FoRM: Preserving First-Order Real-Manifold Geometry for Generalizable AI-Generated Image Detection

Anonymous authors
Paper under double-blind review

Abstract

As generative models continue to evolve rapidly, generalization to unseen generators has become a central challenge in generalizable AI-generated image detection. Recent reconstruction-based, real-centric envelope learning methods construct near-real fake samples to characterize the outer boundary of the real manifold, offering a promising approach to cross-generator generalization. However, these methods do not sufficiently guarantee a key prerequisite of envelope learning—the stability of the real manifold enclosed by the envelope. Once the real manifold becomes distorted during training, even a well-characterized outer boundary cannot compensate for the resulting discriminative bias caused by the degradation of its internal structure. To address this limitation, we propose FoRM, a first-order real-manifold preservation framework. Specifically, FoRM performs pointwise feature anchoring against a frozen visual foundation model to establish zeroth-order consistency in real representations. Nevertheless, pointwise consistency alone is insufficient to determine the local shape and geometric directions of sample neighborhoods. We therefore introduce Local Directional Jacobian Alignment (LJ), which matches the local responses and their magnitudes between the teacher and detector along natural image variation directions. LJ extends pointwise consistency to local first-order geometric consistency, thereby mitigating local geometric contraction and distortion of the real manifold. Furthermore, FoRM introduces distributionally diverse real samples to prevent real-side optimization from persistently concentrating on homogeneous local regions, enhancing the structural stability of the real manifold from a data-centric perspective. It also employs heterogeneous reconstruction mechanisms to construct diverse near-real fake samples to strengthen outer-boundary supervision. By jointly modeling the internal geometry of the real manifold and its surrounding discriminative envelope, FoRM achieves competitive cross-generator generalization performance across 12 benchmarks.

1 Introduction

As generative models (Sohl-Dickstein et al., 2015; Goodfellow et al., 2014) produce increasingly realistic visual content, reliable AI-generated image detection is becoming increasingly important for maintaining the authenticity and trustworthiness of digital content. However, generative architectures (Brock et al., 2019; Ho et al., 2020; Karras et al., 2018; 2019; Rombach et al., 2022) continue to evolve rapidly, making it impossible to cover all future generation methods during detector training. Therefore, enabling detectors to generalize to unseen generators, namely achieving cross-generator generalization, has become a central challenge in generalizable AI-generated image detection (Shen et al., 2025; Cui et al., 2026; Corvi et al., 2022).

Traditional detection methods (Wang et al., 2020; Frank et al., 2020; Liu et al., 2020; Tan et al., 2023b; Yao et al., 2027) typically extract frequency and texture cues from a limited set of fake training images, making them prone to shortcut learning based on incidental factors such as semantics or overfitting to artifacts specific to known generators, thereby limiting generalization to unseen fake images. To alleviate this issue, recent methods (Wang et al., 2023; Chen et al., 2024) have redesigned supervision at the data level, gradually establishing a real-centric envelope learning paradigm. Using

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.

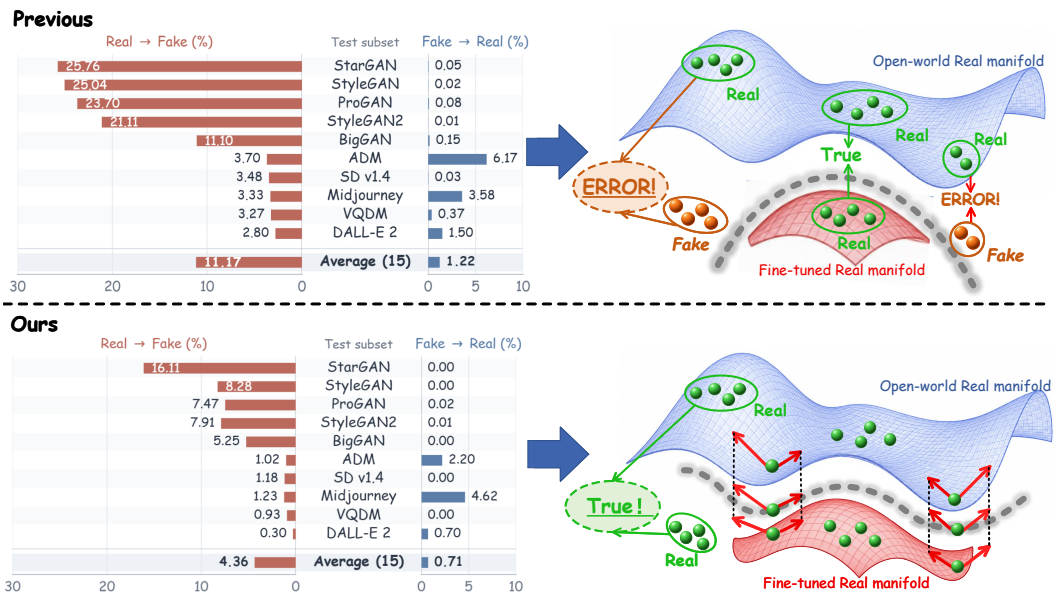


Figure 1: Motivation and core idea of FoRM. The top and bottom rows correspond to DDA and FoRM, respectively. Left: Misclassification rates for real and generated images. FoRM substantially reduces the misclassification of real images as generated. Right: Illustration of how distortion of the real manifold causes out-of-domain real samples to fall outside the discriminative envelope.

real images as a reference, they construct near-real fake samples through reconstruction (Kingma & Welling, 2022; Wang et al., 2023) or inversion (Lin et al., 2026) that closely match the originals in content and semantics while introducing small deviations near the real distribution. These hard negatives provide outer-boundary supervision, encouraging the model to learn a compact discriminative envelope enclosing the real manifold and detect generated images based on whether the input deviates from the learned real region, thereby mitigating overfitting to specific generative traces and shortcut learning. Following this line of research, existing studies (Chen et al., 2025a; Liu et al., 2025; 2026) primarily focus on constructing near-real fake samples that lie closer to real images and exhibit more diverse directions of deviation, improving the characterization of the boundary around the real manifold and forming a more comprehensive discriminative envelope.

However, existing envelope learning relies on an important prerequisite: as increasingly effective near-real fake samples are used to characterize the outer envelope boundary, the real manifold serving as the internal reference must remain well structured. We further ask whether this prerequisite always holds. The answer is no. As shown in the upper-left panel of Fig. 1, experimental results for DDA, a state-of-the-art method within the existing envelope learning paradigm, show that fake accuracy is already close to saturation, reaching approximately 99%, whereas the main loss in overall performance comes from substantially lower real accuracy. A similar pattern emerges during training: fake accuracy saturates early, while real accuracy continues to fluctuate and deteriorate. These observations indicate that sufficient outer-boundary supervision does not necessarily imply stable real-side representations. Detector fine-tuning may cause local contraction or distortion of the real manifold, leaving some out-of-domain real images outside the discriminative envelope and causing them to be misclassified, as illustrated in the upper-right panel of Fig. 1. Therefore, when the real manifold serving as the reference for the envelope becomes distorted, even sufficient coverage of the surrounding region by the fake-side boundary cannot compensate for the discriminative bias caused by the degradation of the internal real manifold structure.

This leads to a further question: once the fake-side boundary has been sufficiently characterized, how can a well-structured and stable real manifold be maintained during envelope learning? The real-world distribution of real images cannot be exhaustively represented by a finite set of training samples, whereas large-scale pretrained visual foundation models have learned more generalizable real representations from diverse natural images. A natural approach is therefore to

freeze such a model and use it as a reference teacher for the real manifold, employing feature distillation to encourage the detector to inherit its real representation structure while adapting to the forgery detection task. However, existing distillation methods typically focus on consistency in output probabilities or sample-level features (Hinton et al., 2015), and such pointwise constraints are insufficient for manifold preservation. From the perspectives of function approximation and local differential geometry, matching two representation mappings at sampled points merely requires them to have similar function values at these discrete locations, establishing zeroth-order consistency. Yet, constraints on pointwise values do not uniquely determine the local behavior of the mappings in the neighborhoods of these samples. Even if two representation mappings coincide exactly at the same anchor points, they may still exhibit different first-order differential structures around these points, yielding different directions of local geometric variation and, consequently, different local manifold geometries. As illustrated in the left panel of Fig. 2, with pointwise anchoring alone, the fine-tuned model may pass through the same feature anchors as the frozen teacher while extending in directions that deviate substantially from the teacher manifold in their neighborhoods. Therefore, constraining feature values alone is insufficient to preserve a stable local structure of the real manifold. This issue is particularly pronounced when real training images come from a single source and follow a relatively homogeneous distribution. Homogeneous real samples repeatedly concentrate anchoring supervision on a few high-density regions of the real manifold, gradually biasing the detector's task adaptation toward common patterns in the training distribution. This may reduce the broader real representations learned during pretraining to a local real manifold closely tied to the current training source, thereby exacerbating real-side shortcut learning (Geirhos et al., 2020).

To address these issues, we propose FoRM, a first-order real-manifold preservation framework for generalizable AI-generated image detection. Specifically, we first use a frozen pretrained visual foundation model to perform pointwise feature anchoring for the fine-tuned detector exclusively on real images, preserving the zeroth-order positions of real representations. Building on this, we introduce Local Directional Jacobian Alignment (LJ), which aligns the Jacobian responses of the teacher and detector along natural directions of variation in real images. This further constrains their local differential behavior in the neighborhoods of real samples, extending zeroth-order feature consistency at discrete points to first-order local geometric consistency of the real manifold, thereby mitigating local contraction and structural distortion during detector fine-tuning. Meanwhile, we introduce distributionally heterogeneous real samples to break the persistent reinforcement of local patterns associated with a single training source, alleviating the tendency of real representations to contract toward high-density regions of the training distribution. We further retain the advantages of near-real envelope learning by constructing near-real fake samples through diverse reconstruction mechanisms, enriching the characterization of the boundary surrounding the real manifold. In this way, FoRM jointly models the internal real manifold and the surrounding discriminative envelope, establishing a more reliable real-centric detection framework. Extensive experiments demonstrate that FoRM achieves competitive cross-generator generalization performance across 12 large-scale benchmarks.

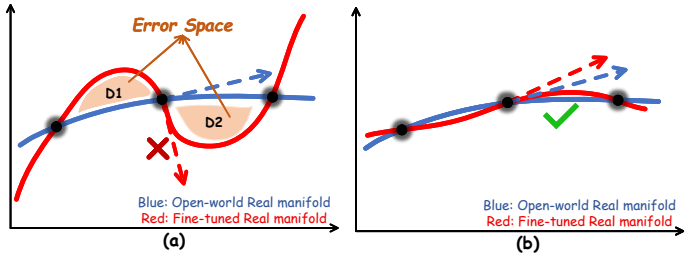


Figure 2: Illustration of pointwise anchoring and first-order local geometry preservation. (a) Enforcing feature consistency only at anchor points may still allow local geometric directions to deviate within their neighborhoods. (b) FoRM further constrains the directions of local geometric variation of the teacher and student around anchor points through Local Directional Jacobian Alignment. Black dots denote shared anchor points, and dashed arrows indicate local geometric directions.

2 Motivation

Challenges in Preserving Local Manifold Geometry. Pointwise feature anchoring can enforce consistency between teacher and student representations on a finite set of real samples, but it is in-

sufficient to determine the local geometry in their neighborhoods (Romero et al., 2015). To illustrate this limitation, we consider a local planar curve described by a differentiable function $y = f(x)$ and construct the following counterexample. Given a teacher function f_T and distinct anchor points $\{x_i\}_{i=1}^n$, we construct the student function as:

$$f_S(x) = f_T(x) + c\prod_{i=1}^{n}(x - x_i), \qquad c \neq 0. \tag{1}$$

At each anchor point, we have

$$f_S(x_i) = f_T(x_i), \qquad i = 1, \ldots, n. \tag{2}$$

However, the difference between their derivatives at each anchor point is

$$f'_S(x_i) - f'_T(x_i) = c\prod_{\substack{j=1 \\ j\neq i}}^{n}(x_i - x_j) \neq 0. \tag{3}$$

Therefore, even when the two mappings coincide exactly at all anchor points, their local tangent directions can still differ. For high-dimensional representation mappings, feature consistency likewise does not imply consistency of the corresponding Jacobians. Thus, zeroth-order pointwise constraints are insufficient to guarantee first-order local geometric preservation. Additional constraints on responses along multiple local geometric directions are needed to constrain local geometric deviations of real representations during detector fine-tuning.

Challenges of Real-Data Homogeneity. Semantic diversity in real training images does not necessarily imply diversity in low-level structural statistics. Existing envelope learning methods primarily strengthen outer-boundary supervision by diversifying the reconstruction mechanisms and directions of deviation of near-real fake samples, while statistical coverage on the real side may remain limited. Similar texture, sharpness, and frequency-domain characteristics may cause the model to repeatedly rely on low-level statistics shared by real training images, giving rise to real-side shortcut learning and increasing the risk of local contraction or distortion of real representations.

To investigate this issue, we perform five independent random samplings of real training images from the widely used MS-COCO dataset. As shown in Fig. 3, the gradient magnitude, edge density, and high-frequency energy statistics of all five samples fall within the upper quantiles of the reference distributions (Recht et al., 2019), with little variation across samples, indicating consistent homogeneity in their low-level statistical distributions. We therefore naturally extend the diversity principle from the fake side to the real side by introducing real samples with heterogeneous low-level statistics, breaking the persistent reinforcement of similar local patterns under homogeneous supervision.

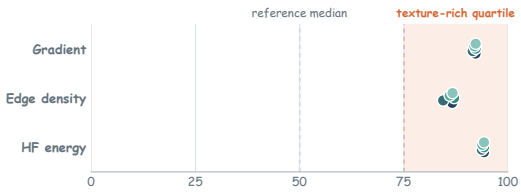


Figure 3: Bias in the low-level feature distributions of MS-COCO real training images. Across five independent random samples, gradient magnitude, edge density, and high-frequency energy consistently fall within the upper quantiles of the reference distributions.

3 METHOD

FoRM aims to preserve both the real manifold structure and the surrounding discriminative boundary within the real-centric envelope learning framework. To this end, we adopt a frozen visual foundation model as a geometric reference for real representations and constrain the student model to preserve this structure during adaptation to the detection task. Specifically, the student encoder f_θ is based on the DINOv3 ViT (Zhang et al., 2022) architecture and undergoes parameter-efficient fine-tuning via LoRA to extract forensic representations from input images x_i, where $i \in \{\text{real}, \text{fake}\}$. A frozen DINOv3 ViT serves as the teacher encoder f_T and is applied only to real images x_{real}. Its features $f_T(x_{\text{real}})$ provide a stable geometric reference for the student's real representations. Building on this, we preserve positional consistency in real representations through pointwise feature anchoring and further constrain their first-order variation within local neighborhoods through Local Directional Jacobian Alignment. This maintains the local geometric stability of the internal real manifold while learning the surrounding discriminative envelope, as illustrated in Fig. 4.

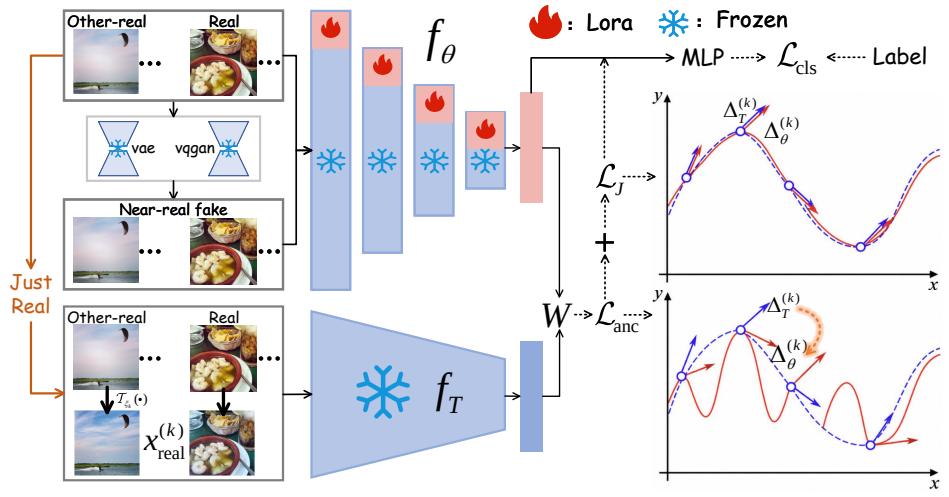


Figure 4: Overview of FoRM. Offline VAE and VQGAN reconstructions provide diverse near-real fakes for envelope supervision, while heterogeneous real samples alleviate real-side homogeneity. A frozen teacher guides pointwise anchoring and Local Directional Jacobian Alignment to preserve real-manifold geometry during real/fake discrimination.

3.1 Real-Side Pointwise Feature Anchoring.

To retain a positional reference for pretrained real representations while adapting the detector to the forgery detection task, we impose pointwise feature anchoring on the student using the frozen teacher. Specifically, for a real image x_{real}, the student and teacher extract the final-layer CLS representations $f_\theta(x_{\text{real}})$ and $f_T(x_{\text{real}})$, respectively. The student features are projected into the teacher feature space through a learnable linear mapping W. We normalize the alignment error by the squared norm of the teacher features to formulate a relative squared L_2 anchoring objective:

$$\mathcal{L}_{\text{anc}} = \mathbb{E}_{x_{\text{real}} \sim P_R} \left[\frac{\| W f_\theta(x_{\text{real}}) - f_T(x_{\text{real}}) \|_2^2}{\max\left(\| f_T(x_{\text{real}}) \|_2^2, \epsilon_{\text{anc}} \right)} \right], \tag{4}$$

where P_R denotes the real-image sampling distribution actually used during training, and $\epsilon_{\text{anc}} = 10^{-6}$ is a numerical stability constant.

3.2 Real-Side Local Directional Jacobian Alignment

Although pointwise anchoring constrains feature positions at real samples, the directions of local variation and response magnitudes in their manifold neighborhoods may still deviate from the pretrained geometry. To mitigate local contraction and geometric distortion of the real manifold during detector fine-tuning, we further propose LJ. By probing local geometric directions of the real manifold through natural variations around real images, LJ preserves the first-order variation structure of the teacher and student in real-image neighborhoods.

Specifically, for a real image x_{real}, we start from its view without additional degradation and construct n locally transformed views:

$$x_{\text{real}}^{(k)} = \mathcal{T}_{\xi_k}(x_{\text{real}}), \qquad k = 1, \ldots, n. \tag{5}$$

The natural transformation $\mathcal{T}_{\xi_k}(\cdot)$ comprises operations such as brightness, contrast, saturation, and hue adjustments, as well as mild blurring. The parameter ξ_k controls the selection and magnitude of these operations, providing multiple directions of local variation for the same real image.

We then map the student features into the teacher feature space and normalize both sets of features to compare their responses in a shared coordinate system:

$$\begin{aligned} \bar{f}_\theta(u) &= \mathcal{N}\left(W f_\theta(u) \right), \\ \bar{f}_T(u) &= \mathcal{N}\left(f_T(u) \right), \quad u \in \left\{ x_{\text{real}}, x_{\text{real}}^{(1)}, \ldots, x_{\text{real}}^{(n)} \right\}. \end{aligned} \tag{6}$$

where u denotes a real image or any of its neighboring views, and $\mathcal{N}$ denotes L_2 normalization.

Based on these normalized representations, we use feature displacements between neighboring views and the original image to approximate the first-order responses of the representation mappings along different local directions, thereby constraining the first-order local geometry of the real manifold. The student and teacher feature displacements are computed as

$$\begin{aligned} \Delta_\theta^{(k)} &= \bar{f}_\theta(x_{\text{real}}^{(k)}) - \bar{f}_\theta(x_{\text{real}}), \\ \Delta_T^{(k)} &= \bar{f}_T(x_{\text{real}}^{(k)}) - \bar{f}_T(x_{\text{real}}), \end{aligned} \tag{7}$$

where $\Delta_\theta^{(k)}$ and $\Delta_T^{(k)}$ denote the local response displacements of the student and teacher, respectively, under the k-th natural variation. Under local smoothness assumptions, a first-order Taylor expansion with respect to the input displacement $\delta x^{(k)} = x_{\text{real}}^{(k)} - x_{\text{real}}$ yields

$$\Delta_a^{(k)} = J_{\bar{f}_a}(x_{\text{real}})\, \delta x^{(k)} + \mathcal{O}\left(\left\| \delta x^{(k)} \right\|_2^2 \right), \qquad a \in \{\theta, T\}, \tag{8}$$

where $J_{\bar{f}_a}(x_{\text{real}})$ is the Jacobian of the normalized representation mapping evaluated at the real image, and $\mathcal{O}(\|\delta x^{(k)}\|_2^2)$ denotes the second-order remainder. Thus, aligning local feature displacements approximately constrains the first-order responses of the teacher and student along the sampled directions.

We then separately constrain the direction and magnitude of the responses. Directional alignment preserves the directions of local geometric variation along the real manifold, while magnitude alignment mitigates mismatches in local response strength even when the directions are aligned. The corresponding losses are computed as

$$\ell_{\text{dir}}^{(k)} = 1 - \cos\left(\Delta_\theta^{(k)}, \Delta_T^{(k)} \right), \tag{9}$$

$$\ell_{\text{mag}}^{(k)} = \text{SmoothL1}\Big(\log\left(\left\| \Delta_\theta^{(k)} \right\|_2 + \epsilon \right), \log\left(\left\| \Delta_T^{(k)} \right\|_2 + \epsilon \right) \Big), \tag{10}$$

where ϵ is a numerical stability constant, and $\text{SmoothL1}(\cdot, \cdot)$ denotes the smooth L_1 loss.

Finally, we formulate the real-side Local Directional Jacobian Alignment loss as

$$\mathcal{L}_J = \frac{1}{|\mathcal{V}|} \sum_{(x_{\text{real}}, k) \in \mathcal{V}} \left[\ell_{\text{dir}}^{(k)} + \beta\, \ell_{\text{mag}}^{(k)} \right], \tag{11}$$

where $\mathcal{V}$ is the set of valid pairs of real images and local variation indices, $|\mathcal{V}|$ denotes its cardinality, and β controls the relative strength of the magnitude constraint. By matching the directions and magnitudes of the local responses of the teacher and student, $\mathcal{L}_J$ extends pointwise feature anchoring to first-order geometric constraints within anchor neighborhoods, preserving the local geometry of variation in real representations and mitigating excessive contraction and geometric distortion of the real manifold.

3.3 Joint Optimization

To preserve the geometry of real representations within the discriminative envelope while learning its boundary, we apply pointwise feature anchoring and LJ exclusively on the real side. Together, they constrain representation positions and first-order variations within their neighborhoods, yielding the real-side local geometry preservation objective:

$$\mathcal{L}_{\text{real}} = \lambda_{\text{anc}} \mathcal{L}_{\text{anc}} + \lambda_J \mathcal{L}_J, \tag{12}$$

where λ_{anc} and λ_J control the strengths of pointwise anchoring and local first-order geometric constraints, respectively.

Furthermore, to reduce the persistent concentration of real-side supervision on homogeneous regions (Zhong et al., 2025), we introduce a heterogeneous real-sample set (Other-real), denoted as $\mathcal{D}\text{het}$. Detailed experimental results are presented in the experimental section. During training, with probability $p\text{real}$, we sample an image from $\mathcal{D}\text{het}$ to replace the original real image and apply $\mathcal{L}\text{real}$ as the corresponding constraint.

Table 1: Cross-generator detection accuracy (%). Higher is better. Bold and underlined values indicate the best and second-best results in each column, respectively; ties are marked equally. Abbreviations: D2 = DALL-E 2, GLD = GLIDE, MJ = Midjourney, VQ = VQDM, BGAN = BigGAN, Cyc = CycleGAN, Gau = GauGAN, PGAN = ProGAN, SD14/SD15 = Stable Diffusion 1.4/1.5, SGr = StarGAN, SGen/SGen2 = StyleGAN/StyleGAN2, WKG = Wukong, and Avg. = average. WFR follows the original benchmark notation.

Method	ADM	D2	GLD	MJ	VQ	BGAN	Cyc	Gau	PGAN	SD14	SD15	SGr	SGen	SGen2	WFR	WKG	Avg.
NPR	43.8	20.0	41.2	53.4	48.4	53.1	76.6	42.2	58.7	55.1	55.0	67.4	57.9	54.6	58.8	57.4	52.7
UnivFD	62.5	50.0	61.3	55.1	76.9	87.5	96.9	98.8	**99.4**	55.6	55.7	95.1	80.0	69.4	69.2	61.1	73.4
FatFormer	80.2	68.5	91.1	54.4	88.0	**99.2**	**99.5**	**99.1**	98.5	67.5	67.2	99.4	**98.0**	**98.8**	88.3	75.6	85.8
SAFE	49.5	49.5	53.0	49.0	50.2	52.2	51.9	50.0	50.0	49.7	49.8	50.1	50.0	50.0	49.8	50.3	50.3
C2P-CLIP	71.6	52.3	73.5	56.6	73.7	98.4	96.8	98.8	99.3	77.5	76.9	**99.6**	93.1	79.4	94.8	79.4	82.6
AIDE	52.9	51.1	60.2	49.8	69.3	70.1	93.6	60.6	89.0	51.6	51.0	72.1	66.5	59.0	80.6	54.5	64.5
DDA	89.5	94.6	89.6	95.6	76.6	91.0	72.5	92.7	92.8	98.7	98.6	72.7	87.8	90.2	52.1	98.8	87.1
Envelope-learning paradigm methods																	
DRCT	79.9	89.2	89.2	85.5	88.6	81.4	91.0	93.8	71.1	91.4	91.0	53.0	62.7	63.8	73.9	90.8	81.0
Aligned	51.6	52.0	55.6	96.2	72.1	51.2	49.5	50.8	50.7	**99.7**	**99.6**	53.8	52.7	51.6	50.0	**99.6**	64.8
DDA*	95.1	97.8	95.6	96.5	98.2	94.4	91.3	91.0	88.1	98.2	98.3	87.1	87.5	89.4	85.9	98.6	93.3
REM*	92.9	96.1	93.9	91.7	93.1	93.5	91.0	96.2	94.4	95.6	95.9	88.8	88.6	89.8	79.6	96.0	92.3
Ours-setting 1	**98.4**	**99.5**	**98.0**	**97.1**	**99.5**	97.4	96.3	97.6	96.3	99.4	99.3	91.9	95.9	96.0	**97.5**	99.4	**97.5**
Ours-setting 2	98.0	99.2	97.9	96.7	99.3	96.4	95.3	96.3	95.0	99.3	99.2	89.8	94.	94.6	96.3	99.3	96.7

On the fake side, to improve the characterization of the envelope surrounding the real manifold, we reconstruct real images using VAE and VQGAN (Yu et al., 2022) to generate near-real fake samples x_{fake} with different patterns of deviation. These samples provide diverse boundary supervision around the real manifold, reducing the dependence of the discriminative envelope on a single reconstruction mechanism. During training, we select a reconstruction corresponding to the current real image from either model with equal probability:

$$x_{\text{fake}} = \mathcal{R}_m(x_{\text{real}}), \qquad m \in \{\text{VAE}, \text{VQGAN}\}, \tag{13}$$

where $\mathcal{R}_m(\cdot)$ denotes the reconstruction operation performed by model m.

Finally, FoRM jointly optimizes the real/fake classification objective and the real-side geometry preservation objective:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{real}}, \tag{14}$$

where $\mathcal{L}_{\text{cls}}$ is the binary cross-entropy classification loss, and λ_{cls} is its weight. The classification objective learns the surrounding discriminative boundary from real images and their reconstructed counterparts, while $\mathcal{L}_{\text{real}}$ constrains the geometry of real-side representations, thereby jointly optimizing the real manifold structure and the surrounding discriminative envelope.

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Implementation Details. Following the experimental protocols of existing methods, we adopt DINOv3-L (Zhang et al., 2022; Hu et al., 2021) as the visual backbone. All experiments are conducted on a single NVIDIA RTX 3090 GPU. Further details of the experimental settings are provided in the Appendix. Following DDA, we use MS-COCO (Lin et al., 2015) as the source of the original real training images. To alleviate the homogeneity of low-level structural statistics in the real training data, we construct heterogeneous real samples (Other-real) from two data sources (Torralba & Efros, 2011). Our motivation analysis shows that images in MS-COCO real data generally exhibit relatively high gradient magnitude, edge density, high-frequency energy, and sharpness, with real samples concentrated in similar regions of these low-level statistical measures. Accordingly, following the principle of distributional heterogeneity, we select real images with low gradient magnitude, low edge density, low high-frequency energy, and relatively low sharpness (Geirhos et al., 2022). This selection breaks the persistent reinforcement of local patterns induced by a single homogeneous data distribution and alleviates the potential realness shortcuts it may introduce.

Table 3: Cross-dataset detection accuracy (%). Abbreviations: AIGCDet = AIGCDetectionBenchmark, GI = GenImage, D2M = DRCT-2M, EGen = EvalGEN, SB = Synthbuster, FSyn = ForenSynths, SWild = SynthWildx, WRF = WildRF, SRF = SocialRF, CAI = CommunityAI, BFO = BFree-Online, and Cham = Chameleon. Avg. denotes the average.

Method	AIGCDet	GI	D2M	EGen	SB	FSyn	SWild	WRF	SRF	CAI	BFO	Cham	Avg.
NPR	53.1	51.5	37.3	2.9	50.0	47.9	49.8	63.5	53.3	55.1	49.5	59.9	47.8
UnivFD	72.5	64.1	61.8	15.4	67.8	77.7	52.3	55.3	54.6	51.4	49.0	50.7	56.1
FatFormer	85.0	62.8	52.2	45.6	56.1	90.0	52.1	58.9	56.9	51.9	50.0	51.2	59.4
SAFE	50.3	50.3	59.3	1.1	46.5	49.7	49.1	57.2	58.4	54.5	50.5	59.2	48.8
C2P-CLIP	81.4	74.4	59.2	38.9	68.5	**92.0**	57.1	59.6	56.4	51.0	50.0	51.1	61.6
AIDE	63.6	61.2	64.6	19.1	53.9	59.4	48.8	58.4	57.3	53.7	53.1	63.1	54.7
DDA	87.8	91.7	98.1	97.2	90.1	81.4	90.9	90.3	79.6	88.6	**95.1**	82.4	89.4
DRCT	81.4	84.7	90.5	77.8	84.8	73.9	55.1	50.6	67.5	76.7	55.7	56.6	71.3
Aligned	66.6	79.0	95.5	68.0	77.4	53.9	78.8	80.1	73.5	71.9	68.5	71.0	73.7
DDA*	93.3	96.9	98.2	**99.3**	89.7	81.8	90.3	92.8	83.8	89.5	87.5	87.5	90.9
REM*	92.3	94.1	96.0	96.6	88.9	85.7	93.2	92.7	85.6	84.0	82.5	82.5	89.5
Ours-setting 1	**97.5**	**98.6**	**99.2**	98.6	**97.5**	90.4	**94.0**	**95.3**	**88.4**	**91.8**	91.5	**92.8**	**94.6**
Ours-setting 2	96.7	98.3	98.8	98.0	93.2	88.1	92.9	92.3	85.6	90.3	90.5	89.3	92.8

The first setting follows the real-image prior constructed from YFCC100M in SAID (Zhong et al., 2025). Specifically, we use KonIQ-10k (Hosu et al., 2020), which is derived from YFCC100M (Thomee et al., 2016), and select heterogeneous real samples according to the aforementioned low-level statistical criteria, obtaining approximately 3,000 images. The second setting performs selection and resampling exclusively within the original MS-COCO training set, allowing us to evaluate the effect of statistical heterogeneity itself without introducing external images. Using the same selection criteria, approximately 1,500 qualifying images can be obtained from MS-COCO. Both settings use identical selection metrics and training procedures.

Table 2: Real/fake classification accuracy (%) on each sub-dataset. R and F denote RealAcc and FakeAcc, respectively. S1 and S2 denote Ours-setting 1 and Ours-setting 2.

Sub.	DDA*		REM*		S1		S2	
	R	F	R	F	R	F	R	F
ADM	96.3	93.8	91.6	94.1	**99.0**	**97.8**	98.6	97.4
D2	97.2	98.5	96.3	95.8	**99.7**	**99.3**	99.5	99.3
GLD	96.5	94.7	92.7	95.1	**98.7**	**97.4**	98.4	97.1
MJ	96.7	**96.4**	92.2	91.2	**98.8**	95.4	98.2	95.5
VQ	96.7	99.6	91.9	94.3	**99.1**	**100.0**	98.7	100.0
BGAN	88.9	99.9	88.0	99.0	**94.8**	**100.0**	93.6	100.0
Cyc	83.3	99.3	89.3	92.7	**92.8**	**99.7**	91.7	99.6
Gau	81.9	100.0	92.6	99.7	**95.2**	**100.0**	94.0	99.9
PGAN	76.3	99.9	88.8	99.9	**92.5**	**100.0**	91.1	99.8
SD14	96.5	100.0	92.1	99.1	**98.8**	**100.0**	98.4	100.0
SD15	96.8	99.8	92.5	99.2	**98.8**	**99.9**	98.4	99.8
SGr	74.2	100.0	78.9	98.8	**83.9**	**100.0**	78.5	100.0
SGen	75.0	100.0	82.6	94.5	**91.7**	**100.0**	87.4	100.0
SGen2	78.9	**100.0**	86.9	92.8	**92.1**	**100.0**	87.8	100.0
WFR	71.8	**100.0**	85.1	74.0	**96.5**	98.4	94.7	98.3
WKG	97.2	99.9	92.5	99.5	**98.8**	**100.0**	98.5	99.6

Datasets. We evaluate the model on cross-dataset benchmarks, including AIGCDetectionBenchmark (Zhong et al., 2024), GenImage (Zhu et al., 2023), DRCT-2M (Chen et al., 2024), ForenSynths (Wang et al., 2020), Synthbuster (Bammey, 2024), EvalGEN (Chen et al., 2025b), SynthWildx (Cozzolino et al., 2024) , WildRF (Cavia et al., 2024), CommunityAI (Li et al., 2025b), SocialRF (Li et al., 2025b), BFree-Online (Guillaro et al., 2025), and Chameleon (Yan et al., 2025). These benchmarks cover images generated by diffusion models, GANs, autoregressive models, and unknown generation mechanisms, as well as real-world data collected from social media and the Internet. Following the evaluation protocol of DDA (Chen et al., 2025a), we apply JPEG compression with a quality factor of $Q = 96$ to the generated images in GenImage, ForenSynths, and AIGCDetectionBenchmark to mitigate format bias. Notably, the model is trained only once and then directly evaluated on all datasets using the same model, without any dataset-specific fine-tuning, threshold adjustment, or hyperparameter optimization.

Comparative Methods. We use accuracy (Acc) as the evaluation metric and consistently apply 0.5 as the classification threshold. The compared methods cover three representative categories: methods based on frequency or low-level artifacts, including NPR (Tan et al., 2023a), SAFE (Li et al., 2025a), and AIDE (Yan et al., 2025); methods based on pretrained visual representations, including UnivFD (Ojha et al., 2024), FatFormer (Liu et al., 2024), and C2P-CLIP (Tan et al., 2024); and methods based on reconstruction or real-manifold envelope learning, including DRCT (Chen et al., 2024), AlignedForensics (Rajan et al., 2025), DDA (Chen et al., 2025a), and REM (Liu et al., 2025). To reduce the influence of differences in backbone architectures and training configurations, we additionally reproduce DDA and REM using the same DINOv3-L backbone, denoted as DDA* and REM*, respectively. Since the source code of REM is not publicly available, REM* is fully implemented according to the algorithm described in the paper, with only the backbone replaced by DINOv3-L for consistency.

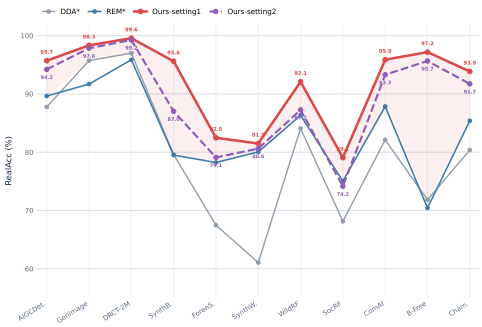


Figure 5: Real-image accuracy curves of DDA*, REM*, and the two FoRM settings across 11 cross-dataset benchmarks. EvalGEN is excluded because it contains no real images. FoRM achieves higher and more stable real-side accuracy on most datasets.

4.2 Results

We first evaluate the methods on AIGCDetectionBenchmark, with the results reported in Table 1. FoRM achieves the best overall accuracy, reaching state-of-the-art performance. To further analyze the performance gap between the real and generated sides, we select two representative envelope-learning methods, DDA* and REM*, and report their real and fake accuracies on different generator subsets, together with those of FoRM, in Table 2. As shown, all three methods achieve relatively high fake accuracy on most generators, while the main differences in overall accuracy arise from real accuracy. This suggests that distortion of the real manifold enclosed by the envelope is a major limiting factor. In contrast, FoRM improves real accuracy through real-side geometric constraints while maintaining high fake accuracy, thereby achieving superior overall performance.

The model is trained only once and then directly evaluated on 12 large-scale cross-dataset benchmarks, without any dataset-specific fine-tuning, threshold adjustment, or parameter optimization. As shown in Table 3, FoRM achieves the highest average accuracy. To further analyze the cross-domain stability of real representations and the effectiveness of local geometric preservation, we plot the real-image accuracy curves of all methods across different datasets, as shown in Fig. 5. Compared with representative real-centric envelope-learning methods, both FoRM settings achieve higher and more stable real accuracy on most datasets. This indicates that real-side geometric constraints can effectively mitigate local degradation of the real manifold during detector fine-tuning, thereby improving cross-domain generalization. More results are provided in the Appendix.

4.3 Ablation Study and Robustness Evaluation

To validate the contributions of the individual components, we conduct ablation studies on pointwise feature anchoring, Local Directional Jacobian Alignment, heterogeneous real samples, and heterogeneous reconstruction mechanisms. We further perform robustness experiments under various image degradations to evaluate the model's adaptability to common perturbations. Detailed experimental settings and results are provided in the Appendix.

4.4 Conclusion and Limitations

FoRM improves cross-generator and cross-dataset detection performance through real-side local geometry preservation and heterogeneous supervision. However, it still relies on a pretrained visual teacher and offline reconstruction samples. Reducing this reliance and extending the method to a broader range of generation mechanisms remain open problems.

LARGE LANGUAGE MODEL USAGE STATEMENT

Large language models were used to assist with code implementation and language polishing. The authors reviewed and verified all generated content and take full responsibility for the paper.

REFERENCES

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.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. URL https://arxiv.org/abs/1809.11096.

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.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24ay.html.

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, 2025a. URL https://arxiv.org/abs/2505.14359.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=C39ShJwtD5.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models, 2022. URL https://arxiv.org/abs/2211.00680.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

Manni Cui, Ruiqi Liu, Dianyuan Zou, Ziheng Qin, Jingrui Xu, ZiAn Wang, Jianglan Wei, Han Zhou, Yu Liu, Yan Wang, et al. Globalforge: Towards robust ai-generated image detection. *arXiv preprint arXiv:2607.14684*, 2026.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition, 2020. URL https://arxiv.org/abs/2003.08685.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022. URL https://arxiv.org/abs/1811.12231.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

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 Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 18685–18694, June 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. URL https://arxiv.org/abs/1710.10196.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

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, 2025a.

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*, 2025b.

Huanqi Lin, Jinghui Qin, Xiaoqi Wu, Tianshui Chen, and Zhijing Yang. Revisiting dire: towards universal ai-generated image detection. *Neural Networks*, 193:108084, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.108084. URL https://www.sciencedirect.com/science/article/pii/S0893608025009645.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

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

Rui-Qi Liu, Yi Han, Bo Sun, Houxin He, Yuxuan Zhou, and Yan Wang. M 2 ea: Multi-vae manifold envelope alignment for challenging ai-generated image detection. *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. URL https://api.semanticscholar.org/CorpusID:287675827.

Ruiqi Liu, Yi Han, Zhengbo Zhang, Liwei Yao, Zhiyuan Yan, Jialiang Shen, ZhiJin Chen, Boyi Sun, Lubin Weng, Jing Dong, Yan Wang, and Shu Wu. Beyond artifacts: Real-centric envelope modeling for reliable ai-generated image detection, 2025. URL https://arxiv.org/abs/2512.20937.

Zhengzhe Liu, Xiaojuan Qi, and Philip Torr. Global texture enhancement for fake face detection in the wild, 2020. URL https://arxiv.org/abs/2002.00133.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models, 2024. URL https://arxiv.org/abs/2302.10174.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images, 2025. URL https://arxiv.org/abs/2410.11835.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL https://arxiv.org/abs/1902.10811.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. URL https://arxiv.org/abs/1412.6550.

Jialiang Shen, Jiyang Zheng, Yunqi Xue, Huajie Chen, Yu Yao, Hui Kang, Ruiqi Liu, Helin Gong, Yang Yang, Dadong Wang, et al. Dino-detect: A simple yet effective framework for blur-robust ai-generated image detection. *arXiv preprint arXiv:2511.12511*, 2025.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Chuangchuang Tan, Huan Liu, 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, 2023a.

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 (CVPR)*, pp. 12105–12114, June 2023b.

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, 2024. URL https://arxiv.org/abs/2408.09647.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73, January 2016. ISSN 1557-7317. doi: 10.1145/2812802. URL http://dx.doi.org/10.1145/2812802.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2020. URL https://arxiv.org/abs/1912.11035.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection, 2025. URL https://arxiv.org/abs/2406.19435.

Liwei Yao, Sen Niu, Xiaomei Feng, Bofeng Zhang, Ruiqi Liu, and Kefeng Li. All patches are not equal: Focusing on a single exposed patch for ai-generated image detection. *Information Processing & Management*, 64(1):105058, 2027.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan, 2022. URL https://arxiv.org/abs/2110.04627.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. URL https://arxiv.org/abs/2203.03605.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2024. URL https://arxiv.org/abs/2311.12397.

Nan Zhong, Mian Zou, Yi-Ran Xu, Zhenxing Qian, Xinpeng Zhang, Baoyuan Wu, and Kede Ma. Self-supervised ai-generated image detection: A camera metadata perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:6192–6207, 2025. URL https://api.semanticscholar.org/CorpusID:283672339.

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, 2023.

A Appendix

A.1 Summary of Contributions

The main contributions of this work are summarized as follows:

1. We revisit the data-centric, real-centric envelope learning paradigm and identify a key limitation: existing methods emphasize modeling the outer boundary while overlooking the integrity and stability of the real manifold enclosed within it.
2. We propose FoRM, which centers on Local Jacobian Alignment (LJ). LJ extends conventional pointwise feature anchoring from zeroth-order feature consistency to first-order local geometry preservation, better maintaining the structure of the real manifold during detector fine-tuning.
3. We introduce distributionally heterogeneous real samples to mitigate local representation bias and real-manifold contraction induced by homogeneous real data, thereby improving the stability of real-side representations.
4. On the fake side, we build on and strengthen the boundary-modeling principle of existing envelope learning methods. We use heterogeneous reconstruction mechanisms to generate near-real fake samples with more diverse distributions, improving coverage of the envelope surrounding the real manifold.

A.2 Related work

Traditional Detectors. Traditional AI-generated image detection methods (Wang et al., 2020; Tan et al., 2023b; Liu et al., 2020; Yao et al., 2027) primarily learn discriminative differences between real and generated images, including low-level forgery traces such as frequency, texture, and gradient cues. Subsequent studies have leveraged pretrained visual representations to improve cross-generator generalization. For example, UnivFD (Ojha et al., 2024) performs classification based on pretrained features, while FatFormer (Liu et al., 2024) enhances forensic representations through a forgery-aware adapter. However, real/fake classification supervision may still cause detectors to rely on content or data-source differences that are spuriously correlated with the labels, as well as artifacts specific to known generators, thereby limiting generalization to unseen generators.

Reconstruction- and Real-Centric Envelope-Learning Detectors. From the perspective of envelope-based reconstruction learning, these methods reduce shortcut learning on the fake side by aligning real and fake data and constructing hard fake samples close to the real distribution, thereby providing supervision for the discriminative envelope surrounding the real manifold. DRCT constructs (Chen et al., 2024) hard negatives through high-quality diffusion reconstruction, encouraging

envelope learning to focus on real/fake differences closer to the real distribution. DDA (Chen et al., 2025a) further performs dual-domain alignment in both the pixel and frequency domains, reducing the dependence of the envelope boundary on content and frequency biases. REM (Liu et al., 2025) explicitly introduces real-centric envelope modeling, generating near-real samples through feature perturbations during reconstruction and learning a discriminative boundary enclosing the real manifold. M²EA (Liu et al., 2026) uses multiple VAE reconstructions to diversify near-real samples and improve the characterization of the envelope boundary. These methods primarily strengthen outer-boundary supervision by optimizing fake-side samples and data alignment. FoRM further focuses on the local geometric stability of real representations inside the envelope during fine-tuning and incorporates heterogeneous real samples to jointly improve internal structure preservation and outer-boundary learning.

A.3 Experiments

A.3.1 Detailed Experimental Settings

We use DINOv3-L as the visual backbone and initialize the student encoder with its pretrained weights. The DINOv3-L backbone parameters in the student remain frozen, and only the LoRA parameters are trained. Specifically, LoRA modules are inserted into the q_{proj}, k_{proj}, and v_{proj} layers of all 24 Transformer blocks (layers 0–23), with rank $r = 8$ and scaling factor $\alpha = 8$. Thus, task adaptation of the student is performed solely through these LoRA parameters. The real-side teacher encoder is a copy of the DINOv3-L backbone used to initialize the student; all of its parameters remain frozen, and it receives no gradient updates during training. The teacher is used only to provide a representational reference for real images. For each clean real crop, LJ constructs $K = 4$ independent local variation directions and is applied only to the real-image branch. The random seed is set to 0. The probability of introducing a heterogeneous real sample (Other-real) is set to $p_{\mathrm{real}} = 0.2$. At this probability, the original real image is replaced during training, and its corresponding reconstructed fake is selected accordingly to preserve the real–fake pairing.

The training input resolution is 336×336. We use random cropping during training and center cropping during validation and testing. When an input image is smaller than the required size, we apply the padding strategy used in the DDA baseline protocol. We use the Adam optimizer with an initial learning rate of 1×10^{-4} and a cosine annealing learning-rate scheduler. The batch size is 16, the number of gradient accumulation steps is 4, and the total training budget is 15,000 steps.

To construct near-real fake samples, we use the pretrained VAE (Kingma & Welling, 2022) from Stable Diffusion 2.1 for VAE reconstruction and the CompVis pretrained VQGAN-vq-f8 model for VQGAN (Yu et al., 2022) reconstruction. Both reconstruction models remain frozen. During training, a near-real fake sample corresponding to the current real image is selected with equal probability from the VAE and VQGAN reconstructions, providing discriminative supervision with complementary patterns of deviation around the real manifold.

A.3.2 Comparison Methods

Traditional Non-Envelope-Based Detection Methods

- **NPR (Tan et al., 2023a)** detects generated images using neighboring-pixel relationship features introduced by the upsampling process of generative models.
- **SAFE (Li et al., 2025a)** uses image transformations to improve detector generalization across different generative models.
- **AIDE (Yan et al., 2025)** is included as a representative recent method for generalizable AI-generated image detection.
- **UnivFD (Ojha et al., 2024)** builds a generated-image detector using general-purpose representations from a pretrained vision model.
- **FatFormer (Liu et al., 2024)** uses a forgery-aware Transformer to learn generalizable visual representations.
- **C2P-CLIP (Tan et al., 2024)** introduces class-common prompts into CLIP to improve generalization in generated-image detection.

Table 4: Accuracy (%) on the DRCT-2M benchmark. Bold and underlined values indicate the best and second-best results in each column, respectively; ties are marked equally. Ref. = SDXL-Refiner, T = Turbo, LCM1.5 = LCM-SDv1.5, LCMXL = LCM-SDXL, C = Ctrl, and Avg. = average. The DR suffix follows the original benchmark naming. An asterisk denotes a reproduced comparison variant.

Method	LDM	SD1.4	SD1.5	SD2	SDXL	Ref.	SD-T	SDXL-T	LCM1.5	LCMXL	SD1-C	SD2-C	SDXL-C	SD1-DR	SD2-DR	SDXL-DR	Avg.
NPR	33.0	29.1	29.0	35.1	33.2	28.4	27.9	27.9	29.4	30.2	28.4	28.3	34.7	67.9	67.4	66.1	37.3
UnivFD	85.4	56.8	56.4	58.2	63.2	55.0	56.5	53.0	54.5	65.9	68.0	65.4	75.9	64.6	56.2	53.9	61.8
FatFormer	55.9	48.2	48.2	48.2	48.2	48.3	48.2	48.2	48.3	50.6	49.7	49.9	59.8	66.3	60.6	56.0	52.2
SAFE	50.3	50.1	50.0	50.0	49.9	50.1	50.0	50.0	50.1	50.0	49.9	50.0	54.7	98.2	98.5	97.3	59.3
C2P-CLIP	83.0	51.7	51.7	52.9	51.9	64.6	51.7	50.6	52.0	66.1	56.9	54.7	77.8	67.2	57.1	56.7	59.2
AIDE	64.4	74.9	75.1	58.5	53.5	66.3	52.8	52.8	70.0	54.3	65.9	53.6	53.9	95.3	73.3	69.0	64.6
Envelope-learning paradigm methods																	
DDA	99.2	98.9	99.0	98.3	98.0	96.8	97.9	94.8	95.9	98.2	98.7	99.0	99.4	99.0	99.5	96.3	98.1
DRCT	96.7	96.3	96.3	94.9	96.2	93.5	93.4	92.9	91.2	95.0	95.6	92.7	92.0	94.1	69.6	57.4	90.5
AlignedForensics	**99.9**	**99.9**	**99.9**	**99.6**	90.2	81.3	**99.7**	89.4	**99.7**	90.0	**99.9**	99.2	87.6	**99.9**	**99.8**	92.6	95.5
DDA*	98.5	98.4	98.3	98.0	98.1	97.6	97.9	97.7	98.3	98.3	98.5	98.4	98.5	98.4	98.3	**97.7**	98.2
REM*	97.7	96.1	96.0	95.4	91.9	96.9	96.0	95.6	96.3	97.4	97.4	97.5	97.6	96.8	97.4	90.4	96.0
Ours-setting-1	99.8	99.5	99.4	98.9	**99.4**	**98.8**	99.5	**98.7**	99.4	**99.7**	99.8	**99.7**	**99.7**	99.8	99.5	95.4	**99.2**

Reconstruction-Based and Real-Centric Envelope-Learning Methods

These methods use the real-image distribution as a reference and provide discriminative supervision in the neighborhood of the real manifold through data alignment, reconstruction, or near-real fake sample construction. This enables them to characterize the decision boundary surrounding the real distribution.

- **DRCT (Chen et al., 2024)** uses real images and their diffusion reconstructions to provide contrastive supervision between content-matched samples, strengthening boundary characterization near the real manifold.
- **AlignedForensics (Rajan et al., 2025)** aligns real images with latent-diffusion-generated images so that supervision focuses on authenticity differences between semantically corresponding samples, providing additional boundary supervision near the real distribution.
- **DDA*** (Chen et al., 2025a) uses dual data alignment to strengthen the modeling of relationships between real and generated samples, providing more comprehensive discriminative supervision in the neighborhood of the real distribution. For a fair comparison, we follow its original training procedure and replace only the backbone with DINOv3-L.
- **REM*** (Liu et al., 2025) constructs a discriminative envelope centered on the real manifold and uses near-real fake samples to characterize the boundary surrounding the real distribution. Since the official REM code is not publicly available, we reproduce the method based on the description in its paper. For a fair comparison, we replace only the backbone with DINOv3-L.

A.3.3 Detailed Experimental Results

This section reports fine-grained results across the 12 evaluation benchmarks. For benchmarks comprising multiple subsets or generators, we provide detection performance for each one to illustrate generalization across different generation mechanisms and sources of real images. The results are presented in Table. 4-Table. 9.

To further analyze the performance bottlenecks of real-centric envelope learning, we report RealAcc and FakeAcc for DDA*, REM*, and FoRM on each benchmark and its subsets. The results are presented in Table 10–Table 14. DDA* and REM* achieve high FakeAcc on several subsets, yet their RealAcc remains substantially lower, becoming the main constraint on overall performance. This suggests that continued strengthening of fake-side outer-boundary supervision alone is insufficient to adequately address real-side misclassification. Envelope learning must also account for the structural stability of the real manifold that serves as its internal reference. In contrast, FoRM improves RealAcc across multiple benchmarks while maintaining or improving FakeAcc, supporting the design principle of jointly optimizing real-side geometry preservation and outer-boundary modeling.

Table 5: Accuracy (%) on the GenImage benchmark. Bold and underlined values indicate the best and second-best results in each column, respectively; ties are marked equally. MJ = Midjourney, GLD = GLIDE, WKG = Wukong, BGAN = BigGAN, and Avg. = average. An asterisk denotes a reproduced comparison variant.

Method	MJ	SD1.4	SD1.5	ADM	GLD	WKG	VQDM	BGAN	Avg.
NPR	53.4	55.1	55.0	43.8	41.2	57.4	48.4	57.7	51.5
UnivFD	55.1	55.6	55.7	62.5	61.3	61.1	76.9	84.4	64.1
FatFormer	52.1	53.6	53.8	61.4	65.5	60.9	72.5	82.2	62.8
SAFE	49.0	49.7	49.8	49.5	53.0	50.3	50.2	50.9	50.3
C2P-CLIP	56.6	77.5	76.9	71.6	73.5	79.4	73.7	85.9	74.4
AIDE	58.2	77.2	77.4	50.4	54.6	70.5	50.8	50.6	61.2
Envelope-learning paradigm methods									
DDA	95.6	98.7	98.6	89.5	89.6	98.7	76.5	86.5	91.7
DRCT	82.4	88.3	88.2	76.9	86.1	87.9	85.4	87.0	84.7
AlignedForensics	**97.5**	**99.7**	**99.6**	52.4	57.6	**99.6**	75.0	50.6	79.0
DDA*	96.5	98.2	98.3	<u>95.1</u>	<u>95.6</u>	98.6	<u>98.2</u>	<u>94.4</u>	<u>96.9</u>
REM*	91.7	95.6	95.9	92.9	93.9	96.0	93.1	93.5	94.1
Ours-setting-1	<u>97.1</u>	<u>99.4</u>	<u>99.3</u>	**98.4**	**98.0**	<u>99.4</u>	**99.5**	**97.4**	**98.6**

Table 6: Accuracy (%) on the ForenSynths benchmark. Bold and underlined values indicate the best and second-best results in each column, respectively; ties are marked equally. BGAN = BigGAN, Cyc = CycleGAN, DF = DeepFake, Gau = GauGAN, PGAN = ProGAN, SDark = SeeingDark, SGr = StarGAN, SGen = StyleGAN, SGen2 = StyleGAN 2, WFR = WhichFaceIsReal, and Avg. = average. An asterisk denotes a reproduced comparison variant.

Method	BGAN	CRN	Cyc	DF	Gau	IMLE	PGAN	SAN	SDark	SGr	SGen	SGen2	WFR	Avg.
NPR	53.1	0.4	76.6	35.7	42.2	5.3	58.7	48.4	63.6	67.4	57.9	54.6	58.8	47.9
UnivFD	87.5	55.7	<u>96.9</u>	69.4	<u>98.8</u>	68.1	**99.4**	58.2	62.2	95.1	80.0	69.4	69.2	77.7
FatFormer	**99.3**	72.1	**99.5**	**93.0**	**99.3**	72.1	98.4	70.8	81.9	<u>99.4</u>	**98.1**	**98.9**	88.3	90.1
SAFE	52.2	50.0	51.9	50.1	50.0	50.0	50.0	50.9	41.1	50.1	50.0	50.0	49.8	49.7
C2P-CLIP	<u>98.4</u>	**93.3**	96.8	<u>92.6</u>	<u>98.8</u>	**93.2**	<u>99.3</u>	63.2	**94.7**	**99.6**	93.1	79.4	<u>94.8</u>	**92.1**
AIDE	70.1	12.2	93.6	53.2	60.6	15.9	89.0	55.3	44.2	72.1	66.5	59.0	80.6	59.4
Envelope-learning paradigm methods														
DDA	91.0	<u>87.0</u>	72.5	76.5	92.7	<u>89.7</u>	92.8	**94.7**	58.6	72.7	87.8	90.2	52.1	81.4
DRCT	81.4	78.4	91.0	51.5	93.8	82.6	71.1	84.9	72.2	53.0	62.7	63.8	73.9	73.9
AlignedForensics	51.2	50.4	49.5	71.7	50.8	49.7	50.7	67.6	51.4	53.8	52.7	51.6	50.0	53.9
DDA*	94.4	52.4	91.3	72.8	91.0	52.4	88.1	90.2	80.6	87.1	87.5	89.4	85.9	81.8
REM*	93.5	74.1	91.0	77.0	96.2	74.4	94.4	90.2	76.9	88.8	88.6	89.9	79.6	85.7
Ours-setting-1	97.4	75.1	96.3	74.4	97.6	75.1	96.3	<u>93.3</u>	<u>87.8</u>	92.0	<u>95.9</u>	<u>96.0</u>	**97.5**	<u>90.4</u>

A.3.4 Ablation Study

To analyze the contribution of each component, we conduct ablation studies on AIGCDetection-Benchmark. This benchmark covers multiple types of generators and real images from different sources, allowing us to assess the role of each component across a broad range of data distributions. Table 15 compares combinations of real-side pointwise feature anchoring ($\mathcal{L}_{\text{anc}}$), Local Directional Jacobian Alignment (LJ), heterogeneous real sampling (HRS), and heterogeneous reconstruction (HRec), reporting ACC, RealAcc, and FakeAcc separately. The results show that adding LJ to pointwise anchoring improves real-image recognition, demonstrating the additional contribution of first-order local geometric constraints beyond zeroth-order feature consistency. HRS and HRec provide further benefits through real-side sampling and outer-boundary supervision, respectively, and exhibit a degree of complementarity when combined. The full configuration achieves the highest ACC (97.46%) and RealAcc (95.70%) while maintaining high FakeAcc (99.23%), supporting the

Table 7: Accuracy (%) on the Synthbuster benchmark. Bold and underlined values indicate the best and second-best results in each column, respectively; ties are marked equally. D2 = DALL-E 2, D3 = DALL-E 3, FF = Firefly, GLD = GLIDE, MJ = Midjourney, and Avg. = average. An asterisk denotes a reproduced comparison variant.

Method	D2	D3	FF	GLD	MJ	SD1.3	SD1.4	SD2	SDXL	Avg.
NPR	51.1	49.3	46.5	48.5	52.8	51.4	51.8	46.0	52.8	50.0
UnivFD	83.5	47.4	89.9	53.3	52.5	70.4	69.9	75.7	68.0	67.8
FatFormer	59.4	39.5	60.3	72.7	44.4	53.7	54.0	52.3	69.1	56.1
SAFE	58.0	9.9	10.3	52.2	56.7	59.4	59.1	53.0	59.5	46.5
C2P-CLIP	55.6	63.2	59.5	86.7	52.9	75.2	76.7	69.2	77.7	68.5
AIDE	34.9	33.7	24.8	65.0	57.5	74.1	73.7	53.2	68.4	53.9
Envelope-learning paradigm methods										
DDA	86.3	90.0	91.9	76.5	93.5	92.9	92.7	93.3	93.5	90.1
DRCT	77.2	86.6	84.1	82.6	73.7	86.6	86.6	83.2	71.3	84.8
AlignedForensics	50.2	48.9	51.7	53.5	**98.7**	**98.8**	**98.8**	**98.6**	97.3	77.4
DDA*	89.6	89.8	89.8	88.8	89.8	89.8	89.8	89.8	89.8	89.7
REM*	88.5	89.6	88.0	86.8	89.6	89.5	89.3	89.6	89.7	88.9
Ours-setting-1	**97.6**	**97.8**	**97.7**	**95.6**	97.8	97.8	97.8	97.8	**97.8**	**97.5**

Table 8: Accuracy (%) on the WildRF benchmark. Bold and underlined values indicate the best and second-best results in each column, respectively; ties are marked equally. FB = Facebook, Red = Reddit, Tw = Twitter, and Avg. = average. An asterisk denotes a reproduced comparison variant.

Method	FB	Red	Tw	Avg.
NPR	78.1	61.0	51.3	63.5
UnivFD	49.1	60.2	56.5	55.3
FatFormer	54.1	68.1	54.4	58.9
SAFE	50.9	74.1	37.5	57.2
C2P-CLIP	54.4	68.4	55.9	59.6
AIDE	57.8	71.5	45.8	58.4
Envelope-learning paradigm methods				
DDA	93.1	86.4	91.5	90.3
DRCT	46.6	53.1	55.2	50.6
AlignedForensics	89.4	69.1	81.8	80.1
DDA*	93.8	90.7	93.8	92.8
REM*	92.5	91.3	94.3	92.7
Ours-setting-1	**95.3**	**93.1**	**97.4**	**95.3**

design principle of jointly optimizing real-side geometry preservation, heterogeneous sampling, and outer-boundary modeling.

A.3.5 Robustness Evaluation

To evaluate the model's robustness to common image processing operations, we apply JPEG compression with the quality factor reduced from 100 to 60, image resizing with scaling factors ranging from 0.5 to 2.0, and Gaussian blur with σ ranging from 0 to 2.0, without retraining and with the test threshold fixed. As shown in Fig. 6, FoRM maintains relatively stable detection performance under all three transformations. Its performance declines only gradually as the compression quality decreases, and it is only slightly affected by changes in the scaling factor. Under strong blur, although all methods experience performance degradation, FoRM still performs well and outperforms most competing methods.

Table 9: Accuracy (%) on the EvalGEN benchmark. Bold and underlined values indicate the best and second-best results in each column, respectively; ties are marked equally. Flx = Flux, Inf. = Infinity, Omni = OmniGen, and Avg. = average. An asterisk denotes a reproduced comparison variant.

Method	Flx	GoT	Inf.	NOVA	Omni	Avg.
NPR	0.7	0.2	6.5	4.7	2.2	2.9
UnivFD	4.0	9.2	15.7	8.3	39.6	15.4
FatFormer	9.9	47.9	44.7	98.3	27.3	45.6
SAFE	1.0	0.5	1.9	0.6	1.6	1.1
C2P-CLIP	8.7	49.6	35.3	86.4	14.5	38.9
AIDE	17.9	24.7	3.4	16.3	33.4	19.1
Envelope-learning paradigm methods						
DDA	89.9	99.5	<u>97.8</u>	99.5	99.5	97.2
DRCT	72.5	81.4	77.9	84.6	72.5	77.8
AlignedForensics	32.0	72.3	74.0	84.8	77.0	68.0
DDA*	**97.5**	<u>99.8</u>	**100.0**	<u>99.8</u>	<u>99.7</u>	**99.3**
REM*	90.3	98.8	94.8	**100.0**	99.2	96.6
Ours-setting-1	<u>93.1</u>	**99.9**	**100.0**	**100.0**	**99.8**	<u>98.6</u>

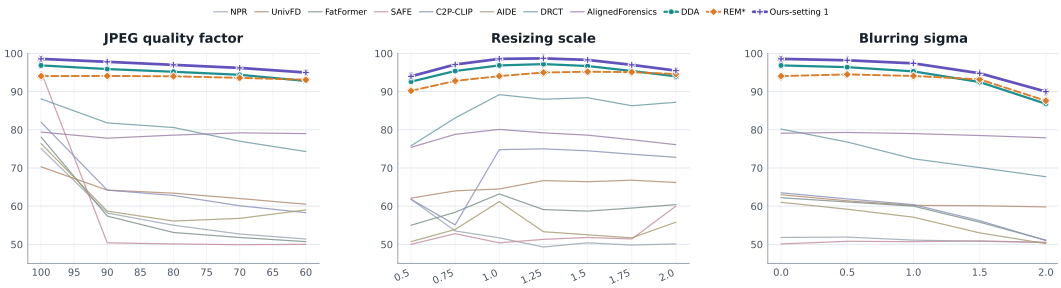


Figure 6: Robustness comparison under different image processing conditions, including JPEG compression, image resizing, and Gaussian blur. All methods are evaluated without retraining and with a fixed test threshold. The horizontal axis indicates the processing strength, and the vertical axis indicates detection accuracy. FoRM maintains more stable performance overall under the three benign transformations.

Table 10: Real and fake classification accuracy (%) on DRCT-2M. R and F denote RealAcc and FakeAcc, respectively. Bold and underlined values indicate the best and second-best results among the three methods for each subdataset and metric, respectively. Avg. denotes the reported average. The identical RealAcc values across subdatasets arise because all subdatasets use the same real-image source (Chen et al., 2024).

Subdataset	DDA*		REM*		Ours-setting-1	
	R	F	R	F	R	F
LDM	97.0	99.9	95.9	99.5	**99.6**	**100.0**
SDv1.4	97.0	**99.7**	95.9	96.4	**99.6**	99.5
SDv1.5	97.0	**99.6**	95.9	96.2	**99.6**	99.2
SDv2	97.0	**98.9**	95.9	94.9	**99.6**	98.2
SDXL	97.0	99.1	95.9	87.9	**99.6**	**99.2**
SDXL-Refiner	97.0	**98.3**	95.9	97.9	**99.6**	98.1
SD-Turbo	97.0	98.8	95.9	96.2	**99.6**	**99.3**
SDXL-Turbo	97.0	**98.3**	95.9	95.3	**99.6**	97.9
LCM-SDv1.5	97.0	**99.6**	95.9	96.7	**99.6**	99.3
LCM-SDXL	97.0	99.5	95.9	99.0	**99.6**	**99.8**
SDv1-Ctrl	97.0	**100.0**	95.9	99.0	**99.6**	100.0
SDv2-Ctrl	97.0	99.8	95.9	99.2	**99.6**	**99.9**
SDXL-Ctrl	97.0	**99.9**	95.9	99.3	**99.6**	99.9
SDv1-DR	97.0	99.8	95.9	97.8	**99.6**	**100.0**
SDv2-DR	97.0	**99.7**	95.9	99.0	**99.6**	99.4
SDXL-DR	97.0	**98.4**	95.9	84.9	**99.6**	91.2
Avg.	97.0	**99.3**	95.9	96.2	**99.6**	98.8

Table 11: Real and fake classification accuracy (%) on GenImage. R and F denote RealAcc and FakeAcc, respectively. Bold and underlined values indicate the best and second-best results among the three methods for each subdataset and metric, respectively. Avg. denotes the reported average.

Subdataset	DDA*		REM*		Ours-setting-1	
	R	F	R	F	R	F
Midjourney	96.7	**96.4**	92.2	91.2	**98.8**	95.4
SDv1.4	96.5	100.0	92.1	99.1	**98.8**	**100.0**
SDv1.5	96.8	99.8	92.5	99.2	**98.8**	**99.9**
ADM	96.3	93.8	91.6	94.1	**99.0**	**97.8**
GLIDE	96.5	94.7	92.7	95.1	**98.7**	**97.3**
Wukong	97.2	99.9	92.5	99.5	**98.8**	**100.0**
VQDM	96.7	99.6	91.9	94.3	**99.1**	**100.0**
BigGAN	88.9	99.8	88.0	99.0	**94.8**	**100.0**
Avg.	95.7	98.0	91.7	96.4	**98.3**	**98.8**

Table 12: Real and fake classification accuracy (%) on ForenSynths. R and F denote RealAcc and FakeAcc, respectively. Bold and underlined values indicate the best and second-best results among the three methods for each subdataset and metric, respectively. Avg. denotes the reported average.

Subdataset	DDA*		REM*		Ours-setting-1	
	R	F	R	F	R	F
BigGAN	88.9	99.8	88.0	99.0	**94.8**	**100.0**
CRN	4.9	100.0	48.8	99.4	**50.3**	**100.0**
CycleGAN	83.3	99.3	89.3	92.7	**92.8**	**99.7**
DeepFake	**84.1**	61.4	66.1	87.9	58.3	**90.6**
GauGAN	81.9	100.0	92.6	99.7	**95.2**	**100.0**
IMLE	4.9	100.0	48.8	99.9	**50.3**	**100.0**
ProGAN	76.3	99.9	88.8	99.9	**92.5**	**100.0**
SAN	82.5	**97.3**	86.0	94.1	**89.0**	**97.3**
SeeingDark	70.6	**90.6**	75.0	78.9	**85.0**	**90.6**
StarGAN	74.2	100.0	78.9	98.8	**83.9**	**100.0**
StyleGAN	75.0	100.0	82.6	94.5	**91.7**	**100.0**
StyleGAN 2	78.9	**100.0**	86.9	92.8	**92.1**	**100.0**
WFR	71.8	**100.0**	85.1	74.0	**96.5**	98.4
Avg.	67.5	96.0	78.2	93.2	**82.5**	**98.2**

Table 13: Real and fake classification accuracy (%) on Synthbuster. R and F denote RealAcc and FakeAcc, respectively. Bold and underlined values indicate the best and second-best results among the three methods for each subdataset and metric, respectively. Avg. denotes the reported average. The identical RealAcc values across subdatasets arise because all subdatasets use the same real-image source.

Subdataset	DDA*		REM*		Ours-setting-1	
	R	F	R	F	R	F
DALL-E 2	<u>79.6</u>	**99.5**	79.5	<u>97.5</u>	**95.6**	**99.5**
DALL-E 3	<u>79.6</u>	**100.0**	79.5	99.6	**95.6**	<u>99.9</u>
Firefly	<u>79.6</u>	**99.9**	79.5	96.4	**95.6**	<u>99.7</u>
GLIDE	<u>79.6</u>	**97.9**	79.5	94.0	**95.6**	<u>95.5</u>
Midjourney	<u>79.6</u>	**100.0**	79.5	99.6	**95.6**	<u>99.9</u>
SD 1.3	<u>79.6</u>	**100.0**	79.5	<u>99.5</u>	**95.6**	**100.0**
SD 1.4	<u>79.6</u>	**100.0**	79.5	<u>99.1</u>	**95.6**	**100.0**
SD 2	<u>79.6</u>	**100.0**	79.5	<u>99.6</u>	**95.6**	**100.0**
SDXL	<u>79.6</u>	**100.0**	79.5	<u>99.9</u>	**95.6**	**100.0**
Avg.	<u>79.6</u>	**99.7**	79.5	98.4	**95.6**	<u>99.4</u>

Table 14: Real and fake classification accuracy (%) on WildRF. R and F denote RealAcc and FakeAcc, respectively. Bold and underlined values indicate the best and second-best results among the three methods for each subdataset and metric, respectively. Avg. denotes the reported average.

Subdataset	DDA*		REM*		Ours-setting-1	
	R	F	R	F	R	F
Facebook	<u>87.5</u>	**100.0**	<u>87.5</u>	<u>97.5</u>	**93.8**	96.9
Reddit	82.1	<u>99.3</u>	<u>84.7</u>	97.9	**86.7**	**99.6**
Twitter	82.6	**99.6**	<u>86.8</u>	<u>98.2</u>	**95.8**	<u>98.2</u>
Avg.	84.1	**99.7**	<u>86.3</u>	97.8	**92.1**	<u>98.2</u>

Table 15: Ablation study of the proposed modules. Lanc denotes real-side pointwise feature anchoring, LJ denotes Local Directional Jacobian Alignment, HRS denotes heterogeneous real-sample sampling, and HRec denotes heterogeneous VAE/VQGAN reconstruction.

Lanc	LJ	HRS	HRec	ACC (%)	Real ACC (%)	Fake ACC (%)
×	×	×	×	93.3	87.8	98.9
✓	×	×	×	94.3	90.1	98.5
×	×	✓	×	94.4	90.8	97.9
×	×	×	✓	93.7	88.3	99.1
✓	✓	×	×	95.6	92.4	98.8
✓	×	✓	×	95.5	92.7	98.2
✓	×	×	✓	94.9	90.6	99.2
×	×	✓	✓	95.3	92.0	98.5
✓	✓	✓	×	96.7	94.8	98.6
✓	✓	×	✓	96.2	93.0	99.3
✓	×	✓	✓	96.1	93.4	98.7
✓	✓	✓	✓	**97.5**	**95.7**	**99.2**