GENERALIZABLE AI-GENERATED IMAGE DETECTION VIA NEIGHBORHOOD DEPENDENCY REPRESENTATION AND HIERARCHICAL DEPENDENCY MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing AI-generated image detectors often degrade on unseen generators because synthesis traces exhibit substantial variations across generation paradigms. Most existing methods focus on generator-specific artifacts or model forensic evidence from a single perspective, limiting their ability to capture transferable synthesis traces. We observe that although artifact appearances differ, different generators consistently alter image dependency structures across spatial and spectral scales. Based on this observation, we propose a hierarchical dependency modeling framework for generalizable AI-generated image detection. Specifically, we introduce the Neighborhood Dependency Representation (NDR), which reconstructs each pixel from its local context and extracts residual dependency variations to reduce semantic bias and reveal synthesis-related irregularities. Building upon NDR, the local subband dependency branch and global spectral dependency branch capture complementary local and global dependency structures, respectively. A bidirectional confidence guided fusion module further adaptively integrates these representations for robust classification. Extensive experiments on AIGCDetectBenchmark, ForenSynths, and WildRF demonstrate superior cross-generator and real-world generalization, showing the effectiveness of dependency-based forensic representation learning.

1 INTRODUCTION

With the rapid development of generative models, AI-generated images have achieved remarkable improvements in visual quality, text-image alignment, and content diversity Xie et al. (2025); Liu et al. (2026). From GANs Goodfellow et al. (2014) to diffusion models Ho et al. (2020) and large-scale text-to-image systems Rombach et al. (2022); Saharia et al. (2022), modern generators can now synthesize highly realistic visual content. While these advances benefit content creation, they also facilitate misinformation, identity impersonation, and the fabrication of digital evidence. Reliable AI-generated image detection has therefore become a critical problem in multimedia forensics and AI security.

For AI-generated image detection, the central challenge is no longer merely recognizing images produced by generators seen during training, but detecting images from previously unseen generative models. Although existing detectors perform well in-domain, their performance often deteriorates substantially under cross-generator evaluation Wang et al. (2020); Chai et al. (2020); Ojha et al. (2023); Guillaro et al. (2025); Wang et al. (2026). Existing studies generally attribute this degradation to semantic shortcuts or generator-specific artifacts Ye et al. (2024); Zhou et al. (2025). Consequently, recent efforts seek more transferable forensic evidence from frequency-domain anomalies Tan et al. (2024a); Yan et al. (2026); Alam et al. (2025), reconstruction errors Wang et al. (2023), and local statistics Liang et al. (2025). These methods have significantly improved robustness by moving beyond high-level semantic cues.

However, we argue that the limitation of current detectors lies not in the lack of transferable artifacts, but in how these artifacts are represented. Existing methods typically characterize synthesis traces from a single forensic perspective He et al. (2026), such as local statistics Tan et al. (2024b); Liang et al. (2025), reconstruction errors, or spectral anomalies Durall et al. (2020); Yan et al. (2026).

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.

In reality, artifact appearances vary substantially across different generators due to diverse synthesis mechanisms. Nevertheless, the synthesis process consistently perturbs the natural dependency structures of images across multiple spatial and spectral scales. Local neighborhood relationships exhibit irregular pixel dependencies and texture inconsistencies, while global frequency distributions reveal long-range spectral correlations and periodic structures. As illustrated in Fig. 1, modeling only one of these dependency characteristics inevitably captures only a partial view of synthesis traces, limiting cross-generator generalization.

Our key insight is that what generalizes is not a specific artifact appearance, but the dependency irregularities underlying artifact formation. Based on this view, we reformulate generalizable AI-generated image detection as dependency-oriented forensic representation learning. Instead of learning directly from RGB images, we first construct a Neighborhood Dependency Representation (NDR), measuring each pixel's deviation from a neighborhood-reconstructed contextual baseline. This representation suppresses dominant semantic content while explicitly exposing synthesis-induced dependency irregularities. Building upon NDR, we develop a hierarchical dependency modeling framework. A Local Subband Dependency Branch (LSDB) captures localized dependency irregularities through directional wavelet subbands, whereas a Global Spectral Dependency Branch (GSDB) models long-range spectral dependencies and contextual correlations in the Fourier domain. Finally, a Bidirectional Confidence Guided Fusion (BCGF) module adaptively integrates complementary dependency cues, enabling robust forensic representations across diverse generators.

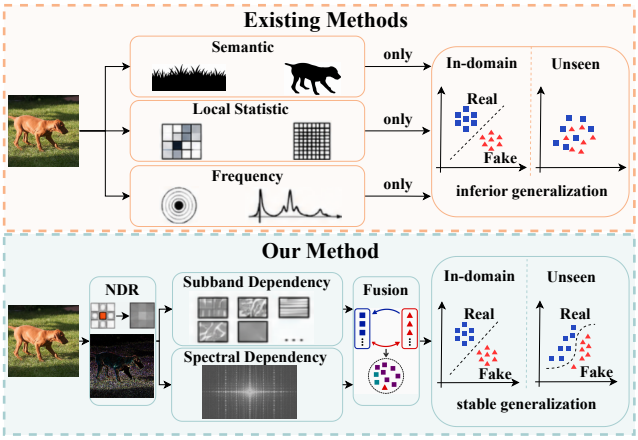

Figure 1: Comparison between existing methods and ours. Existing methods typically model isolated forensic cues, whereas ours captures and adaptively integrates multi-scale dependency disruptions.

Our contributions are summarized as follows.

- We introduce a dependency-oriented perspective for generalizable AI-generated image detection, arguing that transferable forensic evidence lies in synthesis-induced dependency irregularities rather than generator-specific artifact appearances.
- We develop a hierarchical dependency modeling framework that jointly captures local and global dependency irregularities through complementary spectral representations, together with a bidirectional confidence guided fusion strategy for adaptive dependency integration.
- Extensive experiments on AIGCDetectBenchmark, ForenSynths, and WildRF demonstrate strong cross-generator and real-world generalization over state-of-the-art methods.

2 Related Work

2.1 Spectral Artifact-Based Detection

Spectral artifact-based methods seek transferable forensic evidence from frequency-domain deviations introduced by generation operations such as upsampling, decoding, and denoising Zhou et al. (2023). Early studies reveal abnormal spectral distributions in GAN-generated images Frank et al. (2020); Durall et al. (2020), while F3Net Qian et al. (2020) and FreqNet Tan et al. (2024a) learn frequency-aware representations for forgery detection and cross-model generalization, respectively. Synthbuster Bammey (2024) identifies diffusion-generated images by analyzing frequency artifacts in the Fourier transform of image residuals. Recent methods further extend spectral modeling to diverse generators: SPAI Karageorgiou et al. (2025) supports arbitrary-resolution inputs, STAL

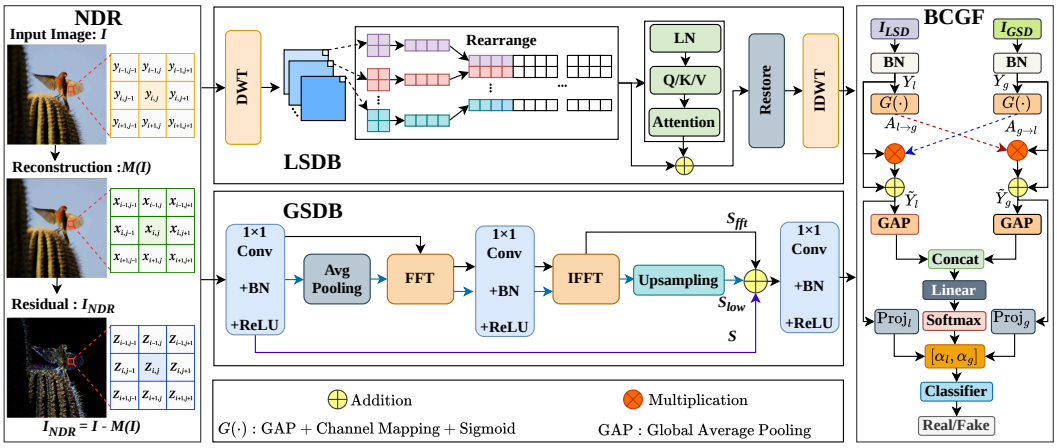

Figure 2: Overview of the proposed framework. NDR derives $I_{\text{NDR}} = I - M(I)$ to attenuate semantic interference. LSDB and GSDB capture local subband dependencies and global spectral dependencies, respectively, while BCGF adaptively integrates the complementary representations for real/fake classification.

Li et al. (2026) exploits spectral-tail uplift through frequency-domain auxiliary supervision, and SpecXNet Alam et al. (2025) jointly captures spatial and spectral traces. Despite their effectiveness, spectral methods focus on global frequency distributions, providing only one perspective of synthesis-induced dependency irregularities while failing to explicitly model multi-scale, localized dependency structures.

2.2 Reconstruction Error-Based Detection

Reconstruction-based methods distinguish images through discrepancies introduced by inversion or regeneration Qi et al. (2026). DIRE Wang et al. (2023) uses diffusion reconstruction errors; DRCT Chen et al. (2024) learns discriminative features through reconstruction contrast; and LaRE^2 Luo et al. (2024) models reconstruction errors in the latent space. Recent studies further refine these cues: FIRE Chu et al. (2025) analyzes frequency-decomposed reconstruction errors, DEUA Huang et al. (2025) models diffusion epistemic uncertainty, and DySy-Det Jin et al. (2026) combines localized reconstruction errors with reconstruction-path consistency. Although reconstruction discrepancies can reveal synthesis-related deviations, these methods rely on diffusion models or reconstruction processes, increasing computational cost and introducing model dependence. Moreover, the extracted evidence is intrinsically tied to the adopted reconstruction prior, making its forensic representation dependent on the reconstruction process itself rather than directly modeling synthesis-induced dependency irregularities shared across different generators.

2.3 Local Statistics-Based Detection

Local statistics-based methods seek transferable traces from gradients, neighborhood relationships, textures, and entropy patterns. LGrad Tan et al. (2023) learns gradient artifacts; NPR Tan et al. (2024b) exposes upsampling-induced neighborhood inconsistencies; MLEP Yuan et al. (2025) models multi-granularity local entropy; FerretNet Liang et al. (2025) efficiently captures local pixel dependencies; Pixel-level Mapping Zhou et al. (2026) suppresses semantic bias to emphasize high-frequency evidence; and DCCT Zhong et al. (2026) models camera-induced color-correlation statistics through demosaicing-guided training. These studies demonstrate that local dependency irregularities provide more transferable forensic evidence than source-related semantic cues. Nevertheless, most methods characterize dependency from a single local statistical perspective, emphasizing predefined neighborhood relationships or handcrafted statistical patterns. Such representations are insufficient to capture the hierarchical nature of synthesis-induced dependency irregularities, which simultaneously manifest through localized interactions and global spectral organizations.

3 PROPOSED METHOD

The overall framework is illustrated in Fig. 2. Given an input image I, we first construct a Neighborhood Dependency Representation (NDR), which suppresses dominant semantic content while exposing synthesis-induced dependency irregularities. Building upon this representation, the proposed hierarchical dependency modeling framework characterizes complementary forensic evidence from both local and global perspectives through the Local Subband Dependency Branch (LSDB) and the Global Spectral Dependency Branch (GSDB), respectively. Finally, the Bidirectional Confidence Guided Fusion module (BCGF) adaptively calibrates and integrates the two dependency representations for authenticity classification.

3.1 NEIGHBORHOOD DEPENDENCY REPRESENTATION

Generalizable AI-generated image detection requires representations that preserve transferable dependency irregularities while reducing interference from image semantics. Directly learning from RGB images is challenging because object structures, textures, and scene semantics often dominate feature responses, making subtle synthesis-induced dependency variations difficult to capture. To address this issue, we introduce the NDR module, a parameter-free transformation that explicitly measures each pixel's deviation from its neighborhood-supported contextual baseline. Rather than emphasizing high-frequency residuals alone, NDR provides a dependency-oriented forensic representation for subsequent hierarchical dependency modeling.

Given an input image $I \in \mathbb{R}^{C\times H\times W}$, let $I_c(i,j)$ denote the pixel value at location (i,j) in the c-th channel, and let $\Omega^k_{i,j}$ denote a $k \times k$ neighborhood centered at (i,j). The center pixel is excluded from contextual reconstruction, i.e., replaced with zero, and the neighborhood-supported baseline is obtained through channel-wise median aggregation:

$$M_c(i,j) = \mathrm{Median}\left(\{0\} \cup \{I_c(u,v)\}_{(u,v)\in\Omega^k_{i,j}\setminus\{(i,j)\}}\right). \tag{1}$$

The median operation provides a robust estimate of the local contextual structure by reducing the influence of isolated high-frequency responses. As a result, $M(I)$ represents the image content supported by the surrounding neighborhood and provides a stable reference for measuring relational deviations. NDR is then defined as:

$$I_{\mathrm{NDR}} = I - M(I). \tag{2}$$

Through neighborhood reconstruction and residualization, I_{NDR} characterizes each pixel's deviation from its contextual baseline. This representation attenuates dominant semantic responses and emphasizes irregular textures, disrupted pixel relationships, and synthesis-induced dependency variations. The resulting I_{NDR} establishes a shared forensic representation for subsequent local–global dependency modeling.

3.2 DUAL-SPECTRAL DEPENDENCY MODELING

The dependency irregularities preserved by NDR exhibit complementary characteristics across different spatial and spectral scales. Local dependency irregularities are reflected by directional texture inconsistencies and cross-subband interactions, whereas global dependency irregularities manifest as long-range spectral correlations and distributional deviations. To jointly characterize these complementary forensic cues, we develop a hierarchical dependency modeling framework consisting of the LSDB and GSDB.

3.2.1 LOCAL SUBBAND DEPENDENCY BRANCH

LSDB models localized dependency irregularities by jointly organizing spatial neighborhoods and directional wavelet subbands. Unlike conventional frequency modeling that primarily captures spectral responses, LSDB explicitly preserves the spatial correspondence among neighboring regions and wavelet subbands, enabling localized dependency structures to be characterized across complementary frequency components.

Specifically, Haar wavelet decomposition, subband concatenation, and local-window rearrangement are compactly formulated as

$$I_W = \mathcal{P}\left(\mathrm{Cat}\left(\mathcal{W}(I_{\mathrm{NDR}})\right)\right), \tag{3}$$

where $\mathcal{W}(\cdot)$ decomposes the representation into a low-frequency approximation and multiple directional high-frequency components. $\mathrm{Cat}(\cdot)$ concatenates these subbands, and $\mathcal{P}(\cdot)$ rearranges them into local windows. This organization enables spatial neighborhoods and cross-subband associations to be modeled jointly.

The resulting subband representation is enhanced through local window attention:

$$\widehat{I}_W = I_W + \mathcal{A}_w(\mathrm{LN}(I_W)), \tag{4}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization and $\mathcal{A}_w(\cdot)$ denotes the wavelet subband association operation. The enhanced representation is then restored to the spatial domain:

$$I_{\mathrm{LSD}} = \mathcal{W}^{-1}\left(\mathcal{P}^{-1}(\widehat{I}_W)\right). \tag{5}$$

Consequently, I_{LSD} captures localized dependency irregularities, directional texture inconsistencies, and fine-grained cross-subband associations, providing discriminative forensic evidence at local spatial-spectral scales.

3.2.2 Global Spectral Dependency Branch

Complementary to LSDB, GSDB models dependency irregularities from a global spectral perspective. Instead of focusing on localized interactions, GSDB captures image-wide spectral correlations, periodic responses, and long-range contextual dependencies, providing complementary forensic evidence beyond local dependency modeling.

We first obtain the base feature S from I_{NDR} through a lightweight channel mapping composed of convolution, batch normalization, and nonlinear activation. The full-resolution feature is transformed into the Fourier domain, where its real and imaginary responses are jointly calibrated before inverse transformation:

$$S_{\mathrm{fft}} = \mathcal{F}^{-1}\left(\Phi(\mathcal{F}(S))\right). \tag{6}$$

Here, $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the Fourier and inverse Fourier transforms, respectively, while $\Phi(\cdot)$ denotes the learnable complex spectral calibration function. This path models global spectral dependencies and highlights periodic artifacts associated with generative synthesis.

A parallel low-frequency path further captures dependencies on coarse structural context:

$$S_{\mathrm{low}} = \mathrm{Up}\left(\mathcal{F}^{-1}\left(\Phi(\mathcal{F}(\mathrm{AAP}(S)))\right)\right), \tag{7}$$

where $\mathrm{AAP}(\cdot)$ and $\mathrm{Up}(\cdot)$ denote adaptive average pooling and upsampling, respectively. The branch output aggregates the base feature, full-resolution spectral response, and low-frequency structural context:

$$I_{\mathrm{GSD}} = \phi_1(S + S_{\mathrm{fft}} + S_{\mathrm{low}}), \tag{8}$$

where $\phi_1(\cdot)$ denotes the output channel mapping. The resulting representation characterizes image-wide dependency irregularities through global spectral responses and long-range contextual correlations, complementing the localized dependency structures learned by LSDB.

3.3 Adaptive Fusion and Optimization

Local and global dependency representations provide complementary forensic evidence whose relative importance varies across synthesis models and image content. Rather than assuming fixed contributions from the two branches, BCGF performs bidirectional calibration followed by input-adaptive integration, allowing the detector to dynamically emphasize the dependency representation that provides stronger forensic evidence for each input.

Given I_{LSD} and I_{GSD}, they are first stabilized via batch normalization to yield Y_l and Y_g. Each branch estimates a channel-wise confidence descriptor that guides the calibration of its counterpart. The mutually calibrated features, $\widetilde{Y}_l$ and $\widetilde{Y}_g$, are formulated as:

$$\widetilde{Y}_l = Y_l + Y_l \odot G_l(Y_g), \quad \widetilde{Y}_g = Y_g + Y_g \odot G_g(Y_l), \tag{9}$$

where $\odot$ denotes element-wise multiplication, and the gating functions $G_l(\cdot)$ and $G_g(\cdot)$ sequentially apply global average pooling (GAP), channel mapping, and Sigmoid activation. Through this

Table 1: Accuracy (%) comparison on the ForenSynths dataset. Bold and underlined values denote the best and second-best results on the mean metric, respectively.

Method	Venue	Big-GAN	CRN	Cycle-GAN	Deep-fake	Gau-GAN	IMLE	Pro-GAN	SAN	Seeing-Dark	Star-GAN	Style-GAN	Style-GAN2	WFIR	mAcc.
UniFD	CVPR2023	95.1	56.6	98.3	68.6	99.5	69.1	99.8	56.6	62.2	95.7	84.9	74.9	87.1	80.7
LGrad	CVPR2023	83.0	50.7	85.4	58.1	72.5	50.8	99.8	50.9	63.1	99.6	94.8	96.0	63.5	74.5
NPR	CVPR2024	84.4	50.0	96.1	80.2	82.5	50.0	99.8	69.2	62.5	99.3	97.7	98.4	65.8	79.7
FreqNet	AAAI2024	90.5	59.0	95.8	88.9	93.4	59.1	99.6	71.9	65.6	85.7	90.2	87.9	49.1	79.8
SPAI	CVPR2025	68.9	72.5	68.1	52.1	75.4	62.8	88.3	50.5	91.4	54.8	72.8	73.5	79.5	70.1
CO-SPY	CVPR2025	95.3	61.3	98.1	83.4	98.4	61.3	100.0	74.0	72.8	89.0	92.2	93.8	64.2	<u>83.4</u>
MLEP	NeurIPS2025	87.1	50.0	98.3	52.4	82.0	50.0	99.5	71.9	50.0	100.0	99.6	99.9	79.4	78.5
DFFreq	TIFS2026	87.1	50.0	97.9	64.2	85.1	50.0	99.8	96.4	41.8	94.9	99.8	99.9	52.2	78.4
Ours	–	92.9	51.6	95.8	78.1	93.7	51.6	99.5	83.1	80.3	96.7	95.9	94.5	88.9	**84.8**

Table 2: Average precision (%) comparison on the ForenSynths dataset. Bold and underlined values denote the best and second-best results on the mean metric, respectively.

Method	Venue	Big-GAN	CRN	Cycle-GAN	Deep-fake	Gau-GAN	IMLE	Pro-GAN	SAN	Seeing-Dark	Star-GAN	Style-GAN	Style-GAN2	WFIR	mAP
UniFD	CVPR2023	99.3	96.6	99.8	81.8	100.0	98.6	100.0	78.8	63.8	99.4	97.6	97.9	97.3	<u>93.1</u>
LGrad	CVPR2023	90.8	63.5	94.0	67.9	79.3	69.6	100.0	51.4	59.1	100.0	99.9	99.9	65.9	80.1
NPR	CVPR2024	87.8	50.0	98.5	82.4	85.5	50.0	100.0	71.6	60.9	99.9	99.8	99.9	65.3	80.9
FreqNet	AAAI2024	96.0	74.7	99.6	94.4	98.6	77.9	100.0	80.1	62.4	99.8	99.7	99.5	51.1	87.2
SPAI	CVPR2025	71.9	79.9	70.5	74.5	79.2	70.6	99.4	42.7	96.3	98.4	97.1	98.8	90.8	82.3
CO-SPY	CVPR2025	99.0	78.8	99.7	92.7	99.5	78.7	100.0	86.0	70.1	99.0	99.7	100.0	81.5	91.1
MLEP	NeurIPS2025	93.6	54.7	99.3	93.7	87.9	57.5	100.0	67.0	51.1	100.0	100.0	100.0	87.2	84.0
DFFreq	TIFS2026	95.1	68.3	98.8	71.8	96.7	68.7	100.0	99.6	55.1	99.2	100.0	100.0	46.8	84.6
Ours	–	97.8	74.7	99.4	86.7	99.0	74.8	100.0	89.0	99.4	99.9	98.7	98.8	97.8	**93.5**

bidirectional interaction, global spectral context calibrates local subband responses, while localized dependency evidence refines the global spectral representation.

Adaptive fusion weights $[\alpha_l, \alpha_g]$ are subsequently computed by applying Softmax to the concatenated GAP descriptors of $\widetilde{Y}_l$ and $\widetilde{Y}_g$. The final unified representation is formulated as

$$I_F = \alpha_l P_l(\widetilde{Y}_l) + \alpha_g P_g(\widetilde{Y}_g), \tag{10}$$

where $P_l(\cdot)$ and $P_g(\cdot)$ project the respective calibrated features into a shared dimensional space. BCGF therefore combines bidirectional calibration with input-adaptive weighting, allowing local and global evidence to reinforce each other while emphasizing the representation that provides stronger forensic support for each input.

The aggregated representation I_F is finally fed into a lightweight classifier $f(\cdot)$ for authenticity prediction. The entire framework is optimized end-to-end using binary cross-entropy loss with logits.

4 Experiments

4.1 Setup

Datasets. We follow the protocols established in Wang et al. (2020); Tan et al. (2024a;b) and use four semantic classes (car, cat, chair, and horse) from the ForenSynths dataset Wang et al. (2020) for training. To evaluate the generalization capability of the proposed method across diverse synthesis paradigms and real-world scenarios, we conduct testing on three datasets: AIGCDetectBenchmark Zhong et al. (2023), ForenSynths Wang et al. (2020), and WildRF Cavia et al. (2024). Detailed dataset descriptions are provided in the Supplementary Materials.

Implementation details. The proposed method is implemented in PyTorch and trained from scratch using the Adam optimizer with an initial learning rate of 1×10^{-4} and betas $(\beta_1, \beta_2) = (0.9, 0.999)$. Our model is trained for 30 epochs with a batch size of 128. The learning rate is multiplied by 0.9 every ten epochs. During training, images are randomly cropped to 256×256 and horizontally flipped, while center cropping is adopted for evaluation. All experiments are conducted on a single NVIDIA GeForce RTX 5090 GPU.

Table 3: Accuracy (%) comparison on AIGCDetectBenchmark. Bold and underlined values denote the best and second-best results on the mean metric, respectively.

Method	ADM	DALL-E2	Glide	Mid-journey	VQDM	Big-GAN	Cycle-GAN	Gau-GAN	Pro-GAN	SDXL	SDv 1.4	SDv 1.5	Star-GAN	Style-GAN	Style-GAN2	WFIR	Wukong	mAcc.
UniFD	66.9	50.7	62.5	56.3	85.4	95.1	98.3	99.5	99.8	60.1	63.8	63.5	95.7	84.9	74.9	87.1	71.0	77.4
LGrad	71.1	78.0	66.8	68.4	73.5	83.0	85.4	72.4	99.8	77.0	74.0	74.2	99.6	94.8	96.1	63.5	71.8	79.4
NPR	69.7	64.9	78.3	77.8	78.1	84.4	96.1	82.5	99.8	78.5	78.6	78.9	99.3	97.7	98.4	65.8	76.1	82.6
FreqNet	83.3	55.1	81.6	69.8	81.6	90.5	95.8	93.4	99.6	78.3	64.2	64.9	85.7	90.2	87.9	49.1	57.7	78.2
SPAI	58.3	44.9	45.6	43.3	60.8	68.9	68.1	75.4	88.3	51.5	49.1	49.3	54.8	72.8	73.5	79.5	48.0	60.7
CO-SPY	82.2	87.9	88.7	77.1	86.4	95.3	98.1	98.4	100.0	88.8	72.6	72.6	89.0	92.2	93.8	64.2	67.1	85.6
MLEP	73.6	71.0	80.7	78.3	79.8	87.1	98.3	82.0	99.5	79.4	80.4	80.2	100.0	99.6	99.9	79.4	80.0	85.3
DFFreq	94.2	96.8	95.2	94.7	96.5	87.1	97.9	85.1	99.8	64.2	94.2	94.5	94.9	99.8	99.9	52.2	93.5	<u>90.6</u>
Ours	82.4	96.3	93.1	90.5	91.0	92.9	95.8	93.7	99.5	96.1	87.5	87.7	96.7	95.9	94.5	88.9	84.6	**92.2**

Table 4: Average precision (%) comparison on AIGCDetectBenchmark. Bold and underlined values denote the best and second-best results on the mean metric, respectively.

Method	ADM	DALL-E2	Glide	Mid-journey	VQDM	Big-GAN	Cycle-GAN	Gau-GAN	Pro-GAN	SDXL	SDv 1.4	SDv 1.5	Star-GAN	Style-GAN	Style-GAN2	WFIR	Wukong	mAP
UniFD	87.1	63.0	84.3	74.6	96.7	99.3	99.8	100.0	100.0	88.0	86.6	86.2	99.4	97.6	97.9	97.3	91.3	91.1
LGrad	80.1	95.0	81.6	77.8	81.8	90.8	94.0	79.3	100.0	85.8	80.5	80.4	100.0	99.9	99.9	65.8	77.0	86.4
NPR	74.6	76.7	85.7	85.4	81.2	87.8	98.5	85.5	100.0	87.7	84.0	84.6	99.9	99.8	99.9	65.3	80.5	86.9
FreqNet	91.4	54.5	88.8	78.9	89.6	96.0	99.6	98.6	100.0	84.8	74.3	75.6	99.8	99.7	99.5	51.1	66.9	85.3
SPAI	60.6	42.9	43.2	41.0	64.0	71.9	70.5	79.2	99.4	52.1	50.4	50.5	98.4	97.1	98.8	90.8	48.6	68.2
CO-SPY	93.8	95.7	96.7	91.2	96.4	99.0	99.7	99.5	100.0	97.8	89.8	89.8	99.0	99.7	100.0	81.5	85.5	<u>95.0</u>
MLEP	84.8	85.9	92.2	85.7	85.6	93.6	99.3	87.9	100.0	88.8	89.3	89.7	100.0	100.0	100.0	87.2	89.0	91.7
DFFreq	97.3	99.2	98.7	97.8	98.3	95.1	98.8	96.7	100.0	71.8	97.6	98.0	99.2	100.0	100.0	46.8	96.9	93.6
Ours	91.9	99.1	97.6	95.4	95.2	97.8	99.4	99.0	100.0	99.0	92.8	93.2	99.9	98.7	98.8	97.8	91.1	**96.9**

Compared methods. We compare our method with UniFD Ojha et al. (2023), LGrad Tan et al. (2023), NPR Tan et al. (2024b), FreqNet Tan et al. (2024a), SPAI Karageorgiou et al. (2025), CO-SPY Cheng et al. (2025), MLEP Yuan et al. (2025), and DFFreq Yan et al. (2026). All methods compared in this study were trained or fine-tuned using the same limited CNNDet/ProGAN training set Wang et al. (2020). Following existing works Tan et al. (2024b); Ojha et al. (2023), we report classification accuracy (Acc) and average precision (AP) as evaluation metrics.

4.2 Quantitative Analysis

We compare our method with representative AI-generated image detectors on three benchmarks evaluating cross-generator and real-world generalization.

Evaluation on ForenSynths. As shown in Tables 1 and 2, our method achieves the best mean performance, with 84.8% Acc. and 93.5% AP, outperforming the runner-up in mean accuracy by 1.4 percentage points. Notably, it achieves the best performance on WFIR and the highest AP on SeeingDark, indicating that the learned dependency cues remain effective under challenging real-world and low-light conditions. The overall gains across multiple subsets also suggest that our method does not rely on a single type of generation artifact, but captures more transferable forensic evidence across diverse generators. Similar to the compared methods, however, performance on CRN and IMLE remains relatively limited. One possible reason is that these generators exhibit different dependency characteristics that are more difficult to characterize with the current local–global modeling strategy, leaving room for more fine-grained dependency modeling in future work.

Evaluation on AIGCDetectBenchmark. As shown in Tables 3 and 4, our method achieves the best mean performance, with 92.2% Acc. and 96.9% AP, surpassing the corresponding runner-up results by 1.6 and 1.9 percentage points, respectively. It also obtains the best results on SDXL and WFIR while remaining competitive across diverse GAN- and diffusion-based subsets. Although it does not rank first on every generator, its consistently strong performance demonstrates more balanced cross-family generalization. These results validate the effectiveness of NDR and complementary local–global modeling in capturing transferable dependency disruptions across generators.

Method	Venue	Facebook	Reddit	Twitter	mAcc.
UniFD	CVPR2023	49.1	60.2	55.9	55.1
LGrad	CVPR2023	67.5	63.3	45.6	58.8
NPR	CVPR2024	77.8	65.1	52.3	65.0
FreqNet	AAAI2024	57.8	62.7	51.8	57.4
SPAI	CVPR2025	49.1	53.7	50.5	51.1
CO-SPY	CVPR2025	52.2	65.5	53.4	57.0
MLEP	NeurIPS2025	55.6	74.3	51.4	60.4
DFFreq	TIFS2026	62.5	79.5	55.8	<u>65.9</u>
Ours	–	65.9	70.4	65.0	**67.1**

Table 5: Accuracy (%) comparison on WildRF. Bold and underlined values denote the best and second-best results on mAcc., respectively.



Figure 3: Visualization of original images and corresponding NDR maps on SDXL and Midjourney. For each subset, the left image is real and the right image is AI-generated.

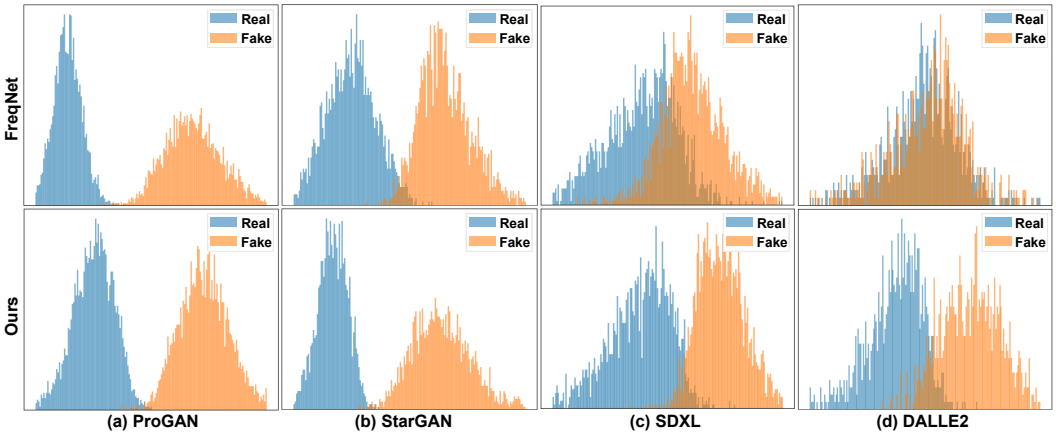

Figure 4: Logit distributions of FreqNet and our method on (a) ProGAN, (b) StarGAN, (c) SDXL, and (d) DALLE2. Blue and orange histograms denote real and fake samples, respectively.

Evaluation on WildRF. As shown in Table 5, our method achieves the highest mean accuracy of 67.1% on real-world social-media images, outperforming the runner-up DFFreq by 1.2 percentage points. It also obtains the highest accuracy of 65.0% on Twitter, while maintaining competitive performance on Facebook and Reddit. These results demonstrate balanced cross-platform generalization under compression, resampling, and platform-specific distribution shifts.

4.3 Qualitative Analysis

We visualize logit distributions and NDR maps to analyze cross-generator discriminability and dependency patterns.

Visualization of NDR Maps. Figure 3 visualizes original images and their corresponding NDR maps on SDXL and Midjourney. Compared with RGB images, NDR reduces dominant semantic responses while retaining local texture-dependency responses. Real images exhibit more continuous and fine-grained residual patterns in natural regions such as sandy ground and snow, whereas the generated train and jellyfish show weaker or less continuous responses in grass and water backgrounds. This contrast indicates that NDR exposes local dependency disruptions associated with over-smoothed or insufficiently detailed synthetic textures.

Visualization of Logit Distributions. Figure 4 compares FreqNet and our method on representative GAN- and diffusion-based generators. Both methods clearly separate real and fake samples on the GAN-based ProGAN and StarGAN subsets. However, when generalized to diffusion-based generators, FreqNet exhibits substantial distribution overlap, particularly on DALLE2, indicating that its learned frequency cues transfer less reliably across generator families. In contrast, our method maintains clearer decision margins on both SDXL and DALLE2. This improvement shows that

Table 6: Component ablation on AIGCDetectBenchmark. "✓" and "×" denote used and removed components, respectively. Bold and underlined values indicate the best and second-best results.

NDR	LSDB	GSDB	BCGF	mAcc. (%)	mAP (%)
✓	✓	✓	×	<u>87.2</u>	93.2
✓	×	✓	×	62.1	73.5
✓	✓	×	×	86.1	93.2
×	✓	✓	✓	84.8	<u>93.4</u>
✓	×	×	×	78.2	89.5
✓	✓	✓	✓	**92.2**	**96.9**

Table 7: Ablation results with different NDR neighborhood sizes on AIGCDetectBenchmark.

Size ($k \times k$)	mAcc. (%)	mAP (%)
3×3	<u>91.5</u>	<u>96.3</u>
5×5	**92.2**	**96.9**
7×7	87.3	94.0
9×9	78.9	93.5

NDR and complementary local–global modeling robustly capture transferable dependency disruptions and produce more generator-agnostic forensic representations. Additional visualizations are provided in the Supplementary Materials.

4.4 Ablation Studies

We conduct ablation experiments on the AIGCDetectBenchmark dataset to evaluate the effectiveness of each component and analyze the influence of the NDR neighborhood size.

Ablation of Components. As shown in Table 6, the full model achieves 92.2% mAcc. and 96.9% mAP. Removing NDR reduces the two metrics by 7.4 and 3.5 points, respectively, demonstrating that suppressing dominant semantic content is important for exposing more transferable dependency irregularities. Removing BCGF causes drops of 5.0 and 3.7 points, confirming the effectiveness of bidirectional calibration and adaptive fusion in exploiting complementary local–global cues. With the same NDR input and without BCGF, LSDB achieves 86.1% mAcc. and 93.2% mAP, substantially outperforming GSDB at 62.1% and 73.5%. This indicates that local subband cues provide the primary discriminative evidence, while global spectral cues offer complementary context that becomes more effective when jointly modeled with local features. NDR alone reaches 78.2% mAcc. and 89.5% mAP, further confirming that neighborhood dependency residuals already contain useful forensic information, while hierarchical dual-spectral modeling further strengthens their discriminability and generalization.

Ablation of NDR Neighborhood Size. As shown in Table 7, the 3×3 neighborhood provides limited contextual support for stable local reconstruction, while excessively large neighborhoods introduce broader structural and semantic information that may dilute subtle dependency irregularities. In particular, performance gradually degrades when the neighborhood size increases beyond 5×5, indicating that overly broad context weakens the locality of NDR. The 5×5 neighborhood achieves the best performance, providing sufficient contextual information while preserving localized dependency cues, and is therefore adopted as the default setting.

5 Conclusion

In this paper, we presented a dependency-oriented hierarchical modeling framework for generalizable AI-generated image detection. We introduced the NDR, which suppresses dominant semantic content while exposing synthesis-induced dependency irregularities through neighborhood contextual reconstruction. Building upon this representation, LSDB and GSDB jointly capture complementary local and global dependency structures, while BCGF performs bidirectional calibration and input-adaptive integration to construct robust forensic representations. Extensive experiments demonstrate superior cross-generator and real-world generalization, validating the effectiveness of hierarchical dependency modeling for learning transferable forensic evidence.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, grammar and spelling correction, and formatting checks. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REFERENCES

Inzamamul Alam, Md Tanvir Islam, and Simon S. Woo. SpecXNet: A dual-domain convolutional network for robust deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11667–11676. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3755707.

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. doi: 10.1109/OJSP.2023.3337714.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world, 2024. URL https://arxiv.org/abs/2406.09398.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer Vision – ECCV 2020*, volume 12371 of *Lecture Notes in Computer Science*, pp. 103–120. Springer, 2020. doi: 10.1007/978-3-030-58574-7_7.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639. PMLR, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13455–13465, 2025.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12830–12839, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7890–7899, 2020.

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 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3247–3258. PMLR, 2020.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18685–18694, 2025.

Qinghui He, Haifeng Zhang, Qiao Qin, Bo Liu, Xiuli Bi, and Bin Xiao. Diversity over uniformity: Rethinking representation in generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40407–40417, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Yingsong Huang, Hui Guo, Jing Huang, Bing Bai, and Qi Xiong. Diffusion epistemic uncertainty with asymmetric learning for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17097–17107, 2025.

Fanli Jin, Feng Lin, Gaojian Wang, Tong Wu, and Zhisheng Yan. DySy-Det: A synergistic framework with dynamic reconstruction-path consistency for AI-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35571–35579, 2026. doi: 10.1609/aaai.v40i42.40868.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18706–18717, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Xingyi Li, Jiahui Zhang, Yiheng Li, Yun Cao, and Wenhao Wang. Spectral tail auxiliary learning for ai-generated image detection. *arXiv preprint arXiv:2605.22751*, 2026.

Shuqiao Liang, Jian Liu, Renzhang Chen, and Quanlong Guan. FerretNet: Efficient synthetic image detection via local pixel dependencies. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. DiverseGRPO: Mitigating mode collapse in image generation via diversity-aware GRPO. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1864–1873, 2026.

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

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

Xinyi Qi, Kai Ye, Chengchun Shi, Ying Yang, Jin Zhu, and Hongyi Zhou. A difference-in-difference approach to detecting AI-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42964–42975, 2026.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision – ECCV 2020*, volume 12357 of *Lecture Notes in Computer Science*, pp. 86–103. Springer, 2020. doi: 10.1007/978-3-030-58610-2_6.

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*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 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. doi: 10.1609/aaai.v38i5.28310.

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.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020.

Xinyuan Wang, Yingxin Lai, Zhiming Luo, and Zhihui Liu. PPM-CLIP: Probabilistic prompt modeling for generalizable AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21316–21325, 2026.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025.

Jiazhen Yan, Ziqiang Li, Fan Wang, Ziwen He, and Zhangjie Fu. Dual frequency branch framework with reconstructed sliding windows attention for AI-generated image detection. *IEEE Transactions on Information Forensics and Security*, 21:2313–2325, 2026. doi: 10.1109/TIFS.2026.3664000.

Wei Ye, Xinan He, and Feng Ding. Decoupling forgery semantics for generalizable deepfake detection. In *Proceedings of the British Machine Vision Conference*. BMVA, 2024.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop, 2015. URL `https://arxiv.org/abs/1506.03365`.

Lin Yuan, Xiaowan Li, Yan Zhang, Jiawei Zhang, Hongbo Li, and Xinbo Gao. MLEP: Multi-granularity local entropy patterns for generalized AI-generated image detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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.

Nan Zhong, Yiran Xu, and Mian Zou. Color matters: Demosaicing-guided color correlation training for generalizable AI-generated image detection. *arXiv preprint arXiv:2601.22778*, 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. doi: 10.1609/aaai.v40i42.40927.

Yue Zhou, Bing Fan, Pradeep K. Atrey, and Feng Ding. Exposing deepfakes using dual-channel network with multi-axis attention and frequency analysis. In *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, pp. 169–174. Association for Computing Machinery, 2023. doi: 10.1145/3577163.3595103.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable AIGC image detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

The appendix is organized as follows:

- Section A provides detailed descriptions of the training and evaluation datasets.
- Section B presents the implementation details, including the training configuration, input processing, model settings, and evaluation protocol.
- Section C provides additional qualitative visualizations of the neighborhood reconstruction and the resulting NDR maps across different generative models.
- Section D presents comparative t-SNE visualizations of MLEP and our method on representative subsets of AIGCDetectBenchmark.

A Detailed Dataset Descriptions

This section describes the composition, source domains, synthesis families, and image resolutions of the datasets used for training and evaluation. These datasets cover conventional CNN-based generators, modern diffusion and text-to-image systems, and real-world images collected from social-media platforms.

Training Split. Following the four-class training protocol adopted in prior studies Tan et al. (2024a;b), we use the car, cat, chair, and horse categories from the ForenSynths training data Wang et al. (2020). The fake images are generated by ProGAN Karras et al. (2018), while the corresponding real images are collected from LSUN Yu et al. (2015). The real and fake subsets are balanced. All fake training images therefore originate from a single generator, while the semantic content covers four object categories.

ForenSynths. ForenSynths was constructed to evaluate cross-model generalization across different synthesis architectures, source datasets, and generation tasks Wang et al. (2020). The original benchmark contains eleven synthesis sources: ProGAN, StyleGAN, BigGAN, CycleGAN, StarGAN, GauGAN, CRN, IMLE, Seeing In The Dark (SITD), SAN, and DeepFake. Our evaluation additionally includes StyleGAN2, which was evaluated by Wang et al. after its release, and WhichFaceIsReal (WFIR), following later evaluation protocols.

ProGAN, StyleGAN, StyleGAN2, and BigGAN represent unconditional or class-conditional image generation. CycleGAN, StarGAN, and GauGAN represent image-to-image translation or semantic image synthesis. CRN and IMLE represent non-adversarial synthesis methods based on perceptual objectives, while SITD and SAN represent low-level image restoration tasks. DeepFake contains synthesized facial regions blended with real image content, whereas WFIR contains high-resolution real and synthetic face images.

The corresponding source domains include LSUN, ImageNet, CelebA, COCO, GTA, FaceForensics++, FFHQ, raw camera images, and standard super-resolution benchmarks. ForenSynths therefore covers diverse architectures, semantic domains, and image synthesis tasks.

AIGCDetectBenchmark. AIGCDetectBenchmark contains seventeen generative sources and extends AI-generated image detection evaluation from conventional GAN-based synthesis to diffusion and commercial text-to-image systems Zhong et al. (2023). Its subsets include ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, WFIR, ADM, GLIDE, Midjourney, Stable Diffusion v1.4, Stable Diffusion v1.5, VQDM, Wukong, DALL-E 2, and SDXL.

The GAN-based images and WFIR are obtained from the CNNSpot collection Wang et al. (2020). The DALL-E 2 and SDXL images were curated by the benchmark authors, while the remaining images were obtained from GenImage Zhu et al. (2023). For some text-to-image generators, including DALL-E 2 and Midjourney, image labels are used as prompts to generate the synthetic images. The number of real and fake images is balanced for each generative source.

The benchmark covers unconditional and conditional GANs, pixel-space diffusion, latent diffusion, vector-quantized diffusion, and commercial text-to-image systems. Its nominal image resolutions range from 256×256 to 1024×1024, introducing variations in generation architecture, semantic domain, image scale, and synthesis pipeline.

WildRF. WildRF is an in-the-wild deepfake detection benchmark collected from Reddit, X (Twitter), and Facebook Cavia et al. (2024). Unlike simulated benchmarks containing images directly

obtained from generative models and standard vision datasets, WildRF contains images distributed through social-media platforms. It therefore captures variations in resolution, aspect ratio, file format, compression, resizing, editing transformations, semantic content, and generation technique.

The authors manually collected real images using keywords and hashtags associated with authentic, non-manipulated content, such as `#photography`, `#nature`, `#nofilter`, and `#streetphotography`. Fake images were collected using terms associated with AI-generated or manipulated visual content, including `#deepfake`, `#AIart`, `#midjourney`, `#stablediffusion`, `#dalle`, and `#aigenerated`. The dataset contains balanced real and fake subsets from all three platforms and reflects the distribution shifts and platform-dependent processing encountered in real-world online environments.

B Experimental Details

Training Configuration. The proposed framework is implemented in PyTorch and trained from scratch on a single NVIDIA GeForce RTX 5090 GPU. We use the Adam optimizer with an initial learning rate of 1×10^{-4} and $(\beta_1, \beta_2) = (0.9, 0.999)$. The model is trained for 30 epochs with a batch size of 128. The learning rate is multiplied by 0.9 after every ten epochs. A fixed random seed of 100 is used for reproducibility.

Input Processing. During training, each image is randomly cropped to 256×256 and randomly horizontally flipped. During evaluation, a 256×256 center crop is used without random flipping. The resulting images are converted to tensors and normalized using channel means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$.

Model Settings. The NDR module adopts a 5×5 neighborhood for contextual reconstruction. The center pixel is excluded from the neighborhood support, and channel-wise median aggregation is used to obtain the reconstructed baseline. The LSDB employs a one-level Haar wavelet decomposition and performs local dependency modeling with 4×4 windows. All model components are jointly optimized end-to-end without ImageNet-pretrained weights.

Optimization and Evaluation. The classifier produces a single authenticity logit and is optimized using binary cross-entropy loss with logits. Real and fake images are assigned labels 0 and 1, respectively. At inference time, the logit is converted into a probability using the sigmoid function, and a threshold of 0.5 is used to compute classification accuracy. Average precision is computed directly from the predicted probabilities.

C Additional NDR Visualizations

Figure 5 presents additional qualitative visualizations on StyleGAN, WFIR, Stable Diffusion v1.4, and GLIDE, covering both GAN-based and diffusion-based generators. For each subset, the upper and lower examples correspond to real and AI-generated images, respectively, while the three columns show the original image, the neighborhood-reconstructed image $M(I)$, and the resulting NDR map I_{NDR}. The neighborhood reconstruction preserves dominant image content while smoothing local pixel variations, and subtracting this contextual baseline suppresses dominant color and semantic responses, thereby highlighting fine-grained texture, boundary, and local dependency variations that are less apparent in the original RGB images.

D Comparative Feature Visualizations

Figure 6 compares the t-SNE feature distributions of MLEP Yuan et al. (2025) and our method on GauGAN, StarGAN, SDXL, and WFIR. Compared with MLEP, our method generally produces more distinguishable real–fake feature distributions, particularly on GauGAN and SDXL, where MLEP exhibits substantial inter-class overlap while our features form more distinct regions. Similar trends are observed on StarGAN and WFIR, indicating that the proposed representation preserves more class-discriminative forensic cues across different synthesis families and semantic domains. Since t-SNE is a nonlinear dimensionality-reduction technique, these visualizations serve as qualitative evidence complementary to the quantitative evaluation.

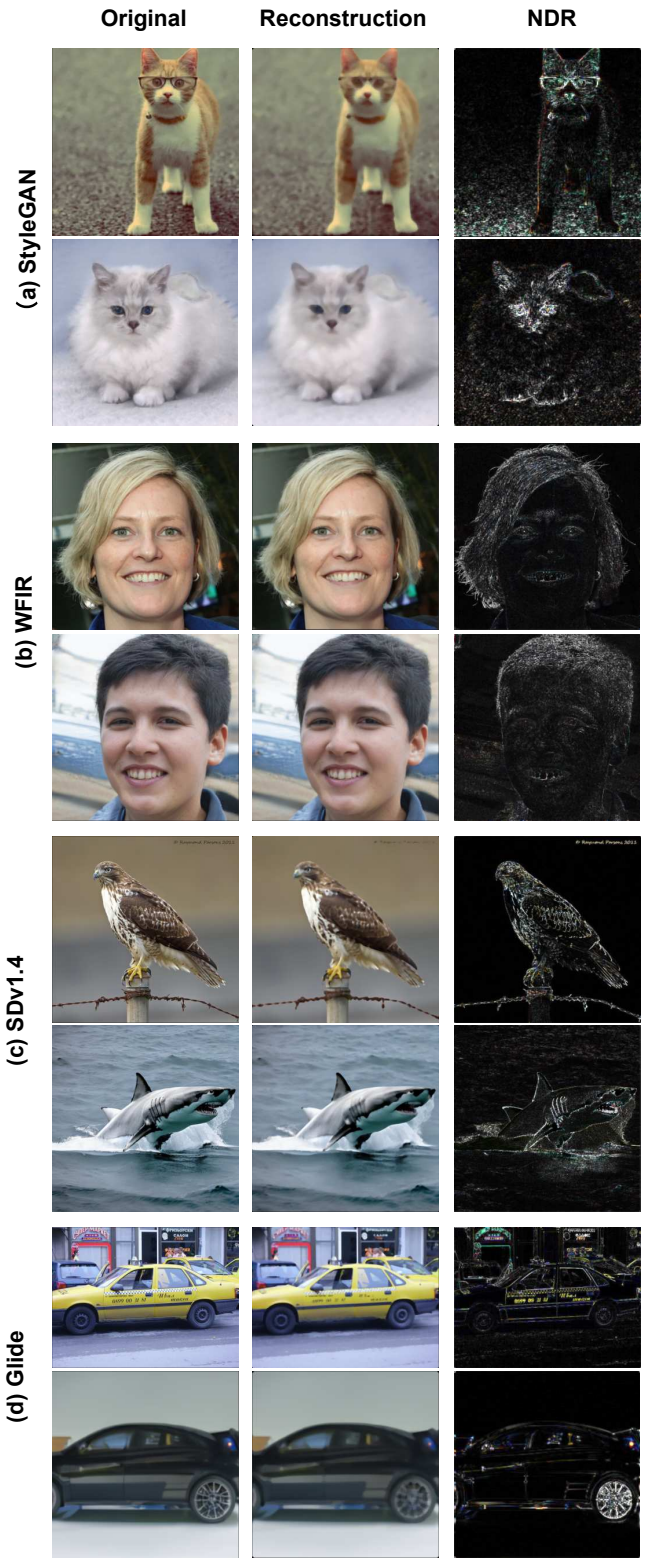

Figure 5: Additional visualizations of original images, neighborhood-reconstructed images, and NDR maps on (a) StyleGAN, (b) WFIR, (c) Stable Diffusion v1.4, and (d) GLIDE. For each subset, the upper and lower examples correspond to a real and an AI-generated image, respectively, while the three columns show the original image, neighborhood reconstruction, and NDR map.

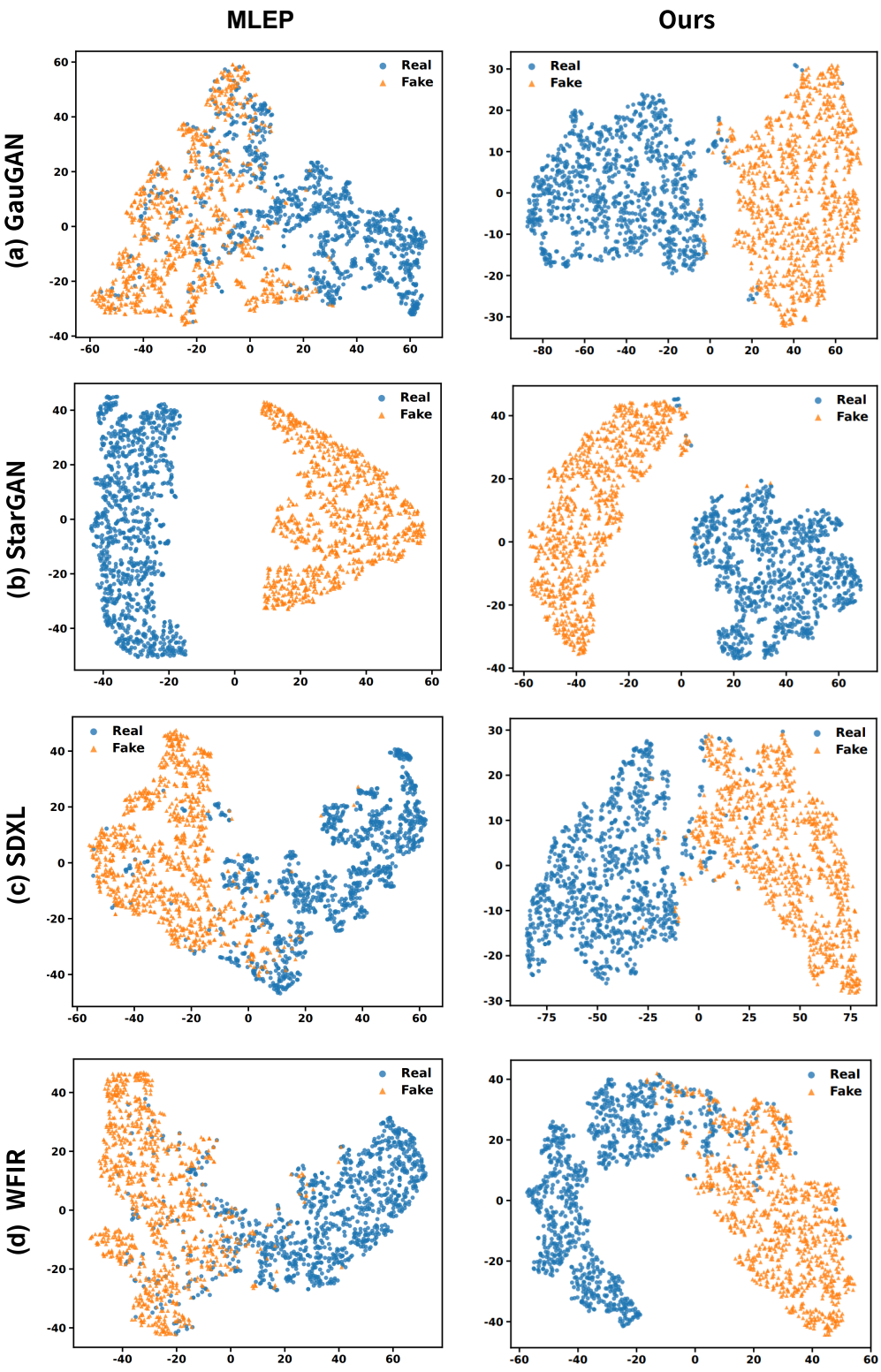

Figure 6: Comparative t-SNE visualizations of MLEP and our method on four subsets of AIGCDetectBenchmark: (a) GauGAN, (b) StarGAN, (c) SDXL, and (d) WFIR. The four rows correspond to the four subsets, while the left and right columns show the feature distributions produced by MLEP and our method, respectively. Blue circles denote real images, while orange triangles denote AI-generated images.