 2019] achieved high-quality image synthesis through adversarial training mechanisms. However, their generation capability was constrained by the distribution of training data and susceptible to mode collapse. Subsequent models, including variational autoencoders and autoregressive models [Razavi *et al.*, 2019; Esser *et al.*, 2021; Ramesh *et al.*, 2022], extended the scope of generation tasks. Yet, these methods suffered from limitations in either generation speed or image resolution. More recently, diffusion models [Dhariwal and Nichol, 2021; Nichol *et al.*, 2021; Rombach *et al.*, 2022; Gu *et al.*, 2022] have emerged as the dominant architecture due to their iterative denoising paradigm, achieving breakthroughs in fidelity, diversity, and training stability. Given the diverse forgery patterns introduced by different generation mechanisms (e.g., adversarial training vs. denoising diffusion), this paper aims to design a general detection framework capable of identifying images generated by mainstream generative models.

2.2 Generalized Generated Image Detection

Early detection methods focused on identifying GAN-specific artifacts in spatial and frequency domains [Wang *et al.*, 2020; Frank *et al.*, 2020; Chandrasegaran *et al.*, 2021]. However, these detectors suffer significant performance drops when applied to diffusion models [Corvi *et al.*, 2023]. Consequently, research has shifted toward improving cross-model generalization. To address the challenge of cross-model generalization, recent studies have explored the use of pre-trained feature spaces and reconstruction errors for detection. For instance, pre-trained-based methods [Ojha *et al.*, 2023; Wu *et al.*, 2023; Cozzolino *et al.*, 2024; Khan and Dang-Nguyen, 2024; Wu *et al.*, 2025; Zhang *et al.*, 2025] demonstrate that the pre-trained feature space of CLIP offers superior adaptability to unseen generative models. Concurrently, reconstruction-based methods have gained traction.

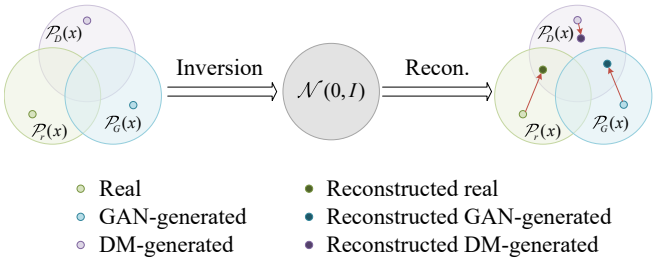


Figure 2: Visualization of the differences between real, GAN-generated, and DM-generated samples in the diffusion-based reconstructed representation space. $p_r(x)$ represents the distribution of real samples, $p_G(x)$ represents the distribution of GAN-generated samples, and $p_D(x)$ represents the distribution of DM-generated samples. After reconstruction, both real samples and GAN-generated samples are mapped into $p_D(x)$, making it difficult to simultaneously detect GAN- and DM-generated samples.

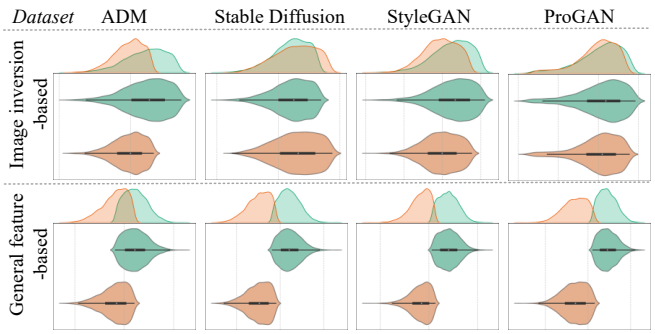


Figure 3: Visualization of error distributions for different reconstruction methods, including image inversion-based methods (e.g., DIRE and general feature-based methods (Ours), across various datasets. The violin plot illustrates the overall error distribution range. A smaller overlapping area indicates a higher distinguishability between real and generated images.

DIRE [Wang *et al.*, 2023] utilized diffusion reconstruction error for detection, though it suffers from high computational costs due to multi-step inversion. To improve efficiency and robustness, subsequent works introduced optimizations: $LaRE^2$ [Luo *et al.*, 2024] proposed single-step sampling, Fakeinversion [Cazenavette *et al.*, 2024] leveraged inversion noise maps, and DRCT [Chen *et al.*, 2024] employed contrastive training with hard samples. AEROBLADE [Ricker *et al.*, 2024] further showed that simple autoencoder reconstruction error is also effective. While these methods provide valuable insights into reconstruction-based detection, they still face challenges in terms of efficiency and generalization.

3 Methodology

3.1 Preliminaries

Recent studies [Wang *et al.*, 2023; Luo *et al.*, 2024; Cazenavette *et al.*, 2024; Chen *et al.*, 2024; Ricker *et al.*, 2024; Chu *et al.*, 2025] have shown that reconstruction-based strategies can be effective for detecting images generated by diffusion models. The reconstruction process typically involves two stages. First, an input image x is mapped to a latent noise variable x_T through an inversion process, where x_T follows a standard Gaussian distribution $\mathcal{N}(0, I)$. Subsequently, a multi-step denoising procedure is applied to x_T to produce a reconstructed image $\hat{x}$. These reconstruction-based approaches rely on the observation that diffusion-generated images can often be reconstructed more faithfully by their corresponding generative models than real images.

Based on this principle, mainstream detection methods employ pre-trained diffusion models (e.g., ADM or Stable Diffusion) to reconstruct input images and utilize the resulting error maps as discriminative features. While effective under matched-generation settings, such image-level reconstruction methods exhibit two fundamental limitations. First, reconstructing images in the pixel space inevitably introduces redundant semantic and spatial information, which may obscure subtle generation-specific discrepancies. Second, detection performance is highly sensitive to the alignment between the reconstruction model and the underlying image generation process. As a result, diffusion-based reconstruction methods often struggle to generalize to images generated by GANs or other non-diffusion models. Existing reconstruction-based detection methods are therefore primarily tailored to diffusion-generated images, relying on reconstruction processes that implicitly assume a close alignment between the reconstruction model and the image generation mechanism. Motivated by these, we analyze the failure modes of image-level reconstruction under cross-model settings and seek a more generalizable reconstruction paradigm that is less dependent on specific generative processes.

3.2 Analysis of General Feature Representation

As illustrated in Fig. 2, the reconstructed outputs are constrained to lie within the sample distribution of the diffusion model. While this property benefits the reconstruction of diffusion-generated images, it introduces an inherent limitation for general-purpose detection. Specifically, real images are also projected into the diffusion model's generative distribution during reconstruction, resulting in a distributional shift away from their original data manifold. Consequently, real images tend to exhibit larger reconstruction errors compared to diffusion-generated images. However, this assumption does not extend naturally to images generated by GANs. When GAN-generated images undergo diffusion-based reconstruction, they are likewise forced into the diffusion model's distribution, leading to substantial distributional shifts that are not meaningfully distinguishable from those observed for real images. As a result, diffusion-based reconstruction methods struggle to simultaneously discriminate between real images, GAN-generated images, and diffusion-generated images, fundamentally limiting their cross-model generalization capability.

Reconstructing representations in a general-purpose feature space provides a viable solution to this limitation. To empirically validate this observation, we conduct comparative experiments and analyze the distributions of reconstruction errors, as illustrated in Fig. 3. When using ADM as the reconstruction model (e.g., DIRE), image-level reconstruction errors exhibit substantial variability across datasets

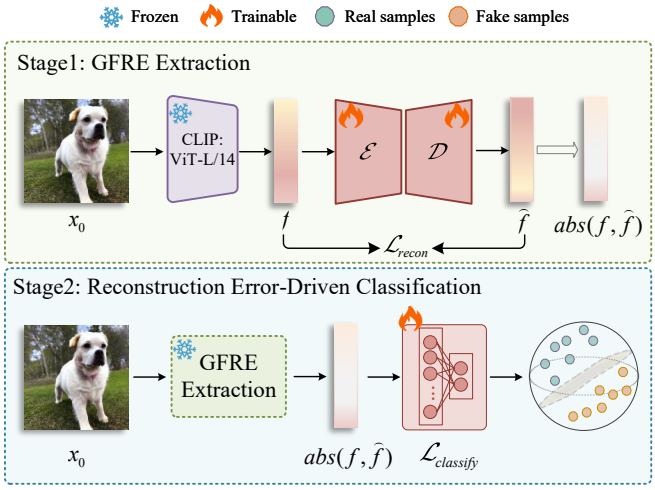


Figure 4: Overview of our method. In the first stage, we extract GFRE from the general representation space using a trainable autoencoder. In the second stage, we compute the reconstruction error by comparing the features of the input image x_0 with the reconstructed features obtained from the pre-trained autoencoder. Finally, an MLP classifier is employed to distinguish real / fake.

such as Stable Diffusion, ProGAN, and StyleGAN. In particular, the separation between real and generated images becomes notably weak for GAN-based datasets, indicating that image-level reconstruction is highly sensitive to the underlying generative model and susceptible to interference from redundant spatial and semantic information. In contrast, we adopt a feature-level reconstruction paradigm that discards the built-in feature extractor of diffusion models and instead leverages general-purpose representations extracted by large-scale pre-trained models (e.g., CLIP). Reconstruction errors are computed directly in this representation space. As shown in Fig. 3, feature-level reconstruction errors exhibit a more compact distribution, while the separation between real and generated images is consistently enhanced across both GAN- and diffusion-generated samples. This empirical evidence supports the effectiveness of general feature representations in capturing generation-induced discrepancies and highlights their advantage in enabling robust cross-model detection.

3.3 Overall Framework

In this part, we present our proposed method, which introduces a novel paradigm for detecting generated images by integrating large-scale pre-trained models with a lightweight autoencoder to capture subtle yet discriminative artifacts between real and generated images. As illustrated in Fig. 4, our framework operates in two stages: (1) GFRE Extraction and (2) Reconstruction Error-Driven Classification.

GFRE Extraction

In this stage, we extract the GFRE from a general-purpose representation space to construct a generator-agnostic forgery characterization. Given an input image x, we obtain its semantic representation using a frozen CLIP image encoder Φ:

$$f = \Phi(x), \quad f \in \mathbb{R}^d. \tag{1}$$

These representations are assumed to lie on a latent manifold $\mathcal{M} \subseteq \mathbb{R}^d$ shaped by natural image statistics learned by large-scale pre-training. Empirically, we observe that real images exhibit higher local variability in this representation space, whereas synthesized images tend to concentrate in more structured and lower-variability regions induced by generative priors.

To characterize how a given representation conforms to this underlying structure, we employ a lightweight autoencoder $\mathcal{A} = \{\mathcal{E}, \mathcal{D}\}$ defined as

$$\mathcal{E} : \mathbb{R}^d \to \mathbb{R}^k, \qquad \mathcal{D} : \mathbb{R}^k \to \mathbb{R}^d, \qquad k < d. \tag{2}$$

The reconstructed representation is given by

$$\hat{f} = \mathcal{D}(\mathcal{E}(f)). \tag{3}$$

Optimize the autoencoder using the reconstruction objective

$$\mathcal{L}_{\text{recon}} = |f - \hat{f}|_1, \tag{4}$$

which encourages the network to approximate the identity mapping on regions where the representation distribution is dense. As a result, the autoencoder learns a local approximation of the dominant subspace of the representation manifold $\mathcal{M}$, while deviations from this subspace are less faithfully reconstructed.

Define GFRE as the element-wise reconstruction residual

$$r = |f - \hat{f}|, \tag{5}$$

which quantifies the extent to which a representation deviates from the locally learned structure. To interpret this residual, consider the local tangent subspace $T_f\mathcal{M}$ at f and the associated projection operator $P_{T_f\mathcal{M}}$. When viewed as a nonlinear manifold approximator, the autoencoder primarily preserves variations within $T_f\mathcal{M}$, while reconstruction errors are dominated by components orthogonal to this subspace:

$$|f - \hat{f}| \approx \left|(I - P_{T_f\mathcal{M}})(f - \mu_f)\right|, \tag{6}$$

where μ_f denotes a local mean of neighboring representations.

Let Σ_{real} and Σ_{fake} denote the local covariance matrices of representations corresponding to real and generated images, respectively. Empirically, we observe that

$$\lambda_{\max}(\Sigma_{\text{real}}) > \lambda_{\max}(\Sigma_{\text{fake}}), \tag{7}$$

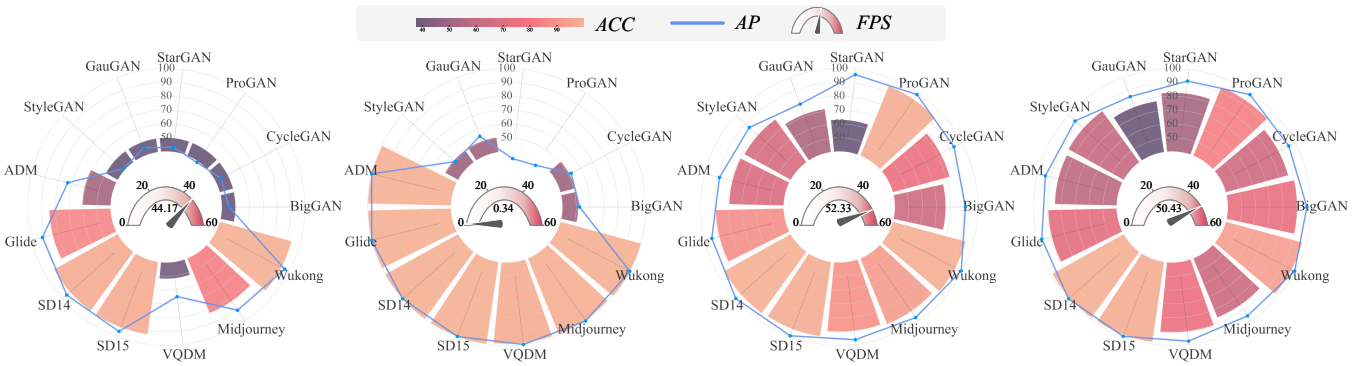
indicating that real images exhibit higher variability along directions that are less aligned with the dominant subspace captured by the autoencoder. Consequently, the expected reconstruction error satisfies

$$\mathbb{E}_{x \in \mathcal{X}_{\text{real}}}\left[|f - \hat{f}|_1\right] > \mathbb{E}_{x \in \mathcal{X}_{\text{fake}}}\left[|f - \hat{f}|_1\right]. \tag{8}$$

In practice, this behavior is consistently observed across different generative models:

$$r(x_{\text{real}}) > r(x_{\text{DM}}) \approx r(x_{\text{GAN}}), \tag{9}$$

demonstrating that GFRE captures representation stability differences between natural and synthesized images, and provides a robust and generator-agnostic detection signal.



(a) AEROBLADE (b) DIRE (c) DRCT (d) GFRE (Ours)

Figure 5: **Comparison of reconstruction-based detection methods under cross-source settings.** All methods are trained on the Stable Diffusion v1.4 dataset. Bar colors indicate classification accuracy (*ACC*), blue curves denote average precision (*AP*), and the inner gauges report inference speed in frames per second (*FPS*).

Reconstruction Error–Driven Classification

After training the autoencoder, its parameters are fixed and the reconstruction residual r is used as the discriminative representation. As discussed earlier, reconstruction errors of real and generated images exhibit consistently different statistical behaviors in the learned representation space (see Fig. 3), which provides a principled basis for downstream decision making.

In the second stage, we construct a reconstruction error dataset $\mathcal{D}_s = \{(r_i, y_i)\}_{i=1}^{N}$, where each residual vector r_i is obtained from the original dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ via the mapping $x_i \mapsto r_i$, and $y_i \in \{0, 1\}$ denotes whether x_i is real or generated.

Rather than classifying images directly from high-dimensional visual representations, we employ a lightweight MLP classifier $\mathcal{H}$ to model the class-conditional distributions of reconstruction errors. Given a residual r_i, the classifier outputs a posterior probability

$$p_i = \mathcal{H}(r_i), \tag{10}$$

indicating the likelihood that the input image is generated. The classifier is trained using the standard binary cross-entropy objective:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]\,. \tag{11}$$

This two-stage formulation decouples representation modeling from decision learning, allowing the classifier to focus solely on the statistical separation induced by GFRE while preserving computational efficiency.

4 Experiments

4.1 Experimental Setup

Training Datasets. Reconstruction-based detection methods have been shown to be particularly effective for images generated by diffusion models. To ensure fair comparison, we follow prior works and train diffusion-based detectors using images generated by Stable Diffusion v1.4 from the GenImage [Zhu *et al.*, 2024]. In addition, we also train detectors on the ProGAN training set [Wang *et al.*, 2020], which is a commonly adopted setting in existing GAN detection benchmarks. These two training sources enable a comprehensive evaluation under both diffusion- and GAN-based training regimes.

Evaluation Metrics. Following prior works [Wang *et al.*, 2020; Liu *et al.*, 2022], we adopt average precision (*AP*) and classification accuracy (*ACC*) as the primary evaluation metrics. *ACC* is computed using a decision threshold of 0.5.

Implementation Details. All experiments are implemented in PyTorch and conducted on NVIDIA H100 PCIe GPUs. We adopt the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 512, and apply early stopping to mitigate overfitting. CLIP:ViT-L/14 is used as a frozen backbone for feature extraction. In the first stage, the autoencoder is trained solely to model the dominant structure of the representation space using features extracted from Stable Diffusion v1.4 generated images, without accessing real images or class labels. In the second stage, reconstruction errors are computed for both real images and generated images and used to train a lightweight classifier for final decision making.

4.2 Comparison to Reconstruction-based Methods

To comprehensively evaluate the effectiveness of the proposed approach, we conduct a focused comparison with representative reconstruction-based detectors. Specifically, we select three state-of-the-art methods within this category as baselines: DIRE [Wang *et al.*, 2023], DRCT [Chen *et al.*, 2024], and AEROBLADE [Ricker *et al.*, 2024].

Figure 5 presents a comprehensive comparison of reconstruction-based detectors trained on Stable Diffusion v1.4 and evaluated on test sets covering both GAN-based and diffusion-based generators. The results reveal a pronounced source dependence in existing methods. AEROBLADE exhibits high inference efficiency, as reflected by its high FPS, but performs close to random guessing on most GAN datasets, with *ACC* values around 50%, while achieving higher accuracy on diffusion-generated images. DIRE shows the opposite behavior: it attains near-perfect *ACC* and *AP* on diffusion models but degrades severely on GAN-generated images, indicating strong reliance on the diffusion inversion

Method	Generative Adversarial Networks						Perceptual loss		Low-level vision		Faceswap	Diffusion Models							*ACC*
	Pro-GAN	Big-GAN	Cycle-GAN	Star-GAN	Style-GAN	Gau-GAN	CRN	IMLE	SITD	SAN	Deepfake	ADM	GLIDE	SDV1.4	SDV1.5	VQDM	Midj	Wukong	avg.
CNNDet	**99.99**	71.18	87.59	94.60	89.95	81.44	**85.87**	85.87	**91.39**	50.00	51.69	60.20	57.86	50.82	50.88	56.20	50.77	51.13	69.18
DCTDet	99.36	81.97	78.76	94.62	78.01	80.57	60.04	60.13	33.33	50.00	63.29	64.66	55.43	40.02	40.38	78.95	46.89	41.54	65.57
Fusing	99.90	77.32	87.01	97.04	85.21	76.95	85.17	88.54	80.28	54.56	53.76	56.50	57.15	51.04	51.34	55.09	52.12	51.67	69.43
LNP	99.09	88.05	91.63	**100.00**	95.89	79.70	50.78	50.81	68.89	43.61	53.99	**94.53**	87.84	68.58	68.06	77.89	79.92	73.89	76.29
LGrad	99.83	82.90	85.28	99.60	94.66	72.49	50.49	50.50	49.72	44.47	56.42	66.96	62.75	67.20	67.91	67.03	65.69	63.67	69.31
Univfd	99.81	95.08	98.33	95.75	84.93	**99.47**	56.59	69.11	62.22	56.62	68.57	66.94	62.53	63.76	63.57	85.43	56.24	71.06	75.33
NPR	99.98	83.33	91.41	98.65	**96.91**	78.36	50.00	50.00	58.89	66.44	78.17	70.06	80.47	82.55	83.01	63.79	81.61	80.00	77.42
VIB-Net	99.88	**95.35**	**98.75**	97.75	89.02	99.33	72.72	83.95	67.78	68.95	**82.35**	71.28	69.59	69.04	68.82	86.79	58.03	74.65	80.78
DIRE	59.18	51.68	60.75	62.11	57.85	82.69	64.69	64.68	45.00	48.32	60.15	44.69	59.14	47.53	47.02	73.83	53.13	50.09	57.36
AEROBLADE	50.01	50.05	50.00	50.23	50.00	50.00	72.82	76.30	83.06	76.85	67.90	60.62	84.47	92.67	92.93	51.90	82.85	93.82	68.69
FakeInversion	97.26	53.48	68.89	61.26	62.55	65.01	54.81	58.23	72.50	50.46	58.22	52.73	52.71	49.65	49.58	50.80	52.96	49.83	58.94
Ours	97.49	92.50	95.80	86.59	86.52	96.37	82.80	**91.07**	69.72	**85.84**	67.10	82.46	**87.89**	**95.12**	**94.83**	**90.34**	**86.95**	**94.28**	**87.98**

Table 1: **The *ACC* values of different methods trained on GAN source images.** Data in bold represents the best, while data underlined represents the second best.

Method	Generative Adversarial Networks						Perceptual loss		Low-level vision		Faceswap	Diffusion Models							*AP*
	Pro-GAN	Big-GAN	Cycle-GAN	Star-GAN	Style-GAN	Gau-GAN	CRN	IMLE	SITD	SAN	Deepfake	ADM	GLIDE	SDV1.4	SDV1.5	VQDM	Midj	Wukong	avg.
CNNDet	**100.00**	85.99	94.93	99.04	99.82	90.80	**98.89**	**98.92**	**97.50**	69.19	87.60	76.94	74.02	59.76	60.18	70.72	58.90	57.41	82.26
DCTDet	99.99	93.62	84.78	99.49	88.98	82.86	82.27	69.78	35.76	49.50	70.77	63.73	54.73	38.51	38.42	86.02	47.25	40.44	68.16
Fusing	**100.00**	90.76	95.50	99.82	99.48	88.32	90.32	96.03	71.13	77.34	71.13	74.85	77.45	65.31	65.62	75.42	69.91	64.53	81.83
LNP	99.98	95.20	98.14	**100.00**	99.56	83.25	56.75	94.62	96.92	45.29	57.72	**99.25**	**96.45**	89.77	89.30	95.54	94.39	91.16	87.96
LGrad	**100.00**	90.76	94.02	99.98	99.81	79.29	61.65	67.06	38.91	45.10	71.71	71.84	75.96	70.91	71.73	70.23	71.43	66.51	74.83
Univfd	**100.00**	**99.27**	99.80	99.37	97.56	**99.98**	96.59	98.61	63.84	78.81	81.76	87.13	84.26	86.56	86.19	96.65	74.61	91.34	90.13
NPR	**100.00**	90.18	98.73	99.99	**99.97**	78.83	50.01	50.01	60.06	73.82	89.24	82.22	93.82	93.42	93.58	68.42	94.29	87.99	83.59
VIB-Net	**100.00**	99.11	**99.83**	99.49	98.50	99.93	89.38	95.14	66.90	88.78	**92.25**	86.89	88.09	87.10	86.62	96.32	75.09	90.85	91.13
DIRE	69.76	51.92	64.76	67.98	71.40	88.18	94.61	92.78	37.98	54.89	67.90	44.99	58.11	47.46	47.06	85.18	55.55	51.25	63.99
AEROBLADE	45.24	44.63	39.08	43.51	45.63	41.45	73.07	74.76	83.25	81.74	76.16	73.18	91.93	95.85	96.10	65.20	90.61	96.52	69.88
FakeInversion	99.65	64.23	85.61	67.30	79.29	86.12	80.16	86.40	83.60	61.23	60.85	65.65	68.51	49.16	49.52	56.06	62.74	47.78	69.66
Ours	99.70	97.99	99.38	96.02	95.05	99.39	88.53	96.34	73.23	**94.10**	75.25	91.58	94.96	**99.63**	**99.51**	**97.08**	**94.46**	**99.12**	**93.96**

Table 2: **The *AP* values of different methods trained on GAN source images.**

and denoising process. DRCT partially alleviates this issue by achieving more balanced performance across datasets; however, its accuracy on several GAN subsets remains notably lower than on diffusion-based subsets, and its inference speed is still constrained by diffusion-based reconstruction.

In contrast, GFRE demonstrates a substantially more consistent performance profile across all evaluated datasets while maintaining high inference efficiency. On GAN-generated images, GFRE achieves an average *ACC* of 86.05% and an average *AP* of 93.53%, outperforming DRCT by 8.20% in *ACC* and 2.60% in *AP*. When averaged over all evaluated datasets, GFRE attains 89.42% *ACC* and 95.77% *AP*, compared to 84.72% *ACC* and 93.94% *AP* achieved by DRCT, the strongest competing reconstruction-based baseline. Notably, GFRE sustains a high *FPS* comparable to lightweight methods such as AEROBLADE, while significantly outperforming them in detection accuracy.

4.3 Cross-Model Comparison

In this section, we compare our method with state-of-the-art detectors trained on ProGAN and evaluated across diverse generative models. While existing methods perform well when training and testing data are drawn from the same distribution, real-world scenarios require robustness to unseen generative processes, which remains a challenging problem.

Tables 1 and 2 report the *ACC* and *AP* of different methods on a wide range of datasets. Most methods achieve strong performance on GAN-generated images but exhibit a noticeable performance degradation on unseen subsets, particularly those generated by diffusion models. This trend suggests that detectors trained on GAN data may rely on generation-specific characteristics that do not transfer well to diffusion-based images, leading to reduced generalization performance under cross-model evaluation. Although UnivFD and NPR demonstrate relatively more stable behavior across generative models, their performance remains limited on high-quality diffusion datasets.

In contrast, GFRE consistently achieves higher accuracy in distinguishing real and generated images across both in-domain and cross-model settings. Compared to the strongest competing baseline, GFRE improves *ACC* and *AP* by 7.20% and 2.83%, respectively, when averaged over all datasets. Under a stricter cross-model evaluation that excludes GAN-generated images, GFRE further improves *ACC* by 12.87% and *AP* by 5.03%. These results demonstrate that GFRE offers improved robustness to unseen generative models and more reliable generalization across diverse generators.

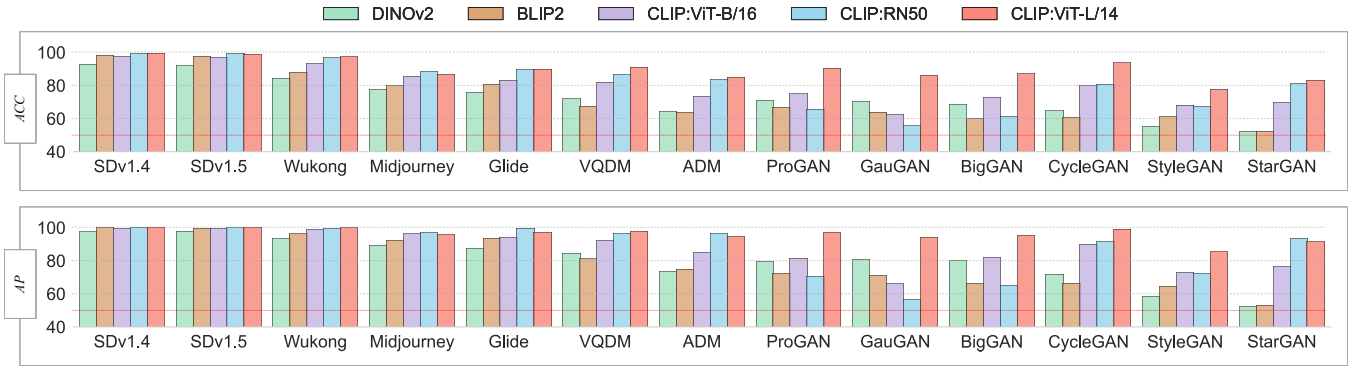


Figure 6: **Ablation study on different types of general feature extractors.** All methods are trained on the SDv1.4 dataset, and the results are averaged over the GAN source and DM source. CLIP:ViT-L/14 achieves superior performance.

4.4 Results on Different Types of Models

To evaluate the impact of different general feature extractors on detection performance, we compare several representative pre-trained models using the Stable Diffusion v1.4 subset of GenImage for training. Specifically, we evaluate the self-supervised visual representation DINOv2 [Oquab *et al.*, 2023], the multimodal-aligned BLIP2 [Li *et al.*, 2023], and CLIP [Radford *et al.*, 2021] with three image encoder variants (ResNet-50, ViT-B/16, and ViT-L/14). As shown in Fig. 6, all feature extractors achieve strong performance on seen diffusion-generated datasets, indicating that reconstruction error computed in a general representation space provides a robust detection signal independent of specific generator architectures. On unseen datasets, however, clear performance differences emerge across representation types. Among the evaluated models, CLIP-based representations consistently outperform DINOv2 and BLIP2, with CLIP:ViT-L/14 achieving the best overall results in both *ACC* and *AP*. This observation suggests that representations learned from large-scale vision–language supervision are particularly effective for capturing generation-induced discrepancies that generalize across different generative model families. In contrast, representations learned solely from visual objectives or reconstruction within diffusion-specific feature spaces exhibit limited transferability to images generated by unseen models. These results further support the advantage of adopting general-purpose representations for feature-level reconstruction in cross-model detection.

4.5 Robustness to Unseen Perturbations

In real-world scenarios, generated images are often subject to post-processing operations. These perturbations may partially suppress generation-specific traces, posing challenges for detectors that rely heavily on low-level visual cues. To evaluate robustness under such conditions, we assess detection performance under two common post-processing operations: JPEG compression and Gaussian blur. As shown in Fig. 7, GFRE consistently maintains strong detection performance across all compression and blur levels in terms of both *ACC* and *AP*. In contrast, NPR exhibits a pronounced performance degradation as perturbation strength increases, particularly under JPEG compression. DRCT shows relatively stable behavior under moderate perturbations but degrades under stronger compression and blur. This behavior suggests that detectors relying primarily on fragile low-level visual patterns may be vulnerable to post-processing operations that remove or distort such cues.

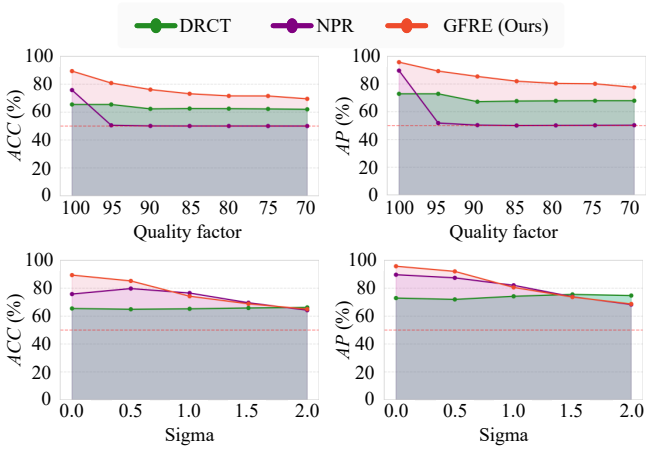


Figure 7: **Robustness evaluation under unseen post-processing perturbations.** Detection performance under JPEG compression (top row) and Gaussian blur (bottom row). All methods were trained on Stable Diffusion v1.4.

5 Discussion and Conclusion

In this paper, we propose GFRE to improve the generalization and efficiency of reconstruction-based generated image detection. By performing reconstruction in a general representation space rather than relying on generator-specific image or latent reconstructions, GFRE reduces source dependence and provides a more transferable detection signal across diverse generative models. Extensive experiments demonstrate that GFRE achieves a higher detection accuracy with a substantially lower computational cost than existing reconstruction-based methods, highlighting the effectiveness of feature-level reconstructability for practical detection. In addition, GFRE remains sensitive to severe post-processing and other distribution shifts that may distort feature representations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62376046, 62406047, U24B20182, 62172067, 62536002, and 62561160098, in part by the New Chongqing Youth Innovation Talent Program under Grant CSTB2025YITP-QCRCX0074, and in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJQN202400608.

References

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

[Cazenavette *et al.*, 2024] 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.

[Chandrasegaran *et al.*, 2021] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at fourier spectrum discrepancies for cnn-generated images detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7200–7209, 2021.

[Chen *et al.*, 2024] 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.

[Chu *et al.*, 2025] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12830–12839, 2025.

[Corvi *et al.*, 2023] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Cozzolino *et al.*, 2024] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024.

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

[Esser *et al.*, 2021] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[Frank *et al.*, 2020] 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.

[Gu *et al.*, 2022] 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.

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

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

[Khan and Dang-Nguyen, 2024] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1006–1015, 2024.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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

[Luo *et al.*, 2024] 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.

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

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

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Ramesh *et al.*, 2022] 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.

[Razavi *et al.*, 2019] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[Ricker *et al.*, 2024] 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.

[Rombach *et al.*, 2022] 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.

[Wang *et al.*, 2020] 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.

[Wang *et al.*, 2023] 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.

[Wu *et al.*, 2023] Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800*, 2023.

[Wu *et al.*, 2025] Shiyu Wu, Jing Liu, Jing Li, and Yequan Wang. Few-shot learner generalizes across ai-generated image detection. In *Forty-second International Conference on Machine Learning*, 2025.

[Zhang *et al.*, 2025] Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal ai-generated image detection by variational information bottleneck network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23828–23837, 2025.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[Zhu *et al.*, 2024] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024.