NOVAD: VARIABLE NATIVE OBSERVATION FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated images appear at widely varying resolutions and aspect ratios, yet fine-grained forensic cues are better preserved at native scale. Resizing large images can suppress these cues, motivating detectors to encode large images as local views at native scale. However, exhaustively processing all native regions makes inference cost grow with image area. This raises a natural question: can a detector preserve access to native detail without requiring full spatial coverage? We introduce **NOVAD** (**N**ative **O**bservation with **V**ariable **A**llocation for **D**etection), a detector that learns to operate on subsets of native local views together with a global view. During training, NOVAD is exposed to different numbers and spatial configurations of local views, teaching a shared detector to combine global features with local evidence from changing view subsets. These subsets reuse features encoded once per view, enabling multi-set supervision without repeated view encoding. Geometry-aware fusion combines local detail, whole-image context, and spatial information, allowing one model to handle changing view sets, resolutions, and aspect ratios. At inference, attribution from the global-only prediction guides local tile selection before local encoding, while the inference budget determines how many tiles are processed. The same model operates across inference budgets without retraining. Experiments show that NOVAD retains strong detection performance with only a few native local views. Trained on OpenFake, it achieves **98.44%** mean accuracy across Chameleon, CommunityAI, SocialRF, and WildRF, and **98.05%** unweighted mean accuracy across the eight resolution strata of HiRes-50K.

1 INTRODUCTION

Progress from generative adversarial networks (GANs) (Goodfellow et al., 2014) to latent diffusion models (Rombach et al., 2022) has expanded image synthesis toward photorealistic, high-resolution outputs. At the same time, practical detection has moved beyond controlled generator outputs. Images encountered on social platforms may be selected, edited, resized, and recompressed before circulation, and can arrive at widely varying resolutions and aspect ratios (Yan et al., 2025a; Li et al., 2025b; Mu et al., 2026). Reliable detection must therefore generalize not only across generators, but also across the image sizes and processing conditions encountered in practice.

Early detectors learned synthesis artifacts with CNNs or analyzed frequency domain fingerprints, showing that visually plausible generated images could still carry detectable traces (Wang et al., 2020; Zhang et al., 2019; Frank et al., 2020). Later work emphasized transferable pretrained representations: UnivFD used frozen vision-language features (Ojha et al., 2023), and subsequent studies demonstrated strong detection with simple classifiers on modern visual encoders (Zhou et al., 2026; Cai et al., 2026; Abdullah et al., 2026; Lee et al., 2026). Yet representation quality alone does not eliminate the effects of input preprocessing. Resizing a high-resolution image to a fixed-size input can suppress fine-grained forensic cues, whereas cropping preserves native detail only within the observed region (Cavia et al., 2024; Konstantinidou et al., 2025).

Recent detectors therefore process high-resolution images through native-scale local regions, using global-local aggregation or spectral processing to preserve fine detail (Mu et al., 2026; Karageorgiou et al., 2025). Such approaches address an important question of *how* native evidence should be preserved. A separate question is *how much* native evidence must actually be observed. Encoding every

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.

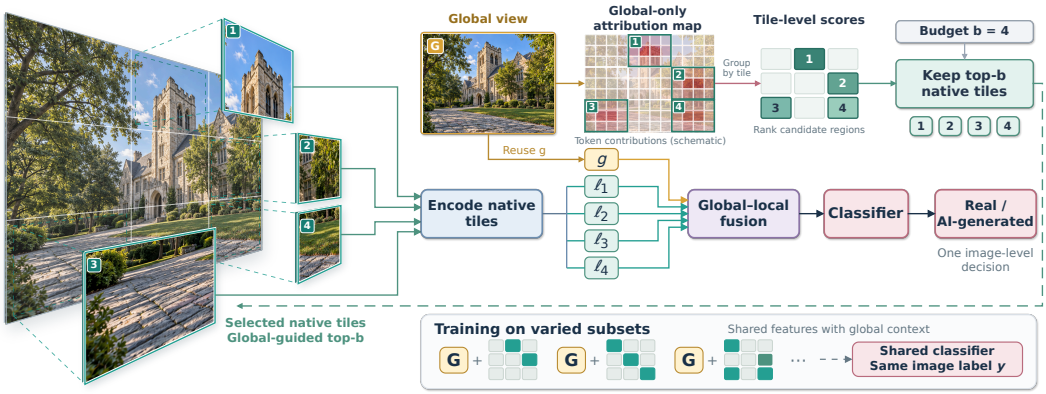

Figure 1: **Variable native observation for AI-generated image detection.** NOVAD combines a global view with a budget-controlled subset of native local views. Training exposes one detector to different numbers and spatial configurations of local views, while inference uses the global-only prediction to prioritize which native regions to observe.

native region maintains full spatial coverage, but makes the cost of a prediction grow with image area. Preserving access to native detail and deciding how much of it to process are therefore distinct problems. This raises a central question: can a detector retain the benefits of native-resolution evidence without requiring exhaustive native-resolution observation?

We present **NOVAD** (**N**ative **O**bservation with **V**ariable **A**llocation for **D**etection), which combines a global view with a variable number of native local views, controlling how much native evidence is observed without reducing its resolution. During training, the same detector is supervised under changing observation conditions, learning to combine global context with different numbers and spatial configurations of native local views. Shared features are reused during partial set supervision, which explicitly trains the detector to operate with incomplete local observations. The fusion module combines spatial information, global image context, and native detail, allowing the model to interpret evidence across varying view sets, image resolutions, and aspect ratios.

At inference, attribution from the global-only prediction guides local tile selection before local encoding, while the observation budget determines how many tiles are processed. The global view thus provides both context for detection and guidance for selecting local evidence. The same trained detector can operate across different observation budgets without retraining.

Trained on OpenFake, NOVAD achieves 98.44% mean accuracy across four in-the-wild benchmarks and 98.05% mean accuracy across the eight resolution strata of HiRes-50K. Budget sweeps show strong performance with few local views, and selection guided by global-only attribution outperforms random sampling. These results show that native-resolution detection need not require exhaustive native observation. Our contributions are as follows:

- **Variable native observation.** We train one detector on partial view sets varying in number and spatial configuration, reusing encoded features across supervised subsets.
- **Geometry-aware global-local fusion.** We combine spatial geometry, global context, and native local evidence in a shared fusion module that accommodates changing view sets, image resolutions, and aspect ratios.
- **Global guided tile selection.** Attribution from the global-only prediction guides local tile selection before local encoding, prioritizing informative native regions under fixed budgets.

2 Related Work

Representation-based detection. Early detectors exploited CNN features and frequency domain artifacts (Wang et al., 2020; Zhang et al., 2019; Frank et al., 2020), followed by methods exploiting neighboring-pixel relationships and spectral representations for improved cross-generator transfer (Tan et al., 2024b;a). UnivFD established the transferability of frozen CLIP features (Ojha et al., 2023), motivating methods that preserve pretrained knowledge while adapting to forensic discrimination. Subsequent methods adapt pretrained vision-language models through prompt learning,

constrained parameter updates, or test-time adaptation (Tan et al., 2025; Yan et al., 2025b; Li et al., 2026). Studies of simple classifiers on modern encoders likewise emphasize representation and readout choices (Zhou et al., 2026; Cai et al., 2026). NOVAD builds on such transferable representations but focuses on structuring and fusing global and native local evidence.

Supervision and data for generalization. Training strategies improve generalization by diversifying supervision and reducing shortcuts. In image forensics, DRCT uses real/reconstructed pairs for contrastive learning (Chen et al., 2024); B-Free combines self-conditioned reconstruction, inpainting, and content augmentation (Guillaro et al., 2025); DDA aligns real/synthetic pairs in pixel and frequency domains, while OMAT constructs on-manifold adversarial examples by optimizing diffusion initial noise (Chen et al., 2025b; Zhou et al., 2025). PPL (Yang et al., 2026) combines randomized patch reconstruction and patchwise contrastive learning to reduce dominant-patch reliance.

Complementing these training strategies, broader datasets support generalization through greater diversity: Community Forensics collects examples from thousands of community-created generators (Park & Owens, 2025), and OpenFake expands generator and provenance coverage (Livernoche et al., 2025). NOVAD complements these data-centric approaches by training the same detector under varying local observations, with both limited-source and diverse training sets.

Native-resolution detection. Resizing can weaken local forensic traces, motivating crop and patch-based processing (Cavia et al., 2024; Konstantinidou et al., 2025). TextureCrop selects texture rich regions, LaDeDa aggregates local descriptors, and AIDE combines high and low frequency patch evidence with semantic information (Konstantinidou et al., 2025; Cavia et al., 2024; Yan et al., 2025a). SPAI aggregates patch level spectral evidence for arbitrary resolution detection (Karageorgiou et al., 2025), while CO-SPY fuses semantic features with VAE-derived artifact features (Cheng et al., 2025). HiDA-Net combines global context with native tiles but relies on broad local coverage at inference (Mu et al., 2026). These methods focus on preserving local detail; NOVAD asks how much native evidence must be observed and where the observation budget should be allocated. It explicitly trains one detector on varying partial native observations. At inference, attribution from the global-only prediction guides native region selection before local encoding, so the observation budget bounds local encoding without requiring full spatial coverage.

3 METHODS

We consider image level binary classification of an input $I \in \mathbb{R}^{H\times W\times 3}$, with $y = 0$ for real images and $y = 1$ for generated images. NOVAD fuses a global view with selected native local views. During training, multiple local view subsets share the same encoded features and supervise a common detector (Figure 2). At inference, global-only attribution selects native regions for encoding under a given budget; their features are then fused with the global view.

3.1 DYNAMIC NATIVE VIEWS

We construct native local views dynamically from the input geometry. After augmentation or an external crop, the input is partitioned into a nonoverlapping grid of rectangular native views $\mathcal{R}(I) = \{R_i\}$. Here, native resolution refers to the pixel scale of this observed input. We choose a compact grid that satisfies a fixed per view patch capacity $T_{\max}$ and an aspect ratio bound, adapting the grid to the input resolution and shape. Each local view is cropped directly from its corresponding region of the input without resizing, then padded at the bottom and right to patch aligned dimensions. At inference, selected local views always retain this native pixel scale and are never resized to meet the observation budget. A global view $G(I)$ preserves the input aspect ratio and is downscaled only when needed to fit the same patch capacity. Section 4.1 specifies $T_{\max}$.

The per view capacity determines how much spatial context can interact within the backbone, whereas the selected set S determines how broadly native detail is observed across the input. The global view is always retained and is excluded from local view counts and native coverage. We define native coverage as $C(I, S) = |\bigcup_{i\in S} R_i|/(HW)$, excluding padding. A shared encoder F yields patch tokens; a within-view pooler ρ maps them to one vector. With $E = \rho \circ F$,

$$g = E(G(I)), \qquad \ell_i = E(I|_{R_i}). \tag{1}$$

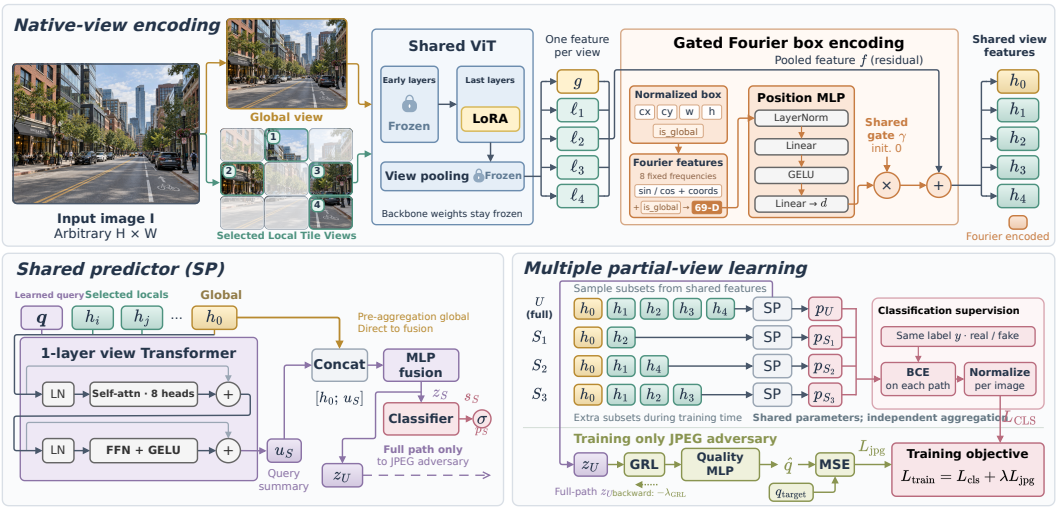

Figure 2: **NOVAD training overview.** A shared ViT encodes global and native local views, and gated Fourier box encodings add geometry to the pooled features. The full retained set and sampled partial sets reuse these features and are independently fused by shared aggregation and prediction modules. Every set retains the global view.

Fusion operates on one pooled vector per view rather than jointly on all input patch tokens. We retain the pre-pooling global patch features $F(G(I))$ for the attribution step in Section 3.4.

3.2 Geometry-aware attention fusion

Because the number and geometry of observed views vary across predictions, fusion must operate on changing view sets while retaining their spatial meaning. We fuse the retained view features jointly and add geometry to each view feature. Pooled appearance features alone do not specify where a view was observed, so we explicitly encode its relative position and extent. Each view has geometry $b_i = (c_x, c_y, w, h, t)$, normalized to the input, with t indicating the global view. For $x = (c_x, c_y, w, h)$ we use

$$\psi(b) = [x; \{\sin(2\pi 2^k x), \cos(2\pi 2^k x)\}_{k=0}^{7}; t], \qquad P(b) = \gamma\,\mathrm{MLP}(\psi(b)). \tag{2}$$

The sinusoidal functions act elementwise on x, yielding a 69-dimensional Fourier representation that passes through LayerNorm and a two-layer GELU MLP. A shared scalar gate γ, initialized to zero, initially suppresses the geometry embedding so that training starts from appearance features alone. The augmented features are $h_0 = g + P(b_0)$ and $h_i = \ell_i + P(b_i)$.

Normalized coordinates avoid position embeddings tied to a specific tile grid and describe relative location and extent as view layouts change. They do not directly encode absolute pixel scale or by themselves imply scale invariance. For externally cropped inputs, all views and coordinates are recomputed relative to the input crop; no unavailable coordinates from the original image are used.

A learned query q and one Transformer self-attention layer A summarize the selected global-local set, followed by direct global concatenation:

$$u_S = [A([q, h_0, \{h_i\}_{i\in S}])]_q, \quad z_S = \phi([h_0; u_S]), \quad p_S = \sigma(w^\top \mathrm{LN}(z_S) + c). \tag{3}$$

Here $s_S = w^\top \mathrm{LN}(z_S) + c$ is the fake class logit and $p_S = \sigma(s_S)$ its probability; ϕ applies LayerNorm, a linear projection, and GELU. Training also uses classifier dropout. The classifier receives global context through both local view attention and direct concatenation, retaining global-local interaction and a direct global pathway. Section 4.3 ablates these pathways.

3.3 Partial set supervision

Accepting a variable length view set does not by itself teach the detector to classify with missing local evidence. A model supervised only on the full training pool may rely on evidence absent at a smaller inference budget. We therefore supervise multiple subsets of the same pool, making partial local evidence an explicit training condition.

We form a training pool U of K local views, retaining the full grid when it fits the training limit. Otherwise, we select spatially dispersed regions from a randomized starting location. We use "full pool" to denote all views retained in U; when the training limit is active, this pool need not cover the full input. The pool bounds encoding cost; its views and the global view are encoded once.

For each image, we supervise the full pool together with up to $J_{\max}$ partial sets of different sizes. Specifically, for $K > 1$, we draw $J_I = \min(J_{\max}, K-1)$ distinct local view counts k_j from $\{1, \ldots, K-1\}$. For each selected count, we uniformly sample $S_j \subset U$ subject to $|S_j| = k_j$. Subsets may overlap; each retains the global view. For $K = 1$, we set $J_I = 0$ and supervise only the full pool prediction. With image label y, the objective is

$$\mathcal{L}_{\text{cls}}(I, y) = \frac{\text{BCE}(p_U, y) + \beta \sum_{j=1}^{J_I} \text{BCE}(p_{S_j}, y)}{1 + \beta J_I}. \tag{4}$$

We average over images, using $J_{\max} = 3$ and $\beta = 1$. Normalization gives each image the same total classification weight regardless of the number of sampled subsets. Each subset is formed *before* fusion and aggregated independently with shared parameters, preventing omitted views from influencing the corresponding prediction while reusing the same encoded features. Extra predictions therefore incur only additional fusion and classification.

These subsets represent different observation conditions for the same deployed classifier, rather than separate predictors. Training also applies compression adversarial regularization to discourage the full pool representation from encoding JPEG quality.

3.4 Budgeted inference with global-guided selection

Global-guided selection. The global-only prediction provides a spatial signal for deciding which native regions to observe. We first apply the shared predictor with $S = \varnothing$, so the attention sequence is $[q, h_0]$, and obtain the fake class logit $s_{\text{fake}} = s_{\varnothing}$. For the global spatial patch features $Z = F(G(I))$, we compute signed gradient-times-activation scores:

$$a_j = \sum_d Z_{jd} \frac{\partial s_{\text{fake}}}{\partial Z_{jd}}. \tag{5}$$

We sum over feature channels and retain the sign: positive attribution supports the fake class logit under this first-order approximation. The spatial map excludes special classification tokens.

Let $\mathcal{C}(I)$ index the native regions available for inference time selection. This candidate pool is distinct from the training pool U: it indexes the full native grid unless the evaluation protocol restricts the candidates before ranking. The local view budget limits the number of selected candidates.

Each candidate $i \in \mathcal{C}(I)$ is projected onto the spatial grid of the global view. Let $\widetilde{R}_i$ denote the projected region and P_j the spatial support of global patch j, clipped to the unpadded image. We assign each candidate an overlap-weighted mean score and select the highest-scoring regions:

$$r_i = \frac{\sum_j |\widetilde{R}_i \cap P_j| \, a_j}{\sum_j |\widetilde{R}_i \cap P_j|}, \qquad S_b = \text{TopK}_{i \in \mathcal{C}(I)}(r_i, b). \tag{6}$$

Here b is the local view budget, excluding the global view; all candidates are retained if $|\mathcal{C}(I)| < b$. We refer to this as the *Gradient: fake* selection rule. It measures contributions to the global image level prediction rather than assigning local fake probabilities, and requires neither test labels nor a separately trained selector. Ranking conventions are given in Appendix A.

Final prediction and selection cost. We encode only the selected native tiles and reuse the global feature for the final prediction p_{S_b}. Attribution requires one backward pass from the global-only logit through the classifier, aggregator, and pooler to the global patch features. These features are treated as the attribution input: gradients are computed with respect to them but are not propagated through the encoder, and no model parameters are updated. The attribution map is computed once per image and reused across inference budgets. Thus, native regions are selected before any local encoding, rather than scored after encoding every candidate tile.

The budget determines how many local views can be observed, while the selection rule determines which ones. A zero local view budget uses the global-only prediction directly. With no local view

Table 1: Acc (%) on four in-the-wild domains, averaged equally over real/fake classes. Published rows follow Cai et al. (2026, Table 2).

Method	Chameleon	CommunityAI	SocialRF	WildRF	Avg
AIDE (Yan et al., 2025a)	57.4	54.2	56.0	58.4	56.50
OMAT (Zhou et al., 2025)	62.9	63.4	60.7	67.4	63.60
DDA (Chen et al., 2025b)	82.4	84.7	83.2	90.4	85.18
SigLIP2-Linear (Tschannen et al., 2025)	85.9	86.6	80.5	79.0	83.00
MetaCLIP2-Linear (Chuang et al., 2025)	93.0	94.0	80.0	72.8	84.95
PE-Linear (Bolya et al., 2025)	<u>95.9</u>	97.1	86.1	83.6	90.68
DINOv3-Linear (Siméoni et al., 2026)	91.4	94.8	<u>94.3</u>	<u>96.1</u>	<u>94.15</u>
PE-SPC (Cai et al., 2026)	94.7	<u>97.2</u>	89.3	94.8	94.00
Ours, SD1.4 training	**96.3**	96.9	**95.4**	**98.8**	**96.84**
Ours, OpenFake training	**98.4**	**99.0**	**97.4**	**99.1**	**98.44**

cap or candidate pool restriction, the model processes the full native grid. One trained detector therefore serves all budgets without retraining or budget-specific models. Section 4.4 reports budget sweeps and compares selection rules.

Complexity. Let $N = |S_b|$ be the number of selected local views and let each view contain T tokens. Since backbone attention is applied independently within each view, its attention cost is $O((N + 1)T^2)$, including the global view. After pooling each view to one token, the fusion Transformer operates on $N + 2$ tokens, with attention cost $O((N + 2)^2)$. The overall attention complexity is therefore $O\big((N + 1)T^2 + (N + 2)^2\big)$. For fixed T, the dominant backbone cost grows linearly with the number of selected views.

4 Experiments

4.1 Implementation details

Datasets. We evaluate NOVAD on established in-the-wild, high-resolution, and cross-generator benchmarks. Chameleon is taken from AIDE (Yan et al., 2025a), CommunityAI and SocialRF from AIGIBench (Li et al., 2025b), and WildRF from LaDeDa (Cavia et al., 2024). We follow the four-domain grouping of Cai et al. (2026). HiRes-50K provides high-resolution evaluation (Mu et al., 2026), while GenImage and DRCT-2M provide cross-generator tests (Zhu et al., 2023; Chen et al., 2024). We treat GenImage SD1.4, the full GenImage training set, OpenFake (Livernoche et al., 2025), and DRCT SD1.4 as distinct training settings. DRCT uses MS COCO as its real image source (Chen et al., 2024; Lin et al., 2014). Appendix B lists the published benchmark compositions.

Model and optimization. The main recipe uses PE-Core-L14-336 (Bolya et al., 2025) with a frozen pooler, rank-8 LoRA on the last four blocks (Hu et al., 2022), $T_{\max} = 576$ patches per view, and one eight-head aggregation layer. Trainable modules use a learning rate 10^{-2} and LoRA uses 10^{-4}, with an effective batch 128 in FP32. Training uses shared resize, blur, and JPEG augmentations.

Main comparison protocol. All main NOVAD comparisons use global-guided gradient selection with up to eight local views in addition to the global view. Budget sweeps vary this cap; ablations follow their table settings. At threshold 0.5, Acc averages real/fake class accuracies (balanced accuracy) on the four domains, following published comparisons; elsewhere it denotes sample accuracy. Bold/underline mark the best/second distinct scores. For method comparisons, each NOVAD variant is ranked against published methods (No VAE excludes +VAE rows); published methods are ranked with the strongest NOVAD variant.

4.2 Comparison with Published Methods

Four In-the-wild benchmarks. Table 1 compares NOVAD with published detectors on four challenging in-the-wild OOD benchmarks: Chameleon, CommunityAI, SocialRF, and WildRF. These datasets contain real-world images collected outside controlled training distributions and therefore provide a demanding test of generalization. NOVAD trained on OpenFake achieves 98.4%, 99.0%, 97.4%, and 99.1% accuracy on the four benchmarks, respectively, for an average of 98.44%. For a training-source-matched comparison, NOVAD trained on SD1.4 achieves a 96.84% average. Among the displayed reference methods, DINOv3-Linear has the highest average accuracy at 94.15%; even taking the best reported reference result independently on each domain yields an average of 95.88%.

Table 2: HiRes-50K Acc (%) by long edge stratum (Mu et al., 2026). Baselines are trained on all-GenImage; Avg is the mean of the eight strata.

Long edge (px)	0–900	900–1200	1200–1500	1500–2000	2000–2500	2500–3000	3000–5000	>5000	Avg
CNNSpot (Wang et al., 2020)	58.46	56.37	58.73	63.14	67.42	60.90	66.30	58.90	61.28
FreDect (Frank et al., 2020)	63.67	55.63	57.31	58.90	58.16	67.78	57.23	51.06	58.72
UnivFD (Ojha et al., 2023)	67.00	58.35	62.20	65.95	66.05	58.20	58.15	60.05	61.99
DIRE (Wang et al., 2023)	59.11	64.25	62.10	66.66	75.84	63.40	69.31	62.18	65.36
AIDE (Yan et al., 2025a)	65.87	57.29	49.90	58.23	51.88	65.31	61.04	42.16	56.46
DRCT (Chen et al., 2024)	65.30	66.19	68.78	68.85	68.78	75.79	69.65	54.03	67.17
TextureCrop (Konstantinidou et al., 2025)	57.45	52.81	59.83	63.17	65.31	55.56	70.65	69.49	61.78
RINE (Koutlis & Papadopoulos, 2024)	65.21	60.30	63.60	63.02	58.60	61.30	61.87	63.77	62.21
SPAI (Karageorgiou et al., 2025)	79.76	76.50	74.01	73.72	71.80	65.50	69.31	64.41	71.88
HiDA-Net (Mu et al., 2026)	82.98	81.39	78.16	78.99	88.26	80.18	82.88	69.84	80.34
Ours	**96.75**	**95.31**	**94.67**	**95.38**	**95.12**	**94.24**	**93.91**	**91.99**	**94.67**
Ours, OpenFake training	**97.93**	**99.11**	**98.84**	**95.95**	**97.81**	**98.20**	**98.84**	**97.72**	**98.05**

Table 3: GenImage Acc (%) after SD1.4 training (Zhu et al., 2023). No VAE and +VAE are separate training variants; Avg is the mean of the eight generators.

Method	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg
Swin-T (Liu et al., 2021)	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot (Wang et al., 2020)	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec (Zhang et al., 2019)	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net (Qian et al., 2020)	50.1	99.9	**99.9**	49.9	50.0	**99.9**	49.9	49.9	68.7
UnivFD (Ojha et al., 2023)	93.9	96.4	96.2	71.9	85.4	94.3	81.6	90.5	88.8
NPR (Tan et al., 2024b)	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
FreqNet (Tan et al., 2024a)	89.6	98.8	98.6	66.8	86.5	97.3	75.8	81.4	86.9
FatFormer (Liu et al., 2024)	92.7	**100.0**	**99.9**	75.9	88.0	**99.9**	**98.8**	55.8	88.9
DRCT (Chen et al., 2024)	91.5	95.0	94.4	79.4	89.1	94.6	90.0	81.6	89.5
AIDE (Yan et al., 2025a)	79.4	99.7	99.8	78.5	91.8	98.7	80.3	66.9	86.9
Effort (Yan et al., 2025b)	82.4	99.8	99.8	78.7	93.3	97.4	91.7	77.6	90.1
SAFE (Li et al., 2025a)	95.3	99.4	99.3	82.1	96.3	98.2	96.3	97.8	95.6
C2P-CLIP (Tan et al., 2025)	88.2	90.9	97.9	**96.4**	99.0	98.8	96.5	**98.7**	95.8
HiDA-Net/No VAE (Mu et al., 2026)	97.8	98.4	98.3	86.2	98.0	98.4	95.6	96.2	96.1
HiDA-Net/+VAE (Mu et al., 2026)	94.2	99.1	99.2	92.7	**99.1**	98.0	96.9	97.8	97.1
Ours/No VAE	**98.1**	99.6	99.5	85.5	96.4	99.6	98.1	97.5	**96.8**
Ours/+VAE	96.0	98.6	98.4	95.1	98.0	98.6	98.4	97.6	**97.6**

Table 4: Acc (%) on sixteen DRCT-2M subsets (Chen et al., 2024). Ctrl and DR denote the source table's ControlNet and diffusion reconstruction groups; DRCT/SDv2 uses SDv2 training.

Method	LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL Refiner	SD Turbo	SDXL Turbo	LCM SDv1.5	LCM SDXL	SDv1 Ctrl	SDv2 Ctrl	SDXL Ctrl	SDv1 DR	SDv2 DR	SDXL DR	Avg
CNNSpot (Wang et al., 2020)	**99.9**	99.9	**99.9**	97.6	66.3	86.6	86.2	72.4	98.3	61.7	98.0	85.9	82.8	60.9	51.4	50.3	81.1
F3Net (Qian et al., 2020)	**99.9**	99.8	99.8	88.7	55.9	87.4	68.3	63.7	97.4	55.0	98.0	72.4	82.0	65.4	50.4	50.3	77.1
Conv-B (Liu et al., 2022)	**99.9**	**100.0**	**99.9**	95.8	64.4	82.0	80.8	60.8	99.2	62.3	99.8	83.4	73.3	61.7	51.8	50.4	79.1
UnivFD (Ojha et al., 2023)	98.3	96.2	96.3	93.8	91.0	93.9	86.4	85.9	90.4	89.0	90.4	81.1	89.1	52.0	51.0	50.5	83.5
DRCT/SDv1.4 (Chen et al., 2024)	**99.9**	99.9	**99.9**	96.3	83.9	85.6	91.9	70.0	99.7	78.8	**99.9**	95.0	81.2	**99.9**	95.4	75.4	90.8
DRCT/SDv2 (Chen et al., 2024)	99.7	98.6	98.5	**99.9**	96.1	98.7	99.6	83.3	98.5	93.8	96.7	**99.9**	97.7	93.9	**99.9**	90.4	96.6
HiDA-Net/No VAE (Mu et al., 2026)	98.7	98.8	98.8	98.7	98.8	98.8	97.7	97.6	98.6	98.8	98.8	98.4	98.2	90.4	74.8	71.2	94.8
HiDA-Net/+VAE (Mu et al., 2026)	98.8	98.9	98.9	98.0	99.0	98.8	98.5	98.8	98.4	98.9	98.5	98.7	98.4	99.0	97.4	**94.9**	98.4
Ours/No VAE	98.9	99.8	99.5	96.8	**99.4**	**99.9**	**99.7**	**99.9**	**99.8**	**99.9**	99.7	99.8	**99.9**	84.2	79.2	66.6	95.2
Ours/+VAE	99.8	99.7	**99.9**	99.5	**99.4**	**99.7**	**99.9**	**99.5**	**99.8**	**99.5**	**99.9**	99.5	**99.8**	99.0	94.7	90.2	**98.8**

Under the matched SD1.4 training source, NOVAD therefore improves the four-domain average by 2.69 percentage points over the strongest single reference method.

High-resolution images. Table 2 evaluates resolution robustness on HiRes-50K across eight long edge strata. Under the matched all-GenImage training setting, NOVAD achieves a 94.67% average accuracy, outperforming the strongest published baseline, HiDA-Net, by over 14%. The advantage becomes even larger at the highest resolutions: for images with long edges above 5000 pixels, NOVAD retains 91.99% accuracy, compared with 69.84% for HiDA-Net. The decrease from the smallest to the largest stratum is 4.76 points under all-GenImage training. With OpenFake training, accuracy ranges from 95.95% to 99.11% across strata, with a 98.05% average; the smallest and largest strata differ by only 0.21 points (97.93% versus 97.72%). These results show that NOVAD preserves strong detection performance at high resolutions, consistent with its use of native local observations without exhaustive high-resolution processing.

GenImage. Table 3 evaluates transfer across all eight GenImage generators after training on SD1.4. NOVAD achieves an average accuracy of 96.8% without VAE augmentation, which increases to 97.6% with VAE augmentation. The largest gain is on ADM, from 85.5% to 95.1%, while BigGAN accuracy changes from 97.5% to 97.6%. VAE augmentation improves the overall average but not every generator. Both variants show strong generalization from SD1.4 to diverse generator families.

Table 5: Partial set supervision: four-domain mean Acc (%). Both models share the OpenFake protocol.

Local cap	1	2	4	8	16
Full pool only (J=0)	97.95	97.86	97.74	97.68	97.51
Partial sets (J=3)	98.41	98.41	98.41	98.44	98.44

Table 6: Global view ablation: mean Acc (%) on four domains. GenImage SD1.4 training.

Local cap	1	2	4	8
No global view	79.81	83.80	89.05	90.80
Global by concatenation only	77.35	83.56	88.71	90.19
Global in attention (ours)	96.59	96.74	96.78	96.84

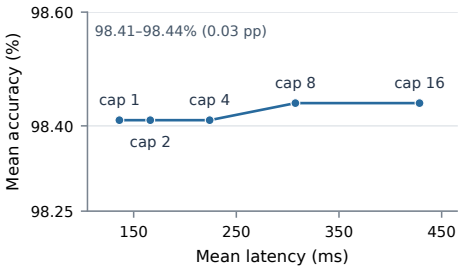

Figure 3: Accuracy-latency tradeoff on four-domain evaluation. OpenFake training; batch-1 end-to-end mean latency.

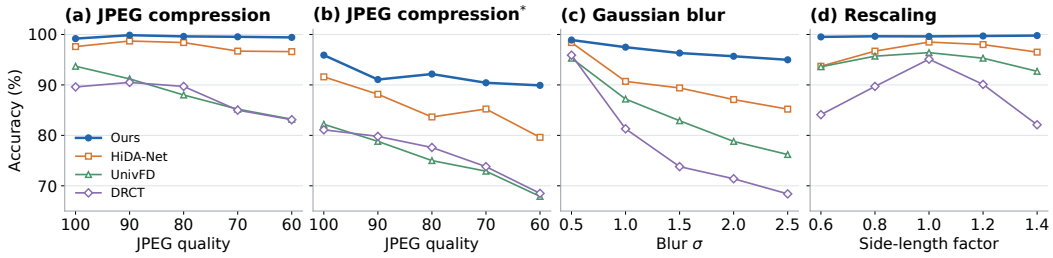

Figure 4: Acc (%) under JPEG recompression, Gaussian blur, and rescaling. Panels (a), (c), and (d) use the SDv1.4 subset; (*) averages over the eight GenImage generators.

DRCT-2M. Table 4 further evaluates transfer on DRCT-2M. Without VAE augmentation, NOVAD achieves between 96.8% and 99.9% accuracy across the ten text-to-image subsets and three ControlNet subsets. The three diffusion reconstruction subsets, SDv1-DR, SDv2-DR, and SDXL-DR, are more challenging, with accuracies ranging from 66.6% to 84.2%. These subsets reconstruct real images using diffusion models (Chen et al., 2024), making them different from standard text-to-image generation. Pairing each real training image with its VAE reconstruction improves accuracy on these subsets to between 90.2% and 99.0%, and raises the average across all sixteen subsets from 95.2% to 98.8%, compared with 98.4% for HiDA-Net with VAE augmentation.

4.3 Robustness and Ablation Study

Partial set supervision and observation budget. Table 5 examines how performance changes with the local view cap. The partial set model achieves a four-domain mean accuracy of 98.41% at caps of one, two, and four, and 98.44% at caps of eight and sixteen. In comparison, the full pool model decreases from 97.95% at cap one to 97.51% at cap sixteen. These results show that partial set supervision maintains stable performance across the evaluated observation budgets.

Global and local branches. Table 6 compares three ways of incorporating global context under the same SD1.4 training protocol. At a local view cap of one, placing the global view in attention achieves 96.59%, compared with 79.81% without a global view and 77.35% when global context is concatenated only after aggregation. At cap eight, the corresponding accuracies are 96.84%, 90.80%, and 90.19%. The attention-based variant changes by only 0.25 percentage points across the four budgets, whereas the model without a global view gains 10.99 points from cap one to cap eight. Each setting uses a fixed checkpoint across budgets. This indicates that direct interaction between global and local evidence is particularly important when only a few native regions are observed.

Auxiliary components. Table 10 in Appendix E compares the full 69-dimensional Fourier box encoding with raw coordinates and with geometry removed entirely, obtaining four-domain accuracies of 98.44%, 98.29%, and 98.06%, respectively.

Robustness to postprocessing. We evaluate the GenImage-trained model without VAE augmentation under JPEG recompression, Gaussian blur, and rescaling. Figure 4 reports SDv1.4 results and JPEG evaluation averaged over all eight GenImage generators. NOVAD outperforms the baselines across the tested perturbation levels and is particularly stable under rescaling.

Table 7: Four-domain Acc (%) for each view selection rule at fixed observation budgets.

Selection rule	Global-only	1 view	2 views	4 views	8 views
Random	95.06	96.31	96.42	96.54	96.51
Global attention	95.06	96.53	96.65	96.70	96.65
Gradient: fake	95.06	**96.59**	**96.74**	**96.78**	**96.84**

Training resolution and external crops. Appendix E evaluates two distinct interventions. Table 11 varies the training area cap while holding source images, random seed, and augmentation fixed. Table 12 compares center and off-center crops at fixed retained area fractions. Cropping and reducing the view budget affect the input differently: cropping removes scene content, whereas a smaller observation budget preserves the full image but processes fewer native regions.

4.4 Accuracy and inference cost

NOVAD maintains nearly unchanged accuracy over the evaluated observation budgets. On the OpenFake four-domain evaluation, the same checkpoint achieves 98.41% mean accuracy at local view caps of one, two, and four, and 98.44% at caps of eight and sixteen (Table 5). Thus, in this setting, most of the available detection performance is already recovered from a small number of native observations. The actual number of processed local views can be lower when an image contains fewer candidate regions than the specified cap.

Figure 3 makes the accuracy–latency tradeoff explicit: cap one achieves 98.41% at 136.10 ms, while cap sixteen achieves 98.44% at 428.45 ms. The accuracy gain is only 0.03 percentage points despite more than tripling latency. Table 13 in Appendix E reports latency, memory, valid patches, actual view counts, and native coverage across the observation budgets. Mean end-to-end latency is 136.10 ms at cap one, 307.54 ms at cap eight, and 1076.32 ms for the full grid. Table 14 provides a separate controlled-input comparison with HiDA-Net (Mu et al., 2026): cap two is faster at input sides of 896 and 1344 pixels, but not at 224 or 448 pixels; full grid NOVAD is slower at all four sizes. The benefit therefore comes from limiting native observations, not from uniformly faster full grid processing. The cap bounds local encoding, while end-to-end latency also includes input preparation and other fixed costs.

Selecting informative views. A small observation budget also makes the choice of local views important. Table 7 isolates this effect using a single SD1.4 checkpoint and changing only the inference time selection rule. The rules rank the same pool of up to 16 candidate tiles at local view caps of one, two, four, and eight. Random selection provides a task-agnostic reference. Global attention instead ranks candidates using attention weights produced by the encoder's frozen pooler on the global view. Our gradient-based rule uses signed gradient-times-activation attribution of the global-only fake logit (Section 3.4) and prioritizes regions that most support the fake prediction. Attribution requires a single backward pass through the pooler, aggregator, and classifier, without differentiating through the encoder, and is computed once per image regardless of the final view budget.

Adding even one native local view with gradient-based selection improves accuracy over the global-only prediction, from 95.06% to 96.59%, indicating that native detail provides complementary evidence. The gradient-based rule has the highest accuracy at every evaluated budget, exceeding random selection by 0.24–0.33 percentage points in this comparison. At cap eight, it reaches 96.84%, compared with 96.51% for random selection and 96.65% for global attention. These modest gains support using the detector's own signal to allocate a limited observation budget.

5 Conclusion

Our results show that native resolution detection does not require exhaustive observation of all native regions. NOVAD combines global context with selected native views using attribution based selection and fusion with geometric information, allowing a single model to operate across different observation budgets without retraining. Trained on OpenFake, NOVAD achieves 98.44% mean accuracy across four in-the-wild domains and 98.05% across the eight resolution strata of HiRes-50K. Selective observation may still miss decisive local evidence, while pooling each view into a single vector limits fine grained interactions across views. Future work should further improve view selection and evidence fusion under fixed budgets.

AI USE STATEMENT

Generative AI assistance was used for manuscript editing, citation and metadata checking, comparison table review, figure review, and compilation checks. The authors are responsible for the experimental results, the protocol decisions, and the verification of all claims.

REPRODUCIBILITY STATEMENT

Section 3 describes view construction, aggregation, the training objective, and inference. Section 4.1 states the evaluation protocols and metrics, and Appendices A and B summarize the experimental configurations and dataset information. Code will be released together with the paper.

REFERENCES

Ahmed Abdullah, Nikolas Ebert, and Oliver Wasenmüller. TAP into the patch tokens: Leveraging vision foundation model features for AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 379–389, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/SAFE/html/Abdullah_TAP_into_the_Patch_Tokens_Leveraging_Vision_Foundation_Model_Features_CVPRW_2026_paper.html.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/57bc0a850255e2041341bf74c7e2b9fa-Abstract-Conference.html.

Weihan Cai, Hao Tan, Zichang Tan, Jun Wan, and Xinping Gao. Unleashing the potential of vision-language models for generalizable ai-generated image detection. *arXiv preprint arXiv:2608.04935*, 2026. URL https://arxiv.org/abs/2608.04935v2. Version 2; preprint.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024. URL https://arxiv.org/abs/2406.09398v1. Version 1; preprint.

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. URL https://proceedings.mlr.press/v235/chen24ay.html.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. AlignGemini: Generalizable AI-generated image detection through task-model alignment. *arXiv preprint arXiv:2512.06746*, 2025a. URL https://arxiv.org/abs/2512.06746v2. Version 2, revised January 2026; preprint.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/991c9f1799cfeebb4217baaacc462d86-Abstract-Conference.html.

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. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Cheng_CO-SPY_Combining_Semantic_and_Pixel_Features_to_Detect_Synthetic_Images_CVPR_2025_paper.pdf.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason E. Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie,

Scott Yih, Shang-Wen Li, and Hu Xu. Meta CLIP 2: A worldwide scaling recipe. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/449fb670956a93b3d4d95167f72093e1-Abstract-Conference.html.

Manni Cui, Ruiqi Liu, Zijian Yu, Hao Tan, Zibo Wei, Zian Wang, Ziheng Qin, Huijia Zhu, Weiqiang Wang, Jun Lan, and Shu Wu. Learning continuous source responses for generalizable AI-generated image detection. *arXiv preprint arXiv:2609.14316*, 2026. URL https://arxiv.org/abs/2609.14316v1. Version 1; preprint.

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. URL https://proceedings.mlr.press/v119/frank20a.html.

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, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

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 (CVPR)*, pp. 18685–18694, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Guillaro_A_Bias-Free_Training_Paradigm_for_More_General_AI-generated_Image_Detection_CVPR_2025_paper.html.

Lawrence Han. Attention to detail: Global-local attention for high-resolution AI-generated image detection. *arXiv preprint arXiv:2601.00141*, 2026. URL https://arxiv.org/abs/2601.00141v1. Version 1; preprint.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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 (CVPR)*, pp. 18706–18717, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Karageorgiou_Any-Resolution_AI-Generated_Image_Detection_by_Spectral_Learning_CVPR_2025_paper.html.

Despina Konstantinidou, Christos Koutlis, and Symeon Papadopoulos. TextureCrop: Enhancing synthetic image detection through texture-based cropping. In *Proceedings of the Winter Conference on Applications of Computer Vision Workshops*, pp. 1459–1468, 2025. URL https://openaccess.thecvf.com/content/WACV2025W/SynRDinBAS/html/Konstantinidou_TextureCrop_Enhancing_Synthetic_Image_Detection_through_Texture-based_Cropping_WACVW_2025_paper.html.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pp. 394–411. Springer, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/9160_ECCV_2024_paper.php.

Seunghyun Lee, Byoungkwon Kim, Jaehyun Nam, Kyungmin Lee, and Jinwoo Shin. SSAFE: Simple and strong AI-generated image detection via frozen vision encoders. *arXiv preprint arXiv:2606.08634*, 2026. URL https://arxiv.org/abs/2606.08634v1. Version 1; preprint.

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. ACM, 2025a. doi: 10.1145/3690624.3709392. URL https://dl.acm.org/doi/10.1145/3690624.3709392.

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*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Li_Towards_Generalizable_AI-Generated_Image_Detection_via_Image-Adaptive_Prompt_Learning_CVPR_2026_paper.pdf.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fb693c67f61e5321746ffce8b6fdd2d0-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Forgery-aware_Adaptive_Transformer_for_Generalizable_Synthetic_Image_Detection_CVPR_2024_paper.html.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

Victor Livernoche, Akshatha Arodi, Andreea Musulan, Zachary Yang, Adam Salvail, Gaétan Marceau Caron, Jean-François Godbout, and Reihaneh Rabbany. OpenFake: An open dataset and platform toward real-world deepfake detection. *arXiv preprint arXiv:2509.09495*, 2025. URL https://arxiv.org/abs/2509.09495v2. Version 2; preprint.

Lianrui Mu, Haoji Hu, Xingze Zou, Jianhong Bai, and Jiaqi Hu. No pixel left behind: A detail-preserving architecture for robust high-resolution AI-generated image detection. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/820e42f39773c6cbbd875553db45658f-Abstract-Conference.html.

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*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.html.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Park_Community_Forensics_Using_Thousands_of_Generators_to_Train_Fake_Image_CVPR_2025_paper.html.

Shengbo Qi, Hongyi Fang, Benjia Zhou, and Rui Mao. PatchHead: Learning spatial patch evidence for generalizable AI-generated image detection. *arXiv preprint arXiv:2608.09223*, 2026. URL https://arxiv.org/abs/2608.09223v1. Version 1; preprint.

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. doi: 10.1007/978-3-030-58610-2_6. URL https://link.springer.com/chapter/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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

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. URL https://ojs.aaai.org/index.php/AAAI/article/view/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 (CVPR)*, pp. 28130–28139, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tan_Rethinking_the_Up-Sampling_Operations_in_CNN-based_Generative_Network_for_Generalizable_CVPR_2024_paper.html.

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. doi: 10.1609/aaai.v39i7.32772. URL https://ojs.aaai.org/index.php/AAAI/article/view/32772.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786v1. Version 1; preprint.

Kutub Uddin, Nusrat Tasnim, and Khalid Malik. PE-Mamba: Bidirectional selective layer aggregation for AI-generated image detection. *arXiv preprint arXiv:2608.07999*, 2026. URL https://arxiv.org/abs/2608.07999v1. Version 1; preprint.

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 (CVPR)*, pp. 8695–8704, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_CNN-Generated_Images_Are_Surprisingly_Easy_to_Spot..._for_Now_CVPR_2020_paper.html.

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. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_DIRE_for_Diffusion-Generated_Image_Detection_ICCV_2023_paper.html.

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*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b0303773962ea1b5394c3a83cc7dd066-Abstract-Conference.html.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 2025b. URL https://proceedings.mlr.press/v267/yan25b.html.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, and Xi Li. All patches matter, more patches better: Enhance AI-generated image detection via panoptic patch learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c5ee2a08fbe743b171b0b4b2bdfd6f86-Abstract-Conference.html.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2019. doi: 10.1109/WIFS47025.2019.9035107. URL https://ieeexplore.ieee.org/document/9035107.

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. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2c2be72635379ec8600eea13df4743e3-Abstract-Conference.html.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, and Bin Li. Simplicity prevails: The emergence of generalizable AIGI detection in visual foundation models. *arXiv preprint arXiv:2602.01738*, 2026. URL https://arxiv.org/abs/2602.01738v2. Version 2; preprint.

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, pp. 77771–77782, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html.

Table 8: Configurations and provenance of the reported measurements.

Item	Configuration or protocol
Four-domain SD1.4 results	Table 1: 96.3 / 96.9 / 95.4 / 98.8; reported four-domain mean 96.84%.
Global routing	Table 6: the same GenImage SD1.4 training protocol without VAE; each checkpoint is fixed across local view budgets.
Main recipe	PE-Core-L14-336, a frozen pooler, rank-8 LoRA on the last four encoder blocks, $T_{\max} = 576$, and one eight-head aggregation layer (Section 4.1).
Partial set supervision	Shared encoded features; independent fusion of the full pool and strict subsets, each retaining the global view. Classification losses are normalized per image, with $J_{\max} = 3$ and $\beta = 1$.
Training area cap	Table 11: image identities, seed, and augmentation are fixed across arms.
OpenFake training	A source pool of 50,000 images, with 20,000 images resampled per epoch; this is the training pool, not the size of the full OpenFake dataset.
Latency protocols	Table 13: end-to-end timing. Table 14: prepared-input detector timing. Appendix E specifies their different scopes.
Result provenance	Manifests, split hashes, checkpoint selection rules, and runnable configurations accompany the code release.

A Evidence provenance and configuration

Table 8 summarizes the reported configurations and protocols. Section 4.1 specifies the main training recipe; the ablation tables identify their corresponding training settings.

Augmentation and trainable modules. Pair-shared augmentation applies resizing, blur, and JPEG compression in that order. Virtual real/fake pairs share transformation plans rather than semantic content. In the main recipe, gradients from all supervised view sets update the shared fusion and classification modules and the rank-8 LoRA adapters on the encoder's last four blocks (Hu et al., 2022); the pretrained encoder weights and within-view pooler remain frozen.

Observation sets. The full training pool U contains the local views retained for an image, not necessarily its entire native grid. The global view is present in the full pool prediction and every supervised partial set. At inference, the local view cap excludes the global view and bounds the number of selected native tiles; actual view counts can be smaller when the image has fewer candidates. Local tiles are cropped without resizing, as described in Section 3.1.

Attribution and ranking conventions. Gradient-times-activation scores retain their signs without rectification or additional normalization. Special classification tokens are excluded from the spatial map, and patch supports are clipped to the unpadded global view. Candidates with zero overlap receive a score of zero; ties are resolved by candidate index. Selection precedes local encoding, and the global feature is reused for the final prediction.

Table 9: Published benchmark inventories, not per-run evaluation counts.

Benchmark	Real	Fake	Total	Source
Chameleon	14,863	11,170	26,033	AIDE (Yan et al., 2025a)
CommunityAI	6,000	6,000	12,000	AIGIBench (Li et al., 2025b)
SocialRF	3,000	3,000	6,000	AIGIBench (Li et al., 2025b)
WildRF	1,250	1,250	2,500	LaDeDa (Cavia et al., 2024)
HiRes-50K	25,284	25,284	50,568	HiDA-Net (Mu et al., 2026)

Table 10: Four-domain mean Acc (%) for each geometry encoding.

Geometry variant	Acc (%)
Fourier box encoding (ours)	**98.44**
Raw box coordinates	<u>98.29</u>
No geometry	98.06

B Dataset versions

Table 9 reports benchmark compositions from the original papers, rather than per-run evaluation manifests. For HiRes-50K, the counts of 25,284 real and 25,284 fake images follow from the reported total of 50,568 images and the 1:1 real/fake balance (Mu et al., 2026, Section 5 and Appendix C.2). These published inventory counts should not be interpreted as the class counts of every reported evaluation run.

GenImage and DRCT manifests are separate. DRCT plain training uses MS COCO real images (Chen et al., 2024; Lin et al., 2014) with SD1.4-generated images, distinct from reconstruction training. The four-domain mean covers Chameleon, CommunityAI, SocialRF, and WildRF; HiRes-50K is reported separately as the unweighted mean over its eight resolution strata.

C Sources of published comparisons

The published reference rows in Table 1 follow PE-SPC (Cai et al., 2026, Table 2), consistent with the table caption. The baseline entries in Tables 2, 3, and 4 were imported from the corresponding comparison tables in HiDA-Net (Mu et al., 2026); original methods remain cited in each row. DRCT (Chen et al., 2024) is the source of the DRCT-2M benchmark, not the sole source of all later baseline results. Training variants and averaging conventions follow the individual table captions; these published entries are not reruns under the NOVAD training recipe.

D Additional published comparisons

Additional published work evaluates related settings. IAPL reports 96.7% mean Acc over eight GenImage sources after SD1.4 training, with test-time token adaptation (Li et al., 2026, Table 3). AlignGemini evaluates Chameleon and WildRF (Chen et al., 2025a, Table 1). GLASS reports 97.533% mean Acc over five ViT-B/16 runs on its mixed-source test split (Han, 2026, Table 5). IAPL and AlignGemini share benchmarks with Section 4.2, while GLASS uses its own split. Their reported scores therefore provide context rather than an additional protocol-matched ranking.

Concurrent work improves detection along other axes: Panoptic Patch Learning supervises patch tokens, PatchHead reads spatial patch tokens, PE-Mamba aggregates across encoder layers, SSAFE pairs frozen encoders with data curation, and CuRe supervises continuous real/generated mixtures (Yang et al., 2026; Qi et al., 2026; Uddin et al., 2026; Lee et al., 2026; Cui et al., 2026). NOVAD focuses on a different question: training one detector under partial native observation and selecting native regions before local encoding.

Table 11: Four-domain mean Acc (%) by post-resize training area cap.

Train cap	Cap 1	Cap 4	Cap 16	Full
448^2	98.21	97.94	97.14	96.72
720^2	**98.30**	**98.22**	98.02	97.80
1024^2	98.11	98.10	**98.11**	**98.09**
1536^2	98.24	98.21	97.99	97.83
None	97.84	97.95	98.04	97.97

Table 12: Four-domain mean Acc (%) under external crops at fixed retained area fractions.

Retained area	100%	75%	50%	25%
Center	**98.44**	**98.31**	**97.94**	**96.30**
Off-center	**98.44**	98.28	**97.94**	95.92

Table 13: Inference cost across local view budgets on the four-domain evaluation. Latency is batch-1 end-to-end; coverage excludes the global view.

Detector	Local views	Coverage (%)	Local valid patches	Latency mean (ms)	Latency p95 (ms)	Peak mem (GB)
Ours, cap 1	1.00	14.25	499.91	136.10	359.77	1.41
Ours, cap 2	2.00	28.42	999.58	166.18	396.38	1.44
Ours, cap 4	3.83	48.15	1924.48	224.03	462.97	1.47
Ours, cap 8	6.52	64.77	3346.53	307.54	599.62	1.47
Ours, cap 16	10.25	79.49	5365.05	428.45	878.22	1.47
Ours, full grid	29.05	100.00	16053.50	1076.32	5184.15	1.47

Table 14: Batch-1 detector latency (ms) on an RTX 3090 at controlled input sizes. Source decoding and resizing are excluded.

Input side (px)	224	448	896	1344
HiDA-Net, full grid	29.55	68.65	209.31	447.00
Ours, cap 2	36.46	76.87	108.36	124.20
Ours, full grid	36.85	117.05	316.99	609.93

E GEOMETRY, TRAINING RESOLUTION, CROPPING, AND INFERENCE COST

Geometry encoding. Table 10 compares the 69-dimensional Fourier box encoding, raw coordinates, and no geometry, with four-domain mean accuracies of 98.44%, 98.29%, and 98.06%, respectively.

Training resolution and external crops. Table 11 varies training area caps across rows, holding source images, seed, and augmentation fixed. Columns specify inference time local view caps; Full uses the full native grid. Neither cap is the per view patch capacity $T_{\max}$ (Section 3.1). Table 12 evaluates center and off-center crops at fixed retained area fractions, recomputing the global view and boxes from each crop. Retained area describes the external crop, not selected tile coverage. Cropping removes scene content; reducing the local view budget preserves the input but processes fewer native regions.

End-to-end inference cost. Table 13 reports batch-1 mean and p95 latency, peak memory, local valid patches, actual local view counts, and native coverage for one checkpoint across budgets. Timing includes decoding, view construction, and detector processing. Counts, patches, and coverage exclude the global view; latency and memory cover the complete detector. A cap bounds, rather than fixes, the local view count. Coverage excludes padding (Section 3.1).

Controlled-input latency. Table 14 compares NOVAD and HiDA-Net (Mu et al., 2026) on an RTX 3090 in FP32, with TF32 disabled. Each entry uses one decoded image resized to the target dimensions, two warm-up passes, and five timed passes. Timing covers detector execution, excluding source decoding and resizing. Cap two is faster than HiDA-Net at 896 and 1344 pixels, but not at 224 or 448 pixels; full grid NOVAD is slower at all four sizes. The two cost tables thus have different timing scopes.