

Enabling Supervised Learning of Generative Signatures for Generalized AI-Generated Images Detection

Jianwei Fei^{1*} Yunshu Dai² Xiaoyu Zhou³ Zhihua Xia³ Alessandro Piva¹

¹ Department of Information Engineering, University of Florence, Florence, Italy

² School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen, China

³ College of Cyber Security, Jinan University, Guangzhou, China

Abstract

Extracting reliable generative traces in generated images is critical for AI-generated images (AIGIs) detection. However, a fundamental challenge exists: AIGIs inherently contain generative traces with no trace-free counterpart available, making supervised extraction of these artifacts infeasible. In this work, we overcome this through a surrogate supervision framework. We design a dynamic reconstructor that simulates diverse generative traces on real images through stochastically varied architectures and parameters. The reconstruction residuals serve as supervision to train an extractor that learns to isolate traces, i.e., generative signatures (GenSign). A detector then fuses extracted GenSign with RGB features to distinguish real images from AIGIs. Our key insight is that sufficient architectural diversity in simulation enables effective transfer to real-world generators, resolving the absence of ground truth GenSign. Extensive experiments across four benchmarks demonstrate state-of-the-art generalization, confirming that our simulation-based learning paradigm is capable of extracting general and transferable forensic features. Code available at <https://github.com/jumpycat/GenSign>

1. Introduction

The detection of AI-generated images (AIGIs) has become a critical task in visual forensics, as generative models are now capable of producing synthetic images with remarkable realism, posing threats to media authenticity and public trust [2, 9, 10, 26, 29]. Given the rapid and continuous evolution of generative techniques, it is essential to ensure that AIGI detectors generalize well to unseen or novel models in order to build practical and reliable detection systems.

In response, various strategies have been explored to enhance the generalization of AIGI detectors across unseen

generative models [7, 12, 42, 45, 48]. Among these, extracting intrinsic traces left by generative models has emerged as a promising solution [22, 38, 39, 51]. Recent studies have provided evidence that generative models leave distinctive and detectable traces in their outputs. For instance, Yu *et al.* [50] and Marra *et al.* [28] showed that Generative Adversarial Networks (GANs) have unique fingerprints even under minor training variations, e.g. different initialization seeds. Their findings reveal that these fingerprints are *instance-specific*¹ rather than architecture-specific. Beyond GANs, recent work has confirmed that diffusion models similarly leave instance-specific traces, observable as spectral anomalies in the noise residuals [7, 14, 43]. These results show that the generative traces persist across generated images and can serve as reliable forensic cues.

However, accurately extracting generative traces remains challenging due to the lack of a ground-truth trace or trace-free counterpart for supervision. Existing methods typically treat these traces as noise-like residuals, often concentrated in high-frequency components, and rely on hand-crafted filters [48] or CNN-based denoisers [7]. Since these extractors are optimized for generic denoising rather than isolating model-specific artifacts, the resulting residuals cannot reliably capture generative signals left in AIGIs.

To this end, we introduce a surrogate training strategy based on the proposed Dynamic-Architecture Image Reconstructor (DAIR), which reconstructs real images using stochastically varied architectures to simulate AIGIs generated by various generative models. Using the reconstruction residuals as ground-truth, we train an extractor to isolate artifacts, i.e., generative signature (GenSign), from image content. A key insight is that GenSign extracted from real images differs substantially from that of AIGIs, providing strong cues for distinguishing real and AIGIs. We integrate the extractor into an AIGI detector that leverages both

¹Instance denotes a unique model obtained through a particular training process. Models with identical architecture are considered distinct instances if trained with different conditions (e.g., random seeds or data).

*Corresponding author, fei.jianwei@163.com

GenSign and RGB features. DAIR also serves as a self-augmentation module during the detector training to enrich training data with diverse simulations. Our simulation-to-reality strategy allows us to generalize to unseen generative models, achieving state-of-the-art performance in cross-model AIGI detection. Our main contributions are:

- We propose a Dynamic-Architecture Image Reconstructor equipped with Context-Conditioned Dynamic Convolution that simulates diverse generative signatures on real images, thereby enabling supervised learning of signature extraction with ground-truth references.
- We design a dual-stream detector that integrates learned generative signatures with RGB features, achieving generalized AIGI detection that remains agnostic to both generative architectures and image content.
- We conduct extensive experiments on four AIGI detection benchmarks, demonstrating state-of-the-art generalization across different image generative models.

2. Related Work

2.1. Generative Traces in AIGIs

AIGIs contain distinctive generative traces that serve as intrinsic cues distinguishing generated images from real ones. Yu *et al.* [50] and Marra *et al.* [28] introduced *model fingerprints*, demonstrating that GANs leave distinct traces in generated images. These fingerprints are unique, even minor variations in architecture, training, or optimization yield different traces [13, 15]. Frank *et al.* [14] revealed that GANs leave frequency-domain artifacts due to upsampling operations [11, 14]. However, these traces are brittle to post-processing [32] and fail on the diffusion models (DMs) [8]. Recent work redefines fingerprints as intrinsic, content-independent properties. Song *et al.* [38] interpreted fingerprints as imperfections arising from inaccurate manifold approximation during generation, and further generalized this perspective through Riemannian geometric [37].

2.2. AIGIs Detection

2.2.1. Generalization Gap and Content Bias

Early detection methods achieved high accuracy in closed-set scenarios by exploiting low-level artifacts in the spatial [30, 39] or frequency domains [14, 33]. However, these detectors suffer from severe performance degradation in open-world settings, when confronted with images from novel, previously unseen generative architectures [15, 32]. While Wang *et al.* [44] observed some transferability among different GANs, this generalization capability collapses when transferring across distinct model families, such as from GANs to DMs [31, 55].

Another challenge identified by Zheng *et al.* [53] is the issue of *semantic artifacts*. Detectors often overfit to content biases (e.g., specific object classes in training data)

rather than learning true generative traces. This leads to a dilemma where performance drops significantly when the semantic domain shifts, even if the generative model remains constant. Consequently, a robust detector must simultaneously decouple forensic features from image semantics while maintaining generalization across diverse generative architectures.

2.2.2. Generalized AIGI Detection

Foundation Model-based Detectors. Some recent methods exploit the rich features of large-scale vision-language foundation models to avoid the training of specific forensic models [1, 19, 52]. UnivFD [31] demonstrated that the frozen CLIP features possess generalization capabilities for AIGI detection, with simple k-NN or linear probing outperforming trained specialized detectors. This paradigm has seen wide adoption, encompassing explicit real/fake concept injection [42], universal video forgery detection [21], and the integration of multi-modal cues or adapter modules for various forensic tasks [20, 25, 48].

Reconstruction and Training-Free Methods. An alternative paradigm exploits the reconstruction bias of generative models, hypothesizing that AIGIs are more easily reconstructed by their source models than real images. DIRE [45] pioneered this approach using DDIM inversion error, with subsequent works refining this via latent space analysis [27] or text-guided inversion [4]. However, as these methods require access to specific models, training-free approaches have emerged to enhance practicality. These methods focus on universal structural anomalies, such as autoencoder reconstruction errors [35], aiming to capture fundamental generative traces without supervision.

3. Method

Our methodology aims to address the challenge: **how to learn universal generative signatures in a supervised manner for detecting AIGIs?** We hypothesize that the generative traces of AIGIs can be isolated via residual computation, *i.e.*, subtracting generated images from their idealized trace-free version. However, direct separation is infeasible, as this trace-free ground truth does not exist for AIGIs. Therefore, we formulate a surrogate supervision strategy and develop a three-stage framework (see Fig. 1) to address the lack-of-supervision problem.

How to Create the Supervision? To overcome the lack-of-supervision problem, we train a Dynamic-Architecture Image Reconstructor (DAIR) $\mathcal{R}$. It reconstructs real images through stochastically varied architectures and parameters, thereby imparting diverse simulated GenSign on real images I (§3.1). The resulting reconstruction residual, $R_{gt} = I - \mathcal{R}(I)$, serves as our surrogate ground-truth.

How to Learn a Generalizable Signature Extractor? Enabled by the surrogate supervision from DAIR, our sec-

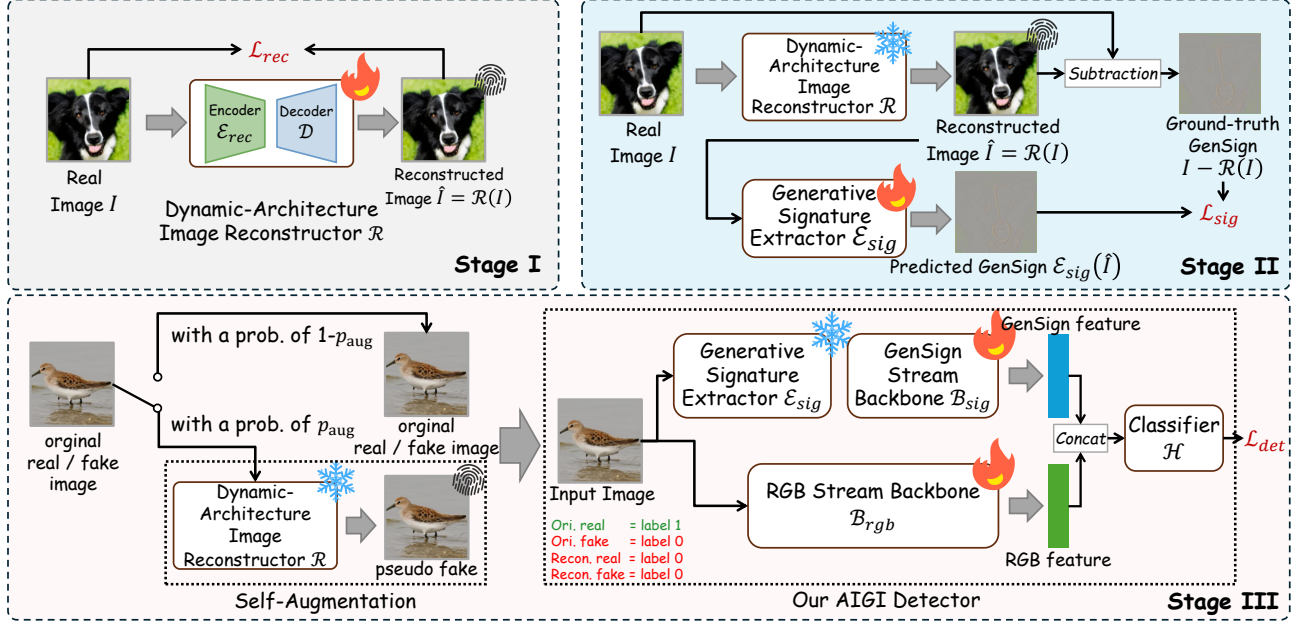


Figure 1. Overview of our three-stage framework for generative signature extraction and AIGI detection.

ond stage is to train a Generative Signature Extractor $\mathcal{E}_{sig}$. We task it to predict the residual R_{gt} , thereby learning to isolate GenSign from AIGIs (§3.2). The high diversity of $\mathcal{R}$ enables $\mathcal{E}_{sig}$ to generalize well.

How to Build a Generalized AIGI Detector? Finally, with $\mathcal{E}_{sig}$, we construct a dual-stream AIGI detector $\mathcal{M}$, which fuses RGB image I with the extracted GenSign $\mathcal{E}_{sig}(I)$ for real/fake classification (§3.3). To ensure $\mathcal{M}$ relies on GenSign rather than content biases, the frozen $\mathcal{R}$ is re-purposed as a self-augmentation module. This enriches the training distribution with additional generative patterns and forces the model to learn GenSign as the discriminative cue. We describe the details as follows.

3.1. Dynamic-Architecture Image Reconstructor

We propose DAIR $\mathcal{R}$ to synthesize diverse surrogate ground truths. It consists of an encoder $\mathcal{E}_{rec}$ and a dynamic decoder $\mathcal{D}$ whose architecture and parameters are randomly varied at each forward pass. It is trained to reconstruct real images $I \in \mathbb{R}^{H \times W \times 3}$ under these stochastic configurations, thereby simulating a wide range of generative models.

Specifically, the encoder $\mathcal{E}_{rec}$ is a hierarchical convolutional neural network (CNN) with L downsampling stages, which produces feature maps $\{F_i\}_{i=1}^L$ at different scales. The decoder $\mathcal{D}$ introduces architectural and parametric diversity. In each forward pass, it randomly selects a starting feature map F_s (where $s \sim \text{Uniform}(1, \dots, L)$) and reconstructs the image $\hat{I}$ through a series of randomized upsampling stages. Each stage consists of:

- **Up-sampling Cell** performs $2 \times$ upsampling by randomly

sampling an operation from a predefined pool $\mathcal{O}_{up} = \{\text{Bilinear}, \text{Nearest-Neighbor}, \text{Bicubic}, \text{Pixel Shuffle}\}$.

- **Decoder Cell** is the basic module with a randomized internal architecture. It samples a normalization layer from $\mathcal{O}_{norm} = \{\text{BatchNorm}, \text{InstanceNorm}, \text{GroupNorm}\}$, an activation from $\mathcal{O}_{act} = \{\text{GELU}, \text{SiLU}, \text{LeakyReLU}\}$, and a proposed context-conditioned dynamic convolution, then executes them in a randomly permuted order.

To simulate parameter-level variations, we introduce Context-Conditioned Dynamic Convolution (CCDC). CCDC maintains K trainable base kernels $\{\mathbf{W}_i \in \mathbb{R}^{C_{out} \times C_{in} \times k \times k}\}_{i=1}^K$, where C_{out}, C_{in}, k , k indicates that each kernel has C_{out} output channels (filters), C_{in} input channels, and a spatial size of $k \times k$. Given an input feature h , a router network $g(\cdot)$ generates filter-wise combination weights $\mathbf{A} = g(h) \in \mathbb{R}^{K \times C_{out}}$ via global average pooling and an MLP. The dynamic kernel is computed as:

$$\mathbf{W}_{dyn}[c, :, :, :] = \sum_{i=1}^K \mathbf{A}[i, c] \cdot \mathbf{W}_i[c, :, :, :], \quad (1)$$

where $\mathbf{A}[i, c]$ is the weight for the c -th filter of i -th base kernel, and $\mathbf{W}_i[c, :, :, :]$ denotes the c -th filter of base kernel $\mathbf{W}_i$. This enables decoders sharing the same architecture to instantiate distinct parameters for different inputs, facilitating the learning of instance-level GenSign.

The reconstructor $\mathcal{R}$ is trained end-to-end on real images $I \in \mathcal{I}_{real}$ to minimize a reconstruction loss $\mathcal{L}_{rec}$, which combines Mean Squared Error (MSE) and the Learned Perceptual Image Patch Similarity (LPIPS) metric:

$$\mathcal{L}_{rec} = \mathbb{E}_{I \in \mathcal{I}_{real}} \left[\lambda_1 \|I - \hat{I}\|_2^2 + \lambda_2 \cdot \text{LPIPS}(I, \hat{I}) \right], \quad (2)$$

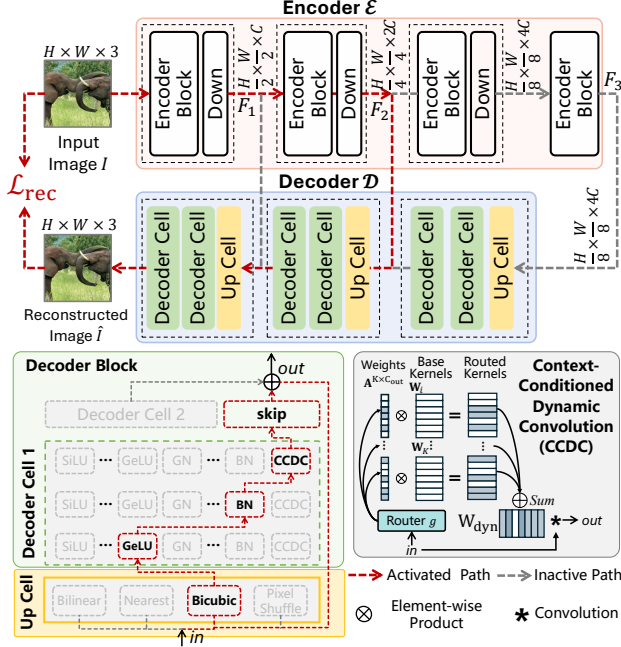


Figure 2. Architecture overview of the DAIR $\mathcal{R}$.

where $\hat{I} = \mathcal{R}(I) = \mathcal{D}(\mathcal{E}_{rec}(I))$. Once trained, the residual $R_{gt} = I - \mathcal{R}(I)$ serves as the surrogate ground-truth that our next stage, the Signature Extractor, will learn to predict.

3.2. Generative Signature Extractor

With DAIR $\mathcal{R}$ trained and frozen, we train the Generative Signature Extractor $\mathcal{E}_{sig}$ to predict the generative signature from images, with R_{gt} as the ground-truth supervisory signal. $\mathcal{E}_{sig}$ is implemented as a fully convolutional network (FCN) that takes an image I as input and outputs the predicted signature $R_{pred} = \mathcal{E}_{sig}(\mathcal{R}(I))$. During training, $\mathcal{E}_{sig}$ is optimized to approximate the frozen reconstructor’s residual R_{gt} by minimizing the mean squared error (MSE):

$$\mathcal{L}_{sig} = \mathbb{E}_{I \in \mathcal{I}_{real}} [\|\mathcal{E}_{sig}(\mathcal{R}(I)) - (I - \mathcal{R}(I))\|_2^2]. \quad (3)$$

Through this process, $\mathcal{E}_{sig}$ learns to isolate the generative signature embedded in images. Once trained, it can extract the corresponding signature from an unseen input, such as a synthetic image I_{fake} , producing $\mathcal{E}_{sig}(I_{fake})$ that effectively reveals its generative traces. This frozen, pre-trained $\mathcal{E}_{sig}$ now serves as the core forensic analysis component for our final AIGI detector.

3.3. GenSign-Based AIGI Detector

With the extractor $\mathcal{E}_{sig}$, we construct a dual-stream detector $\mathcal{M}$ that jointly leverages the extracted GenSign and RGB image. The detector comprises two backbones, $\mathcal{B}_{rgb}$ and $\mathcal{B}_{sig}$, and a classifier $\mathcal{H}$. The RGB Stream processes the RGB image I and outputs feature $f_{rgb} = \mathcal{B}_{rgb}(I) \in \mathbb{R}^{D_{rgb}}$.

For the Signature Stream, the input image I is passed through the frozen, pre-trained extractor $\mathcal{E}_{sig}$ to obtain the GenSign $R = \mathcal{E}_{sig}(I)$. R is then fed into a backbone $\mathcal{B}_{sig}$ to produce the feature $f_{sig} = \mathcal{B}_{sig}(R) \in \mathbb{R}^{D_{sig}}$. The features f_{rgb} and f_{sig} are concatenated $f = [f_{rgb} \oplus f_{sig}]$, and passed to $\mathcal{H}$ which predicts the final logit $y = \mathcal{H}(f)$.

Self-Augmentation for Generalization. We re-purpose the frozen DAIR $\mathcal{R}$ as a self-augmentation module. This is essential for ensuring the detector $\mathcal{M}$ uses subtle traces rather than overfitting to content biases. Specifically, during the training of the detector $\mathcal{M}$, a subset of training images $I_{aug} \subset \{\mathcal{I}_{real} \cup \mathcal{I}_{fake}\}$ is randomly selected with probability p_{aug} and reconstructed as $\hat{I} = \mathcal{R}(I_{aug})$. Then, the final label assignment follows a rule:

Original real I_{real}	$\rightarrow$ label = 1 (real)
Original fake I_{fake}	$\rightarrow$ label = 0 (fake)
Reconstructed real $\mathcal{R}(I_{real})$	$\rightarrow$ label = 0 (fake)
Reconstructed fake $\mathcal{R}(I_{fake})$	$\rightarrow$ label = 0 (fake)

All reconstructed images $\hat{I} = \mathcal{R}(I_{aug})$, regardless of their original labels, are assigned a fake label, i.e., 0, as the reconstruction process intentionally introduces model-specific artifacts. This strategy enriches the training distribution and compels the detector to distinguish fakes by generative traces rather than content or dataset bias. The detector $\mathcal{M}$ is trained using Binary Cross-Entropy (BCE):

$$\mathcal{L}_{det} = -\mathbb{E}_{I \in \mathcal{I}} [y(I) \cdot \log \mathcal{M}(I) + (1 - y(I)) \cdot \log(1 - \mathcal{M}(I))], \quad (4)$$

where $\mathcal{M}(I) \in [0, 1]$ is the predicted probability that image I is real, and $y(I)$ denotes the final label assignment.

4. Experiment

4.1. Experimental Setup

Datasets. We adopt the same settings to prior works such as ForenSynths [44] and GenImage [55], using images from ProGAN [18] (Table 1, 3) or SDv1.4 [36] (Table 2) as the training data. To evaluate generalization to unseen generative models, we conduct experiments on four widely used benchmarks, including **UniversalFakeDetect** [31], **AIGCDetectBenchmark** [54] **AIGIBenchmark** [23], and **GenImage** [55]. These datasets encompass AIGIs from diverse generative models, including GANs, diffusion models, commercial text-to-image systems, and face manipulation models, alongside real images from various sources.

Models. The GenSign Backbone $\mathcal{B}_{sig}$ is based on EfficientNet-B0. The RGB Backbone $\mathcal{B}_{rgb}$ used CLIP ViT-L/14 [34], and we fine-tuned only the attention layers using Low-Rank Adaptation with a rank of $r = 4$, keeping all other parameters frozen. The classifier $\mathcal{H}$ is an MLP with 256 neurons, using LeakyReLU with a negative slope of 0.2 as the activation function. We set $K=4$ for CCDC.

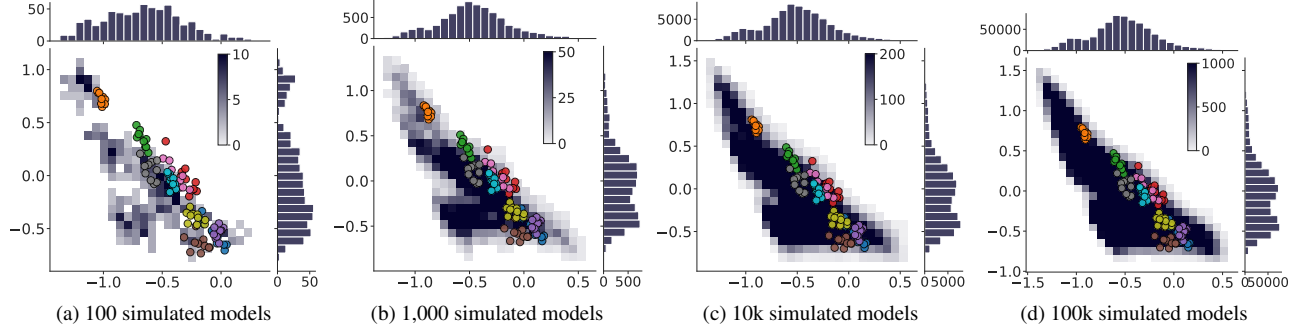


Figure 3. **PCA visualization of the spectral features from real and simulated generative models.** Colored points denote real generative models, while the dark blue density map represents simulated models generated by $\mathcal{R}$, with the number increasing from 100 to 100K. ● AMD, ● BigGAN, ● Glide, ● MidJourney, ● SDv1.5, ● Imagen3, ● FLUX-dev, ● StyleGAN3, ● SD3, ● Dalle2.

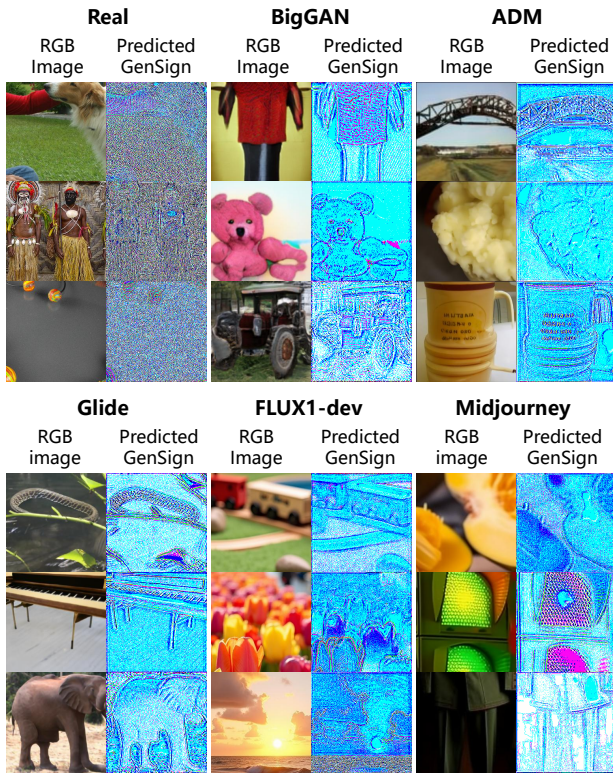


Figure 4. **Extracted GenSign of real images and AIGIs.** Each pair shows the image I and GenSign $\mathcal{E}_{sig}(I)$ (scaled to $[0, 255]$).

Implementation Details. We train all models with a batch size of 8, Adam optimizer, and a learning rate of 3×10^{-4} . Images are resized to 256×256 , and for the CLIP, we apply a center crop of 224×224 to match its input requirements. We set λ_1 and λ_2 as 1.0 and 0.25. We report average precision (AP) and classification accuracy (Acc).

Complexity. Although training $\mathcal{R}$ and $\mathcal{E}_{sig}$ on a single NVIDIA A100 costs $\sim 49h$ and $\sim 41h$, respectively, they serve as universal, foundational components that can be

frozen and reused. The detector $\mathcal{M}$ requires only $\sim 7h$ to train, with inference at $\sim 49ms$ per image.

4.2. From Simulation to Reality

Generative Model Distribution Coverage Analysis. We study whether DAIR can generalize to real-world generative models. To this end, we collect N groups of real images, each containing 10 images. Each group is independently reconstructed by $\mathcal{R}$, producing N sets of reconstructed images that simulate N distinct generative models. For each real generative model, we sample 100 images and divide them into 10 groups of 10 images each.

Following Corvi *et al.* [7], we compute the magnitude spectrum of the residual of each image, perform radial averaging, and then take the mean spectrum within each 10-image group. All spectral features are projected into a two-dimensional space for visualization. As illustrated in Fig. 3, we make two key observations: (1) Real-world generative models exhibit substantial overlap with simulated ones. (2) Expanding the simulated model set from 100 to 100K enlarges the simulated distribution. This demonstrates that DAIR effectively covers real generative models, providing the foundation for simulation-based generalization.

GenSign Visualization. Fig. 4 visualizes the extracted GenSign across real and synthetic images from diverse generative models. We observe a clear distinction between the two categories: GenSign of real images resembles random noise with no discernible spatial structure, whereas GenSign of AIGIs exhibits large, coherent patches and distinct spatial patterns. These visual differences confirm that extractor $\mathcal{E}_{sig}$ successfully isolates model-specific generative artifacts that are absent in real images.

4.3. Generalization Evaluation

4.3.1. Performance on UniversalFakeDetect Benchmark

To evaluate the generalization capability of our proposed method, we conduct experiments on the UniversalFakeDe-

Table 1. **Generalization Evaluation on UniversalFakeDetect** (Average Precision, %). The detector is trained on ProGAN 20 classes.

Methods	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	Mean
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
CNN-Spot [44]	100.0	93.47	84.50	99.54	89.49	98.15	89.02	73.75	59.47	98.24	98.40	73.72	70.62	71.00	70.54	80.65	84.91	82.07	70.59	83.58
Patchfor [5]	80.88	72.84	71.66	85.75	65.99	69.25	76.55	76.19	76.34	74.52	68.52	75.03	87.10	86.72	86.40	85.37	83.73	78.38	75.67	77.73
F3Net [33]	<u>99.96</u>	84.32	69.90	99.72	56.71	100.0	78.82	52.89	46.70	63.39	64.37	70.53	73.76	81.66	74.62	89.81	91.04	90.86	71.84	76.89
UniFD [31]	100.0	98.13	94.46	86.66	99.25	99.53	91.67	78.54	67.54	83.12	91.06	79.24	95.81	79.77	95.93	93.93	95.12	94.59	88.45	90.14
LGrad [39]	100.0	93.98	90.69	99.86	79.36	<u>99.98</u>	67.91	59.42	51.42	63.52	69.61	87.06	99.03	99.16	99.18	93.23	95.10	94.93	97.23	86.35
FreqNet [40]	99.92	99.63	96.05	<u>99.89</u>	99.71	<u>98.63</u>	<u>99.92</u>	94.42	74.59	80.10	75.70	96.27	96.06	100.0	62.34	<u>99.80</u>	99.78	96.39	77.78	91.95
NPR [41]	100.0	99.53	94.53	99.94	88.82	100.0	84.41	<u>97.95</u>	99.99	50.16	50.16	98.26	99.92	<u>99.91</u>	99.92	99.87	99.89	99.92	99.26	92.76
FatFormer [25]	100.0	100.0	<u>99.98</u>	99.75	100.0	100.0	97.99	97.94	81.21	<u>99.84</u>	<u>99.93</u>	91.99	<u>99.81</u>	99.09	99.87	99.13	99.41	99.20	99.82	98.16
Ours	99.74	<u>99.96</u>	100.0	98.11	100.0	99.94	99.98	98.38	<u>95.10</u>	100.0	99.98	99.91	99.63	99.68	99.59	98.26	99.98	99.97	<u>99.78</u>	99.37

Table 2. **Cross-Method Evaluation on AIGCDetectBenchmark** (Average Precision, %). The detector is trained on ProGAN 20 classes.

Methods	ProGAN	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN	StyleGAN2	WFIR	ADM	Glide	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	DALLE2	SDXL	Mean
CNNSpot [44]	100.0	99.83	85.99	94.94	99.04	90.82	99.48	99.85	75.67	72.28	66.24	61.20	61.56	68.83	57.34	53.51	72.62	79.95
FreDect [14]	<u>99.99</u>	88.98	93.62	84.78	99.49	82.84	82.54	55.85	61.77	52.92	46.09	37.83	37.76	85.10	39.58	38.20	49.45	66.87
Fusing [17]	100.0	99.50	90.70	95.50	<u>99.80</u>	88.30	99.60	93.30	94.10	77.50	70.00	65.40	65.70	75.60	64.60	68.12	79.41	83.95
LNP [24]	100.0	99.27	94.54	89.52	100.0	84.54	<u>99.70</u>	42.75	93.37	92.76	<u>86.92</u>	96.34	96.00	94.91	95.33	<u>98.26</u>	87.75	91.29
LGrad [39]	100.0	98.31	92.93	95.01	100.0	95.43	97.89	57.99	72.95	80.42	71.86	62.37	62.85	77.47	62.48	82.55	80.03	81.80
DIRE [45]	99.08	91.74	75.25	80.56	99.34	72.15	88.29	60.13	85.84	78.35	61.86	49.87	49.52	54.57	55.38	74.48	53.97	72.38
UniFD [31]	100.0	97.56	99.27	<u>99.80</u>	99.37	99.98	97.90	96.73	86.81	83.81	74.00	86.14	85.84	<u>96.53</u>	91.07	63.04	67.59	89.73
PatchCraft [54]	100.0	98.96	99.42	85.26	100.0	81.33	97.74	95.26	93.40	<u>94.04</u>	96.48	99.06	99.06	96.26	97.54	99.56	99.89	96.07
Ours	100.0	99.98	99.74	99.99	100.0	<u>99.93</u>	99.99	<u>99.81</u>	<u>93.53</u>	97.80	81.17	99.66	99.21	99.41	99.55	97.45	<u>99.42</u>	98.04

test (UniFD) benchmark, trained on ProGAN 20 classes. The results are summarized in Table 1. Our method achieves a mAP of 99.37%, yielding nearly perfect AP across most GAN-based generators and maintaining strong detection ability on diffusion models with over 98% AP consistently. This surpasses the best compared method, FatFormer [25], by 1.21% in mean AP, confirming that our method captures generalizable forensic cues shared among different AIGIs rather than overfitting to specific artifacts.

4.3.2. Performance on AIGCDetectBenchmark

Our evaluation on the AIGCDetectBenchmark (Table 2), which consists of 17 distinct generative models, further confirms the generalization of our method. We obtain the highest overall mean AP of 98.04%, clearly surpassing the strong PatchCraft baseline (96.07%). Beyond the state-of-the-art performance, the key finding from this benchmark is the **exceptional consistency**. It achieves near-perfect detection (over 99% AP) on 12 of the 17 generators and maintains remarkable performance on all others. Its lowest score, 81.17% on Midjourney, is still highly effective and significantly better than most competitors. For instance, LNP, despite a high mean, suffers a critical failure on WFIR (42.75%), and UniFD drops significantly on DALLE2 (63.04%) and SDXL (67.59%). This uniformly high performance across varied architectures validates that our framework learns universal representations, rather than

overfitting to the artifacts of a specific model family.

4.3.3. Performance on AIGIBenchmark

We further conduct generalization evaluations on AIGIBenchmark (Table 3), which features a diverse set of 17 recent generative models that were selected to avoid overlap with other benchmarks. Our method achieves a mean AP of 89.96%. This result represents a substantial leap over existing state-of-the-art methods, outperforming the next-best competitor, AIDE (75.36%), by a significant margin of 14.6%. Notably, this benchmark highlights a critical weakness in many existing detectors: a failure to generalize to localized forgeries. Most baselines, including AIDE, DFFreq, and NPR, exhibit a severe performance collapse on face-swapping models (*e.g.*, *BlendFace*, *InSwap*, *FaceSwap*, *SimSwap*), with AP dropping close to random-chance levels. In contrast, our method demonstrates superior generalization to local forgeries, achieving 66.37% (*BlendFace*), 87.10% (*InSwap*), 88.57% (*FaceSwap*), and 75.09% (*E4S*). These results indicate that our method not only achieves state-of-the-art overall performance but also maintains strong generalization on different forgery types.

4.3.4. Performance on GenImage Benchmark

We further evaluate generalization on the GenImage benchmark [55], following [6, 55]: the model is trained on SDv1.4 and evaluated on all models. As shown in Ta-

Table 3. **Generalization Evaluation on AIGIBenchmark** (Average Precision, %). The detector is trained on ProGAN 20 classes.

Methods	R3GAN	BlendFace	InSwap	FaceSwap	SD3	StyleSwim	InstantID	Infinite_ID	Imagen3	StyleGAN-XL	FLUX1-dev	IP_Adapter	BLIP	SimSwap	E4S	StyleGAN3	PhotoMaker	Mean
CNNSpot [44]	48.40	34.20	44.30	45.30	66.30	88.90	67.10	55.80	50.30	60.00	50.30	50.10	44.70	45.30	35.20	91.60	51.50	54.66
UniFD [31]	98.60	42.70	48.20	<u>51.50</u>	48.60	91.80	79.50	84.20	44.10	<u>95.00</u>	39.60	73.90	68.90	<u>51.50</u>	82.50	79.00	41.40	66.38
LGrad [39]	69.10	39.40	38.50	38.10	71.40	75.00	54.80	34.40	66.90	62.80	70.00	46.30	57.80	38.10	39.40	64.30	35.80	53.06
FreqNet [40]	56.80	34.10	42.10	41.90	84.00	99.30	86.30	74.50	<u>90.80</u>	84.10	87.30	79.90	99.90	41.90	34.70	<u>94.60</u>	74.90	69.84
NPR [41]	75.60	33.30	36.90	38.50	93.50	91.70	75.90	62.80	91.00	75.90	<u>88.00</u>	81.10	97.00	38.50	34.40	89.40	52.00	68.30
DFFreq [47]	77.10	36.60	<u>48.80</u>	45.60	89.30	90.30	85.10	72.60	87.70	81.40	85.40	75.80	95.50	45.60	37.90	92.40	<u>80.90</u>	72.11
LaDeDa [3]	69.20	38.00	41.80	39.60	87.60	86.40	81.80	67.50	84.60	84.20	87.64	82.20	96.70	39.60	42.40	86.30	66.60	69.56
AIDE [48]	<u>95.51</u>	<u>43.80</u>	43.00	46.80	94.10	85.10	89.00	<u>87.90</u>	87.70	87.70	86.60	<u>91.00</u>	96.70	46.80	38.30	77.80	76.60	<u>75.36</u>
Ours	94.80	66.37	87.10	88.57	<u>93.59</u>	<u>97.51</u>	<u>88.72</u>	97.50	79.82	99.05	93.60	93.97	<u>99.78</u>	88.40	<u>75.09</u>	98.15	92.61	89.96

Table 4. **Generalization Evaluation on GenImage**. Following [55], we use SDv1.4 as the training set and other datasets as testing sets. Accuracy (%) is reported following [6].

Methods	Venue	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Mean
CNNSpot [44]	CVPR2020	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
F3Net [33]	ECCV2020	50.1	<u>99.9</u>	99.9	49.9	50.0	99.9	49.9	49.9	68.7
UniFD [31]	CVPR2023	91.5	96.4	96.1	58.1	73.4	94.5	67.8	57.7	79.5
NPR [41]	CVPR2024	81.0	98.2	97.9	76.9	89.8	96.9	84.1	<u>84.2</u>	88.6
FreqNet [40]	AAAI2024	89.6	98.8	98.6	66.8	86.5	97.3	75.8	81.4	86.8
FatFormer [25]	CVPR2024	<u>92.7</u>	100.0	99.9	75.9	88.0	99.9	<u>98.8</u>	55.8	88.9
DRCT [6]	ICML2024	91.5	95.0	94.4	<u>79.4</u>	89.2	94.7	90.0	81.7	89.5
Effort [49]	ICML2025	82.4	99.8	<u>99.8</u>	78.7	93.3	97.4	91.7	77.6	<u>91.1</u>
Ours	-	99.3	99.8	99.7	84.3	<u>90.4</u>	<u>99.8</u>	99.6	99.8	96.6

ble 4, our method achieves the best overall performance with a mean Acc of 96.6%, surpassing the latest Effort [49] (91.1%) and FatFormer [25] (88.9%) by clear margins.

4.4. Ablation Study

Impact of Input Modality and Architecture. We conduct an ablation study to examine the effect of input modality (RGB vs. GenSign) across different architectures. Table 5 reports the performance for each setting. We observe: (1) CNNs trained from scratch using GenSign consistently improve performance over RGB images, demonstrating that GenSign provides more discriminative cues than standard RGB inputs; (2) While CLIP achieves high performance with RGB inputs on UniFD, replacing RGB with GenSign slightly reduces performance on some tasks. This is expected, as CLIP is pretrained to process RGB images rather than GenSign patterns; (3) Our dual-stream detector that combines RGB and GenSign further improves performance across benchmarks, showing that complementary features from different modalities can enhance the generalization.

Comparison with Alternative Fingerprints. To validate the effectiveness of GenSign, we compare it with alternative fingerprint extraction methods under the same training settings. As Table 6 shows, SRM Conv [48],

Table 5. Effect of input modality on detection performance (mAP, %) across different models. All detectors are trained with our self-augmentation strategy ($p_{aug} = 20\%$).

Detector	RGB	GenSign	UniFD	AIGC	AIGI	Mean
ResNet50	✓	✓	87.40	60.15	59.25	68.93
			90.29	78.25	68.03	78.86
EfficientNet-B0	✓	✓	90.21	83.30	69.13	80.88
			90.37	86.96	73.44	83.59
CLIP	✓	✓	98.28	95.98	70.91	88.39
			95.08	95.68	73.04	87.93
Ours (Dual)	✓	✓	99.37	98.04	89.96	95.79

Noiseprint++ [16], and pre-trained denoisers [7] achieve mAPs of 95.27%, 86.18%, and 94.17%, respectively. In contrast, GenSign attains the highest mAP of 95.79%, demonstrating that supervised learning on simulated generative traces yields generalizable forensic features than unsupervised or task-agnostic noise extraction methods.

Robustness Against Image Perturbations. We evaluate robustness against common post-processing operations. Fig. 5 reveals that photometric transformations (brightness, saturation, gamma) cause minimal performance drops,

Table 6. Effect of different feature extraction methods on detection performance (mAP, %). Trained on ProGAN 20 classes.

Input	UniFD	AIGC	AIGI	Mean
Ours with SRM Conv [48]	99.01	97.24	89.57	95.27
Ours with Noiseprint++ [16]	97.03	93.47	68.03	86.18
Ours with Denoiser [7]	98.61	98.29	85.62	94.17
Ours with GenSign	99.37	98.04	89.96	95.79

while spatial distortions and compression induce moderate degradation. Our learned GenSign exhibits slightly higher robustness compared to existing methods.

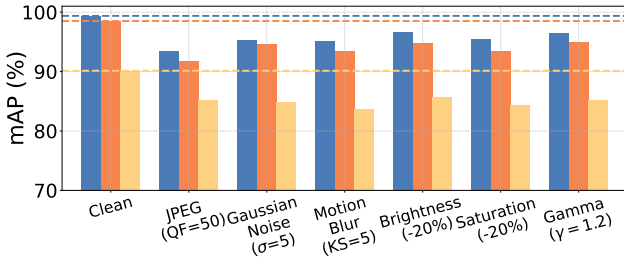


Figure 5. Impact of image perturbations on UniFD. Trained on ProGAN 20 classes. ■ Ours, ■ SDD [46], ■ UniFD [31].

Effect of Self-Augmentation Rate. Fig. 6 demonstrates the critical role of self-augmentation during detector training. Without augmentation (0%), the average AP drops to 88.80%, indicating that models overfit to training distribution artifacts rather than learning generalizable traces. Performance peaks at 20% (95.79%), achieving an optimal balance between preserving the original data distribution and exposing diverse generative patterns.

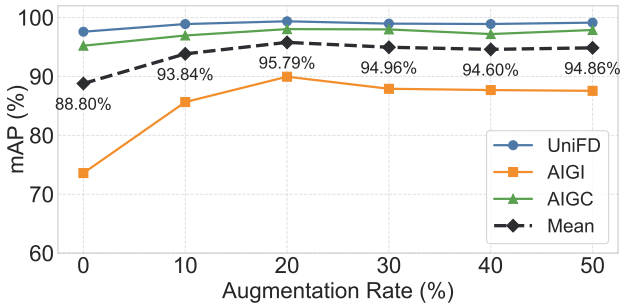


Figure 6. Effect of self-augmentation rate (p_{aug}) on detection generalization (mAP, %). Trained on ProGAN 20 classes.

Ablation Study on Simulation Diversity. To quantitatively verify the benefit of architectural diversity in the reconstructor, we evaluate detection performance by varying the number of downsampling levels in $\mathcal{R}$ (denoted as Levels 1 to 3). A higher number of levels introduces greater

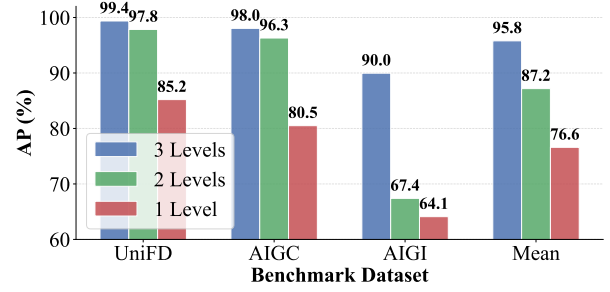


Figure 7. Impact of simulation diversity on detection generalization. We compare the performance using different numbers of downsampling levels in the reconstructor $\mathcal{R}$. 3 Levels denotes the default setting with the highest diversity, while 1 Level represents the most constrained simulation space.

variability in the simulated generative models, thereby expanding the diversity of the synthesized GenSign. As shown in Fig. 7, there is a clear positive correlation between simulation diversity and detection generalization. The model trained with the full 3 levels achieves a mean AP of 95.8%, significantly outperforming the 2-level (87.2%) and 1-level (76.6%) settings. These results confirm that a richer spectrum of simulated GenSign is essential for covering the distribution of real-world generative models.

5. Conclusion

In this work, we address the fundamental challenge of supervised generative signature learning by introducing a surrogate training paradigm that simulates diverse generative artifacts on real images. Our proposed reconstructor enables supervised extraction of model-dependent signatures. The resulting dual-stream detector achieves state-of-the-art generalization across different generative models. Our key finding is that sufficiently diverse architectural simulation enables effective transfer to real-world generators, with learned signatures outperforming other alternatives. Overall, our approach provides a generalized and practical path toward reliable AI-generated image detection.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant numbers 625B2187, U23B2023, and 62472199, in part by the Guangdong Key Technologies R&D Program Project under grant number 2026B0101100003, and in part by the Basic and Applied Basic Research Foundation of Guangdong Province (2025A1515011097). This work is also supported by the Engineering Research Center of Trustworthy AI, Ministry of Education. This work is also supported by the China Scholarship Council 202406380227.

References

- [1] Giulia Bertazzini, Daniele Baracchi, Dasara Shullani, Isao Echizen, and Alessandro Piva. Dragon: A large-scale dataset of realistic images generated by diffusion models. *arXiv preprint arXiv:2505.11257*, 2025. 2
- [2] Luca Bindini, Giulia Bertazzini, Daniele Baracchi, Dasara Shullani, Paolo Frasconi, and Alessandro Piva. Tiny autoencoders are effective few-shot generative model detectors. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2024. 1
- [3] Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024. 7
- [4] 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 (CVPR)*, pages 10759–10769, 2024. 2
- [5] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–120, 2020. 6
- [6] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drc: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 6, 7
- [7] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 973–982, 2023. 1, 5, 7, 8
- [8] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2
- [9] Yunshu Dai, Jianwei Fei, and Fangjun Huang. Idguard: Robust, general, identity-centric poi proactive defense against face editing abuse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11934–11943, 2024. 1
- [10] Yunshu Dai, Jianwei Fei, Fangjun Huang, and Chip Hong Chang. Robust secure swap: Responsible face swap with persons of interest redaction and provenance traceability. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. 1
- [11] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 7890–7899, 2020. 2
- [12] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20270–20280, 2022. 1
- [13] Jianwei Fei, Yunshu Dai, Peipeng Yu, Zhihua Xia, Dasara Shullani, Daniele Baracchi, and Alessandro Piva. One for all: Synthesis-free fingerprint learning for attribution of in-the-wild synthetic images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 35419–35427, 2026. 2
- [14] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3247–3258, 2020. 1, 2, 6
- [15] Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14094–14103, 2021. 2
- [16] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20606–20615, 2023. 7, 8
- [17] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3465–3469, 2022. 6
- [18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 4
- [19] Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C Kot. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *IEEE Transactions on Dependable and Secure Computing*, 2025. 2
- [20] Chenqi Kong, Anwei Luo, Peijun Bao, Haoliang Li, Renjie Wan, Zengwei Zheng, Anderson Rocha, and Alex C Kot. Open-set deepfake detection: a parameter-efficient adaptation method with forgery style mixture. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026. 2
- [21] Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 28050–28060, 2025. 2
- [22] Zhuoling Li, Haoxuan Qu, Jason Kuen, Jiuxiang Gu, Qihong Ke, Jun Liu, and Hossein Rahmani. Diffip: Representation fingerprints for robust ip protection of diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17035–17045, 2025. 1

- [23] Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 4
- [24] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 95–110, 2022. 6
- [25] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10770–10780, 2024. 2, 6, 7
- [26] Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2023. 1
- [27] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lare²: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17006–17015, 2024. 2
- [28] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511, 2019. 1, 2
- [29] Andrea Montibeller, Dasara Shullani, Daniele Baracchi, Alessandro Piva, and Giulia Boato. Bridging the gap: A framework for real-world video deepfake detection via social network compression emulation. In *Proceedings of the 1st on Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media*, pages 29–36, 2025. 1
- [30] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 31:1–7, 2019. 2
- [31] Utkarsh Ojha et al. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 24480–24489, 2023. 2, 4, 6, 7, 8
- [32] Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8245–8257, 2025. 2
- [33] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 86–103, 2020. 2, 6, 7
- [34] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 4
- [35] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 9130–9140, 2024. 2
- [36] 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 (CVPR)*, pages 10684–10695, 2022. 4
- [37] Hae Jin Song and Laurent Itti. Riemannian-geometric fingerprints of generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2
- [38] Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. Manifpt: Defining and analyzing fingerprints of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10791–10801, 2024. 1, 2
- [39] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 12105–12114, 2023. 1, 2, 6, 7
- [40] 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 (AAAI)*, pages 5052–5060, 2024. 6, 7
- [41] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28130–28139, 2024. 6, 7
- [42] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7184–7192, 2025. 1, 2
- [43] Huaming Wang, Jianwei Fei, Yunshu Dai, Lingyun Leng, and Zhihua Xia. General gan-generated image detection by data augmentation in fingerprint domain. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1187–1192, 2023. 1
- [44] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8695–8704, 2020. 2, 4, 6, 7
- [45] 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*

- International Conference on Computer Vision (ICCV)*, pages 22445–22455, 2023. [1](#), [2](#), [6](#)
- [46] Ziye Wang, Minghang Yu, Chunyan Xu, and Zhen Cui. Semantic discrepancy-aware detector for image forgery identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18388–18398, 2025. [8](#)
 - [47] Jiazhen Yan, Ziqiang Li, Ziwen He, and Zhangjie Fu. Generalizable deepfake detection via effective local-global feature extraction. *arXiv preprint arXiv:2501.15253*, 2025. [7](#)
 - [48] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. [1](#), [2](#), [7](#), [8](#)
 - [49] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 70268–70288, 2025. [7](#)
 - [50] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7556–7566, 2019. [1](#), [2](#)
 - [51] Peipeng Yu, Jianwei Fei, Zhihua Xia, Zhili Zhou, and Jian Weng. Improving generalization by commonality learning in face forgery detection. *IEEE Transactions on Information Forensics and Security*, 17:547–558, 2022. [1](#)
 - [52] Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 72925–72943, 2025. [2](#)
 - [53] Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking semantic artifacts for generalized ai-generated image detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:59570–59596, 2024. [2](#)
 - [54] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. [4](#), [6](#)
 - [55] 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*, 36:77771–77782, 2023. [2](#), [4](#), [6](#), [7](#)