Micro-Defects Expose Macro-Fakes: Detecting AI-Generated Images via Local Distributional Shifts

Anonymous Authors[1]

Abstract

Recent generative models can produce images that appear highly realistic, raising challenges in distinguishing real and AI-generated images. Yet existing detectors based on pre-trained feature extractor tend to over-rely on global semantics, limiting sensitivity to the critical micro-defects. In this work, we propose *Micro-Defects expose Macro-Fakes* (MDMF), a local distribution-aware detection framework that amplifies micro-scale statistical irregularities into macro-level distributional discrepancies. To avoid localized forensic cues being diluted by plain aggregation, we introduce a learnable *Patch Forensic Signature* that projects semantic patch embeddings into a compact forensic latent space. We then use *Maximum Mean Discrepancy* (MMD) to quantify distributional discrepancies between generated and real images. Our theory-grounded analysis shows that patch-wise modeling yields provably larger discrepancies when localized forensic signals are present in generated images, enabling more reliable separation from real images. Extensive experiments demonstrate that MDMF consistently outperforms baseline detectors across multiple benchmarks, validating its general effectiveness.

1. Introduction

Deep generative models have made rapid advances in recent years (Ho et al., 2020; Saharia et al., 2022; Podell et al., 2023; Lipman et al., 2022), with diffusion-based architectures enabling the synthesis of highly realistic images from natural language descriptions. Such advances now power widely used platforms, including Stable Diffusion (Rombach et al., 2022), DALL·E (Ramesh et al., 2022), and Midjourney. While this progress has accelerated creative high-quality content generation, it also raises significant concerns regarding misinformation (Zhou et al., 2023), deepfakes (Heidari et al., 2024), and digital forgery (Somepalli et al., 2023). As modern generative models continue to improve in visual fidelity, reliably distinguishing AI-generated images from natural images becomes increasingly challenging and essential, motivating increasing interests in AI-generated image detection (Zhu et al., 2023b; Chen et al., 2024a).

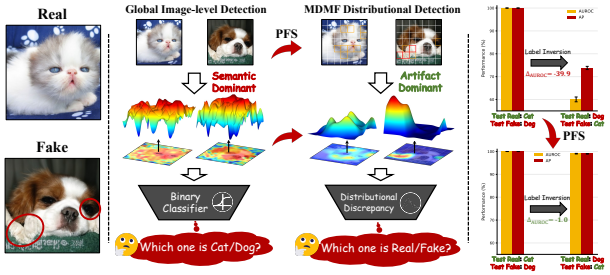


Figure 1. **Intuition behind *Patch Forensic Signature* (PFS).** ***Left:*** a real cat and a generated dog with plausible localized irregularities (highlighted). ***Middle:*** global image-level detection aggregates a **semantic-dominant** representation, inadvertently reducing real/fake detection into semantic recognition (e.g., "cat vs. dog"). PFS maps patch-wise representations into an **artifact-dominant** forensic space, making subtle generation-induced statistical deviations more salient. MDMF thus leverages their distributional discrepancy to answer "real vs. fake". ***Right:*** Label inversion stress test. Global detection suffers a sharp performance drop under inverted labels, whereas PFS remains stable, indicating PFS shifting the decision from semantics to artifacts. (see Section 2.2)

Previous studies have achieved promising progress by exploiting artifacts left by generative processes (Wang et al., 2023; Chen et al., 2024a; Ojha et al., 2023; Zhang et al., 2025b). Most approaches adopt an image-level paradigm and treat detection as global classification, either learning discriminative features with supervision (Chen et al., 2024a; Liu et al., 2024) or measuring deviations in frozen representation spaces (Ojha et al., 2023; He et al., 2024). However, as modern diffusion models increasingly leave *sparse* and *localized* forensic traces (Wang et al., 2024a; 2025), detectors built upon pre-trained representations can over-rely on global semantics, which reduces sensitivity to the micro-scale defects that are most diagnostic of generation. Several recent works have explored patch modeling to capture finer-grained cues (Zhong et al., 2023; Liu et al., 2024; Choi

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

et al., 2025). Nevertheless, when localized evidence is still summarized by plain aggregation, subtle forensic cues can remain diluted and the decision may continue to be driven by semantics rather than generation-induced irregularities. This naturally motivates a fundamental research question: *Can we learn representations that amplify micro-scale statistical irregularities into stable macro-level distributional discrepancies for robust AI-generated image detection?*

In this paper, we propose a distributional detection perspective grounded in localized forensic evidence. Concretely, instead of representing an image with a single global feature vector, we decompose it into local regions and analyze the statistics of their features. This perspective is well matched to modern generators, whose artifacts often manifest as *localized* statistical shifts that are easily suppressed by uniform aggregation into global representations. To operationalize this idea, we introduce the ***Patch Forensic Signature*** **(PFS)**, a learnable patch-level representation tailored for forensic analysis. PFS reparameterizes semantic patch embeddings into a dedicated forensic space that deemphasizes semantic content while preserving, and amplifying subtle statistical irregularities introduced by the generative process (as illustrated in Figure 1 and discussed in Section 2.2).

Based on the Patch Forensic Signature, we propose ***Micro-Defects expose Macro-Fakes*** **(MDMF)**, a distributional detection framework that transforms sparse, localized forensic artifacts into reliable image-level signals. Specifically, MDMF employs *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2012; Liu et al., 2020a) to quantify distributional discrepancy between patch-level PFS representations of test images and those of reference real images (see Section 2.3). The theoretical analysis proves that patch-wise PFS modeling provably amplifies localized defects compared to global aggregation, while the resulting empirical MMD exhibits a positive separation between real and generated images under finite samples (see Section 2.4). This analysis provides a principled explanation for why aggregating localized evidence at the distribution level leads to reliable separation, even when individual artifacts are weak.

We conduct extensive experiments to evaluate the effectiveness and generalization of MDMF. Our evaluation covers widely used benchmarks, including ImageNet (Deng et al., 2009), LSUN-Bedroom (Yu et al., 2015), and GenImage (Zhu et al., 2023b). Across them, MDMF consistently achieves strong and stable detection performance, demonstrating robustness to diverse generative architectures and training paradigms. To further stress-test the method, we conduct case studies on OpenSora-generated videos (Zheng et al., 2024), where many existing detectors degrade substantially, and show that MDMF still identifies stable forensic signals and generalizes to emerging generators in this task. In summary, our main contributions are listed as follows,

- We introduce a new perspective for AI-generated image detection, modeling images as collections of localized visual evidence and revealing that modern generative artifacts manifest as subtle statistical deviations rather than global inconsistencies. (Section 2.2)
- We propose the ***Patch Forensic Signature*** **(PFS)**, a learnable forensic representation that reparameterizes semantic embeddings into a latent space designed to suppress semantic invariances while preserving and amplifying generative artifacts. (Section 2.3)
- We develop ***Micro-Defects expose Macro-Fakes*** **(MDMF)**, a distributional detection framework that aggregates localized forensic evidence through MMD, with theoretical analysis establishing provable separation between real and generated images. Experiments across diverse benchmarks show the effectiveness and generalization of MDMF. (Sections 2.4 and 3)

2. Micro-Defects Expose Macro-Fakes.

Preliminary. Let $\mathbb{P}$ denote the distribution of real images defined on an image space $\mathcal{X} \subset \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the image height, width, and number of channels. Given i.i.d. samples $S_{\mathbb{P}} = \{x_n\}_{n=1}^{N}$ drawn from $\mathbb{P}$, the goal of AI-generated image detection is to determine whether a test image $\tilde{y}$ originates from $\mathbb{P}$ or from an alternative distribution $\mathbb{Q}$ introduced by those generative models.

2.1. Motivation

Recent advances in generative modeling have substantially reduced perceptually salient artifacts. As a result, discrepancies between real and generated images increasingly appear as sparse, localized deviations rather than global inconsistencies (Wang et al., 2024a; 2025). We refer to this regime as *Local Distributional Shifts*. Most existing approaches adopt an image-level paradigm and cast detection as global classification (Ojha et al., 2023; Chen et al., 2024a; Tan et al., 2024). However, these global representations are often dominated by semantic content, which can bias real/fake decisions toward semantic correlations rather than the localized forensic deviations that are most diagnostic of generative process.

We analyze this limitation both conceptually and empirically. Conceptually, Figure 2(a) provides a mechanistic view in which semantic content and generation artifacts jointly contribute to the observed image. Global detectors typically compress the image into a single representation before predicting real/fake, which is often shaped primarily by semantics. As a result, the detector is biased toward semantic correlations rather than the forensic evidence that motivates the real/fake detection. Empirically, we validate this semantic bias using a label inversion toy experiment, as

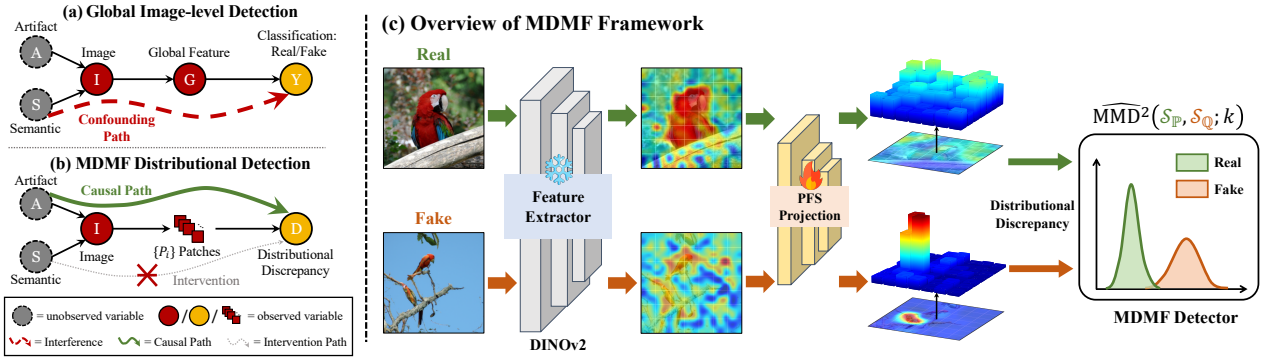


Figure 2. **Motivation and Overview of the MDMF framework.** ***(a)*** Global image-level detection compresses an image into a single feature for real/fake classification, where semantic factors can dominate the decision through a confounding path. ***(b)*** MDMF instead operates on patches and bases the prediction on distributional discrepancy, suppressing semantic interference and aligning the decision with artifact-related signals. ***(c)*** Given real and generated images, a frozen DINOv2 extracts patch representations, which are mapped by the PFS into a forensic space. MDMF then measures the discrepancy between PFS distributions to produce the final detection score.

shown in Figure 1. We train a global image-level real/fake classifier on a confounded split with *real cats* and *generated dogs*, and evaluate it on the inverted split with *real dogs* and *generated cats*. The global classifier exhibits a sharp performance drop under label inversion, indicating its prediction relies heavily on semantic cues instead of artifact evidence.

To mitigate semantic dominance, we seek a representation that weakens the influence of global semantics while retaining artifact-related cues. A natural step is to decompose an image into local patches and operate on the resulting patch representations. As illustrated in Figure 2(b), the patch-wise formulation avoids collapsing the image into a single globally pooled feature, which weakens the semantic shortcut that can confound real/fake prediction under global aggregation. However, generation-induced artifact patterns are diverse and difficult to model explicitly, and patch embeddings from standard visual backbones are still heavily influenced by semantics. This motivates us to learn a patch-wise representation that suppresses semantic dominance while preserving statistical deviations from the generation.

2.2. The Patch Forensic Signature

We introduce the ***Patch Forensic Signature*** **(PFS)**, a learnable representation that reparameterizes semantic patch embeddings into a dedicated forensic space. At a high level, PFS suppresses semantic variation and accentuates generation-induced statistical deviations, yielding signatures that align more closely with artifact-driven evidence. We next formalize PFS by first defining the extracted patch signature field and then specifying the learnable projection.

Patch Signature Field. Let $x \in \mathbb{R}^{H\times W\times C}$ be an input image. We leverage a pre-trained self-supervised vision backbone (e.g., DINOv2 (Oquab et al., 2024)) to decompose the image into a grid of K non-overlapping patch tokens:

$$\mathbf{E}(x) = \{\mathbf{e}_i(x) \in \mathbb{R}^D\}_{i=1}^K, \tag{1}$$

where D is the embedding dimension. While patch-wise modeling weakens semantic shortcuts under global aggregation, $\mathbf{e}_i(x)$ remain largely semantics-oriented, thus generative statistical cues are still not salient in this space. We then introduce a learnable reparameterization into the *forensic space*, defined as a compact latent space where semantic variation is deemphasized while patch-wise statistical deviations become more separable under the detection objective.

Definition 2.1. (**Patch Forensic Signature (PFS)**.) *Given a patch embedding $\mathbf{e}_i(x)$, we define a learnable projection function $\phi_\theta : \mathbb{R}^D \to \mathbb{R}^d$, parameterized by a lightweight Multilayer Perceptron (MLP), to map semantic embeddings into a compact forensic latent space. We refer to the mapped representation as the Patch Forensic Signature (PFS):*

$$\mathbf{z}_i(x) = \phi_\theta(\mathbf{e}_i(x)) \in \mathbb{R}^d. \tag{2}$$

Consequently, the image x is represented by a set of signature vectors $\mathbf{Z}_\theta(x) = [\mathbf{z}_1(x), \ldots, \mathbf{z}_K(x)]^\top \in \mathbb{R}^{K\times d}$. Our later experiments and analysis will show that, under a suitable learning objective (e.g., Eq. 5), this mapping plays a central role by learning to reparameterize patch-level representations into a dedicated forensic space that deemphasizes semantic variation while preserving and amplifying subtle statistical irregularities introduced by the generative process.

2.3. Exploring PFS for Detecting AI-Generated Images

PFS provides patch-wise signatures that emphasize artifact-related statistical cues, yet the resulting evidence remains spatially sparse even in the PFS space. A plain image-level pooling over PFS signatures can still average out these localized cues, making reliable detection difficult for highly realistic generations. This motivates a distributional perspective, where we compare the *distributions* of patch signatures between real and generated images to emphasize subtle statistical irregularities. To operationalize this idea,

Algorithm 1 Training MDMF

Input: Training real images $\mathcal{S}_{\mathbb{P}}^{tr}$, generated images $\mathcal{S}_{\mathbb{Q}}^{tr}$; projection head ϕ_θ; deep kernel k_ω; regularization λ; learning rate η
Initialize $\omega \leftarrow \{\theta_0, \gamma_0\}$
for $t = 1, 2, \ldots, T$ **do**
  Sample mini-batches $\{x_b\}_{b=1}^{B} \sim \mathcal{S}_{\mathbb{P}}^{tr}$ and $\{y_b\}_{b=1}^{B} \sim \mathcal{S}_{\mathbb{Q}}^{tr}$
  Form PFS vectors $\mathbf{Z}_\theta(x_b) \leftarrow [\mathbf{z}_1(x_b), \ldots, \mathbf{z}_K(x_b)]^\top$, $\mathbf{Z}_\theta(y_b) \leftarrow [\mathbf{z}_1(y_b), \ldots, \mathbf{z}_K(y_b)]^\top$
  Compute unbiased MMD $M(\omega) \leftarrow \widehat{\mathrm{MMD}}_u^2(\mathcal{S}_{\mathbb{P}}^{tr}, \mathcal{S}_{\mathbb{Q}}^{tr}; k_\omega)$ using Eqn. 3
  Estimate variance $\hat{\sigma}_{H_1}^2$ using Eqn. 6
  Optimize test-power objective $J_\lambda(\omega) \leftarrow \frac{M(\omega)}{\sqrt{\hat{\sigma}_{H_1}^2 + \lambda}}$ using Eqn. 5
  $\omega \leftarrow \omega + \eta \nabla_{\mathrm{Adam}} J_\lambda(\omega)$
end for
Output: Trained projection head ϕ_{θ^*} and kernel k_{ω^*}

Algorithm 2 Detecting Images with MDMF

Input: Reference real images $\mathcal{S}_{\mathbb{P}}^{re}$; test images $\mathcal{S}^{te}$; trained ϕ_{θ^*}; kernel k_{ω^*}; threshold τ
Build reference PFS vector $\mathbf{Z}_\theta(x)$ from $x \sim \mathcal{S}_{\mathbb{P}}^{re}$
for $\tilde{y} \in \mathcal{S}^{te}$ **do**
  $\mathbf{Z}_\theta(\tilde{y}) \leftarrow [\mathbf{z}_1(\tilde{y}), \ldots, \mathbf{z}_K(\tilde{y})]^\top$
  $S_{\mathrm{MDMF}}(\tilde{y}) \leftarrow \widehat{\mathrm{MMD}}_b^2(\mathcal{S}_{\mathbb{P}}^{re}, \{\tilde{y}\}; k_{\omega^*})$ using Eqn. 7
  $f(\tilde{y}) \leftarrow \mathbb{I}(S_{\mathrm{MDMF}}(\tilde{y}) > \tau)$ using Eqn. 8
end for
Output: Predictions $\{f(\tilde{y})\}$

we adopt the kernel two-sample testing framework via *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2012). MMD quantifies distributional discrepancy through kernel mean embeddings in a *reproducing kernel Hilbert space* (RKHS), where small but systematic deviations across local observations can accumulate into a stable image-level signal (Liu et al., 2020a). Building on PFS and MMD, we establish the ***Micro-Defects expose Macro-Fakes*** **(MDMF)** framework, which transforms sparse patch-level forensic cues into reliable detection scores, as shown in Figure 2 (c).

MMD Formulation. Consider two arbitrary sets of images $\mathcal{S}_{\mathbb{P}} = \{x_n\}_{n=1}^{N} \sim \mathbb{P}$ and $\mathcal{S}_{\mathbb{Q}} = \{y_m\}_{m=1}^{N} \sim \mathbb{Q}$. To measure the distance between distributions $\mathbb{P}$ and $\mathbb{Q}$, we employ an unbiased U-statistic estimator for the squared MMD,

$$\widehat{\mathrm{MMD}}_u^2(\mathcal{S}_{\mathbb{P}}, \mathcal{S}_{\mathbb{Q}}; k) := \frac{1}{N(N-1)} \sum_{i \neq j} H_{ij} \tag{3}$$

$$H_{ij} := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(y_i, x_j),$$

where k denotes the kernel of a RKHS. The similar $\widehat{\mathrm{MMD}}_b^2 := \frac{1}{N^2} \sum_{ij} H_{ij}$ is the squared MMD between the empirical distributions of $\mathcal{S}_{\mathbb{P}}$ and $\mathcal{S}_{\mathbb{Q}}$ (Liu et al., 2020a). According to the null hypothesis testing framework (Gretton et al., 2012), under the null hypothesis $\mathfrak{H}_0 : \mathbb{P} = \mathbb{Q}$, $\widehat{\mathrm{MMD}}_u^2(\cdot)$ should be close to zero, while strictly positive under the alternative hypothesis $\mathfrak{H}_1 : \mathbb{P} \neq \mathbb{Q}$. Leveraging this, we design following optimization and detection protocols.

Optimization Protocol. We construct $\mathcal{S}_{\mathbb{P}}^{tr}$ by aggregating real training images and $\mathcal{S}_{\mathbb{Q}}^{tr}$ from generated training images. We ideally expect to correctly reject $\mathfrak{H}_0$ and derive $\mathfrak{H}_1 : \mathbb{P} \neq \mathbb{Q}$, i.e., $\mathcal{S}_{\mathbb{P}}^{tr}$ and $\mathcal{S}_{\mathbb{Q}}^{tr}$ come from different distributions. To enhance discriminative power, we utilize a deep Gaussian kernel (Liu et al., 2020a) with bandwidth γ for MMD:

$$k_\omega(x, y) = \exp\left(-\frac{\|\mathbf{Z}_\theta(x) - \mathbf{Z}_\theta(y)\|_2^2}{2\gamma^2}\right), \tag{4}$$

Simply maximizing $\widehat{\mathrm{MMD}}_u^2$ can be problematic if the variance of the statistic also increases, leading to unstable gradients. Following the test power maximization principle (Gretton et al., 2012), we optimize the parameters $\omega = \{\theta, \gamma\}$, namely the projection weights in ϕ_θ and kernel bandwidth, to maximize the **regularized test power criterion**:

$$\max_{\omega} \; J_\lambda(\mathcal{S}_{\mathbb{P}}^{tr}, \mathcal{S}_{\mathbb{Q}}^{tr}; k_\omega) = \frac{\widehat{\mathrm{MMD}}_u^2(\mathcal{S}_{\mathbb{P}}^{tr}, \mathcal{S}_{\mathbb{Q}}^{tr}; k_\omega)}{\sqrt{\hat{\sigma}_{H_1}^2 + \lambda}}, \tag{5}$$

where $\hat{\sigma}_{H_1}^2$ is the estimated MMD variance under $\mathfrak{H}_1$:

$$\hat{\sigma}_{H_1}^2 = \frac{4}{N^3} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} H_{ij} \right)^2 - \frac{4}{N^4} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} H_{ij} \right)^2. \tag{6}$$

Detection Protocol. With the learned parameters ω^*, we apply MMD with the biased estimator to detect individual test images by quantifying their PFS distributional deviation from a set of reference images, following recent works (Zhang et al., 2024b; 2025a) demonstrate MMD's effectiveness in single-sample detection. Given a set of reference images $\{x_r\}_{r=1}^{R}$ and a test $\tilde{y}$, we compute the MDMF score:

$$S_{\mathrm{MDMF}}(\tilde{y}) = \widehat{\mathrm{MMD}}_b^2(\mathcal{S}_{\mathrm{ref}}, \{\tilde{y}\}; k_{\omega^*}) = \frac{1}{R^2} \sum_{r,r'=1}^{R} \cdot k_{\omega^*}(x^{(r)}, x^{(r')}) + k_{\omega^*}(\tilde{y}, \tilde{y}) - \frac{2}{R} \sum_{r=1}^{R} k_{\omega^*}(x^{(r)}, \tilde{y}). \tag{7}$$

Hence, we can formalize the following detection model $f(\cdot)$ to determine whether a given input image $\tilde{y}$ is generated:

$$f(\tilde{y}) = \begin{cases} \text{Generated}, & \text{if } S_{\mathrm{MDMF}}(\tilde{y}) > \tau, \\ \text{Real}, & \text{otherwise}, \end{cases} \tag{8}$$

Algorithm 1 and 2 summarize the training and testing pipelines of MDMF. While our method performs detection

by measuring distributional discrepancies via MMD, its effectiveness fundamentally relies on PFS extracting artifact-sensitive patch-level evidence that is often weakened in global image representations (see theoretical analysis in Section 2.4 and detailed empirical analysis in Section 3).

2.4. Theoretical Analysis

In this section, we provide theoretical justification for MDMF's detection mechanism. First, we show that PFS amplifies sparse localized deviations that tends to be diluted in global image-level detection (Propositions 2.4 and 2.5). Second, we establish that MMD on PFS converts this amplified shift into reliable real/fake separation (Proposition 2.6 and Theorem 2.7). We begin with the modeling assumptions.

Assumption 2.2. Real images $\{x_n\}_{n=1}^N$ are i.i.d. sampled from distribution $\mathbb{P}$, and generated images $\{y_m\}_{m=1}^N$ are i.i.d. sampled from distribution $\mathbb{Q}$. Given any real or generated image, we extract K non-overlapping patch embeddings $\{\mathbf{e}_i\}_{i=1}^K \subset \mathbb{R}^D$ using a fixed pre-trained encoder (e.g., DINOv2). Each patch embedding follows a σ_e-sub-Gaussian distribution (Wainwright, 2019) in $\mathbb{R}^D$.

Assumption 2.3 (Sparse Defect Model). For a generated image y, we assume each patch embedding is

$$\mathbf{e}_i(y) = \mathbf{u}_i + a_i s_i \boldsymbol{\mu}_{\text{defect}}, \tag{9}$$

where $\mathbf{u}_i \sim \mathcal{SG}(\mathbf{0}, \sigma_e^2 \mathbf{I}_D)$, $a_i \sim \text{Bernoulli}(\rho)$ indicates whether the patch is defective, and $s_i \in \{+1, -1\}$ is an independent Rademacher variable with $\mathcal{P}(s_i = +1) = \mathcal{P}(s_i = -1) = 1/2$. Hence $\mathbb{E}[\mathbf{e}_i(y)] = \mathbf{0}$ but defective patches elevates second-order energy. For real images, $\mathbf{e}_i(x) = \mathbf{u}_i$.

Assumption 2.2 follows common practice in representation analysis works (Wang et al., 2024b; Zhang et al., 2025a; 2024a), while Assumption 2.3 aligns with sparse-artifact observations in generated images (Wang et al., 2024a; 2025). Under these assumptions, we then establish that PFS amplifies localized defects into a detectable distributional shift.

Proposition 2.4. *Assume ϕ_θ is twice differentiable at $\mathbf{0}$ with Hessian $\mathbf{H}_\phi(\mathbf{0}) \in \mathbb{R}^{d\times D\times D}$. Under Assumption 2.3, let $\Delta_{\text{PFS}} := \mathbb{E}_{\mathbb{Q}}[\phi_\theta(\mathbf{e}_i(y))] - \mathbb{E}_{\mathbb{P}}[\phi_\theta(\mathbf{e}_i(x))]$. Then the leading-order PFS mean shift satisfies*

$$\Delta_{\text{PFS}} \approx \frac{\rho}{2}\, \mathcal{Q}(\boldsymbol{\mu}_{\text{defect}}), \tag{10}$$

where $\mathcal{Q}(\boldsymbol{\mu}) \in \mathbb{R}^d$ denotes the Hessian-induced quadratic form of ϕ_θ evaluated along direction $\boldsymbol{\mu}$, i.e., $\big[\mathcal{Q}(\boldsymbol{\mu})\big]_\ell = \boldsymbol{\mu}^\top \nabla^2 \phi_{\theta,\ell}(\mathbf{0}) \boldsymbol{\mu}$, for $\ell = 1, \ldots, d$. In particular, if $\mathcal{Q}(\boldsymbol{\mu}_{\text{defect}}) \neq 0$, then $\|\Delta_{\text{PFS}}\|_2 > 0$ for any $\rho > 0$.

Proposition 2.5. *Under Assumption 2.3 and Proposition 2.4, we define the global-pooled leading order shift as $\Delta_{\text{global}} := \mathbb{E}_{\mathbb{Q}}[\phi_\theta(\bar{\mathbf{e}}(y))] - \mathbb{E}_{\mathbb{P}}[\phi_\theta(\bar{\mathbf{e}}(x))]$, where $\bar{\mathbf{e}}(x) = \frac{1}{K}\sum_{i=1}^K \mathbf{e}_i(x)$. Then the leading-order shifts satisfy:*

$$\|\Delta_{\text{PFS}}\|_2 \approx K\, \|\Delta_{\text{global}}\|_2 > \|\Delta_{\text{global}}\|_2, \tag{11}$$

Notably, Proposition 2.5 does not imply unbounded gains as the number of patches increases. When finite-sample estimation and patch-resolution effects are taken into account, the patch advantage admits an optimal granularity, as observed in Section 3.3 and analyzed in Appendix A.4. We next quantify how the amplified PFS shift manifests as a measurable population MMD gap between $\mathbb{P}$ and $\mathbb{Q}$.

Proposition 2.6. *Let $k_\omega(\cdot,\cdot)$ be a Gaussian kernel where ω denotes the set of projection weights θ and kernel bandwidth γ. Under Proposition 2.4 and a Gaussian surrogate in PFS space, the population $\widehat{\text{MMD}}^2$ between $\mathbb{P}$ and $\mathbb{Q}$ satisfies:*

$$\widehat{\text{MMD}}^2(\mathbb{P}, \mathbb{Q}; k_\omega) = 2\left(\frac{\gamma^2}{\gamma^2 + 2\sigma_z^2}\right)^{\frac{Kd}{2}} \left[1 - \exp\left(-\frac{K\|\boldsymbol{\Delta}_{\text{PFS}}\|_2^2}{2(\gamma^2 + 2\sigma_z^2)}\right)\right], \tag{12}$$

where σ_z^2 denotes the isotropic proxy variance of the Gaussian surrogate in PFS space. $\widehat{\text{MMD}}^2(\mathbb{P}, \mathbb{Q}; k_\omega)$ is strictly positive for $\|\boldsymbol{\Delta}_{\text{PFS}}\|_2 > 0$ and is monotonically increasing.

Building on Proposition 2.6, we derive the following finite-sample concentration guarantees for practical detection.

Theorem 2.7. *Let $S_r = \{x_i\}_{i=1}^M \overset{i.i.d}{\sim} \mathbb{P}$ be a reference set of real images and $S_t = \{y_j\}_{j=1}^N$ be test images, let $\lambda = \gamma^2 + 2\sigma_z^2$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds:*

(Case I: Real test image). *If $S_t \overset{i.i.d}{\sim} \mathbb{P}$,*

$$\widehat{\text{MMD}}_u^2(S_r, S_t) \leq \underbrace{C_1(\sigma_z, \gamma)\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}}}_{\textit{Finite-sample fluctuation}}. \tag{13}$$

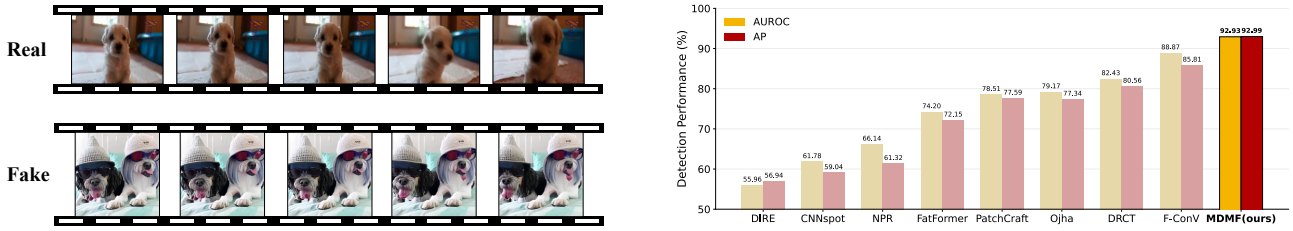
(Case II: Generated test image). *If $S_t \overset{i.i.d}{\sim} \mathbb{Q}$,*

$$\widehat{\text{MMD}}_u^2(S_r, S_t) \geq \underbrace{2\left(\frac{\gamma^2}{\lambda}\right)^{\frac{Kd}{2}} \left[1 - \exp\left(-\frac{K\|\boldsymbol{\Delta}_{\text{PFS}}\|_2^2}{2\lambda}\right)\right]}_{\textit{Artifact-induced signature shift}} - \underbrace{C_2(\sigma_z, \gamma)\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}}}_{\textit{Finite-sample error}}. \tag{14}$$

Interpretation. Theorem 2.7 establishes that the empirical MMD concentrates around its population value with deviation scaling as $O(\sqrt{1/M + 1/N})$. For real test images, the population MMD vanishes and empirical values reflect only finite-sample fluctuations. For generated images, Proposition 2.6 guarantees a positive gap scaling with $\|\boldsymbol{\Delta}_{\text{PFS}}\|_2^2$. When this population separation dominates, real images yield systematically smaller MMD scores than generated ones, justifying reliable detection for AI-generated images.

Table 1. Detection performance (%) on ImageNet. Bold numbers are superior results. We mainly compare training-based methods.

Methods	ADM		ADMG		LDM		DiT		BigGAN		GigaGAN		StyleGAN XL		RQ-Transformer		Mask GIT		Average	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC (↑)	AP (↑)
CNNspot (Wang et al., 2020)	62.25	63.13	63.28	62.27	63.16	64.81	62.85	61.16	85.71	84.93	74.85	71.45	68.41	68.67	61.83	62.91	60.98	61.69	67.04	66.78
Ojha (Ojha et al., 2023)	83.37	82.95	79.60	78.15	80.35	79.71	82.93	81.72	93.07	92.77	87.45	84.88	85.36	83.15	85.19	84.22	90.82	90.71	85.35	84.25
DIRE (Wang et al., 2023)	51.82	50.29	53.14	52.96	52.83	51.84	54.67	55.10	51.62	50.83	50.70	50.27	50.95	51.36	55.95	54.83	52.58	52.10	52.70	52.18
NPR (Tan et al., 2024)	85.68	80.86	84.34	79.79	91.98	86.96	86.15	81.26	89.73	84.46	82.21	78.20	84.13	78.73	80.21	73.21	89.61	84.15	86.00	80.84
PatchCraft (Zhong et al., 2023)	81.83	79.65	70.88	69.36	68.47	65.19	75.38	73.29	99.85	99.26	98.55	97.91	96.33	96.25	91.28	91.47	92.56	92.17	86.13	84.95
DRCT (Chen et al., 2024a)	90.26	90.07	85.74	83.85	90.24	89.88	88.27	89.06	95.87	94.99	86.89	86.12	89.11	88.39	92.38	92.41	94.44	94.47	90.36	89.92
FatFormer (Liu et al., 2024)	91.77	90.36	83.58	83.17	92.58	92.06	86.93	85.14	98.76	98.47	97.65	98.02	97.64	97.57	96.55	95.96	97.65	97.27	93.68	93.11
F-ConV (Zhang et al., 2025b)	**92.74**	91.65	88.51	87.67	88.87	88.47	85.94	84.88	98.94	98.98	98.14	98.72	98.52	98.38	96.79	96.33	95.52	95.38	93.77	93.38
MDMF	92.56	**93.57**	**88.86**	**90.16**	**94.63**	**97.35**	**88.89**	**94.48**	**99.93**	**99.94**	**98.99**	**99.52**	**98.76**	**99.41**	**98.84**	**99.46**	**99.40**	**99.72**	**95.65**	**97.07**



(a) Examples of *Real* and *Fake* videos

(b) Detection Performance on OpenSora

Figure 3. Examples visualization and performance comparison on OpenSora.

3. Experiments

3.1. Experimental Setup

We provide detailed experimental setups in Appendix B.

Datasets. Following previous works (Wang et al., 2020; Zhang et al., 2025b), we evaluate our MDMF on the following benchmarks: **ImageNet** (Deng et al., 2009), **LSUN-Bedroom** (Yu et al., 2015), **GenImage** (Zhu et al., 2023b). To further assess generalization to generators beyond image benchmarks, we additionally conduct a case study on videos generated by **OpenSora** (Zheng et al., 2024). Specifically, we sample 3,275 generated videos and extract 10 frames per video, resulting in 32,750 frames and treat them as generated images. For real data, we sample the same number of natural videos and frames on MSR-VTT (Xu et al., 2016).

Baselines and Evaluation Metrics. We compare our MDMF with the following training-based detection baselines in the main experiments: CNNspot (Wang et al., 2020), Ojha (Ojha et al., 2023), DIRE (Wang et al., 2023), PatchCraft (Zhong et al., 2023), NPR (Tan et al., 2024), DRCT (Chen et al., 2024a), FatFormer (Liu et al., 2024), F-ConV (Zhang et al., 2025b). Following (Zhang et al., 2025b), we adopt the following metrics: (1) average precision (AP); (2) area under the receiver operating characteristic curve (AUROC) and (3) classification accuracy (ACC).

Implementation Details. Following previous studies (Ojha et al., 2023; Liu et al., 2024), we apply random cropping and random horizontal flipping at training, while center cropping at testing, both with no other augmentations. To balance detection performance and efficiency, we adopt DINOv2 ViT-L/14 (Oquab et al., 2024) to extract patch embeddings and pool the patch size to $W = 32$ for PFS computation in main experiments. The projection ϕ_θ and kernel bandwidth γ are jointly trained during optimization.

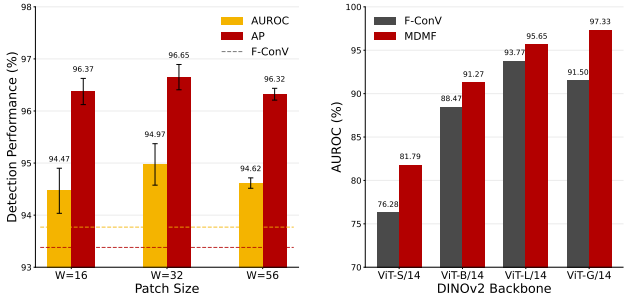
3.2. Main Results

Detection performance comparison with baselines. Table 1 reports detection performance on the ImageNet benchmark across nine generative models spanning diffusion models, GANs, and autoregressive transformers. MDMF demonstrates consistently strong performance across all evaluated generators, indicating robust generalization under diverse generative mechanisms. Notably, MDMF shows particularly strong performance on recent diffusion-based models, which are known to produce highly realistic images with sparse and localized artifacts that challenge existing detectors. These results validate that our PFS distributional modeling effectively captures the subtle, localized forensic signals characteristic of modern generative paradigms. Beyond diffusion models, MDMF also maintains competitive performance on earlier generative paradigms. This consistent behavior further demonstrates MDMF effectively captures generator-agnostic forensic artifacts and amplifies micro-scale defects into robust macro-level detection signals across both emerging and conventional generative models.

Case Study on OpenSora-Generated Content. We further evaluate MDMF on a challenging case study using frames sampled from OpenSora (Zheng et al., 2024), a recent video generation model that is not seen during training. Figure 3(a) shows the advanced diffusion-generated videos with strong temporal consistency introduced by OpenSora, resulting in frames that are globally coherent and largely free of perceptual artifacts. As illustrated in Figure 3(b), while several competitive baselines exhibit notable performance degradation, MDMF still maintains robust detection performance

Table 2. **Ablation study of key components on ImageNet.** Variants without MMD are trained with a BCE objective, while PFS modeling without MMD pluses a lightweight attention head for aggregation. See Appendix C.2 for detailed implementations.

PFS Modeling	MMD Optimization	Average	
		AUROC (↑)	AP (↑)
Global Pooling			
✗	✗	90.14	93.33
✗	✓	86.53	92.18
PFS Modeling			
✓	✗	93.22	95.34
✓	✓	**95.65**	**97.07**



(a) Diff. Patch Size (b) Diff. DINOv2 Backbone

Figure 4. **Further analysis of design choices.** (a) Sensitivity to patch size W; (b) Robustness to DINOv2 backbone variants.

on OpenSora-generated frames. This contrast indicates that the distributional modeling of MDMF captures localized forensic signatures that persist even under substantial domain shifts, enabling effective generalization to emerging video generation paradigms that are unseen during training.

3.3. Ablation and Further Analysis

We provide detailed results and discussions in Appendix C.

Ablation Study of Core Components in MDMF. Table 2 analyzes the contribution of PFS modeling and MMD optimization in MDMF. For variants without MMD, we train binary classifiers with a standard BCE loss. In particular, the PFS w/o MMD variant uses a lightweight attention head to pool patch-wise PFS into an image-level score (details in the Appendix B). Notably, even without MMD optimization, attention aggregated PFS still achieves competitive performance, indicating that the forensic reparameterization already suppresses semantic dominance and highlights generation-related cues. In contrast, the effect of MMD is strongly dependent on the underlying representation. When applied to global pooling, MMD fails to yield performance gains, whereas combining MMD with PFS leads to a clear improvement. This is consistent with MMD serving as a complementary amplifier when patch-wise forensic evidence is preserved in the PFS space, rather than when it is diluted by the semantics-dominant global representations.

Effect of Patch Granularity. To evaluate the effect of patch granularity, we vary the patch size W while keeping other settings fixed. Figure 4(a) shows a non-monotonic dependence on W. While always above baselines, performance improves as W increases from small values and degrades when W becomes overly large. This trend supports Proposition 2.5, which predicts a finite optimal granularity rather than a monotonic behavior. Coarse partitions (e.g., $W = 56$) provide insufficient spatial resolution to capture sparse local shifts, while fine partitions (e.g., $W = 16$) weaken the forensic evidence within each patch and introduce higher variance in distributional comparison. These results indicate that PFS modeling benefits from an intermediate granularity that balances localized sensitivity with reliable estimation.

Robustness to Encoder Architecture. To evaluate sensitivity to the feature extractor, we instantiate MDMF with multiple DINOv2 backbone variants and compare it with F-ConV under the same setting. As shown in Figure 4(b), MDMF consistently achieves higher detection performance across all evaluated encoders, demonstrating robustness to the underlying backbone choice. MDMF maintains an advantage with smaller backbones (e.g., ViT-S/14), and continues to improve as the backbone is scaled up, whereas F-ConV exhibits non-monotonic behavior and degrades on the largest encoder (e.g., ViT-G/14). This contrast suggests that the proposed PFS representation provides robust forensic signals that transfer effectively across encoder scales, leading to a more stable behavior under backbone scaling.

Qualitative Visualization of Localized Forensic Cues. To better understand how MDMF detects highly realistic samples, Figure 5 visualizes representative real images and category-matched generated images produced by ADM, together with Grad-CAM heatmaps from different models where warmer colors indicate a higher predicted probability of being fake. First, we can observe that the generated images exhibit strong semantic coherence and high visual fidelity. Consistently, the global pooling visualization primarily highlights semantically salient regions, such as object boundaries and high-contrast textures, indicating similar patterns for real and generated images and limited sensitivity to sparse local artifacts. In contrast, MDMF produces more localized responses on generated images and assigns higher activation to regions that plausibly contain subtle generative irregularities, while producing more diffuse patterns on real images. This reflects a pronounced distributional discrepancy between real and generated samples in the PFS space, which provides a strong basis for MMD to elaborate a stable image-level detection signal. These visualizations align with our theoretical analysis and quantitative results, suggesting MDMF can surface localized forensic evidence that is suppressed by semantic-dominated global features.

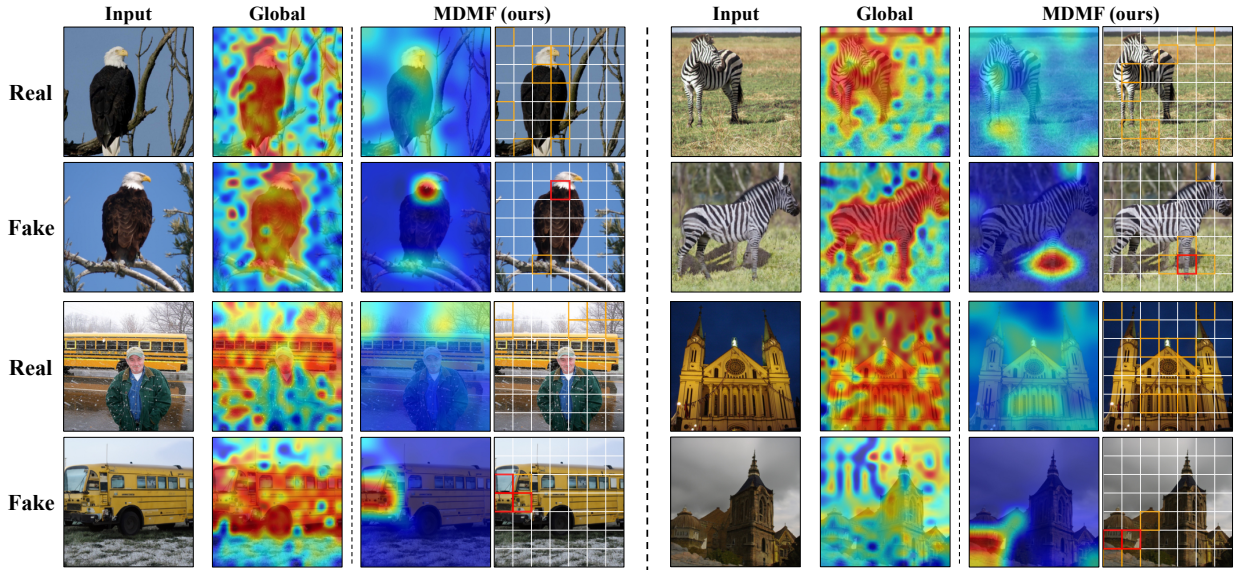


Figure 5. **Qualitative visualization of localized forensic evidence.** We compare representative real images and category-matched generated images with Grad-CAM, where warmer colors indicate higher predicted likelihood of being fake. Global-pooling baseline primarily highlights semantically salient regions with similar patterns for real and generated samples, whereas MDMF shows localized responses on generated images and diffuse activations on real images, consistent with capturing subtle generation-induced irregularities.

4. Related Work

4.1. Generative Models for Image Generation

Early image generation methods, including GANs (Brock, 2018; Karras et al., 2019), VAEs (Kingma & Welling, 2013; Sohn et al., 2015), have established the foundation of modern generative models but often exhibited visible artifacts. Diffusion models have since become the dominant paradigm, achieving strong fidelity (Ho et al., 2020; Saharia et al., 2022). Representative diffusion families include DDPM (Ho et al., 2020; Nichol & Dhariwal, 2021), ADM (Dhariwal & Nichol, 2021), LDM (Rombach et al., 2022), SDXL (Podell et al., 2023), and DiT (Peebles & Xie, 2023). These advances have also enabled widely deployed text-to-image systems like Glide (Nichol et al., 2021), Wukong (Gu et al., 2022), and Midjourney. Recent video-generation systems such as Sora (Brooks et al., 2024) and OpenSora (Zheng et al., 2024) further raise generation quality, which can produce individual frames as challenging synthetic images. As generative models evolve, artifacts become *weak* and *sparse* (Wang et al., 2024a; 2025), which motivates us to amplify localized distributional deviations.

4.2. AI-Generated Image Detection

The rapid improvement of generative models has created an urgent demand for reliable AI-generated image detection. Early detectors mainly train image-level binary classifiers, as exemplified by CNNSpot (Wang et al., 2020). To better generalize across unseen generators, Ojha (Ojha et al., 2023) trains detectors in CLIP space for transfer, DIRE (Wang et al., 2023) uses diffusion reconstruction error as a detection feature. DRCT (Chen et al., 2024a) learns from diffusion reconstructions and contrastive hard samples to enhance robustness, F-ConV (Zhang et al., 2025b) exploits manifold geometry with flow-based extrusion. Motivated by the increasing sparsity of generative artifacts, some methods shift to patch-level evidence. PatchCraft (Zhong et al., 2023) enhances texture traces via smash and reconstruction, FatFormer (Liu et al., 2024) adapts CLIP features with a forgery-aware transformer. However, they still summarize patch evidence via plain aggregation, diluting sparse forensic cues. We instead learn patch forensic signatures and measure the distributional discrepancy for robust detection.

5. Conclusion

In this paper, we present a distributional perspective for AI-generated image detection by modeling an image as a collection of localized visual evidence. Building on this view, we introduce ***Patch Forensic Signature*** **(PFS)**, a learnable forensic representation that reparameterizes semantic embeddings into a latent space to suppress semantic invariances while amplifying generative artifacts. We further propose ***Micro-Defects expose Macro-Fakes*** **(MDMF)**, which measures distributional discrepancy over PFS via MMD to aggregate localized evidence into stable image-level detection signals, and we provide theoretical analysis that establishes the advantage of PFS and the separation between real and generated images. Extensive experiments on multiple benchmarks with detailed ablations and analyses demonstrate the effectiveness and generalization of MDMF.

Impact Statement

This paper advances AI-generated image detection by modeling localized forensic cues and amplifying them via a distributional perspective. The potential positive impact is to support content integrity, provenance, and moderation by helping identify synthetic imagery, which may reduce misuse such as misinformation and fraud. However, detection methods can incentivize evasion and contribute to an arms race between generation and detection. In addition, false positives and false negatives may have downstream costs in real-world settings, especially under distribution shift or post-processing. We therefore recommend careful evaluation across domains and deployment with human oversight, ideally alongside complementary provenance mechanisms such as watermarking or metadata-based verification. Overall, we believe the broader impacts are consistent with established synthetic media detection research.

References

Brock, A. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Chen, B., Zeng, J., Yang, J., and Yang, R. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024a.

Chen, H., Hong, Y., Huang, Z., Xu, Z., Gu, Z., Li, Y., Lan, J., Zhu, H., Zhang, J., Wang, W., et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024b.

Choi, S., Lee, H., and Lee, M. Training-free detection of ai-generated images via cropping robustness. *arXiv preprint arXiv:2511.14030*, 2025.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Gu, J., Meng, X., Lu, G., Hou, L., Minzhe, N., Liang, X., Yao, L., Huang, R., Zhang, W., Jiang, X., et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022.

He, Z., Chen, P.-Y., and Ho, T.-Y. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112*, 2024.

Heidari, A., Jafari Navimipour, N., Dag, H., and Unal, M. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2):e1520, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pp. 6316–6326. PMLR, 2020a.

Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., and Zhao, Y. 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.

Liu, Z., Qi, X., and Torr, P. H. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8060–8069, 2020b.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

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

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6048–6058, 2023.

Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., and Wei, Y. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, K., Zhang, L., and Zhang, J. Detecting human artifacts from text-to-image models. *arXiv preprint arXiv:2411.13842*, 2024a.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

Wang, Y., Zhang, P., Yang, B., Wong, D., Zhang, Z., and Wang, R. Embedding trajectory for out-of-distribution detection in mathematical reasoning. *Advances in Neural Information Processing Systems*, 37:42965–42999, 2024b.

Wang, Y., Chen, X., Xu, X., Ji, S., Liu, Y., Shen, Y., and Zhao, H. Diffdoctor: Diagnosing image diffusion models before treating. *arXiv preprint arXiv:2501.12382*, 2025.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Xu, J., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhang, B., Zhu, J., Wang, Z., Liu, T., Du, B., and Han, B. What if the input is expanded in ood detection? *Advances in Neural Information Processing Systems*, 37:21289–21329, 2024a.

Zhang, S., Song, Y., Yang, J., Li, Y., Han, B., and Tan, M. Detecting machine-generated texts by multi-population aware optimization for maximum mean discrepancy. *arXiv preprint arXiv:2402.16041*, 2024b.

Zhang, S., Lian, Z., Yang, J., Li, D., Pang, G., Liu, F., Han, B., Li, S., and Tan, M. Physics-driven spatiotemporal modeling for ai-generated video detection. *arXiv preprint arXiv:2510.08073*, 2025a.

Zhang, X., Karaman, S., and Chang, S.-F. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019.

Zhang, Y., Nie, J., Tian, X., Gong, M., Zhang, K., and Han, B. Detecting generated images by fitting natural image distributions. *arXiv preprint arXiv:2511.01293*, 2025b.

Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., and You, Y. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Zhong, N., Xu, Y., Li, S., Qian, Z., and Zhang, X. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., and De Choudhury, M. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–20, 2023.

Zhu, M., Chen, H., Huang, M., Li, W., Hu, H., Hu, J., and Wang, Y. Gendet: Towards good generalizations for ai-generated image detection. *arXiv preprint arXiv:2312.08880*, 2023a.

Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., and Wang, Y. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36: 77771–77782, 2023b.

Reproducibility Statement

To facilitate reproducibility, we summarize key experimental details and provide the necessary resources in the submitted supplementary materials.

- **Datasets.** All benchmarks used in this paper are publicly available. We evaluate on ImageNet (Deng et al., 2009), LSUN-Bedroom (Yu et al., 2015), and GenImage (Zhu et al., 2023b) following standard protocols in prior AI-generated image detection works (Zhang et al., 2025b). For our stress test, we construct an OpenSora-generated dataset by sampling videos from the GenVideo's (Chen et al., 2024b) OpenSora (Zheng et al., 2024) subset and extracting frames, and use MSR-VTT (Xu et al., 2016) as the corresponding real-video source (details in Appendix B.1).
- **Assumption.** Our method follows the common *training-based* detection setting adopted by prior detectors (Chen et al., 2024a; Liu et al., 2024), where a detector is trained on a designated training set and then evaluated on multiple generators and benchmarks for generalization. We keep the training pipeline consistent across all experiments.
- **Open source.** We include our source code in the submitted supplementary materials. The release contains training and evaluation scripts and pretrained checkpoints where applicable to reproduce our results.
- **Environment.** Experiments are conducted on a single NVIDIA H200 GPU using Python 3.10.19 and PyTorch 2.9.1. Key hyperparameters (optimizer, learning rate, batch size, epochs, patch granularity, etc.) are reported in Appendix B.3.

A. Theoretical Analysis

A.1. Preliminaries and Modeling Assumptions

This section provides the probabilistic tools and regularity conditions used in our proofs. We formalize (i) sub-Gaussian patch embeddings, (ii) weak spatial dependence across patches (used only when relating patchwise and global pooling), and (iii) second-order regularity of the PFS mapping.

Sub-Gaussian random vectors. A random vector $X \in \mathbb{R}^m$ is called *σ-sub-Gaussian* if for all unit vectors $u \in \mathbb{S}^{m-1}$ and all $t \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\big(t\,u^\top (X - \mathbb{E}[X])\big)\Big] \le \exp\left(\frac{\sigma^2 t^2}{2}\right). \tag{15}$$

We denote by $\mathcal{SG}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ a σ-sub-Gaussian distribution with mean $\boldsymbol{\mu}$ and isotropic proxy covariance $\sigma^2 \mathbf{I}$.

Extracted patch embeddings. Real images are i.i.d. from $\mathbb{P}$ and generated images are i.i.d. from $\mathbb{Q}$. Given an image, we extract K non-overlapping patch embeddings $\{\mathbf{e}_i\}_{i=1}^K \subset \mathbb{R}^D$ from a fixed pre-trained encoder (e.g., DINOv2). Throughout the analysis, we assume each patch embedding is σ_e-sub-Gaussian:

$$\mathbf{e}_i(x) \sim \mathcal{SG}(\mathbf{0}, \sigma_e^2 \mathbf{I}_D), \qquad \mathbf{e}_i(y) \text{ follows the sparse-defect model in Assumption 2.3.}$$

Weak spatial dependence across patches. Within one image, patch embeddings may exhibit spatial correlation. To quantify this, we model $\{\mathbf{e}_i\}_{i=1}^K$ as an α-mixing sequence. Let $\mathcal{F}_1^i$ be the σ-algebra generated by $\{\mathbf{e}_1, \dots, \mathbf{e}_i\}$ and $\mathcal{F}_{i+\ell}^K$ generated by $\{\mathbf{e}_{i+\ell}, \dots, \mathbf{e}_K\}$. The α-mixing coefficient is

$$\alpha(\ell) := \sup_i \sup_{A \in \mathcal{F}_1^i,\, B \in \mathcal{F}_{i+\ell}^K} \big|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\big|. \tag{16}$$

We assume exponential mixing:

$$\alpha(\ell) \le C_\alpha e^{-c_\alpha \ell} \quad \text{for some constants } C_\alpha, c_\alpha > 0. \tag{17}$$

This assumption is only used to control covariance shrinkage after global pooling.

Effective sample size. Define an *effective* patch count

$$\frac{1}{K_{\text{eff}}} := \frac{1}{K} + \frac{2}{K^2}\sum_{\ell=1}^{K-1}(K-\ell)\,\beta(\ell), \tag{18}$$

where $\beta(\ell)$ upper-bounds cross-patch covariance contribution at lag ℓ (e.g., $\beta(\ell) \propto \alpha(\ell)^\eta$ for some $\eta \in (0,1]$ under standard mixing-to-covariance bounds). Under exponential mixing (17), $\sum_{\ell\geq 1}\beta(\ell) < \infty$ and thus $K_{\text{eff}} = \Theta(K)$ (i.e., it scales linearly with K up to constants).

PFS mapping and second-order regularity. Let $\phi_\theta : \mathbb{R}^D \to \mathbb{R}^d$ be the learnable patch forensic signature (PFS) mapping, and write $\phi_\theta = (\phi_{\theta,1}, \ldots, \phi_{\theta,d})$.

Assumption A.1 (Locally Smooth PFS Mapping (Second-order)). There exist constants $L, M, R > 0$ and a neighborhood $\mathcal{E} \subset \mathbb{R}^D$ containing the typical support mass of both real and generated patch embeddings such that for all $\mathbf{e} \in \mathcal{E}$,

$$\|J_\phi(\mathbf{e})\|_{\text{op}} \leq L, \qquad \|\nabla^2\phi_{\theta,\ell}(\mathbf{e})\|_{\text{op}} \leq M \;\text{ for all } \ell = 1,\ldots,d, \tag{19}$$

and the second-order Taylor remainder satisfies, for each ℓ,

$$\left|R_\ell(\mathbf{e})\right| \leq \frac{R}{6}\|\mathbf{e}\|_2^3, \tag{20}$$

where $R_\ell(\mathbf{e})$ is the remainder term in the expansion of $\phi_{\theta,\ell}(\mathbf{e})$ around $\mathbf{0}$.

Remark A.2. Assumption A.1 is mild when ϕ_θ is implemented with smooth activations (e.g., GELU or $\tanh$) and embeddings are ℓ_2-normalized, which effectively restricts $\mathbf{e}$ to a compact region. The Hessian bound ensures the second-order term is controlled, and (20) formalizes that higher-order terms are negligible in the leading-order analysis.

A.2. Proof of Proposition 2.4

Proof. We prove (10) by a second-order Taylor expansion.

Recall the definition of Δ_{PFS}:

$$\Delta_{\text{PFS}} := \mathbb{E}_{\mathbb{Q}}[\phi_\theta(\mathbf{e}_i(y))] - \mathbb{E}_{\mathbb{P}}[\phi_\theta(\mathbf{e}_i(x))]\,.$$

Under Assumption 2.3, we have

$$\mathbf{e}_i(x) = \mathbf{u}_i, \qquad \mathbf{e}_i(y) = \mathbf{u}_i + a_i s_i \boldsymbol{\mu}_{\text{defect}},$$

where $\mathbf{u}_i \sim \mathcal{SG}(\mathbf{0}, \sigma_e^2\mathbf{I}_D)$, $a_i \sim \text{Bernoulli}(\rho)$, and s_i is Rademacher independent of $\mathbf{u}_i, a_i$.

First, we compute the mean of generated patch embeddings:

$$\begin{aligned}\mathbb{E}[\mathbf{e}_i(y)] &= \mathbb{E}[\mathbf{u}_i] + \mathbb{E}[a_i s_i]\,\boldsymbol{\mu}_{\text{defect}} \\ &= \mathbf{0} + \mathbb{E}[a_i]\mathbb{E}[s_i]\,\boldsymbol{\mu}_{\text{defect}} \qquad \text{(by independence of } a_i \text{ and } s_i) \\ &= \rho \cdot 0 \cdot \boldsymbol{\mu}_{\text{defect}} = \mathbf{0},\end{aligned}$$

and similarly $\mathbb{E}[\mathbf{e}_i(x)] = \mathbb{E}[\mathbf{u}_i] = \mathbf{0}$. Hence any leading-order shift cannot arise from the linear (Jacobian) term.

Compute Second-order Taylor expansion of ϕ_θ. Let $\phi_\theta = (\phi_{\theta,1}, \ldots, \phi_{\theta,d})$. For each output coordinate $\ell \in \{1,\ldots,d\}$, since ϕ_θ is twice differentiable at $\mathbf{0}$, a second-order Taylor expansion around $\mathbf{0}$ yields

$$\phi_{\theta,\ell}(\mathbf{e}) = \phi_{\theta,\ell}(\mathbf{0}) + \nabla\phi_{\theta,\ell}(\mathbf{0})^\top\mathbf{e} + \frac{1}{2}\mathbf{e}^\top\nabla^2\phi_{\theta,\ell}(\mathbf{0})\mathbf{e} + R_\ell(\mathbf{e}), \tag{21}$$

where the remainder $R_\ell(\mathbf{e}) = o(\|\mathbf{e}\|_2^2)$ as $\|\mathbf{e}\|_2 \to 0$.

Apply (21) to $\mathbf{e} = \mathbf{e}_i(y)$ and $\mathbf{e} = \mathbf{e}_i(x)$ and take expectations:

$$\mathbb{E}_{\mathbb{Q}}\big[\phi_{\theta,\ell}(\mathbf{e}_i(y))\big] = \phi_{\theta,\ell}(\mathbf{0}) + \nabla\phi_{\theta,\ell}(\mathbf{0})^\top\mathbb{E}[\mathbf{e}_i(y)] + \frac{1}{2}\mathbb{E}\big[\mathbf{e}_i(y)^\top\nabla^2\phi_{\theta,\ell}(\mathbf{0})\mathbf{e}_i(y)\big] + \mathbb{E}[R_\ell(\mathbf{e}_i(y))], \tag{22}$$

$$\mathbb{E}_{\mathbb{P}}\big[\phi_{\theta,\ell}(\mathbf{e}_i(x))\big] = \phi_{\theta,\ell}(\mathbf{0}) + \nabla\phi_{\theta,\ell}(\mathbf{0})^\top\mathbb{E}[\mathbf{e}_i(x)] + \frac{1}{2}\mathbb{E}\big[\mathbf{e}_i(x)^\top\nabla^2\phi_{\theta,\ell}(\mathbf{0})\mathbf{e}_i(x)\big] + \mathbb{E}[R_\ell(\mathbf{e}_i(x))]. \tag{23}$$

Subtracting (23) from (22), the constant terms cancel. Since $\mathbb{E}[\mathbf{e}_i(y)] = \mathbb{E}[\mathbf{e}_i(x)] = \mathbf{0}$, so the linear terms also vanish. Therefore,

$$\Delta_{\mathrm{PFS},\ell} := \mathbb{E}_{\mathbb{Q}}\big[\phi_{\theta,\ell}(\mathbf{e}_i(y))\big] - \mathbb{E}_{\mathbb{P}}\big[\phi_{\theta,\ell}(\mathbf{e}_i(x))\big] = \frac{1}{2}\Big(\mathbb{E}[\mathbf{e}_i(y)^\top H_\ell \mathbf{e}_i(y)] - \mathbb{E}[\mathbf{e}_i(x)^\top H_\ell \mathbf{e}_i(x)]\Big) + \epsilon_\ell, \tag{24}$$

where we write $H_\ell := \nabla^2 \phi_{\theta,\ell}(\mathbf{0})$ and group the remainder difference into

$$\epsilon_\ell := \mathbb{E}[R_\ell(\mathbf{e}_i(y))] - \mathbb{E}[R_\ell(\mathbf{e}_i(x))], \quad \text{which is higher order.}$$

We now compute the difference $\mathbb{E}[\mathbf{e}^\top H_\ell \mathbf{e}]$ under x and y.

Under the real distribution, $\mathbf{e}_i(x) = \mathbf{u}_i$:

$$\mathbb{E}[\mathbf{e}_i(x)^\top H_\ell \mathbf{e}_i(x)] = \mathbb{E}[\mathbf{u}_i^\top H_\ell \mathbf{u}_i].$$

Under the generated distribution, $\mathbf{e}_i(y) = \mathbf{u}_i + a_i s_i \boldsymbol{\mu}_{\mathrm{defect}}$:

$$\begin{aligned}\mathbb{E}[\mathbf{e}_i(y)^\top H_\ell \mathbf{e}_i(y)] &= \mathbb{E}\Big[(\mathbf{u}_i + a_i s_i \boldsymbol{\mu}_{\mathrm{defect}})^\top H_\ell (\mathbf{u}_i + a_i s_i \boldsymbol{\mu}_{\mathrm{defect}})\Big] \\ &= \mathbb{E}[\mathbf{u}_i^\top H_\ell \mathbf{u}_i] + 2\,\mathbb{E}[a_i s_i]\,\mathbb{E}[\boldsymbol{\mu}_{\mathrm{defect}}^\top H_\ell \mathbf{u}_i] + \mathbb{E}[(a_i s_i)^2]\,\boldsymbol{\mu}_{\mathrm{defect}}^\top H_\ell \boldsymbol{\mu}_{\mathrm{defect}}. \end{aligned} \tag{25}$$

Here we used bilinearity of the quadratic expansion and independence to separate expectations in the cross term, so we have

$$\mathbb{E}[a_i s_i] = \mathbb{E}[a_i]\mathbb{E}[s_i] = \rho \cdot 0 = 0,$$

so the cross term in (25) vanishes. Moreover, since $s_i^2 = 1$ and $a_i \in \{0, 1\}$, we have $(a_i s_i)^2 = a_i$ and therefore

$$\mathbb{E}[(a_i s_i)^2] = \mathbb{E}[a_i] = \rho.$$

Plugging these into (25) yields

$$\begin{aligned}\mathbb{E}[\mathbf{e}_i(y)^\top H_\ell \mathbf{e}_i(y)] &= \mathbb{E}[\mathbf{u}_i^\top H_\ell \mathbf{u}_i] + \rho\,\boldsymbol{\mu}_{\mathrm{defect}}^\top H_\ell \boldsymbol{\mu}_{\mathrm{defect}} \\ &= \mathbb{E}[\mathbf{e}_i(x)^\top H_\ell \mathbf{e}_i(x)] + \rho\,\boldsymbol{\mu}_{\mathrm{defect}}^\top H_\ell \boldsymbol{\mu}_{\mathrm{defect}}\end{aligned}$$

Hence,

$$\mathbb{E}[\mathbf{e}_i(y)^\top H_\ell \mathbf{e}_i(y)] - \mathbb{E}[\mathbf{e}_i(x)^\top H_\ell \mathbf{e}_i(x)] = \rho\,\boldsymbol{\mu}_{\mathrm{defect}}^\top H_\ell \boldsymbol{\mu}_{\mathrm{defect}}. \tag{26}$$

Substituting (26) into (24) and ignoring higher-order remainders ϵ_ℓ gives the leading-order approximation

$$\Delta_{\mathrm{PFS},\ell} \approx \frac{\rho}{2}\,\boldsymbol{\mu}_{\mathrm{defect}}^\top \nabla^2 \phi_{\theta,\ell}(\mathbf{0}) \boldsymbol{\mu}_{\mathrm{defect}} = \frac{\rho}{2}\,[\mathcal{Q}(\boldsymbol{\mu}_{\mathrm{defect}})]_\ell, \qquad \ell = 1, \ldots, d.$$

Stacking $\ell = 1, \ldots, d$ proves

$$\Delta_{\mathrm{PFS}} \approx \frac{\rho}{2}\,\mathcal{Q}(\boldsymbol{\mu}_{\mathrm{defect}}),$$

which is exactly (10).

If $\mathcal{Q}(\boldsymbol{\mu}_{\mathrm{defect}}) \neq 0$ and $\rho > 0$, then

$$\Delta_{\mathrm{PFS}} \approx \frac{\rho}{2}\,\mathcal{Q}(\boldsymbol{\mu}_{\mathrm{defect}}) \neq 0,$$

and thus $\|\Delta_{\mathrm{PFS}}\|_2 > 0$, which completes the proof. □

A.3. Proof of Proposition 2.5

Proof. We show that global pooling dilutes the same second-order defect signature by a factor $1/K$, hence $\|\Delta_{\mathrm{PFS}}\|_2 \approx K\|\Delta_{\mathrm{global}}\|_2$ at leading order.

Recall the definition of the global pooled embeddings:

$$\bar{\mathbf{e}}(x) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{e}_i(x), \qquad \bar{\mathbf{e}}(y) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{e}_i(y).$$

Under Assumption 2.3,

$$\mathbf{e}_i(x) = \mathbf{u}_i, \qquad \mathbf{e}_i(y) = \mathbf{u}_i + a_i s_i \boldsymbol{\mu}_{\text{defect}}.$$

Similar to the proof of Proposition 2.4, if we simply adopt the linearity of expectation:

$$\mathbb{E}[\bar{\mathbf{e}}(x)] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}[\mathbf{u}_i] = \mathbf{0}, \qquad \mathbb{E}[\bar{\mathbf{e}}(y)] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}[\mathbf{u}_i + a_i s_i \boldsymbol{\mu}_{\text{defect}}] = \mathbf{0}.$$

Thus, as in Proposition 2.4, the leading-order shift arises from second-order terms. For each coordinate ℓ, we apply the same second-order expansion at $\mathbf{0}$:

$$\phi_{\theta,\ell}(\bar{\mathbf{e}}) = \phi_{\theta,\ell}(\mathbf{0}) + \nabla\phi_{\theta,\ell}(\mathbf{0})^\top \bar{\mathbf{e}} + \frac{1}{2}\bar{\mathbf{e}}^\top H_\ell \bar{\mathbf{e}} + R_\ell(\bar{\mathbf{e}}), \quad H_\ell = \nabla^2\phi_{\theta,\ell}(\mathbf{0}).$$

According to Appendix A.2, taking expectations and subtracting between y and x, the constant and linear terms cancel since $\mathbb{E}[\bar{\mathbf{e}}(y)] = \mathbb{E}[\bar{\mathbf{e}}(x)] = \mathbf{0}$:

$$\Delta_{\text{global},\ell} := \mathbb{E}_{\mathbb{Q}}\big[\phi_{\theta,\ell}(\bar{\mathbf{e}}(y))\big] - \mathbb{E}_{\mathbb{P}}\big[\phi_{\theta,\ell}(\bar{\mathbf{e}}(x))\big] = \frac{1}{2}\Big(\mathbb{E}[\bar{\mathbf{e}}(y)^\top H_\ell \bar{\mathbf{e}}(y)] - \mathbb{E}[\bar{\mathbf{e}}(x)^\top H_\ell \bar{\mathbf{e}}(x)]\Big) + \tilde{\epsilon}_\ell, \tag{27}$$

where $\tilde{\epsilon}_\ell$ collects higher-order remainder differences.

Then we **reduce the pooled quadratic forms to covariances**. Since $\mathbb{E}[\bar{\mathbf{e}}(x)] = \mathbb{E}[\bar{\mathbf{e}}(y)] = \mathbf{0}$, we use $\mathbb{E}[\mathbf{v}^\top A\mathbf{v}] = \operatorname{tr}(A\operatorname{Cov}(\mathbf{v}))$ to write

$$\mathbb{E}[\bar{\mathbf{e}}(y)^\top H_\ell \bar{\mathbf{e}}(y)] - \mathbb{E}[\bar{\mathbf{e}}(x)^\top H_\ell \bar{\mathbf{e}}(x)] = \operatorname{tr}\Big(H_\ell\big(\operatorname{Cov}(\bar{\mathbf{e}}(y)) - \operatorname{Cov}(\bar{\mathbf{e}}(x))\big)\Big). \tag{28}$$

We expand the covariance of the pooled embedding, since $\bar{\mathbf{e}} = \frac{1}{K}\sum_{i=1}^{K}\mathbf{e}_i.$, then

$$\begin{aligned}
\operatorname{Cov}(\bar{\mathbf{e}}) &= \mathbb{E}\Big[\big(\bar{\mathbf{e}} - \mathbb{E}[\bar{\mathbf{e}}]\big)\big(\bar{\mathbf{e}} - \mathbb{E}[\bar{\mathbf{e}}]\big)^\top\Big] \\
&= \mathbb{E}\left[\Big(\frac{1}{K}\sum_{i=1}^{K}(\mathbf{e}_i - \mathbb{E}[\mathbf{e}_i])\Big)\Big(\frac{1}{K}\sum_{j=1}^{K}(\mathbf{e}_j - \mathbb{E}[\mathbf{e}_j])\Big)^\top\right] \\
&= \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\big[(\mathbf{e}_i - \mathbb{E}[\mathbf{e}_i])(\mathbf{e}_j - \mathbb{E}[\mathbf{e}_j])^\top\big] \\
&= \frac{1}{K^2}\sum_{i=1}^{K}\operatorname{Cov}(\mathbf{e}_i) + \frac{1}{K^2}\sum_{i\neq j}\operatorname{Cov}(\mathbf{e}_i, \mathbf{e}_j) \\
&= \frac{1}{K^2}\sum_{i=1}^{K}\operatorname{Cov}(\mathbf{e}_i) + \frac{1}{K^2}\sum_{1\le i<j\le K}\Big(\operatorname{Cov}(\mathbf{e}_i, \mathbf{e}_j) + \operatorname{Cov}(\mathbf{e}_i, \mathbf{e}_j)^\top\Big) \\
&= \frac{1}{K^2}\sum_{i=1}^{K}\operatorname{Cov}(\mathbf{e}_i) + \frac{2}{K^2}\sum_{1\le i<j\le K}\operatorname{Sym}\big(\operatorname{Cov}(\mathbf{e}_i, \mathbf{e}_j)\big),
\end{aligned} \tag{29}$$

where $\operatorname{Sym}(A) := (A + A^\top)/2$. Noting that only the symmetric part of $\operatorname{Cov}(\bar{\mathbf{e}})$ contributes to the second-order expansion, and it only appears through quadratic forms and trace operators, so we replace each $\operatorname{Cov}(\mathbf{e}_i, \mathbf{e}_j)$ by its symmetric part $\operatorname{Sym}(\operatorname{Cov}(\mathbf{e}_i, \mathbf{e}_j))$, yielding a factor of 2 when summing over $i < j$. Using (29) for both x and y, we can write the pooled covariance difference exactly as

$$\begin{aligned}
\operatorname{Cov}(\bar{\mathbf{e}}(y)) - \operatorname{Cov}(\bar{\mathbf{e}}(x)) = &\frac{1}{K^2}\sum_{i=1}^{K}\Big(\operatorname{Cov}(\mathbf{e}_i(y)) - \operatorname{Cov}(\mathbf{e}_i(x))\Big) \\
&+ \frac{2}{K^2}\sum_{1\le i<j\le K}\operatorname{Sym}\Big(\operatorname{Cov}(\mathbf{e}_i(y), \mathbf{e}_j(y)) - \operatorname{Cov}(\mathbf{e}_i(x), \mathbf{e}_j(x))\Big).
\end{aligned} \tag{30}$$

Recall $e_i(x) = \mathbf{u}_i$ and $e_i(y) = \mathbf{u}_i + d_i \boldsymbol{\mu}_{\text{defect}}$ with $d_i = a_i s_i$. Under the stated assumptions $\mathbb{E}[d_i] = 0$ and the independence between $\{d_i\}$ and $\{\mathbf{u}_i\}$, all mixed cross-terms vanish. Based on Assumption 2.3, to complete this proof, we further assume that the signed defect indicators $\{d_i\}_{i=1}^K$ exhibit at most weak spatial dependence, i.e., $|\mathrm{Cov}(d_i, d_{i+\ell})| \le \rho\,\beta(\ell)$ with $\sum_{\ell \ge 1} \beta(\ell) < \infty$. We thus obtain for any $i \neq j$:

$$\mathrm{Cov}(\mathbf{e}_i(y), \mathbf{e}_j(y)) - \mathrm{Cov}(\mathbf{e}_i(x), \mathbf{e}_j(x)) = \mathrm{Cov}(d_i, d_j)\, \boldsymbol{\mu}_{\text{defect}} \boldsymbol{\mu}_{\text{defect}}^\top. \tag{31}$$

Similarly, for the diagonal term we have (cf. Proposition 2.4)

$$\mathrm{Cov}(\mathbf{e}_i(y)) - \mathrm{Cov}(\mathbf{e}_i(x)) = \mathrm{Var}(d_i)\, \boldsymbol{\mu}_{\text{defect}} \boldsymbol{\mu}_{\text{defect}}^\top = \rho\, \boldsymbol{\mu}_{\text{defect}} \boldsymbol{\mu}_{\text{defect}}^\top. \tag{32}$$

Plugging (31)–(32) into (30) yields

$$\mathrm{Cov}(\bar{\mathbf{e}}(y)) - \mathrm{Cov}(\bar{\mathbf{e}}(x)) = \frac{\rho}{K}\, \boldsymbol{\mu}_{\text{defect}} \boldsymbol{\mu}_{\text{defect}}^\top + \frac{2}{K^2} \sum_{1 \le i < j \le K} \mathrm{Cov}(d_i, d_j)\, \boldsymbol{\mu}_{\text{defect}} \boldsymbol{\mu}_{\text{defect}}^\top. \tag{33}$$

Under the mixing decay $|\mathrm{Cov}(d_i, d_{i+\ell})| \le \rho\,\beta(\ell)$ and stationarity, the off-diagonal sum is bounded by

$$\sum_{1 \le i < j \le K} |\mathrm{Cov}(d_i, d_j)| \le \sum_{\ell=1}^{K-1} (K - \ell) \rho\, \beta(\ell),$$

and therefore, in operator norm,

$$\left\| \mathrm{Cov}(\bar{\mathbf{e}}(y)) - \mathrm{Cov}(\bar{\mathbf{e}}(x)) \right\|_{\text{op}} \le \left(\frac{\rho}{K} + \frac{2\rho}{K^2} \sum_{\ell=1}^{K-1} (K - \ell) \beta(\ell) \right) \left\| \boldsymbol{\mu}_{\text{defect}} \boldsymbol{\mu}_{\text{defect}}^\top \right\|_{\text{op}} = \frac{\rho}{K_{\text{eff}}} \left\| \boldsymbol{\mu}_{\text{defect}} \right\|_2^2, \tag{34}$$

where K_{eff} is defined in (18). In particular, if $\{d_i\}$ is independent across patches, then $\mathrm{Cov}(d_i, d_j) = 0$ for $i \neq j$ and the bound is tight with $K_{\text{eff}} = K$. Moreover, under exponential mixing $\sum_{\ell \ge 1} \beta(\ell) < \infty$, we have $K_{\text{eff}} = \Theta(K)$.

Repeating the same second-order Taylor argument as in Proposition 2.4 with $\mathbf{e}$ replaced by $\bar{\mathbf{e}}$ yields, for each coordinate ℓ,

$$\Delta_{\text{global},\ell} \approx \frac{1}{2} \boldsymbol{\mu}_{\text{defect}}^\top \nabla^2 \phi_{\theta,\ell}(\mathbf{0}) \boldsymbol{\mu}_{\text{defect}} \cdot \frac{\rho}{K_{\text{eff}}} = \frac{\rho}{2K_{\text{eff}}} \left[\mathcal{Q}(\boldsymbol{\mu}_{\text{defect}}) \right]_\ell.$$

Stacking $\ell = 1, \ldots, d$ gives

$$\Delta_{\text{global}} \approx \frac{\rho}{2K_{\text{eff}}}\, \mathcal{Q}(\boldsymbol{\mu}_{\text{defect}}).$$

Under the Proposition 2.4, $\Delta_{\text{PFS}} \approx \frac{\rho}{2}\, \mathcal{Q}(\boldsymbol{\mu}_{\text{defect}})$, hence

$$\Delta_{\text{global}} \approx \frac{1}{K_{\text{eff}}} \Delta_{\text{PFS}}.$$

Therefore,

$$\|\Delta_{\text{PFS}}\|_2 \approx K_{\text{eff}} \|\Delta_{\text{global}}\|_2.$$

Under exponential mixing, $K_{\text{eff}} = \Theta(K)$, hence the patch-level shift dominates the global-pooled shift by a factor linear in K up to constants, consistent with (11):

$$\|\Delta_{\text{PFS}}\|_2 \approx K\, \|\Delta_{\text{global}}\|_2 > \|\Delta_{\text{global}}\|_2, \tag{35}$$

□

A.4. Existence of an optimal finite patch number K.

While Proposition 2.5 establishes that, at the population level, patch-wise aggregation amplifies the second-order defect signal relative to global pooling, it does not by itself imply that using arbitrarily many patches is always beneficial. In this subsection, we show that under finite-sample estimation and defect-power dilution at finer patch resolutions, the signal-to-noise ratio admits a finite maximizer. Consequently, the patch advantage saturates beyond a certain granularity, and an optimal finite patch number $K^\star$ necessarily exists.

Corollary A.3. *Assume the setting of Proposition 2.5 and Proposition 2.4. For a K-patch partition, let the per-patch embeddings be $\{\mathbf{e}_i(x)\}_{i=1}^K$ and $\{\mathbf{e}_i(y)\}_{i=1}^K$, and define the* patch-level *population shift*

$$\Delta_{\mathrm{PFS}}(K) := \mathbb{E}_{\mathbb{Q}}[\phi_\theta(\mathbf{e}_i(y))] - \mathbb{E}_{\mathbb{P}}[\phi_\theta(\mathbf{e}_i(x))]\,, \tag{36}$$

which is independent of i by stationarity across patches. Let $\widehat{\Delta}_{\mathrm{PFS}}(K)$ be its empirical estimator constructed from N i.i.d. images per domain,

$$\widehat{\Delta}_{\mathrm{PFS}}(K) := \frac{1}{N}\sum_{n=1}^{N}\Big(\frac{1}{K}\sum_{i=1}^{K}\phi_\theta(\mathbf{e}_{n,i}(y))\Big) - \frac{1}{N}\sum_{n=1}^{N}\Big(\frac{1}{K}\sum_{i=1}^{K}\phi_\theta(\mathbf{e}_{n,i}(x))\Big). \tag{37}$$

Assume further that the defect signature may dilute *with patch refinement: there exists a non-increasing function $g:\mathbb{N}\to\mathbb{R}_+$ and a fixed direction $\boldsymbol{\nu}\in\mathbb{R}^m$ with $\|\boldsymbol{\nu}\|_2 = 1$ such that the defect vector satisfies*

$$\boldsymbol{\mu}_{\mathrm{defect}}(K) = g(K)\,\boldsymbol{\nu}. \tag{38}$$

Let $\mathcal{Q}(\cdot)$ be the Hessian-induced quadratic map defined in Proposition 2.4, and define the defect strength

$$S(K) := \big\|\mathcal{Q}(\boldsymbol{\mu}_{\mathrm{defect}}(K))\big\|_2. \tag{39}$$

Assume exponential α-mixing across patches within each image as in (16)–(17), and let $K_{\mathrm{eff}}(K)$ be defined by (18). Then there exists a finite $K^\star < \infty$ such that the high-probability signal-to-noise ratio

$$\mathrm{SNR}(K) := \frac{\|\Delta_{\mathrm{PFS}}(K)\|_2}{\big\|\widehat{\Delta}_{\mathrm{PFS}}(K) - \Delta_{\mathrm{PFS}}(K)\big\|_2} \tag{40}$$

is non-increasing for all $K \ge K^\star$ (with probability at least $1-\delta$). Moreover, if $g(K) = cK^{-\eta}$ for some $c>0$ and $\eta>0$, then for exponential mixing ($K_{\mathrm{eff}}(K) = \Theta(K)$) we have

$$\mathrm{SNR}(K) \;=\; \tilde{\Theta}\Big(\sqrt{N}\,K^{\frac{1}{2}-2\eta}\Big), \tag{41}$$

and hence $\mathrm{SNR}(K)$ is eventually decreasing whenever $\eta > \frac{1}{4}$, implying a finite maximizer $K^\star$.

Proof. By Proposition 2.4, the leading-order patch-level shift satisfies

$$\Delta_{\mathrm{PFS}}(K) \;\approx\; \frac{\rho}{2}\,\mathcal{Q}(\boldsymbol{\mu}_{\mathrm{defect}}(K)). \tag{42}$$

Taking ℓ_2 norms and using the definition (39) yields

$$\|\Delta_{\mathrm{PFS}}(K)\|_2 \;\approx\; \frac{\rho}{2}\,S(K). \tag{43}$$

In particular, under the dilution model (38), since $\mathcal{Q}(\cdot)$ is quadratic in its argument,

$$S(K) = \big\|\mathcal{Q}(g(K)\boldsymbol{\nu})\big\|_2 = g(K)^2\,\big\|\mathcal{Q}(\boldsymbol{\nu})\big\|_2. \tag{44}$$

For each domain, define the per-image random vector

$$\mathbf{Z}_n^{(y)}(K) := \frac{1}{K}\sum_{i=1}^{K}\phi_\theta(\mathbf{e}_{n,i}(y)), \qquad \mathbf{Z}_n^{(x)}(K) := \frac{1}{K}\sum_{i=1}^{K}\phi_\theta(\mathbf{e}_{n,i}(x)). \tag{45}$$

Then (37) can be written as

$$\widehat{\Delta}_{\mathrm{PFS}}(K) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{Z}_n^{(y)}(K) - \frac{1}{N}\sum_{n=1}^{N}\mathbf{Z}_n^{(x)}(K). \tag{46}$$

By the i.i.d. sampling of images, $\{\mathbf{Z}_n^{(y)}(K)\}_{n=1}^N$ are i.i.d. across n (and similarly for x). Within a fixed image n, dependence across patches is allowed and controlled by α-mixing.

Fix a domain (say y) and suppress (y) in notation. For each coordinate $\ell \in [d]$, define the scalar patch sequence

$$U_i^{(\ell)} := \phi_{\theta,\ell}(\mathbf{e}_i), \qquad i = 1, \dots, K,$$

so that $Z_\ell(K) = \frac{1}{K}\sum_{i=1}^K U_i^{(\ell)}$. By stationarity across patches and the covariance decomposition,

$$\mathrm{Var}(Z_\ell(K)) = \mathrm{Var}\Big(\frac{1}{K}\sum_{i=1}^K U_i^{(\ell)}\Big) = \frac{1}{K^2}\sum_{i=1}^K \mathrm{Var}(U_i^{(\ell)}) + \frac{2}{K^2}\sum_{1\le i<j\le K} \mathrm{Cov}(U_i^{(\ell)}, U_j^{(\ell)}). \tag{47}$$

Under exponential α-mixing, as in Step 4 of the proof of Proposition 2.5, there exists a summable envelope $\beta(t)$ such that

$$\big|\mathrm{Cov}(U_i^{(\ell)}, U_{i+t}^{(\ell)})\big| \le \sigma_\phi^2\,\beta(t), \qquad t \ge 1, \qquad \sum_{t\ge 1}\beta(t) < \infty,$$

where $\sigma_\phi^2 := \sup_\ell \mathrm{Var}(U_i^{(\ell)}) < \infty$. Substituting into (47) and summing by lag yields

$$\mathrm{Var}(Z_\ell(K)) \le \frac{\sigma_\phi^2}{K} + \frac{2\sigma_\phi^2}{K^2}\sum_{t=1}^{K-1}(K-t)\beta(t) = \frac{\sigma_\phi^2}{K_{\mathrm{eff}}(K)}, \tag{48}$$

where $K_{\mathrm{eff}}(K)$ matches (18). Consequently,

$$\mathrm{tr}\big(\mathrm{Cov}(\mathbf{Z}_n(K))\big) = \sum_{\ell=1}^d \mathrm{Var}(Z_\ell(K)) \le \frac{d\,\sigma_\phi^2}{K_{\mathrm{eff}}(K)}. \tag{49}$$

Since $\{\mathbf{Z}_n(K)\}_{n=1}^N$ are i.i.d. across images, we apply a standard vector-valued Bernstein (or equivalently, coordinate-wise Bernstein plus union bound) to obtain, with probability at least $1-\delta$,

$$\Big\|\frac{1}{N}\sum_{n=1}^N \mathbf{Z}_n(K) - \mathbb{E}[\mathbf{Z}_n(K)]\Big\|_2 \;\le\; C_4\sqrt{\frac{\mathrm{tr}(\mathrm{Cov}(\mathbf{Z}_n(K)))\,\log(1/\delta)}{N}} \;\le\; C_5\sqrt{\frac{\log(1/\delta)}{N\,K_{\mathrm{eff}}(K)}}, \tag{50}$$

where $C_4, C_5 > 0$ absorb universal constants and $d\,\sigma_\phi^2$. Applying (50) separately to the real and generated domains and using the triangle inequality, we obtain

$$\big\|\widehat{\Delta}_{\mathrm{PFS}}(K) - \Delta_{\mathrm{PFS}}(K)\big\|_2 \;\le\; C_6\sqrt{\frac{\log(1/\delta)}{N\,K_{\mathrm{eff}}(K)}} \tag{51}$$

with probability at least $1-\delta$.

Combining the signal estimate (43) with the deviation bound (51) yields, on the high-probability event of (51),

$$\mathrm{SNR}(K) = \frac{\|\Delta_{\mathrm{PFS}}(K)\|_2}{\|\widehat{\Delta}_{\mathrm{PFS}}(K) - \Delta_{\mathrm{PFS}}(K)\|_2} \;\gtrsim\; \frac{S(K)}{1}\cdot\sqrt{\frac{N\,K_{\mathrm{eff}}(K)}{\log(1/\delta)}}. \tag{52}$$

Substituting the quadratic scaling (44) gives

$$\mathrm{SNR}(K) \;\gtrsim\; g(K)^2\,\big\|\mathcal{Q}(\boldsymbol{\nu})\big\|_2\sqrt{\frac{N\,K_{\mathrm{eff}}(K)}{\log(1/\delta)}}. \tag{53}$$

If $g(K)$ is non-increasing and $K_{\mathrm{eff}}(K)$ is eventually sublinear or bounded (which occurs when patch dependence strengthens as resolution increases), then the right-hand side of (53) is eventually non-increasing in K, implying the existence of a finite $K^\star$ such that (53) holds for all $K \ge K^\star$.

Assume $g(K) = cK^{-\eta}$ with $\eta > 0$. Then by (44), $S(K) = c^2K^{-2\eta}\|\mathcal{Q}(\boldsymbol{\nu})\|_2$. Under exponential mixing, $K_{\mathrm{eff}}(K) = \Theta(K)$, hence (53) yields

$$\mathrm{SNR}(K) = \tilde{\Theta}\Big(\sqrt{N}\,K^{\frac{1}{2}-2\eta}\Big),$$

which is (41). Therefore, if $\eta > \frac{1}{4}$, then $\frac{1}{2} - 2\eta < 0$ and $\mathrm{SNR}(K)$ is eventually decreasing, so a finite maximizer $K^\star < \infty$ exists. □

A.5. Proof of Proposition 2.6

On the Gaussian surrogate in PFS space. Throughout the analysis, patch embeddings are assumed to be sub-Gaussian, which is sufficient for the Taylor expansions and concentration arguments used in Propositions 2.4, 2.5 and Theorem 2.7. In Proposition 2.6, we additionally adopt a Gaussian surrogate in the PFS space to obtain a closed-form expression for the population MMD under a Gaussian kernel.

This surrogate should be understood as a moment-matched analytic approximation: the true PFS distribution is sub-Gaussian with controlled second-order statistics, and replacing it by a Gaussian with the same mean and isotropic proxy variance preserves the leading-order dependence of the MMD on the mean shift. Importantly, our conclusions rely only on the positivity and monotonic increase of the MMD with respect to $\|\boldsymbol{\Delta}_{\mathrm{PFS}}\|_2$, which holds beyond the exact Gaussian case.

Proof. We prove (12), and the positivity and monotonicity claims in (12).

Recall the definition of our gaussian deep kernel in 4:

$$k_\omega(x,y) = \exp\!\left(-\frac{\|\mathbf{Z}_\theta(x)-\mathbf{Z}_\theta(y)\|_2^2}{2\gamma^2}\right) =: \; k_\gamma\big(\mathbf{Z}_\theta(x),\mathbf{Z}_\theta(y)\big),$$

where $\mathbf{Z}_\theta(x) \in \mathbb{R}^{K\times d}$ is the Patch Signature Field. Here $\|\cdot\|_2$ denotes the entry-wise Euclidean norm of the field (i.e., the ℓ_2 norm after flattening), which coincides with the Frobenius norm on $\mathbb{R}^{K\times d}$. Thus k_γ is a Gaussian RBF kernel on the ambient Euclidean space $\mathbb{R}^{K\times d}$ (equivalently $\mathbb{R}^{Kd}$).

Define the (random) feature fields induced by $\mathbf{Z}_\theta$:

$$\mathbf{X} := \mathbf{Z}_\theta(x), \quad x\sim\mathbb{P}, \qquad \mathbf{Y} := \mathbf{Z}_\theta(y), \quad y\sim\mathbb{Q},$$

and similarly $\mathbf{X}' := \mathbf{Z}_\theta(x')$ for an independent copy $x'\sim\mathbb{P}$ and $\mathbf{Y}' := \mathbf{Z}_\theta(y')$ for an independent copy $y'\sim\mathbb{Q}$. With this notation, the population MMD under k_ω admits the standard expansion

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q};k_\omega) = \mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{X}')\big] + \mathbb{E}\big[k_\gamma(\mathbf{Y},\mathbf{Y}')\big] - 2\,\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{Y})\big], \tag{54}$$

where $\mathbf{X},\mathbf{X}',\mathbf{Y},\mathbf{Y}' \in \mathbb{R}^{K\times d}$ and $k_\gamma(A,B) = \exp\!\Big(-\frac{\|A-B\|_2^2}{2\gamma^2}\Big)$.

To obtain a closed-form expression under the Gaussian kernel k_γ on $\mathbb{R}^{K\times d}$, we adopt the following Gaussian surrogate for the feature fields:

$$\mathbf{X}\sim\mathcal{N}(\mathbf{0},\sigma_z^2\mathbf{I}_{Kd}), \qquad \mathbf{Y}\sim\mathcal{N}(\boldsymbol{\Delta}_{\mathbf{Z}},\sigma_z^2\mathbf{I}_{Kd}),$$

and similarly for independent copies $\mathbf{X}',\mathbf{Y}'$. Here $\boldsymbol{\Delta}_{\mathbf{Z}} := \mathbb{E}[\mathbf{Z}_\theta(y)] - \mathbb{E}[\mathbf{Z}_\theta(x)] \in \mathbb{R}^{K\times d}$ denotes the mean shift of the Patch Signature Field, and $\mathbf{I}_{Kd}$ denotes isotropic covariance under the entry-wise Euclidean structure of $\mathbb{R}^{K\times d}$. In particular, if each patch undergoes the same PFS mean shift $\boldsymbol{\Delta}_{\mathrm{PFS}} \in \mathbb{R}^d$, then $\boldsymbol{\Delta}_{\mathbf{Z}} = \mathbf{1}_K\boldsymbol{\Delta}_{\mathrm{PFS}}^\top$ and hence $\|\boldsymbol{\Delta}_{\mathbf{Z}}\|_2^2 = K\|\boldsymbol{\Delta}_{\mathrm{PFS}}\|_2^2$.

Moreover, if patch embeddings are σ_e-sub-Gaussian (Assumption 2.2) and the mapping ϕ_θ is locally Lipschitz (Assumption A.1) on the typical embedding region $\mathcal{E}$ with constant

$$L_\phi := \sup_{\mathbf{e}\in\mathcal{E}} \|J_\phi(\mathbf{e})\|_{\mathrm{op}} < \infty,$$

then the PFS features admit the sub-Gaussian proxy bound $\sigma_z \le L_\phi\sigma_e$ (hence $\sigma_z^2 \le L_\phi^2\sigma_e^2$). Under this surrogate, the expectations in (54) reduce to Gaussian integrals in $\mathbb{R}^d$. In the following steps, we will compute each expectation in (54) explicitly.

Step 1: Compute the self-terms $\mathbb{E}[k_\gamma(\mathbf{X},\mathbf{X}')]$ and $\mathbb{E}[k_\gamma(\mathbf{Y},\mathbf{Y}')]$. Let $\mathbf{X},\mathbf{X}' \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0},\sigma_z^2\mathbf{I}_{Kd})$ and define $\boldsymbol{\delta} := \mathbf{X}-\mathbf{X}'$. Then $\boldsymbol{\delta}\sim\mathcal{N}(\mathbf{0},2\sigma_z^2\mathbf{I}_{Kd})$ and

$$\begin{aligned}
\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{X}')\big] &= \mathbb{E}_{\boldsymbol{\delta}}\exp\!\left(-\frac{\|\boldsymbol{\delta}\|_2^2}{2\gamma^2}\right) \\
&= \int_{\mathbb{R}^{Kd}} \exp\!\left(-\frac{\|\boldsymbol{\delta}\|_2^2}{2\gamma^2}\right)(2\pi)^{-Kd/2}(2\sigma_z^2)^{-Kd/2}\exp\!\left(-\frac{\|\boldsymbol{\delta}\|_2^2}{4\sigma_z^2}\right)d\boldsymbol{\delta} \\
&= (2\pi)^{-Kd/2}(2\sigma_z^2)^{-Kd/2}\int_{\mathbb{R}^{Kd}}\exp\!\left(-\|\boldsymbol{\delta}\|_2^2\left(\frac{1}{2\gamma^2}+\frac{1}{4\sigma_z^2}\right)\right)d\boldsymbol{\delta}.
\end{aligned} \tag{55}$$

Define

$$A := \left(\frac{1}{\gamma^2} + \frac{1}{2\sigma_z^2}\right)\mathbf{I}_{Kd},$$

so that

$$\|\boldsymbol{\delta}\|_2^2\left(\frac{1}{2\gamma^2} + \frac{1}{4\sigma_z^2}\right) = \frac{1}{2}\,\boldsymbol{\delta}^\top A\boldsymbol{\delta}.$$

Hence

$$\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{X}')\big] = (2\pi)^{-Kd/2}(2\sigma_z^2)^{-Kd/2}\int_{\mathbb{R}^{Kd}}\exp\!\left(-\frac{1}{2}\,\boldsymbol{\delta}^\top A\boldsymbol{\delta}\right)\,d\boldsymbol{\delta}. \tag{56}$$

Using the Gaussian integral identity

$$\int_{\mathbb{R}^{Kd}}\exp\!\big(-\tfrac{1}{2}\,u^\top Au\big)\;du = (2\pi)^{Kd/2}|A|^{-1/2},$$

we obtain

$$\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{X}')\big] = (2\sigma_z^2)^{-Kd/2}\,|A|^{-1/2}. \tag{57}$$

Since $A = c\mathbf{I}_{Kd}$ with

$$c = \frac{1}{\gamma^2} + \frac{1}{2\sigma_z^2} = \frac{\gamma^2 + 2\sigma_z^2}{2\sigma_z^2\gamma^2},$$

we have $|A| = c^{Kd}$ and $|A|^{-1/2} = c^{-Kd/2}$, yielding

$$\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{X}')\big] = (2\sigma_z^2)^{-Kd/2}\left(\frac{2\sigma_z^2\gamma^2}{\gamma^2+2\sigma_z^2}\right)^{Kd/2} = \left(\frac{\gamma^2}{\gamma^2+2\sigma_z^2}\right)^{Kd/2}.$$

Therefore,

$$\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{X}')\big] = \left(\frac{\gamma^2}{\gamma^2+2\sigma_z^2}\right)^{\frac{Kd}{2}}. \tag{58}$$

Let $\mathbf{Y},\mathbf{Y}' \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\Delta}_{\mathbf{Z}},\sigma_z^2\mathbf{I}_{Kd})$ and define $\boldsymbol{\delta}' := \mathbf{Y} - \mathbf{Y}'$. Then $\boldsymbol{\delta}' \sim \mathcal{N}(\mathbf{0}, 2\sigma_z^2\mathbf{I}_{Kd})$ (the means cancel), so the same calculation as Step 1 yields

$$\mathbb{E}\big[k_\gamma(\mathbf{Y},\mathbf{Y}')\big] = \left(\frac{\gamma^2}{\gamma^2+2\sigma_z^2}\right)^{\frac{Kd}{2}}. \tag{59}$$

Step 2: Compute the cross-term $\mathbb{E}[k_\gamma(\mathbf{X},\mathbf{Y})]$. Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0},\sigma_z^2\mathbf{I}_{Kd})$ and $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\Delta}_{\mathbf{Z}},\sigma_z^2\mathbf{I}_{Kd})$ be independent. Define $\boldsymbol{\eta} := \mathbf{X} - \mathbf{Y}$. Then $\boldsymbol{\eta} \sim \mathcal{N}(-\boldsymbol{\Delta}_{\mathbf{Z}}, 2\sigma_z^2\mathbf{I}_{Kd})$ and

$$\begin{aligned}\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{Y})\big] &= \mathbb{E}_{\boldsymbol{\eta}}\exp\!\left(-\frac{\|\boldsymbol{\eta}\|_2^2}{2\gamma^2}\right)\\ &= \int_{\mathbb{R}^{Kd}}\exp\!\left(-\frac{\|\boldsymbol{\eta}\|_2^2}{2\gamma^2}\right)\,(2\pi)^{-Kd/2}(2\sigma_z^2)^{-Kd/2}\exp\!\left(-\frac{\|\boldsymbol{\eta}+\boldsymbol{\Delta}\|_2^2}{4\sigma_z^2}\right)\,d\boldsymbol{\eta},\end{aligned} \tag{60}$$

where $\boldsymbol{\Delta} := \boldsymbol{\Delta}_{\mathbf{Z}}$ for brevity. Expanding $\|\boldsymbol{\eta}+\boldsymbol{\Delta}\|_2^2 = \|\boldsymbol{\eta}\|_2^2 + 2\boldsymbol{\eta}^\top\boldsymbol{\Delta} + \|\boldsymbol{\Delta}\|_2^2$, the exponent becomes

$$\begin{aligned}-\frac{\|\boldsymbol{\eta}\|_2^2}{2\gamma^2} - \frac{\|\boldsymbol{\eta}+\boldsymbol{\Delta}\|_2^2}{4\sigma_z^2} = &-\left(\frac{1}{2\gamma^2}+\frac{1}{4\sigma_z^2}\right)\|\boldsymbol{\eta}\|_2^2\\ &-\frac{1}{2\sigma_z^2}\boldsymbol{\eta}^\top\boldsymbol{\Delta} - \frac{\|\boldsymbol{\Delta}\|_2^2}{4\sigma_z^2}.\end{aligned} \tag{61}$$

As in Step 1, set

$$A := \left(\frac{1}{\gamma^2} + \frac{1}{2\sigma_z^2}\right)\mathbf{I}_{Kd}, \qquad b := \frac{1}{2\sigma_z^2}\boldsymbol{\Delta},$$

so that

$$-\left(\frac{1}{2\gamma^2} + \frac{1}{4\sigma_z^2}\right)\|\boldsymbol{\eta}\|_2^2 - \frac{1}{2\sigma_z^2}\boldsymbol{\eta}^\top\boldsymbol{\Delta} = -\frac{1}{2}\,\boldsymbol{\eta}^\top A\boldsymbol{\eta} - b^\top\boldsymbol{\eta}.$$

Plugging into (60) gives

$$\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{Y})\big] = (2\pi)^{-Kd/2}(2\sigma_z^2)^{-Kd/2}\,\exp\!\left(-\frac{\|\boldsymbol{\Delta}\|_2^2}{4\sigma_z^2}\right)\int_{\mathbb{R}^{Kd}}\exp\!\left(-\frac{1}{2}\,\boldsymbol{\eta}^\top A\boldsymbol{\eta} - b^\top\boldsymbol{\eta}\right)d\boldsymbol{\eta}. \tag{62}$$

Using

$$\frac{1}{2}\,\boldsymbol{\eta}^\top A\boldsymbol{\eta} + b^\top\boldsymbol{\eta} = \frac{1}{2}(\boldsymbol{\eta} + A^{-1}b)^\top A(\boldsymbol{\eta} + A^{-1}b) - \frac{1}{2}b^\top A^{-1}b,$$

we obtain

$$\begin{aligned}\int_{\mathbb{R}^{Kd}}\exp\!\left(-\frac{1}{2}\,\boldsymbol{\eta}^\top A\boldsymbol{\eta} - b^\top\boldsymbol{\eta}\right)d\boldsymbol{\eta} &= \exp\!\left(\frac{1}{2}b^\top A^{-1}b\right)\int_{\mathbb{R}^{Kd}}\exp\!\left(-\frac{1}{2}(\boldsymbol{\eta} + A^{-1}b)^\top A(\boldsymbol{\eta} + A^{-1}b)\right)d\boldsymbol{\eta}\\ &= \exp\!\left(\frac{1}{2}b^\top A^{-1}b\right)\cdot(2\pi)^{Kd/2}|A|^{-1/2}.\end{aligned} \tag{63}$$

Substituting (63) into (62) yields

$$\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{Y})\big] = (2\sigma_z^2)^{-Kd/2}|A|^{-1/2}\,\exp\!\left(-\frac{\|\boldsymbol{\Delta}\|_2^2}{4\sigma_z^2} + \frac{1}{2}b^\top A^{-1}b\right). \tag{64}$$

As before, $A = c\mathbf{I}_{Kd}$ with

$$c = \frac{\gamma^2 + 2\sigma_z^2}{2\sigma_z^2\gamma^2},$$

so

$$(2\sigma_z^2)^{-Kd/2}|A|^{-1/2} = (2\sigma_z^2)^{-Kd/2}c^{-Kd/2} = \left(\frac{\gamma^2}{\gamma^2 + 2\sigma_z^2}\right)^{Kd/2}.$$

Since $A^{-1} = \frac{1}{c}\mathbf{I}_{Kd} = \frac{2\sigma_z^2\gamma^2}{\gamma^2+2\sigma_z^2}\mathbf{I}_{Kd}$ and $b = \frac{1}{2\sigma_z^2}\boldsymbol{\Delta}$, we have

$$\begin{aligned}b^\top A^{-1}b &= \left(\frac{1}{2\sigma_z^2}\boldsymbol{\Delta}\right)^\top\left(\frac{2\sigma_z^2\gamma^2}{\gamma^2 + 2\sigma_z^2}\mathbf{I}_{Kd}\right)\left(\frac{1}{2\sigma_z^2}\boldsymbol{\Delta}\right)\\ &= \frac{\gamma^2}{2\sigma_z^2(\gamma^2 + 2\sigma_z^2)}\|\boldsymbol{\Delta}\|_2^2.\end{aligned} \tag{65}$$

Therefore

$$\begin{aligned}-\frac{\|\boldsymbol{\Delta}\|_2^2}{4\sigma_z^2} + \frac{1}{2}b^\top A^{-1}b &= -\frac{\|\boldsymbol{\Delta}\|_2^2}{4\sigma_z^2} + \frac{\gamma^2}{4\sigma_z^2(\gamma^2 + 2\sigma_z^2)}\|\boldsymbol{\Delta}\|_2^2\\ &= -\frac{\|\boldsymbol{\Delta}\|_2^2}{2(\gamma^2 + 2\sigma_z^2)}.\end{aligned} \tag{66}$$

Combining the prefactor and exponent yields

$$\mathbb{E}\big[k_\gamma(\mathbf{X},\mathbf{Y})\big] = \left(\frac{\gamma^2}{\gamma^2 + 2\sigma_z^2}\right)^{\frac{Kd}{2}}\exp\!\left(-\frac{\|\boldsymbol{\Delta}_\mathbf{Z}\|_2^2}{2(\gamma^2 + 2\sigma_z^2)}\right). \tag{67}$$

Combine the three terms. Substituting (58), (59), and (67) into (54) gives

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q};k_\omega)=2\left(\frac{\gamma^2}{\gamma^2+2\sigma_z^2}\right)^{\frac{Kd}{2}}\left[1-\exp\left(-\frac{\|\mathbf{\Delta_Z}\|_2^2}{2(\gamma^2+2\sigma_z^2)}\right)\right],$$

Since $\mathbf{\Delta_Z}=\mathbf{1}_K\mathbf{\Delta}_{\mathrm{PFS}}^\top$, we have $\|\mathbf{\Delta_Z}\|_2^2=K\|\mathbf{\Delta}_{\mathrm{PFS}}\|_2^2$. Hence,

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q};k_\omega)=2\left(\frac{\gamma^2}{\gamma^2+2\sigma_z^2}\right)^{\frac{Kd}{2}}\left[1-\exp\left(-\frac{K\|\mathbf{\Delta}_{\mathrm{PFS}}\|_2^2}{2(\gamma^2+2\sigma_z^2)}\right)\right],$$

which proves (12).

Positivity and monotonicity in $\|\mathbf{\Delta}_{\mathrm{PFS}}\|_2$. Let $a:=\left(\frac{\gamma^2}{\gamma^2+2\sigma_z^2}\right)^{\frac{Kd}{2}}>0$ and $t:=\|\mathbf{\Delta}_{\mathrm{PFS}}\|_2\geq 0$. Then

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q};k_\omega)=2a\left(1-e^{-t^2/(2(\gamma^2+2\sigma_z^2))}\right).$$

If $t>0$, then $e^{-t^2/(2(\gamma^2+2\sigma_z^2))}\in(0,1)$ and hence $\mathrm{MMD}^2>0$. Moreover,

$$\frac{d}{dt}\left(1-e^{-t^2/(2(\gamma^2+2\sigma_z^2))}\right)=e^{-t^2/(2(\gamma^2+2\sigma_z^2))}\cdot\frac{t}{\gamma^2+2\sigma_z^2}\geq 0,$$

with strict inequality for $t>0$. Hence, positivity and monotonicity are both proved. □

A.6. Proof of Theorem 2.7 (Finite-Sample Detection Guarantee)

Lemma A.4 (Transformation of exponential concentration inequality in (Gretton et al., 2012)). *Let $\widehat{\mathrm{MMD}}_u^2$ denote the unbiased U-statistic estimator and MMD^2 be the population quantity. According to the Theorem 7 in (Gretton et al., 2012), there exists an absolute constant $c>0$ such that for any $\varepsilon>0$,*

$$\Pr\left(\left|\widehat{\mathrm{MMD}}_u^2-\mathrm{MMD}^2\right|>\varepsilon\right)\leq 2\exp\left(-c\,\varepsilon^2\frac{MN}{M+N}\right). \tag{68}$$

Equivalently, for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\left|\widehat{\mathrm{MMD}}_u^2-\mathrm{MMD}^2\right|\leq C\sqrt{\left(\frac{1}{M}+\frac{1}{N}\right)\log\frac{2}{\delta}}, \tag{69}$$

where $C=\frac{1}{\sqrt{c}}$ is an absolute constant (depending only on the kernel bound).

Proof. We show the transformation (68) ⇒ (69) step by step.

According to the Theorem 7 in (Gretton et al., 2012), we want the right-hand side of (68) to be at most δ. So we set

$$2\exp\left(-c\,\varepsilon^2\frac{MN}{M+N}\right)=\delta.$$

Divide both sides by 2 and take log:

$$-c\,\varepsilon^2\frac{MN}{M+N}=\log\frac{\delta}{2}=-\log\frac{2}{\delta}.$$

Multiply by $-\frac{M+N}{cMN}$:

$$\varepsilon^2=\frac{M+N}{cMN}\log\frac{2}{\delta}=\frac{1}{c}\left(\frac{1}{M}+\frac{1}{N}\right)\log\frac{2}{\delta}.$$

Let $C := \frac{1}{\sqrt{c}}$. Then

$$\varepsilon = C\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}}.$$

Plugging this choice of ε back into (68) gives

$$\Pr\left(\left|\widehat{\mathrm{MMD}}_u^2 - \mathrm{MMD}^2\right| > C\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}}\right) \le \delta,$$

which is exactly (69). □

Theorem 2.7. *Let $S_r = \{x_i\}_{i=1}^M \overset{i.i.d}{\sim} \mathbb{P}$ and $S_t = \{y_j\}_{j=1}^N$ be test-image set. Let $\lambda = \gamma^2 + 2\sigma_z^2$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the bounds (13) and (14) hold.*

Proof. By Lemma A.4, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega)\right| \le C\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}}. \tag{70}$$

We will use C_1 and C_2 to allow different absolute constants in the two cases.

Case I: Real test image ($S_t \overset{i.i.d}{\sim} \mathbb{P}$). If $S_r \sim \mathbb{P}$ and $S_t \sim \mathbb{P}$, then the two distributions are identical, hence

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{P}; k_\omega) = 0.$$

$$\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) = \widehat{\mathrm{MMD}}_u^2 - \mathrm{MMD}^2(\mathbb{P}, \mathbb{P}; k_\omega).$$

Therefore, on the event (70),

$$\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) = \left|\widehat{\mathrm{MMD}}_u^2 - \mathrm{MMD}^2(\mathbb{P}, \mathbb{P}; k_\omega)\right| \le C_1\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}},$$

which is exactly (13).

Case II: Generated test image ($S_t \overset{i.i.d}{\sim} \mathbb{Q}$). Define

$$A := \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega), \qquad B := \widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega).$$

Then by construction,

$$\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) = A + B. \tag{71}$$

Since $B \ge -|B|$ always holds; adding A to both sides gives the elementary inequality

$$A + B \;\ge\; A - |B|.$$

Applying this to (71) yields

$$\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) \ge \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega) - \left|\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega)\right|. \tag{72}$$

On the event (70), we have

$$\left|\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega)\right| \le C_2\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}}. \tag{73}$$

By Proposition 2.6 (population MMD under PFS mean shift) and letting $\lambda = \gamma^2 + 2\sigma_z^2$, we have

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega) = 2\left(\frac{\gamma^2}{\lambda}\right)^{\frac{d}{2}}\left[1 - \exp\left(-\frac{K\|\boldsymbol{\Delta}_{\mathrm{PFS}}\|_2^2}{2\lambda}\right)\right]. \tag{74}$$

Substitute (73) and (74) into (72). With probability at least $1-\delta$,

$$\begin{aligned}\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) \geq\ & 2\left(\frac{\gamma^2}{\lambda}\right)^{\frac{d}{2}}\left[1 - \exp\left(-\frac{K\|\boldsymbol{\Delta}_{\mathrm{PFS}}\|_2^2}{2\lambda}\right)\right] \\ & - C_2\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}},\end{aligned} \tag{75}$$

which is exactly (14). This completes Case II and the proof. □

Corollary A.5 (Separation of empirical MMD for real vs. generated images)**.** *Let $S_r = \{x_i\}_{i=1}^M \overset{i.i.d.}{\sim} \mathbb{P}$ be a reference set of real images, and let S_t be a test-image set. Consider the empirical statistic $\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega)$ defined with the deep kernel k_ω.*

Fix any $\delta \in (0,1)$ and define

$$\varepsilon_{M,N}(\delta) := C\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right)\log\frac{2}{\delta}},$$

where C is the constant appearing in Lemma A.4. Then, with probability at least $1-2\delta$, the two statements in Theorem 2.7 hold simultaneously. Consequently, whenever

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega) \ > \ 2\,\varepsilon_{M,N}(\delta), \tag{76}$$

the empirical ordering

$$\widehat{\mathrm{MMD}}_u^2(S_r, S_t^{(\mathbb{Q})}; k_\omega) \ > \ \widehat{\mathrm{MMD}}_u^2(S_r, S_t^{(\mathbb{P})}; k_\omega)$$

holds with probability at least $1-2\delta$.

Proof. We combine the high-probability bounds established in Theorem 2.7 for the real and generated cases.

By Theorem 2.7, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) \leq \varepsilon_{M,N}(\delta) \quad \text{if } S_t \sim \mathbb{P},$$

and with probability at least $1-\delta$,

$$\widehat{\mathrm{MMD}}_u^2(S_r, S_t; k_\omega) \geq \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega) - \varepsilon_{M,N}(\delta) \quad \text{if } S_t \sim \mathbb{Q}.$$

The two deviation events above each fail with probability at most δ. By the union bound, both inequalities hold simultaneously with probability at least $1-2\delta$.

On this event, we have

$$\begin{aligned}&\widehat{\mathrm{MMD}}_u^2(S_r, S_t^{(\mathbb{Q})}; k_\omega) - \widehat{\mathrm{MMD}}_u^2(S_r, S_t^{(\mathbb{P})}; k_\omega) \\ &\quad \geq \big(\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega) - \varepsilon_{M,N}(\delta)\big) - \varepsilon_{M,N}(\delta) = \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega) - 2\varepsilon_{M,N}(\delta).\end{aligned}$$

Therefore, if condition (76) holds, the right-hand side is strictly positive, which yields the claimed empirical ordering.

Interpretation. Since $\varepsilon_{M,N}(\delta) = \mathcal{O}\Big(\sqrt{(1/M + 1/N)\log(1/\delta)}\Big)$, for any fixed population gap $\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k_\omega) > 0$, the separation condition (76) is satisfied once the sample sizes M, N are sufficiently large. □

B. Additional Experiment Setups

B.1. Details of Datasets

B.1.1. Details of Image Benchmarks

ImageNet (Deng et al., 2009). We use the ImageNet real images and their corresponding synthetic counterparts released in the DGM-Eval repository.[1] All images are provided following Stein et al. (2023), and are stored at a resolution of 256×256. The set of generators includes ADM, ADMG, BigGAN, DiT-XL-2, GigaGAN, LDM, StyleGAN-XL, RQ-Transformer, and Mask-GIT.

LSUN-Bedroom (Yu et al., 2015). Real and generated samples for LSUN-Bedroom are also taken from the same DGM-Eval release.[2] The dataset (Stein et al. (2023)) provides images at 256×256; during preprocessing, we apply random cropping to obtain 224×224 inputs. The generated images are produced by ADM, DDPM, iDDPM, StyleGAN, Diffusion-Projected GAN, Projected GAN, and Unleashing Transformers.

GenImage (Zhu et al., 2023b). We additionally adopt GenImage, which is publicly available at:[3] According to Zhu et al. (2023b), the real images are sourced from ImageNet, while the image resolutions vary across subsets. The generative sources covered by GenImage include Midjourney, SD v1.4, SD v1.5, ADM, GLIDE, Wukong, VQDM, and BigGAN.

B.1.2. Details of Video Case Study

OpenSora (Zheng et al., 2024). Recent progress in video generation has substantially improved the realism of synthetic videos, raising new concerns about the trustworthiness of digital media. Since the proprietary model behind Sora[4] is not publicly accessible, we instead employ Open-Sora, an open-source high-fidelity video generation framework with fully released code and model weights as a practical stress test for evaluating generalization. Concretely, we randomly sample 3,275 videos from the OpenSora subset in the GenVideo (Chen et al., 2024b) benchmark. Each video contains 10 frames, yielding 32,750 frames in total, which we treat as the *OpenSora-generated video dataset*. For preprocessing, we follow the same pipeline as in the image benchmarks and apply random cropping to obtain 224×224 inputs.

MSR-VTT (Xu et al., 2016). As natural video data, we use MSR-VTT, a large-scale web video benchmark with diverse content and comprehensive categories, widely adopted for video understanding and video-to-text tasks. For preprocessing, we randomly sample 3,275 videos from MSR-VTT, and then randomly select 10 frames per video, resulting in 32,750 frames in total as the *real* set. We follow the same pipeline as the image benchmarks and apply random cropping to obtain 224×224 inputs.

B.2. Details on Evaluation Metrics

AI-generated image detection is inherently a binary classification task. Let $\mathrm{TP}(t), \mathrm{TN}(t), \mathrm{FP}(t), \mathrm{FN}(t)$ denote the numbers of true positives, true negatives, false positives, and false negatives when thresholding the detector score at t. Accordingly, the true positive rate (TPR) and false positive rate (FPR) are

$$\mathrm{TPR}(t) = \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FN}(t)}, \qquad \mathrm{FPR}(t) = \frac{\mathrm{FP}(t)}{\mathrm{FP}(t) + \mathrm{TN}(t)}.$$

The area under the receiver operating characteristic curve (AUROC). AUROC summarizes the detector's ranking quality by measuring how well positives are separated from negatives across *all* possible thresholds. Formally, it is the area under the ROC curve obtained by plotting $\mathrm{TPR}(t)$ against $\mathrm{FPR}(t)$ as t varies:

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}\big(\mathrm{FPR}^{-1}(u)\big)\, du,$$

where larger AUROC indicates better overall discriminability independent of a specific operating point.

The average precision (AP). Average Precision evaluates precision–recall trade-offs by aggregating precision over different

[1] https://github.com/layer6ai-labs/dgm-eval
[2] https://github.com/layer6ai-labs/dgm-eval
[3] https://github.com/GenImage-Dataset/GenImage
[4] https://openai.com/index/sora/

recall levels, and is commonly used when the positive class may be relatively rare. Let $\mathrm{Precision}(t) = \frac{\mathrm{TP}(t)}{\mathrm{TP}(t)+\mathrm{FP}(t)}$ and $\mathrm{Recall}(t) = \mathrm{TPR}(t)$. AP is defined as the area under the precision–recall (PR) curve:

$$\mathrm{AP} = \int_0^1 \mathrm{Precision}(r)\, dr,$$

where $\mathrm{Precision}(r)$ denotes precision as a function of recall r along the PR curve.

The classification accuracy (ACC). Accuracy reports the fraction of correctly classified samples at a chosen threshold, counting both true positives and true negatives:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}.$$

Unlike AUROC/AP which integrate over thresholds, ACC depends on the selected decision threshold. Following (Ojha et al., 2023), we adopt an automatic thresholding strategy during testing: the decision threshold is chosen to best separate real and AI-generated samples according to the detector scores, i.e., selecting the operating point that yields the strongest class separation on the evaluation split. However, we do not treat ACC as a primary metric for comparison because its value can vary noticeably with the thresholding protocol and the underlying data characteristics (e.g., class prior, domain shift, or how representative a small validation set is if used for calibration). In contrast, AUROC and AP summarize performance across all possible thresholds, making them more robust and better aligned with practical deployment scenarios where the preferred operating point may differ.

B.3. Details of Implementations

Main Experiments. Following prior works (Ojha et al., 2023; Liu et al., 2024), we use random cropping and random horizontal flipping during training, and apply center cropping at test time, without additional augmentations. For the main experiments, we adopt DINOv2 ViT-L/14 (Oquab et al., 2024) as the feature backbone to extract patch embeddings. To balance detection accuracy and efficiency, we pool the patch embeddings with a patch size of $W = 32$ for computing Patch Forensic Signatures (PFS). The PFS projection head ϕ_θ maps each pooled patch embedding (dimension 1024) to a bounded scalar score, using a lightweight feed-forward projection with hidden dimension 256, dropout 0.3, and a final $\tanh$ activation. For training data, we follow the common cross-dataset protocol: ProGAN is used to train models evaluated on ImageNet and LSUN-Bedroom, while SD v1.4 is used as the training set for the GenImage benchmark and our OpenSora case study. For reference data at test time, we use 3k real references that are strictly disjoint from the test split: for ImageNet and GenImage, we sample 3 images per ImageNet training class (3k in total); for LSUN-Bedroom, we randomly sample 3k LSUN real images from a split disjoint from testing; and for the OpenSora case study, we sample 175 MSR-VTT real videos disjoint from testing and extract 10 frames per video (1,750 real frames) as references. During training, we jointly optimize the projection parameters θ and the kernel bandwidth γ. We train for 25 epochs with batch size 256 using AdamW (learning rate 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay 0.01), and initialize the scale parameter with $\sigma = 1.0$. All experiments are conducted on a server with an NVIDIA H200 GPU using Python 3.10.19 and PyTorch 2.9.1.

Figure 1. We compare a *global image-level* baseline and our *PFS-based* pipeline under the label-inversion stress test. ***For data configuration,*** we build the toy benchmark from the ProGAN dataset by selecting the *cat* and *dog* categories from each split. During training, we assign all real samples to cats and all fake samples to dogs, using 18,000 images per class (18,000 real-cat and 18,000 fake-dog). The validation set follows the same configuration with 200 images per class. For testing, we consider two settings. (i) *Matched-label test:* the same label configuration as training, with 200 real-cat and 200 fake-dog images. (ii) *Label-inversion test:* we swap the category composition while keeping the real/fake labels fixed, i.e., 200 real-dog and 200 fake-cat images, to stress-test whether a detector relies on semantic category cues versus generation artifacts. ***For model configuration,*** we use DINOv2 ViT-L/14 as the frozen feature extractor and take the `[CLS]` token as the image representation for global image-level detector. On top of it, we train a lightweight two-layer classification head (hidden dimension 256 with dropout 0.3) to predict a single logit, optimized with binary cross-entropy (BCE) loss. For PFS pipeline, we follow the same patch-wise setup as in the main experiments: the image is partitioned into patches and each patch embedding is mapped into the PFS space via the same projection head as in our main method. To obtain an image-level decision, we additionally learn a lightweight attention (scoring) head with the same hidden dimension and dropout, which outputs a scalar weight/logit for each patch and aggregates patch-level logits into a final image-level logit. The entire model is trained with BCE loss under the same label setting as the Global baseline.

Table 2. For variants *w/o MMD*, we adopt the same model configurations as in the toy experiment (Figure 1): a global baseline that classifies from the DINOv2 `[CLS]` token, and a PFS-based model that aggregates patch-wise PFS scores via a lightweight attention head, both trained with a BCE objective. For *Global + MMD*, we replace the BCE objective with an MMD-based optimization on the *image-level* predictions: we take the global image logit for each sample and compute a one-dimensional, pairwise MMD within each mini-batch between real and AI-generated sets, using it as the training signal. Finally, to further isolate the benefit of *PFS modeling* beyond a particular aggregator, we additionally evaluate alternative patch-level aggregation schemes (e.g., mean/max/top-k pooling) in Appendix C.2, demonstrating that PFS consistently outperforms global pooling under different aggregation choices.

Figure 4. For qualitative localization, we adopt a Grad-CAM-style visualization on the MDMF detector. Given an input image resized to 224×224, we extract DINOv2 ViT-L/14 patch tokens and pool them to a $W{=}32$ patch grid. We then compute patch logits in the learned PFS space, and obtain patch-wise saliency by backpropagating the mean patch logit to the pooled patch embeddings. The final patch importance is computed by combining the patch logit with the gradient magnitude, followed by normalization and resizing to the image resolution for overlay visualization. For the global baseline, we visualize an attention map derived from normalized DINOv2 patch-token magnitudes to provide a comparable heatmap.

C. Additional Experimental Results

C.1. Results on Additional Benchmarks

LSUN-Bedroom. As shown in Table 3, our method maintains consistently strong performance across diverse generators, covering both diffusion-based models and GAN variants, indicating good cross-model generalization. In particular, we achieve the best average AUROC on LSUN-Bedroom, while keeping the average AP highly competitive, suggesting that our detection evidence transfers reliably beyond the training distribution.

Table 3. Detection performance (%) on LSUN-Bedroom. Bold numbers are superior results. We mainly compare training-based methods.

Methods	Models														Average	
	ADM		DDPM		iDDPM		Diffusion GAN		Projected GAN		StyleGAN		Unleashing Transformer			
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC (↑)	AP (↑)
CNNspot	64.83	64.24	79.04	80.58	76.95	76.28	88.45	87.19	90.80	89.94	95.17	94.94	93.42	93.11	84.09	83.75
Ojha	71.26	70.95	79.26	78.27	74.80	73.46	84.56	82.91	82.00	78.42	81.22	78.08	83.58	83.48	79.53	77.94
DIRE	57.19	56.85	61.91	61.35	59.82	58.29	53.18	53.48	55.35	54.93	57.66	56.90	67.92	68.33	59.00	58.59
NPR	75.43	72.60	91.42	90.89	89.49	88.25	76.17	74.19	75.07	74.59	68.82	63.53	84.39	83.67	80.11	78.25
F-ConV	76.59	74.40	93.53	92.16	88.90	86.85	98.10	98.03	97.93	97.81	91.63	90.16	97.31	96.91	92.00	**90.91**
MDMF	74.67	65.21	93.05	90.08	89.10	84.52	99.85	99.74	99.91	99.84	97.96	96.89	99.10	98.59	**93.38**	90.70

GenImage. Table 4 further demonstrates that our method generalizes well to GenImage, which contains heterogeneous sources ranging from proprietary engines (e.g., Midjourney) to various diffusion and GAN models, achieving the best average accuracy among compared methods. We present the results of some baselines reported in (Zhu et al., 2023b), including DeiT-S (Touvron et al., 2021), Swin-T (Liu et al., 2021), Spec (Zhang et al., 2019), F3Net (Qian et al., 2020), GramNet (Liu et al., 2020b), and GenDet (Zhu et al., 2023a). Overall, the strong average performance across such diverse generative sources highlights the robustness of our approach under real-world distribution shifts.

C.2. Full Results of Ablation Study for Core Components in MDMF

Table 5 reports the complete ablation results for the core components of MDMF on ImageNet, including global baselines trained with BCE or MMD, as well as PFS-based variants. Beyond the default attention head aggregation (i.e., PFS-Attn-BCE), we additionally evaluate several alternative ways of aggregating patch logits in the PFS space (mean, max, top-k) to isolate the effect of PFS modeling and aggregation choice.

Two observations consistently emerge and align with our motivation. First, replacing global image-level pooling with PFS-based patch evidence yields a clear improvement across generators, indicating that the cues for AI-generated images are better captured as localized, artifact-sensitive signals rather than a single semantic-dominant representation. This supports the view that modeling an image as a collection of patch-wise forensic evidence provides a stronger and more transferable basis for real/fake detection than relying on global features.

Table 4. AI-generated image detection performance (ACC, %) on GenImage.

Methods	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Average
	Models								
Training Methods									
ResNet-50	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S	55.6	99.9	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNspot	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net	50.1	99.9	99.9	49.9	50.00	99.9	49.9	49.9	68.7
GramNet	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
DIRE	60.2	99.9	99.8	50.9	55.0	99.2	50.1	50.2	70.7
Ojha	73.2	84.2	84.0	55.2	76.9	75.6	56.9	80.3	73.3
NPR	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
FatFormer	92.7	100.0	99.9	75.9	88.0	99.9	98.8	55.8	88.9
GenDet	89.6	96.1	96.1	58.0	78.4	92.8	66.5	75.0	81.6
DRCT	91.5	95.0	94.4	79.4	89.1	94.6	90.0	81.6	89.4
F-Conv	89.3	98.8	98.5	74.9	89.3	95.6	86.7	87.6	90.1
MDMF	83.5	99.4	99.2	79.4	92.4	97.6	89.7	86.6	**91.0**

Second, while simple PFS-space aggregations (mean/max/top-k) already outperform the global baselines, the best performance is achieved when PFS is further coupled with MMD optimization (i.e., ours MDMF). This suggests that the key is not merely the pooling operator, but explicitly learning and comparing the *distributional discrepancy* of patch-level signatures, which amplifies subtle, localized defects into a reliable macro-level detection signal.

Table 5. Detailed detection performance (%) for ablation study. Bold numbers are superior results.

Methods	ADM		ADMG		LDM		DiT		BigGAN		GigaGAN		StyleGAN XL		RQ-Transformer		Mask GIT		Average	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC (↑)	AP (↑)
Global-BCE	86.89	88.01	82.35	83.34	86.53	92.80	80.15	89.00	98.21	98.34	94.07	97.01	93.70	96.81	94.89	97.48	94.46	97.20	90.14	93.33
Global-MMD	83.08	87.01	75.53	80.51	79.60	90.42	72.22	86.23	98.16	98.59	92.22	96.55	91.96	96.42	93.66	97.26	92.36	96.65	86.53	92.18
PFS-Mean-BCE	86.54	87.27	83.53	84.17	90.98	94.85	84.04	90.61	99.39	99.39	97.99	98.74	97.04	98.32	97.63	98.58	98.63	99.07	92.86	94.56
PFS-Max-BCE	83.26	84.82	80.07	81.35	88.89	94.41	81.26	89.92	99.82	99.84	96.27	98.24	95.06	97.65	96.30	98.27	97.98	99.06	90.99	93.73
PFS-Top-5-BCE	86.08	87.19	82.92	83.84	90.99	95.39	83.49	91.14	99.81	99.83	97.19	98.64	95.99	98.09	96.98	98.55	98.38	99.23	92.42	94.66
PFS-Attn-BCE	87.09	88.73	84.11	85.54	91.47	95.76	85.08	92.18	99.90	99.91	97.81	98.97	97.06	98.61	97.69	98.93	98.72	99.41	93.22	95.34
MDMF	**92.56**	**93.57**	**88.86**	**90.16**	**94.63**	**97.35**	**88.89**	**94.48**	**99.93**	**99.94**	**98.99**	**99.52**	**98.76**	**99.41**	**98.84**	**99.46**	**99.40**	**99.72**	**95.65**	**97.07**

C.3. Full Results for the Effect of Patch Granularity

Table 6 reports the full results of the patch-granularity study corresponding to Figure 4(a), where we vary the pooled patch size $W \in \{16, 32, 56\}$ and repeat each setting with five random seeds. Consistent with Figure 4(a), the results exhibit a non-monotonic dependence on W, with intermediate granularity (e.g., $W = 32$) offering the best overall trade-off. The multi-seed breakdown further suggests that overly fine partitions can introduce higher variability in the estimated distributional discrepancy, whereas overly coarse partitions may miss sparse localized artifacts, reinforcing that effective PFS modeling requires a balanced spatial granularity.

C.4. Full Results for the Robustness to Encoder Architecture

Table 7 provides the full per-generator results corresponding to Figure 4(b), comparing MDMF against the training-based baseline F-ConV under different DINOv2 encoder variants (ViT-S/14, ViT-B/14, ViT-L/14, and ViT-G/14), reporting AUROC/AP and their averages. Across all backbones, MDMF consistently improves over F-ConV, indicating that our gains are not tied to a specific feature extractor and transfer reliably across encoder architectures and scales. Notably, MDMF exhibits stable scaling behavior as the backbone grows, suggesting that PFS-based local forensic cues and distributional comparison provide a more backbone-agnostic detection signal than global image-level modeling, which can be more sensitive to semantic representations and thus less stable under architecture changes.

Table 6. Detailed detection performance (%) for patch granularity.

Patch Size	ADM		ADMG		LDM		DiT		BigGAN		GigaGAN		StyleGAN XL		RQ-Transformer		Mask GIT		Average	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC (↑)	AP (↑)
$W = 16$	89.27	91.12	84.45	86.86	92.55	96.46	85.16	92.79	99.93	99.94	98.39	99.27	97.96	99.08	98.27	99.22	98.99	99.56	93.89	96.03
	89.85	91.50	85.41	87.43	92.61	96.47	85.83	93.02	99.92	99.93	98.53	99.33	98.01	99.09	98.35	99.25	99.08	99.59	94.18	96.18
	90.56	92.20	86.07	88.25	93.21	96.78	86.37	93.41	99.92	99.93	98.62	99.37	98.22	99.19	98.42	99.28	99.07	99.58	94.50	96.44
	90.77	92.23	86.26	88.21	93.33	96.81	86.69	93.50	99.93	99.93	98.69	99.39	98.19	99.17	98.45	99.29	99.18	99.63	94.61	96.46
	91.77	92.88	87.67	89.12	94.10	97.11	87.74	93.94	99.94	99.94	98.84	99.45	98.48	99.29	98.69	99.39	99.28	99.67	95.17	96.76
$W = 32$	90.56	92.08	86.14	88.06	93.40	96.80	86.38	93.32	99.93	99.94	98.61	99.36	98.18	99.17	98.58	99.35	99.15	99.62	94.55	96.41
	90.71	92.18	86.39	88.23	93.56	96.86	86.88	93.53	99.94	99.94	98.67	99.38	98.26	99.20	98.53	99.33	99.18	99.62	94.68	96.47
	90.88	92.24	86.73	88.38	93.80	96.97	87.11	93.62	99.94	99.94	98.71	99.40	98.34	99.23	98.59	99.35	99.21	99.64	94.81	96.53
	91.63	92.81	87.75	89.24	93.95	97.06	87.91	94.02	99.93	99.93	98.83	99.44	98.61	99.34	98.75	99.42	99.32	99.68	95.18	96.77
	92.56	93.57	88.86	90.16	94.63	97.35	88.89	94.48	99.93	99.94	98.99	99.52	98.76	99.41	98.84	99.46	99.40	99.72	95.65	97.07
$W = 56$	90.33	91.75	86.11	87.90	93.14	96.25	86.37	92.95	99.93	99.93	98.66	99.37	98.19	99.16	98.52	99.25	99.17	99.61	94.46	96.19
	90.47	92.02	86.25	88.19	93.28	96.60	86.51	93.27	99.93	99.93	98.66	99.30	98.19	99.09	98.52	99.24	99.20	99.64	94.56	96.35
	90.62	92.27	86.40	88.44	93.43	96.85	86.66	93.52	99.93	99.93	98.69	99.40	98.22	99.19	98.55	99.34	99.20	99.64	94.63	96.51
	90.70	91.78	86.48	87.93	93.51	96.28	86.74	92.98	99.93	99.93	98.69	99.40	98.22	99.19	98.55	99.28	99.20	99.64	94.68	96.22
	90.82	91.97	86.60	88.14	93.63	96.55	86.86	93.22	99.93	99.93	98.69	99.37	98.22	99.16	98.55	99.28	99.22	99.66	94.75	96.34

Table 7. Detailed detection performance (%) for different encoder architectures. Bold numbers are superior results.

Backbone	Methods	ADM		ADMG		LDM		DiT		BigGAN		GigaGAN		StyleGAN XL		RQ-Transformer		Mask GIT		Average	
		AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC (↑)	AP (↑)
ViT-S/14	F-ConV	76.97	80.06	70.98	71.46	71.73	71.66	67.51	68.48	86.07	87.25	79.30	80.47	78.71	78.76	78.76	79.64	76.54	77.36	76.28	77.24
	MDMF	**78.68**	**79.90**	**73.46**	**73.94**	**79.29**	**88.35**	**71.95**	**83.38**	**94.69**	**95.42**	**86.07**	**92.60**	**81.23**	**89.44**	**84.11**	**91.46**	**86.66**	**93.19**	**81.79**	**87.52**
ViT-B/14	F-ConV	86.66	86.93	81.16	82.44	85.01	85.36	78.55	79.53	96.07	96.26	90.37	90.42	93.87	94.49	92.41	93.32	92.18	92.18	88.47	88.99
	MDMF	**86.96**	**88.15**	**82.68**	**83.83**	**88.65**	**94.13**	**82.10**	**90.46**	**99.57**	**99.61**	**95.65**	**97.89**	**94.49**	**97.26**	**94.71**	**97.44**	**96.64**	**98.43**	**91.27**	**94.13**
ViT-L/14	F-ConV	**92.74**	91.65	88.51	87.67	88.87	88.47	85.94	84.88	98.94	98.98	98.14	98.72	98.52	98.38	96.79	96.33	95.52	95.38	93.77	93.38
	MDMF	92.56	**93.57**	**88.86**	**90.16**	**94.63**	**97.35**	**88.89**	**94.48**	**99.93**	**99.94**	**98.99**	**99.52**	**98.76**	**99.41**	**98.84**	**99.46**	**99.40**	**99.72**	**95.65**	**97.07**
ViT-G/14	F-ConV	90.90	92.51	85.75	87.70	87.49	89.17	82.49	84.59	96.59	97.08	95.49	96.04	96.38	96.70	93.96	94.97	94.49	95.34	91.50	92.68
	MDMF	**95.64**	**96.26**	**93.20**	**94.10**	**96.65**	**98.39**	**92.38**	**96.30**	**99.99**	**99.99**	**99.55**	**99.79**	**99.39**	**99.72**	**99.43**	**99.74**	**99.73**	**99.88**	**97.33**	**98.24**

C.5. Impact of Reference Size and Runtime Analysis

Table 8 reports the detailed detection performance and runtime on the ImageNet benchmark under different reference set sizes R used in the *test-time* MMD scoring (Eq. 7). Importantly, R is *independent of training*: it only controls the number of reference images used during inference when computing the MDMF score for each test image. The reference images are sampled from the ImageNet training split, strictly disjoint from the test split, to evaluate how R affects detection stability and deployment cost. Overall, MDMF is largely insensitive to R: for a fixed seed, varying R from 1k to 10k yields nearly unchanged AUROC/AP across generators, indicating that the PFS-induced distributional discrepancy can be estimated reliably without requiring a large reference pool. From an efficiency perspective, although each test image must be compared against R references, in practice we precompute and cache the PFS embeddings of the reference set, so the one-off reference encoding cost is amortized and negligible at inference time. The remaining computation reduces to GPU-efficient matrix operations whose arithmetic cost scales linearly with R, but is typically fast relative to feature extraction and is thus weakly reflected in end-to-end inference time, which is dominated by backbone forward passes and instantaneous hardware load. Based on this trade-off, we adopt $R = 3\text{k}$ in our main experiments to balance efficiency with stable performance.

D. Detailed Visualizations

In the main paper, we present qualitative visualizations on ADM (Figure 5/ 6). Here we further report results on images generated by ADMG, LDM, and DiT (Figure 7–9). Across these generators, we observe a consistent trend: the global pooling baseline mainly attends to semantically salient regions (e.g., object contours and high-contrast textures) with similar patterns on real and generated samples, suggesting limited sensitivity to sparse, localized artifacts. In contrast, MDMF produces more localized activations on generated images and comparatively diffuse responses on real images, indicating that patch-wise PFS evidence induces a stronger distributional discrepancy that can be leveraged for robust detection. Overall, these cross-model visualizations support the generalization of MDMF and corroborate our claim that localized forensic cues are suppressed by semantic-dominant global features but become salient under PFS-based distributional modeling.

Table 8. Detailed detection performance (%) and runtime for different reference sizes.

Seed	Ref Size	ADM		ADMG		LDM		DiT		BigGAN		GigaGAN		StyleGAN XL		RQ-Transformer		Mask GIT		Average		Runtime
		AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC (↑)	AP (↑)	
Seed = 0	$R = 1$k	92.47	93.54	88.74	90.11	94.60	97.34	88.81	94.46	99.93	99.94	98.99	99.52	98.75	99.41	98.83	99.46	99.40	99.72	95.61	97.05	718s
	$R = 3$k	92.56	93.57	88.86	90.16	94.63	97.35	88.89	94.48	99.93	99.94	98.99	99.52	98.76	99.41	98.84	99.46	99.40	99.72	95.65	97.07	701s
	$R = 5$k	92.53	93.56	88.82	90.14	94.62	97.35	88.86	94.47	99.93	99.94	98.99	99.52	98.76	99.41	98.84	99.46	99.40	99.72	95.64	97.06	776s
	$R = 7$k	92.56	93.57	88.85	90.16	94.63	97.35	88.89	94.48	99.93	99.94	98.99	99.52	98.76	99.41	98.84	99.46	99.40	99.72	95.65	97.07	616s
	$R = 10$k	92.56	93.57	88.85	90.15	94.63	97.35	88.90	94.48	99.93	99.94	98.99	99.52	98.76	99.41	98.84	99.46	99.40	99.72	95.65	97.07	766s
Seed = 42	$R = 1$k	90.51	92.06	86.09	88.04	93.39	96.80	86.36	93.32	99.93	99.94	98.61	99.36	98.18	99.17	98.58	99.35	99.15	99.62	94.53	96.40	705s
	$R = 3$k	90.56	92.08	86.14	88.06	93.40	96.80	86.38	93.32	99.93	99.94	98.61	99.36	98.18	99.17	98.58	99.35	99.15	99.62	94.55	96.41	734s
	$R = 5$k	90.55	92.07	86.13	88.05	93.39	96.80	86.38	93.32	99.93	99.94	98.61	99.36	98.18	99.17	98.58	99.35	99.15	99.62	94.55	96.41	744s
	$R = 7$k	90.58	92.09	86.17	88.07	93.41	96.81	86.39	93.33	99.93	99.94	98.61	99.36	98.18	99.17	98.59	99.35	99.16	99.62	94.56	96.41	775s
	$R = 10$k	90.55	92.08	86.14	88.05	93.40	96.80	86.37	93.32	99.93	99.94	98.61	99.36	98.18	99.17	98.58	99.35	99.15	99.62	94.55	96.41	784s
Seed = 123	$R = 1$k	90.64	92.15	86.31	88.19	93.53	96.86	86.83	93.51	99.94	99.94	98.66	99.38	98.26	99.20	98.53	99.32	99.17	99.62	94.65	96.46	804s
	$R = 3$k	90.71	92.18	86.39	88.23	93.56	96.86	86.88	93.53	99.94	99.94	98.67	99.38	98.26	99.20	98.53	99.33	99.18	99.62	94.68	96.47	765s
	$R = 5$k	90.70	92.18	86.38	88.22	93.54	96.86	86.86	93.53	99.94	99.94	98.67	99.38	98.26	99.20	98.53	99.33	99.18	99.62	94.67	96.47	820s
	$R = 7$k	90.72	92.18	86.41	88.23	93.56	96.86	86.88	93.53	99.94	99.94	98.67	99.38	98.26	99.20	98.53	99.33	99.18	99.62	94.68	96.48	825s
	$R = 10$k	90.70	92.18	86.38	88.21	93.55	96.86	86.86	93.53	99.94	99.94	98.67	99.38	98.26	99.20	98.53	99.33	99.18	99.62	94.67	96.47	828s
Seed = 456	$R = 1$k	90.86	92.23	86.72	88.37	93.79	96.97	87.11	93.62	99.94	99.94	98.71	99.40	98.34	99.23	98.59	99.35	99.21	99.63	94.81	96.53	809s
	$R = 3$k	90.88	92.24	86.73	88.38	93.80	96.97	87.11	93.62	99.94	99.94	98.71	99.40	98.34	99.23	98.59	99.35	99.21	99.64	94.81	96.53	845s
	$R = 5$k	90.89	92.25	86.74	88.38	93.80	96.97	87.11	93.63	99.94	99.94	98.71	99.40	98.34	99.23	98.59	99.35	99.21	99.64	94.82	96.53	846s
	$R = 7$k	90.88	92.24	86.73	88.37	93.80	96.97	87.11	93.62	99.94	99.94	98.71	99.40	98.34	99.23	98.59	99.35	99.21	99.64	94.81	96.53	824s
	$R = 10$k	90.88	92.24	86.73	88.37	93.80	96.97	87.11	93.62	99.94	99.94	98.71	99.40	98.34	99.23	98.59	99.35	99.21	99.64	94.81	96.53	995s
Seed = 789	$R = 1$k	91.80	92.92	87.99	89.40	94.08	97.11	88.14	94.11	99.93	99.93	98.86	99.46	98.64	99.36	98.77	99.43	99.33	99.68	95.28	96.82	827s
	$R = 3$k	91.63	92.81	87.75	89.24	93.95	97.06	87.91	94.02	99.93	99.93	98.83	99.44	98.61	99.34	98.75	99.42	99.32	99.68	95.18	96.77	878s
	$R = 5$k	91.63	92.82	87.76	89.25	93.96	97.07	87.91	94.02	99.93	99.93	98.83	99.44	98.61	99.34	98.75	99.42	99.32	99.68	95.18	96.77	775s
	$R = 7$k	91.60	92.79	87.71	89.21	93.93	97.05	87.87	94.00	99.93	99.93	98.82	99.44	98.60	99.34	98.75	99.42	99.31	99.68	95.17	96.76	833s
	$R = 10$k	91.60	92.79	87.71	89.21	93.93	97.05	87.86	94.00	99.93	99.93	98.82	99.44	98.60	99.34	98.75	99.42	99.31	99.68	95.17	96.76	925s

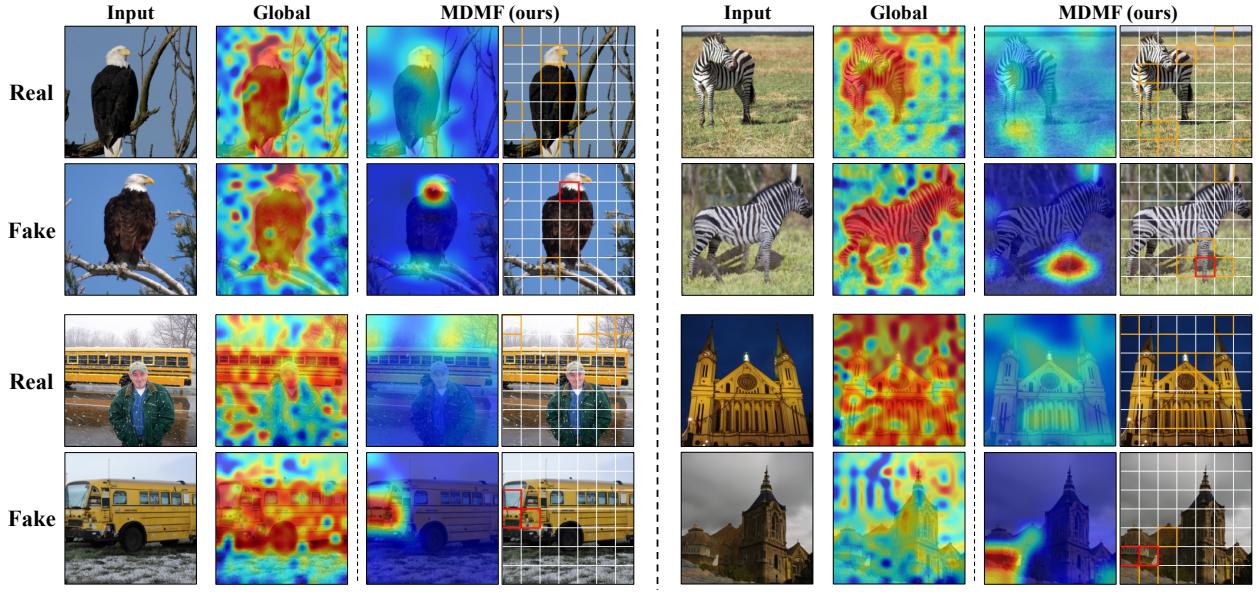


Figure 6. **Qualitative visualization on ADM.** We compare real images and category-matched generated images, visualizing the responses of a global pooling baseline versus MDMF. Warmer colors indicate higher predicted likelihood of being fake.

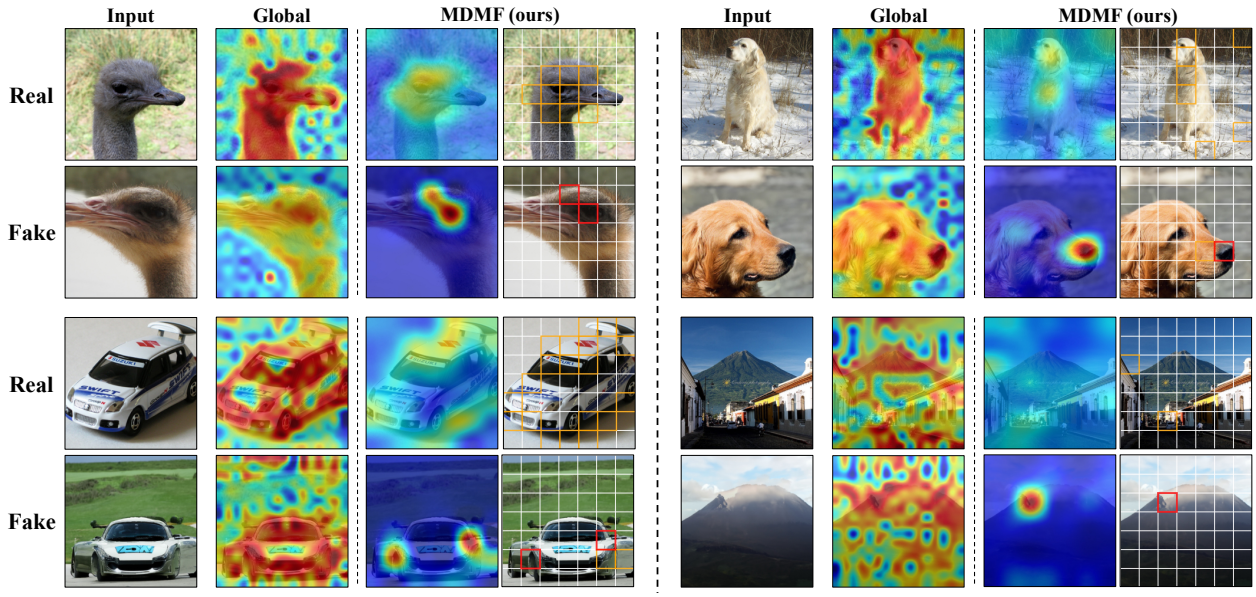


Figure 7. **Additional qualitative visualization on ADMG.** We compare real images and category-matched generated images, visualizing the responses of a global pooling baseline versus MDMF. Warmer colors indicate higher predicted likelihood of being fake.

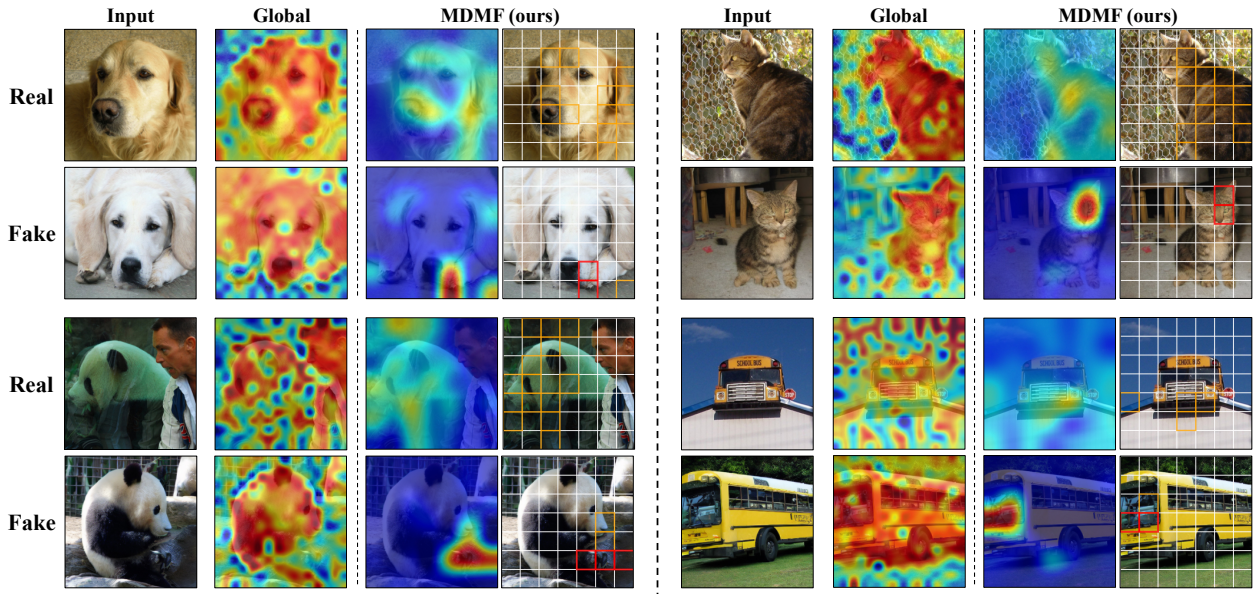


Figure 8. **Additional qualitative visualization on LDM.** We compare real images and category-matched generated images, visualizing the responses of a global pooling baseline versus MDMF. Warmer colors indicate higher predicted likelihood of being fake.

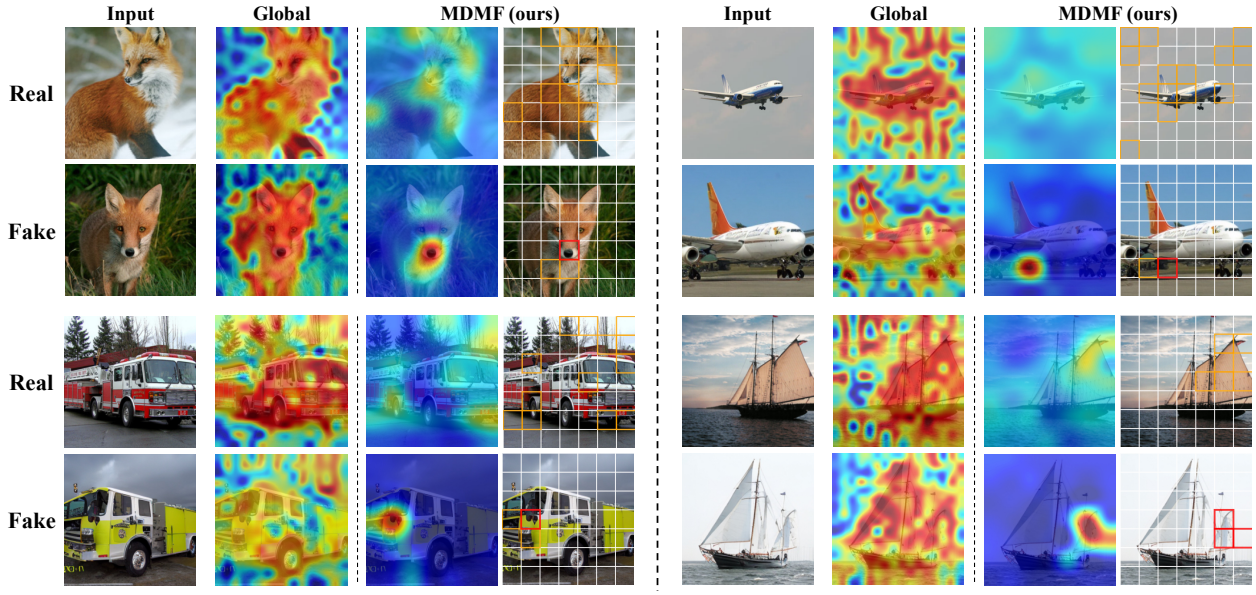


Figure 9. **Additional qualitative visualization on DiT-XL-2.** We compare real images and category-matched generated images, visualizing the responses of a global pooling baseline versus MDMF. Warmer colors indicate higher predicted likelihood of being fake.