FROM PHYSICAL DEFECTS TO DISCRIMINATIVE RESIDUALS: GENERALIZABLE SYNTHETIC IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

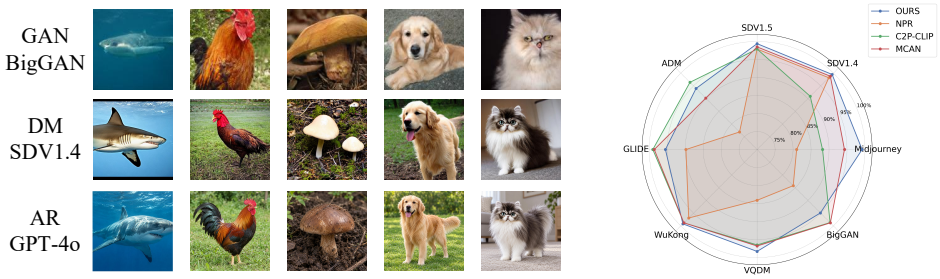
Existing synthetic image detectors often exhibit limited generalization across heterogeneous generative architectures due to their reliance on generator-specific artifacts. To address this limitation, we explore physical discrepancies shared across different generative paradigms to identify forensic cues that remain stable across architectures. Through physically motivated probe experiments, we further identify three shared physical defects across GANs and diffusion models, reflected in systematic discrepancies from natural images under resampling operations, spatial smoothing, and high frequency characteristics. Motivated by these observations, we propose a learnable physical prior decoupling framework that transforms such physical defects into discriminative residual representations. Specifically, we introduce three physical prior simulators, Resample, Blur, and Edge, together with learnable residual modulation to adaptively extract complementary residual cues. These residuals are further enhanced in the high-frequency domain and dynamically fused for classification without additional reconstruction supervision. Extensive experiments using only ProGAN generated images for training show that our method ranks first or second among state-of-the-art methods across multiple benchmarks covering different generative architectures and paradigms, demonstrating strong cross architecture generalization.

1 INTRODUCTION

Recent advances in generative models, including Generative Adversarial Networks (GANs), diffusion models, and Diffusion Transformers (DiTs), have enabled increasingly photorealistic image synthesis Ho et al. (2020); Karras et al. (2017); Peebles & Xie (2023). While enabling diverse creative applications, these advances have also raised societal concerns, including misinformation, deepfakes, and privacy leakage Golda et al. (2024). Synthetic Image Detection (SID) has therefore emerged to distinguish AI-generated content from natural images. However, the diversification of generative paradigms poses a fundamental challenge: forensic cues learned from one generator may not remain reliable on unseen architectures. As shown in Figure 1(a), images generated by different paradigms under identical prompts exhibit distinct characteristics in texture, detail, and edges. Correspondingly, detector performance varies considerably across generators, as shown in Figure 1(b). Learning forensic representations that remain discriminative across unseen generative architectures is therefore central to generalizable SID.

Existing SID methods broadly follow two paradigms. One line of research designs detection-aware modules to automatically learn discriminative artifact representations, exploring frequency characteristics, local-global features, and vision-language representations Qian et al. (2020); Ju et al. (2022); Keita et al. (2025). Another line of research leverages off-the-shelf models or image-processing operations to expose predefined forensic cues before classification Ojha et al. (2023); Tan et al. (2024b). Although these approaches have achieved promising detection performance, they largely rely on identifying discriminative artifacts from the training distribution Li et al. (2025); Yan et al. (2025). Such artifacts can be closely coupled with the synthesis characteristics of particular generators and may vary substantially as generative architectures evolve, resulting in degraded performance on unseen generators. As illustrated in Figure 1, images synthesized by different generative paradigms exhibit noticeable variations in texture, detail, and edge characteristics, accompanied

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Visual comparison of generative models

(b) Radar chart of cross-architecture generalization performance

Figure 1: Cross-architecture variations in synthetic images and detection performance. (a) Images generated by different generative paradigms under identical prompts. (b) Detection accuracy of four detectors across eight generators in GenImage.

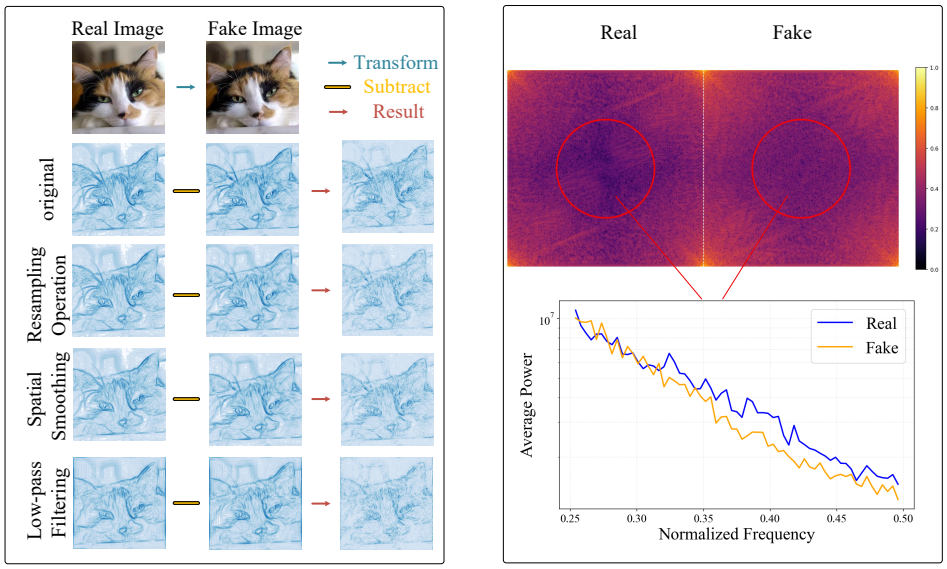
by considerable fluctuations in the cross-architecture performance of existing detectors. These observations motivate the pursuit of forensic cues rooted in generation characteristics shared across heterogeneous generative architectures, rather than artifacts specific to individual generators.

Motivated by this perspective, we explore the underlying image generation process for forensic cues shared across heterogeneous generative architectures. Despite architectural differences, GANs, diffusion models, and DiTs commonly involve operations such as upsampling and convolutional processing. We conduct physically motivated probe experiments to investigate discrepancies between synthetic and natural images. As shown in Figure 2(a), synthetic images exhibit more structured residual responses than natural images under resampling, spatial smoothing, and low-pass filtering. Figure 2(b) compares their log-spectra and RAPS curves, revealing increasingly pronounced discrepancies at higher frequencies, where synthetic images generally exhibit lower average power. While the two curves remain close at low frequencies, they progressively diverge toward higher frequencies. Taken together, these observations reveal systematic differences in physical responses and high-frequency characteristics, providing potentially more architecture-stable forensic cues than generator-specific artifacts.

Building on these observations, we propose a learnable physical prior decoupling framework that explicitly transforms physical defects into discriminative residuals for generalizable synthetic image detection. Specifically, we design three physical prior simulators, namely Resample, Blur, and Edge, each combining a fixed physical operator with learnable residual modulation to preserve prior guidance while adapting to image content. We compute the residual between the input and each simulator output to expose generation-related discrepancies, and further enhance high-frequency information through the high-high (HH) subband of the Discrete Wavelet Transform (DWT) before dynamic fusion and classification. The framework is optimized directly by the classification objective without pixel-level reconstruction constraints, allowing the residual representations to evolve toward task-discriminative patterns. Extensive experiments on 27 distinct generative models across four benchmarks demonstrate strong cross-architecture generalization using only ProGAN-generated images for training. Our main contributions are summarized as follows:

- We investigate the shared generation characteristics of GANs and diffusion models, and identify three physical defects related to resampling, smoothing, and high-frequency characteristics. Physically motivated probe experiments reveal systematic discrepancies between natural and synthetic images, providing the basis for exploiting physical defects as forensic cues.
- We propose a learnable physical prior decoupling framework that transforms physical defects into discriminative residuals. Three physical prior simulators combine fixed physical operators with learnable residual modulation to expose generation-related discrepancies, while high-frequency enhancement and dynamic fusion further integrate complementary residual information for detection.
- We conduct extensive experiments on four benchmarks covering 27 distinct generative models (33 evaluation settings in total) spanning GANs and diffusion models. Using only

ProGAN-generated images for training, our method ranks first or second among state-of-the-art methods across diverse unseen generators, demonstrating strong cross-architecture generalization.



(a)Artifact residuals under different physical priors

(b) Spectral comparison: log-spectrum and RAPS

Figure 2: Empirical evidence of physical defects in synthetic images. (a) Residual responses of real and synthetic images under resampling, spatial smoothing, and low-pass filtering, revealing distinct response patterns under different physical probes. (b) Log-spectrum and RAPS comparisons, showing increasingly pronounced spectral discrepancies toward high frequencies.

2 Related Work

Feature-aware Methods. Feature-aware methods extract forensic cues through predefined transformations before classification. Frequency-domain approaches exploit spectral discrepancies between real and synthetic images: FreqNet Tan et al. (2024a) employs the Fast Fourier Transform to extract high-frequency components from phase and amplitude spectra, while SAFE Li et al. (2025) adopts the Discrete Wavelet Transform for high-frequency feature extraction with data augmentation. Local and residual-based methods focus on pixel dependencies and reconstruction discrepancies. NPR Tan et al. (2024b) constructs artifact representations from neighboring-pixel differences, FerretNet Liang et al. (2026) exposes texture discontinuities through median-filter-based reconstruction, PiD Fu et al. (2025) derives decomposition residuals via color-space transformation and quantization, LOTA Wang et al. (2025) exploits least significant bit-planes for noise patterns, and AEDR Wang et al. (2026) leverages double-reconstruction loss ratios for image attribution. Recent methods integrate multiple forensic cues: MCAN Tan et al. (2026) combines chromatic inconsistency with high-frequency components, CoD Jia et al. (2025) integrates color distribution features with image representations, and Beyond Semantic Features Zhou et al. (2026) maps low-frequency semantic information into high-frequency components to mitigate semantic bias.

Image-aware Methods. Image-aware methods take RGB images as input and rely on deep networks to learn discriminative forensic representations. Early work CNNSpot Wang et al. (2020) employs a CNN classifier with data augmentation for cross-GAN detection, while PPL Yang et al. (2026) introduces the principle that "all patches matter," identifies the "few-patch bias," and mitigates it through patch reconstruction and contrastive learning. More recently, pre-trained vision foundation models have been exploited to improve detection generalization. UniFD Ojha et al. (2023) uses a frozen CLIP visual encoder with a linear classifier, demonstrating the transferability of pre-trained representations. FatFormer Liu et al. (2024) augments CLIP with frequency-aware adapters, IAPL Li et al. (2026) introduces image-adaptive prompt learning with test-time token tuning to adapt to

unseen generators, and C2P-CLIP Tan et al. (2025) injects category-common prompts and employs contrastive learning to embed the concepts of "real" and "fake" into the CLIP encoder.

3 METHODOLOGY

As illustrated in Figure 3, our framework follows an end-to-end pipeline of "physical prior decoupling – residual extraction – frequency-aware fusion – discriminative classification". Given an input image $X \in \mathbb{R}^{3\times H\times W}$, we first feed it in parallel into three physical prior simulators: the Resample Simulator, the Blur Simulator, and the Edge Simulator. Each simulator applies a fixed physical transformation to the input, followed by a learnable residual modulation module for adaptive refinement, producing a prediction map $P \in \mathbb{R}^{3\times H\times W}$. The residual $R = X - P$ is then computed to transform the response induced by each physical prior into a discriminative residual representation.

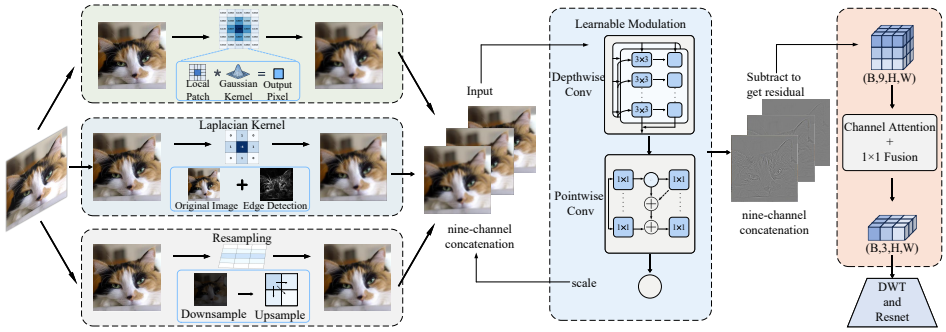


Figure 3: Overall framework of the proposed method. The input image is fed in parallel into three physical prior simulators (Resample, Blur, and Edge) to generate prediction maps. The residuals $R = X - P$ are computed and concatenated into a 9-channel representation, from which the DWT high-frequency (HH) subband is extracted and reweighted by channel attention before 1×1 fusion and ResNet-based binary classification.

3.1 PHYSICAL PRIOR SIMULATORS

Motivated by the physical defects identified in Section 1, we design three physical prior simulators, namely Resample, Blur, and Edge, to characterize the corresponding generation-related discrepancies. Each simulator consists of a fixed physical operator cascaded with a learnable residual modulation module. The fixed operator explicitly encodes the corresponding physical prior, while the learnable module adaptively refines the transformed representation according to image content. The three simulators are designed to capture complementary residual responses associated with resampling, smoothing, and high-frequency edge characteristics, respectively.

Resample Simulator. Despite architectural differences, many generative models involve resolution recovery or upsampling during synthesis or decoding. Such operations can alter local pixel dependencies and frequency characteristics, potentially introducing interpolation-related correlations, aliasing, or grid-like patterns. Accordingly, we design the Resample Simulator to expose discrepancies between real and synthetic images under resampling. Specifically, it applies bilinear downsampling with a scale factor of $\rho = 0.5$, followed by bilinear upsampling to restore the original resolution. Downsampling suppresses fine-scale and high-frequency information, while upsampling restores the resolution, exposing differences in local structure and frequency response. For adaptive refinement before resolution recovery, the learnable residual modulation module is inserted between the two operations:

$$P_{\text{resample}} = \text{Up}\big(\text{Adjust}(\text{Down}(X))\big), \tag{1}$$

where *Down* and *Up* denote fixed bilinear interpolation operations, and $\text{Adjust}(\cdot)$ denotes the learnable residual modulation (Section 3.2). The resulting output P_{resample} is then compared with the original input to obtain the residual $R_{\text{resample}} = X - P_{\text{resample}}$. By suppressing fine-scale informa-

tion through resampling, the resulting residual emphasizes differences in local structure and high-frequency responses between real and synthetic images.

Blur Simulator. Local convolutional processing is widely involved in generative models and can affect the spatial dependencies and local smoothness of synthesized images. Such effects may lead to different responses of real and synthetic images to spatial smoothing, particularly in texture-rich and boundary regions. To probe this discrepancy, the Blur Simulator applies a fixed Gaussian kernel (kernel size 7×7, standard deviation $\sigma = 2.0$) to smooth the input, followed by the learnable modulation module:

$$P_{\text{blur}} = \text{Adjust}\big(G_{7\times7,\sigma=2.0} * X\big), \tag{2}$$

where $*$ denotes convolution and G is the Gaussian kernel. This operation provides a controlled spatial-smoothing probe: it suppresses fine-scale details and local discontinuities, and the resulting residual $R_{\text{blur}} = X - P_{\text{blur}}$ captures how the input responds to this smoothing perturbation. The learnable modulation module then adaptively refines the smoothed representation, allowing the residual to emphasize generation-related differences in local spatial dependencies.

Edge Simulator. Real and synthetic images often exhibit systematic discrepancies in high-frequency edge responses, especially in fine-structure and diagonal regions. To probe such discrepancies, the Edge Simulator applies a fixed Laplacian operator (3×3 kernel) to obtain an edge/high-frequency response and forms an edge-response-modulated representation by injecting the Laplacian response, scaled by a fixed edge enhancement strength α (set to 1.0 by default), into the input, followed by the learnable modulation module:

$$P_{\text{edge}} = \text{Adjust}\big(X + \alpha\, \nabla^2 X\big), \tag{3}$$

where ∇^2 denotes the Laplacian operator with kernel $\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$, and α controls the edge-response strength, set to 1.0 by default. With zero-initialized modulation, $P_{\text{edge}} = X + \alpha\nabla^2 X$, yielding $R_{\text{edge}} = X - P_{\text{edge}} = -\alpha\nabla^2 X$, which explicitly retains the Laplacian edge response. During training, the modulation module adaptively refines this representation under the classification objective, allowing the residual to emphasize generation-related discrepancies in high-frequency edge regions. Rather than assuming that generated images are inherently high-frequency deficient, this simulator provides a physical probe for comparing real and synthetic responses to high-frequency edge perturbations. All three simulators adopt channel-wise processing (i.e., depthwise convolution), where each color channel is transformed independently to prevent cross-channel interference. The prediction P_* is then subtracted from the input X, yielding $R_* = X - P_*$ for subsequent frequency fusion.

3.2 Learnable Residual Modulation and Frequency Fusion

Learnable Residual Modulation. While fixed physical operators provide explicit and deterministic priors, their responses cannot adapt to diverse image contents and generation artifacts. To improve their adaptability, we equip each simulator with a learnable residual modulation module that adaptively refines the physically transformed representation. Let P_{phys} denote the output of a fixed physical operator. For the Resample Simulator, P_{phys} corresponds to the low-resolution intermediate $\text{Down}(X)$ before upsampling. The modulation process is formulated as follows:

$$P'_{\text{phys}} = P_{\text{phys}} + s \cdot \mathcal{F}(P_{\text{phys}}), \tag{4}$$

where $\mathcal{F}$ consists of a 3×3 depthwise convolution cascaded with a 1×1 pointwise convolution, responsible for local spatial correction and cross-channel information fusion, respectively; s is a learnable scalar initialized to 0. This zero-initialization strategy ensures that at the beginning of training, $P'_{\text{phys}} = P_{\text{phys}}$, i.e., the modulation module is strictly equivalent to an identity mapping and the simulator degenerates to the pure fixed physical transformation. As training proceeds, s adaptively evolves under the guidance of the classification loss, gradually enabling the modulation module to refine the fixed physical transformation while preserving its physical prior. The resulting prediction P thus balances determinism and content-adaptivity, and the residual $R = X - P$ captures the discrepancy between the input and the representation produced by the corresponding physical prior simulator.

Residual Extraction and Frequency Fusion. After modulation, the three simulators produce P_{resample}, P_{blur}, and P_{edge}, respectively. Their corresponding residual representations are obtained by subtracting each simulator output from the original input:

$$R_* = X - P_*, \qquad * \in \{\text{resample}, \text{blur}, \text{edge}\}. \tag{5}$$

Applying modulation before residual extraction allows the learnable module to refine each physically transformed representation while retaining the corresponding fixed transformation as an explicit physical anchor. The resulting residuals thus characterize complementary discrepancies in the responses of real and synthetic images to different physical priors. The three residual maps are concatenated along the channel dimension to form a 9-channel representation R_{concat}. Inspired by SAFE Li et al. (2025), which shows that DWT-based high-frequency features provide generalizable forensic cues for synthetic image detection, we apply DWT to R_{concat} and retain its HH subband:

$$R_{\text{HH}} = \text{DWT}_{\text{HH}}(R_{\text{concat}}). \tag{6}$$

Among the DWT subbands, the HH subband primarily captures diagonal high-frequency variations in the residual representation, which helps emphasize subtle generation-related discrepancies in edge and fine-structure regions. R_{HH} is then upsampled to the original spatial resolution and processed by a channel attention module to adaptively reweight the complementary residual responses. Finally, a 1×1 convolution projects the 9-channel representation into a 3-channel feature map, which is fed into the ResNet backbone for binary classification.

4 Experiments

4.1 Experimental Details

Implementation Details. We train all models on a single NVIDIA GPU using AdamW with an initial learning rate of 1e-2, weight decay of 1e-2, batch size of 32, and 30 epochs. We use cosine annealing with 1 epoch of linear warmup. Images are resized to 256×256, with random cropping for training and center cropping for testing. Data augmentation includes horizontal flipping, random rotation, and color jitter. We use label-smoothing cross-entropy with a smoothing factor of 0.1. Following the standard protocol Tan et al. (2025), models are trained only on ProGAN-generated images and their corresponding real images. The model contains 1.44M parameters.

Baselines. We compare our method with several state-of-the-art synthetic image detection methods, including F3Net Qian et al. (2020), SelfBland Shiohara & Yamasaki (2022), Ojha Ojha et al. (2023), LGrad Tan et al. (2023), FreqNet Tan et al. (2024a), NPR Tan et al. (2024b), CO-SPY Cheng et al. (2025), PiD Fu et al. (2025), C2P-CLIP Tan et al. (2025), SAFE Li et al. (2025), FALCON-Net Zhang et al. (2026), and SimLBR Dhakal et al. (2026).

Datasets. We evaluate cross-generator generalization on four benchmarks covering diverse generative architectures. **ForenSynths** Wang et al. (2020) includes eight GAN-based sources: ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, and Deepfake. The **Ojha** benchmark Ojha et al. (2023) includes eight diffusion-based sources from Guided Diffusion, LDM, GLIDE, and DALL-E variants. **Self-Synthesis** includes nine GAN architectures: AttGAN, BEGAN, CramerGAN, InfoMaxGAN, MMDGAN, RelGAN, S3GAN, SNGAN, and STGAN. **GenImage** Zhu et al. (2023) covers GAN and diffusion-based generators, including Midjourney, Stable Diffusion, ADM, GLIDE, WuKong, VQDM, and BigGAN.

Metrics. We use accuracy (ACC) and average precision (AP) as evaluation metrics, with mACC and mAP denoting their averages across generators, respectively.

4.2 Comparison with State-of-the-Art Methods

Cross-GAN-Sources Evaluation on ForenSynths. As shown in Table 1, our method achieves the best overall performance with 97.3% mACC and 99.4% mAP, outperforming recent competitive methods such as PiD (97.1%/98.4%) and SAFE (96.2%/98.8%). Our method maintains consistently high performance across diverse GAN architectures, achieving over 98% ACC on ProGAN, StyleGAN, StyleGAN2, CycleGAN, and StarGAN. Notably, it achieves 96.3% ACC on Deepfake, surpassing PiD (95.4%) and SAFE (93.1%). Although performance varies across individual generators, particularly on BigGAN and GauGAN, our method achieves the highest average performance

Table 1: Intra-architecture evaluation on GAN-based synthetic images from ForenSynths. ACC/AP are reported in %, with their averages (mACC/mAP) in the last columns.

Method	Refer.	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		mACC	mAP
		ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP		
F3Net Qian et al. (2020)	ECCV2020	99.4	**100**	92.6	99.7	88.0	99.8	65.3	69.9	76.4	84.3	**100**	**100**	58.1	56.7	63.5	78.8	80.4	86.1
SelfBland Shiohara & Yamasaki (2022)	CVPR2022	58.8	65.2	50.1	47.7	48.6	47.4	51.1	51.9	59.2	65.3	74.5	89.2	59.2	65.5	93.8	99.3	61.9	66.4
Ojha Ojha et al. (2023)	CVPR2023	99.7	**100**	96.3	98.7	83.9	98.4	90.5	99.1	87.9	99.8	91.4	**100**	89.9	**100**	80.2	90.2	89.9	98.2
LGrad Tan et al. (2023)	CVPR2023	**99.9**	**100**	94.8	99.9	96.0	99.9	82.9	90.7	85.3	94.0	99.6	**100**	72.4	79.3	58.0	67.9	86.2	91.4
FreqNet Tan et al. (2024a)	AAAI2024	99.6	**100**	90.2	99.7	88.0	99.5	90.5	96.0	95.8	99.6	85.7	99.8	93.4	98.6	88.9	**99.4**	91.5	99.0
NPR Tan et al. (2024b)	CVPR2024	99.8	**100**	96.3	99.8	97.3	**100**	87.5	94.5	95.0	99.5	99.7	**100**	86.6	88.8	77.4	86.2	92.4	96.1
CO-SPY Cheng et al. (2025)	CVPR2025	74.7	78.1	63.9	70.2	59.7	63.1	71.6	83.9	58.5	55.8	62.1	94.3	69.6	83.4	65.7	79.7	65.7	76.0
PiD Fu et al. (2025)	ICML2025	99.8	99.9	**99.8**	99.9	–	–	93.7	94.9	**99.7**	**100**	91.3	96.9	**100**	**100**	95.4	97.5	97.1	98.4
C2P-CLIP Tan et al. (2025)	AAAI2025	99.7	**100**	99.4	**100**	–	–	95.3	**99.9**	90.7	99.9	96.6	**100**	95.3	**100**	89.9	97.3	95.2	**99.4**
SAFE Li et al. (2025)	KDD2025	**99.9**	**100**	98.0	99.9	98.6	**100**	89.7	95.9	98.9	99.8	99.9	**100**	91.5	97.2	93.1	97.5	96.2	98.8
FALCON-Net Zhang et al. (2026)	TIFS2026	99.7	**100**	97.6	99.8	97.2	99.7	90.6	96.6	93.4	97.1	**100**	**100**	90.0	97.2	60.4	76.1	91.1	95.8
SimLBR Dhakal et al. (2026)	CVPR2026	99.3	-	87.4	-	84.6	-	**95.5**	-	88.1	-	88.5	-	97.1	-	-	-	91.5	-
OURS		**99.9**	**100**	99.4	**100**	**98.9**	**100**	94.0	98.1	98.4	99.6	99.3	99.9	92.1	98.7	**96.3**	99.1	**97.3**	**99.4**

Table 2: Cross-architecture evaluation on DM-based synthetic images from Ojha. Bold and underline represent the best and second-best performance.

Method	Refer.	Guided		LDM_200		LDM_200_cfg		LDM_100		GLIDE_100_27		GLIDE_50		GLIDE_100_10		DALL-E		mACC	mAP
		ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP		
F3Net Qian et al. (2020)	ECCV2020	53.4	70.8	73.4	83.3	80.7	89.1	74.1	84.0	87.0	94.5	88.5	95.4	88.3	95.4	71.6	79.9	77.1	86.5
SelfBland Shiohara & Yamasaki (2022)	CVPR2022	58.3	63.4	52.6	51.9	51.9	52.6	53.0	54.0	59.4	64.1	64.2	68.3	58.8	63.2	52.4	51.6	56.3	58.6
Ojha Ojha et al. (2023)	CVPR2023	75.7	85.1	90.2	97.1	77.3	88.6	90.5	97.0	90.7	97.2	91.1	97.4	90.1	97.0	89.5	96.8	86.8	94.5
LGrad Tan et al. (2023)	CVPR2023	86.6	**100**	94.2	99.1	95.9	99.2	94.8	99.2	87.4	93.2	90.7	95.1	89.4	94.9	88.5	97.3	90.9	97.2
FreqNet Tan et al. (2024a)	AAAI2024	63.8	68.9	97.5	99.9	97.8	99.9	97.9	99.9	82.6	95.1	85.3	95.5	86.3	95.9	98.8	99.8	88.7	94.3
NPR Tan et al. (2024b)	CVPR2024	75.8	81.0	99.1	99.9	99.0	99.8	99.3	99.9	97.8	**99.7**	98.2	**99.8**	98.2	99.8	94.5	99.5	95.2	97.4
CO-SPY Cheng et al. (2025)	CVPR2025	62.5	86.0	83.1	87.5	85.3	91.0	82.7	86.9	73.5	78.2	69.1	74.6	76.6	81.6	81.8	87.2	76.8	84.1
PiD Fu et al. (2025)	ICML2025	82.6	93.2	**99.9**	**100**	**99.9**	**100**	99.4	99.7	**99.2**	**99.7**	**99.3**	**99.8**	**99.3**	**99.9**	83.5	91.7	95.3	98.0
C2P-CLIP Tan et al. (2025)	AAAI2025	77.8	92.2	99.1	**100**	98.1	98.8	99.0	**100**	94.7	99.3	94.2	99.3	94.4	99.4	98.8	**99.9**	94.5	98.7
SAFE Li et al. (2025)	KDD2025	82.4	95.8	98.8	**100**	98.7	99.9	98.8	**100**	95.8	98.9	96.6	99.2	97.3	99.4	97.5	99.7	95.7	**99.1**
FALCON-Net Zhang et al. (2026)	TIFS2026	**87.4**	94.9	97.0	99.6	97.0	99.6	97.8	99.7	96.2	99.1	92.4	96.8	94.0	97.4	97.0	99.6	94.9	98.3
MCAN Tan et al. (2026)	AAAI2026	70.9	-	98.8	-	94.2	-	98.5	-	97.4	-	97.2	-	97.1	-	98.8	-	94.1	-
OURS		86.0	97.2	99.5	**100**	99.4	**100**	**99.5**	**100**	93.9	98.3	95.2	98.6	96.8	99.0	**99.0**	**99.9**	**96.1**	**99.1**

across all eight sources. These results suggest that the learned residual representations generalize effectively across heterogeneous GAN architectures rather than being specific to the training generator.

Cross-Diffusion-Sources Evaluation on Ojha. To evaluate whether the learned representations generalize beyond GAN-based generators, we test our ProGAN-trained model on the Ojha benchmark comprising eight generative sources. As shown in Table 2, our method achieves the best overall performance with 96.1% mACC and a tied-best 99.1% mAP, clearly outperforming strong baselines such as PiD (95.3%/98.0%) on both metrics and matching the highest mAP of SAFE (95.7%/99.1%). Notably, our method achieves over 99% ACC on all LDM variants and 99.0% on DALL-E. Although Guided Diffusion remains challenging and some methods perform better on individual GLIDE variants, our method maintains strong performance across all eight diffusion sources. These results indicate that the learned residual cues transfer effectively from GAN-based training data to unseen diffusion architectures.

Cross-GAN-Sources Evaluation on Self-Synthesis. To further evaluate generalization across diverse GAN architectures, we test our ProGAN-trained model on the Self-Synthesis benchmark comprising nine unseen GAN generators. As shown in Table 3, our method achieves 98.5% mACC and 99.7% mAP, outperforming PiD (94.7%/98.8%) and FerretNet (77.4%/98.6%) on both metrics. Notably, our method achieves at least 99% ACC on AttGAN, BEGAN, InfoMaxGAN, RelGAN, and STGAN, while attaining 100% AP on eight of the nine generators. Although its performance is relatively lower on S3GAN, our method maintains strong average performance across the nine evaluated sources and attains 99.7% mAP, within 0.1 points of the best result in Table 3. These results further demonstrate that the learned residual cues generalize consistently across unseen GAN architectures beyond the training generator.

Cross-architecture Evaluation on GenImage. To further evaluate generalization across heterogeneous generative paradigms, we evaluate our model on GenImage, which contains both GAN- and diffusion-based generators. As shown in Table 4, our method achieves the best overall performance with 97.5% mACC, outperforming strong baselines such as MCAN (96.9%), C2P-CLIP (95.8%), and SAFE (95.6%). Notably, our method achieves 99.2% ACC on Midjourney and 98.3% on VQD-M, while maintaining at least 94% ACC across all eight generators. Although some competing methods perform better on individual sources, our method maintains consistently strong perfor-

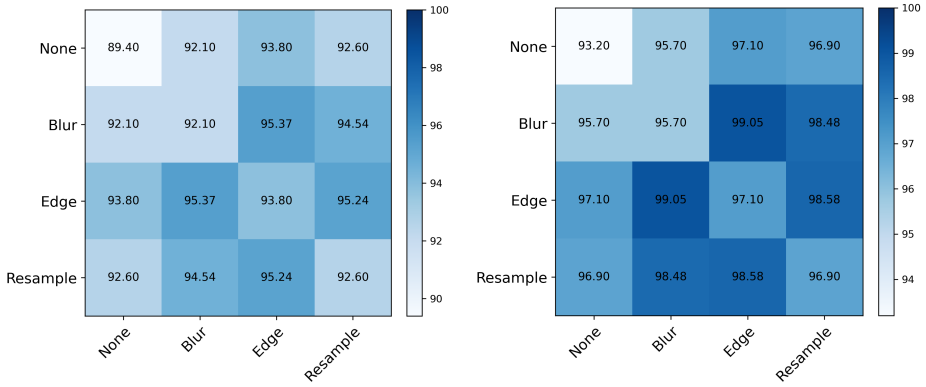
Table 3: Cross-GAN-Sources evaluation on GAN-based synthetic images from Self-Synthesis.

Method	Refer.	AttGAN		BEGAN		CramerGAN		InfoMaxGAN		MMDGAN		RelGAN		S3GAN		SNGAN		STGAN		mACC	mAP
		ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP		
F3Net Qian et al. (2020)	ECCV2020	85.2	94.8	87.1	97.5	89.5	99.8	67.1	83.1	73.7	99.6	98.8	**100**	65.4	70.0	51.6	93.6	60.3	99.9	75.4	93.1
SelfBland Shiohara & Yamasaki (2022)	CVPR2022	63.1	62.4	56.4	57.2	75.1	74.3	79.0	78.7	68.6	69.3	73.6	72.8	53.2	54.1	61.6	62.4	61.2	60.5	65.8	65.7
Ojha Ojha et al. (2023)	CVPR2023	78.5	98.3	72.0	98.9	77.6	99.8	77.6	98.9	77.6	99.7	78.2	98.7	85.2	98.1	77.6	98.7	74.2	97.8	77.6	98.8
LGrad Tan et al. (2023)	CVPR2023	68.6	93.8	69.9	89.2	50.3	54.0	71.1	82.0	57.5	67.3	89.1	99.1	78.5	86.0	78.0	87.4	54.8	68.0	68.6	80.8
FreqNet Tan et al. (2024a)	AAAI2024	89.8	98.8	98.8	**100**	95.2	98.2	94.5	97.3	95.2	98.2	**100**	**100**	88.3	94.3	85.4	90.5	98.8	**100**	94.0	97.5
NPR Tan et al. (2024b)	CVPR2024	83.0	96.2	99.0	99.8	98.7	99.0	94.5	98.3	98.6	99.0	99.6	**100**	79.0	80.0	88.8	97.4	98.0	**100**	93.2	96.6
SAFE Li et al. (2025)	KDD2025	99.4	**100**	99.8	**100**	**99.7**	**100**	99.6	**100**	99.7	**100**	99.5	**100**	**94.5**	**98.6**	98.8	99.98	99.9	**100**	**99.0**	**99.8**
FerretNet Liang et al. (2026)	NeurIPS2026	**100**	99.9	98.2	**100**	66.9	99.9	63.1	98.8	63.3	99.9	99.9	**100**	90.0	91.3	56.6	97.7	59.1	99.9	77.4	98.6
PiD Fu et al. (2025)	ICML2025	**100**	**100**	**99.9**	**100**	95.4	99.4	95.4	99.8	95.4	99.5	**100**	**100**	85.7	91.2	95.4	99.6	85.0	99.9	94.7	98.8
Beyond Semantic Features Zhou et al. (2026)	AAAI2026	99.6	99.9	99.1	**100**	98.9	**100**	**99.8**	**100**	99.4	**100**	99.7	**100**	85.2	91.7	**99.1**	**100**	99.9	**100**	97.9	99.1
OURS		99.1	**100**	99.0	**100**	**99.7**	**100**	99.7	**100**	**99.8**	**100**	99.2	**100**	91.4	97.3	98.9	**100**	**100**	**100**	98.5	99.7

Table 4: Cross-architecture evaluation on GAN and DM-based synthetic images from GenImage. All models are trained exclusively on ProGAN-generated images from ForenSynths.

Method	Refer.	Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	WuKong	VQDM	BigGAN	mACC
F3Net Qian et al. (2020)	ECCV2020	50.1	99.9	**99.9**	49.9	50.0	99.9	49.9	49.9	68.7
GramNet Liu et al. (2020)	CVPR2020	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
DIRE Wang et al. (2023)	ICCV2023	65.8	99.7	99.7	54.5	58.1	99.4	54.3	49.8	72.7
LaRE^2 Luo et al. (2024)	CVPR2024	74.0	**100**	**99.9**	61.7	88.5	**100**	97.2	68.7	86.2
FatFormer Liu et al. (2024)	CVPR2024	93.1	97.0	97.2	82.0	95.0	95.8	88.8	49.9	87.4
NPR Tan et al. (2024b)	CVPR2024	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
DRCT Chen et al. (2024)	ICML2024	91.5	95.0	94.4	79.4	89.2	94.7	90.0	81.7	89.5
SAFE Li et al. (2025)	KDD2025	95.3	99.4	99.3	82.1	96.3	98.2	96.3	97.8	95.6
C2P-CLIP Tan et al. (2025)	AAAI2025	88.2	90.9	97.9	**96.4**	**99.0**	98.8	96.5	98.7	95.8
MCAN Tan et al. (2026)	AAAI2026	94.3	98.8	98.5	90.2	98.6	98.8	96.8	**98.8**	96.9
SimLBR Dhakal et al. (2026)	CVPR2026	75.5	85.5	84.4	85.9	90.1	90.0	93.2	95.5	87.5
OURS		**99.2**	99.5	99.4	94.0	95.5	99.2	**98.3**	94.9	**97.5**

mance across heterogeneous architectures. These results further demonstrate the effectiveness of the learned residual cues for cross-architecture generalization across different generative paradigms. Additional experiments further evaluate the plug-and-play applicability of the proposed method across different SID baselines and its robustness to common image perturbations, as detailed in Appendix A and Appendix B, respectively.



(a) mACC (B) mAP

Figure 4: Ablation of the three physical prior simulators on the Ojha benchmark. Rows and columns denote the active simulators, where "None" indicates no simulator. Each cell reports (a) mACC or (b) mAP (%), with darker colors indicating better performance.

4.3 Ablation Studies

Effectiveness and Complementarity of Physical Prior Simulators. To evaluate the effectiveness and complementarity of the three physical prior simulators, we conduct ablation experiments with individual simulators, pairwise combinations, and their joint application. As shown in Figure 4, each individual simulator consistently improves over the baseline without physical priors (89.40% mACC and 93.20% mAP), with Edge achieving the best individual performance of 93.80% mACC

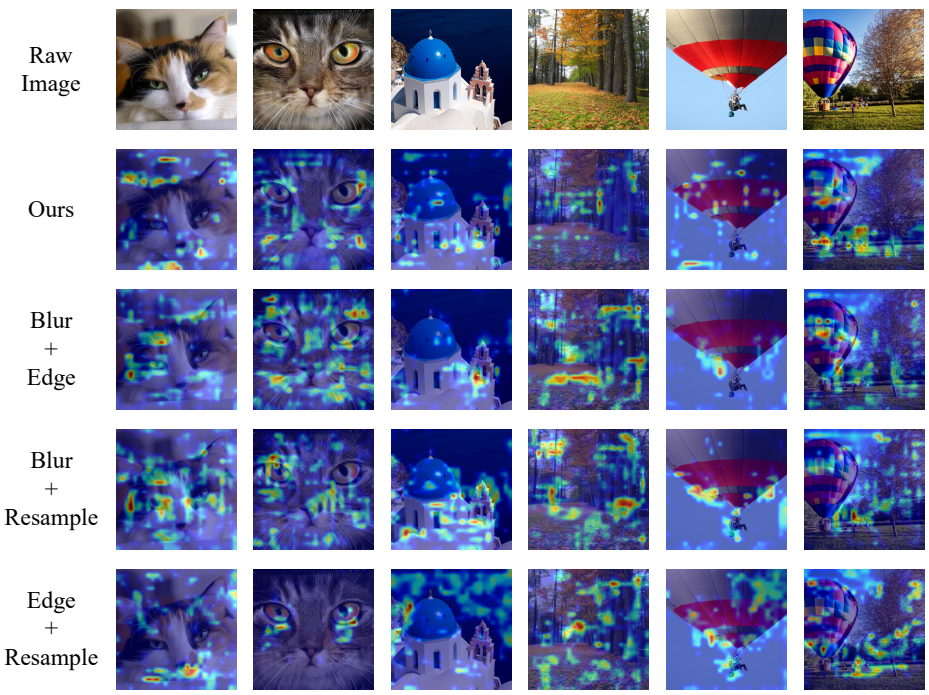


Figure 5: Grad-CAM visualizations of the full model and pairwise simulator combinations. Warmer colors indicate regions receiving higher attention.

and 97.10% mAP. Combining two simulators further improves the performance, where Blur+Edge achieves the best pairwise result of 95.37% mACC and 99.05% mAP. Finally, integrating all three simulators yields the best overall performance of 96.10% mACC and 99.10% mAP. The progressive improvements from individual to combined simulators indicate that the three physical priors capture complementary discrepancy cues rather than redundant information. Further sensitivity analysis of key parameters in three physical prior simulators is provided in Appendix C.1.

Qualitative Analysis of Simulator Complementarity. To further investigate how different physical priors contribute to detection, we employ Grad-CAM to visualize the attention regions induced by different simulator combinations. As shown in Figure 5, different combinations exhibit distinct spatial responses to local textures, object boundaries, and structural regions. Compared with the pairwise configurations, the full model integrates responses over a broader range of artifact-relevant regions while preserving localized activations on fine-grained structures. This suggests that Blur, Edge, and Resample provide complementary spatial cues rather than repeatedly focusing on the same image regions. These qualitative observations are consistent with the performance gains in Figure 4, further supporting the complementary roles of the three physical priors. Additional ablations of framework components are provided in Appendix C.2.

5 Conclusion

In this paper, we investigate cross architecture generalization in synthetic image detection from the perspective of shared physical characteristics in the generation process. Through physically motivated probe experiments, we identify three shared physical defects, reflected in discrepancies from natural images under resampling operations, spatial smoothing, and high frequency characteristics. Building on these observations, we propose a learnable physical prior decoupling framework that transforms physical defects into discriminative residuals by combining physical prior simulators with learnable residual modulation. Extensive experiments across multiple benchmarks show that our method ranks first or second among state-of-the-art methods and generalizes effectively across different generative architectures and paradigms using only ProGAN generated images for training, demonstrating the effectiveness of shared physical characteristics as generalizable forensic cues.

REFERENCES

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, volume 2, pp. 6, 2024.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13455–13465. IEEE, 2025.

Aayush Dhakal, Subash Khanal, Srikumar Sastry, Jacob Arndt, Philipe Dias, Dalton Lunga, and Nathan Jacobs. Simlbr: Learning to detect fake images by learning to detect real images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35472–35482, 2026.

Xinghe Fu, Zhiyuan Yan, Zheng Yang, Taiping Yao, Yandan Zhao, Shouhong Ding, and Xi Li. Pid: Generalized ai-generated images detection with pixelwise decomposition residuals. In *Forty-second International Conference on Machine Learning*, 2025.

Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. Privacy and security concerns in generative ai: A comprehensive survey. *Ieee Access*, 12:48126–48144, 2024.

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

Zexi Jia, Chuanwei Huang, Yeshuang Zhu, Hongyan Fei, Xiaoyue Duan, Zhiqiang Yuan, Ying Deng, Jiapei Zhang, Jinchao Zhang, and Jie Zhou. Secret lies in color: Enhancing ai-generated images detection with color distribution analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13445–13454. IEEE, 2025.

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 *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3465–3469. IEEE, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, David Camacho, and Abdenour Hadid. Bi-lora: A vision-language approach for synthetic image detection. *Expert Systems*, 42(2):e13829, 2025.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 2405–2414, 2025.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21262–21272, 2026.

Shuqiao Liang, Jian Liu, Chen Renzhang, and Quanlong Guan. Ferretnet: Efficient synthetic image detection via local pixel dependencies. *Advances in Neural Information Processing Systems*, 38: 42109–42133, 2026.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10770–10780. IEEE, 2024.

Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8057–8066. IEEE, 2020.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17006–17015. IEEE, 2024.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18699–18708. IEEE, 2022.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12105–12114. IEEE, 2023.

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*, volume 38, pp. 5052–5060, 2024a.

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*, pp. 28130–28139, 2024b.

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*, volume 39, pp. 7184–7192, 2025.

Lei Tan, Shuwei Li, Mohan Kankanhalli, and Robby T Tan. Aggregating diverse cue experts for ai-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9314–9322, 2026.

Chao Wang, Zijin Yang, Yaofei Wang, Weiming Zhang, and Kejiang Chen. Aedr: Training-free ai-generated image attribution via autoencoder double-reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9675–9683, 2026.

Hongsong Wang, Renxi Cheng, Yang Zhang, Chaolei Han, and Jie Gui. Lota: Bit-planes guided ai-generated image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17246–17255. IEEE, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22388–22398. IEEE, 2023.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, et al. All patches matter, more patches better: Enhance ai-generated image detection via panoptic patch learning. In *International Conference on Learning Representations*, volume 2026, pp. 121742–121761, 2026.

Dengyong Zhang, Miao Hu, Jiaxin Chen, Changsheng Chen, Jin Wang, Yun Song, Gaobo Yang, Xin Liao, and Xiangling Ding. Falcon-net: Feature aggregation of local patterns for ai-generated image detection. *IEEE Transactions on Information Forensics and Security*, 21:3842–3855, 2026.

Chenming Zhou, Jiaan Wang, Yu Li, Lei Li, Juan Cao, and Sheng Tang. Beyond semantic features: Pixel-level mapping for generalized ai-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36101–36109, 2026.

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.