Your AI-Generated Image Detector Can Secretly Achieve SOTA Accuracy, If Calibrated

Muli Yang[1], Gabriel James Goenawan[1], Henan Wang[2], Huaiyuan Qin[1], Chenghao Xu[3], Yanhua Yang[3]*, Fen Fang[1], Ying Sun[1], Joo-Hwee Lim[1], Hongyuan Zhu[1]

[1]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[2]Independent Researcher
[3]Xidian University, China
{yangml, goenawan, qinhy, fangf, suny, joohwee, zhuh}@a-star.edu.sg;
nnhhwang@gmail.com; {chx@stu., yanhyang@}xidian.edu.cn

Abstract

Despite being trained on balanced datasets, existing AI-generated image detectors often exhibit systematic bias at test time, frequently misclassifying fake images as real. We hypothesize that this behavior stems from distributional shift in fake samples and implicit priors learned during training. Specifically, models tend to overfit to superficial artifacts that do not generalize well across different generation methods, leading to a misaligned decision threshold when faced with test-time distribution shift. To address this, we propose a theoretically grounded post-hoc calibration framework based on Bayesian decision theory. In particular, we introduce a learnable scalar correction to the model's logits, optimized on a small validation set from the target distribution while keeping the backbone frozen. This parametric adjustment compensates for distributional shift in model output, realigning the decision boundary even without requiring ground-truth labels. Experiments on challenging benchmarks show that our approach significantly improves robustness without retraining, offering a lightweight and principled solution for reliable and adaptive AI-generated image detection in the open world.

Code — https://github.com/muliyangm/AIGI-Det-Calib

1 Introduction

The rapid progress of AI-driven generative models has enabled the creation of highly realistic synthetic images. Modern techniques, from generative adversarial networks (GANs) to diffusion models, now produce photographs and artwork that are often indistinguishable from real images. While these technologies empower creativity in fields such as art, design, and media production, they unintentionally introduce pressing challenges around authenticity, trust, and security. As synthetic media becomes increasingly accessible and indistinguishable from real content, developing reliable methods to detect AI-generated images is not only vital for digital forensics and intellectual property protection, but also foundational to maintaining information integrity in the age of generative AI (Mahara and Rishe 2025).

Most existing methods are developed and evaluated under constrained conditions, often assuming that test data shares the same distribution as the training set (Zhu et al. 2024a; Yang et al. 2025a). In practice, however, detectors trained on forgeries from a specific generative model tend to overfit to superficial artifacts and fail to generalize when exposed to out-of-distribution (OOD) samples, such as those generated by novel architectures or exhibiting unseen statistical properties. A growing body of work has demonstrated that even state-of-the-art detectors, which perform nearly perfectly in-distribution, suffer dramatic performance degradation under distribution shift (Nadimpalli and Rattani 2022; Xiao et al. 2025a; Yan et al. 2025a), underscoring the brittleness of existing AI-generated image detectors in realistic and evolving generative environments.

In our empirical analysis, we identify a consistent failure pattern: even under class-balanced settings, models are significantly more likely to misclassify fake images as real. As illustrated in Fig. 1, we observe that the logits output by detectors on fake images exhibit a global shift, such that the optimal decision boundary no longer lies at zero. This deviation indicates a fundamental mismatch between the model's learned decision threshold and the true distribution at test time, motivating a deeper investigation into its causes. We hypothesize that this bias arises from a "lazy" decision mechanism developed during training (Ojha, Li, and Lee 2023; Rajan and Lee 2025), where the model overly relies on superficial and spurious artifacts that are prevalent in seen fake samples, such as frequency noise or edge inconsistencies. When these artifacts are absent in unseen fake images, the model defaults to classifying them as real. This reliance on non-semantic fake-specific cues severely limits the model's ability to generalize and leads to systematic under-detection of fakes at test time. Further analysis reveals that this behavior can be attributed to two interacting sources of distributional shift: (a) *label prior shift*, where the marginal distribution of real *vs.* fake images differs between training and testing, and (b) *class-conditional input shift*, where the distribution of inputs conditioned on the fake class changes due to new generation techniques. Notably, such shift is particularly prominent in the fake class, as different generative models exhibit coherent and systematic deviations in visual statistics, *e.g.*, texture smoothness, spectral frequency, or semantic integrity. This causes the model's estimated log-likelihood ratios to be consistently distorted, shifting the de-

*Corresponding author.

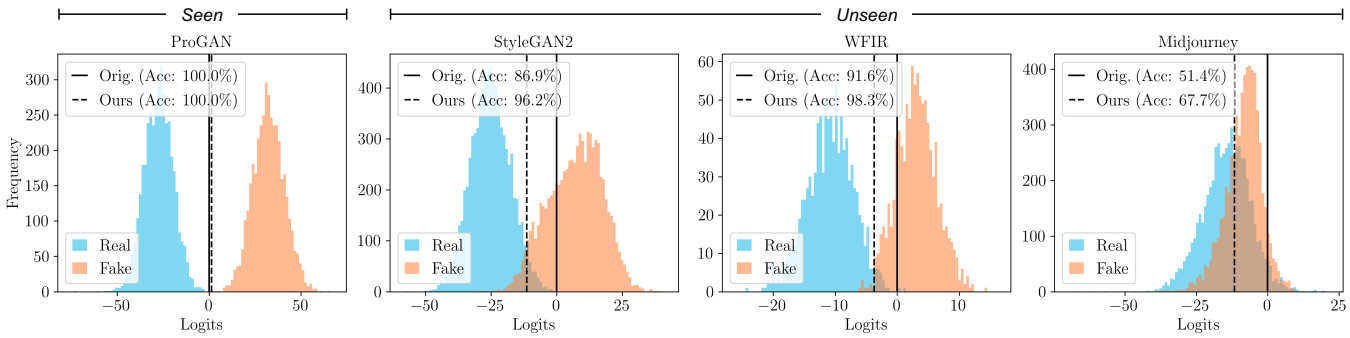

Figure 1: Logit distributions of a popular AI-generated image detector, CNNSpot (Wang et al. 2020), pretrained on ProGAN-generated fake images and evaluated on previously unseen fake images from StyleGAN2, WhichFaceIsReal (WFIR), and Midjourney, reveal a tendency to misclassify these unfamiliar fake samples as real. Our proposed calibration method significantly enhances detection accuracy by adaptively shifting the decision boundary to better align with the skewed data distribution.

cision boundary towards the real class and exacerbating the tendency to misclassify fakes.

To formally model this effect, we adopt a Bayesian decision-theoretic framework and argue that, under distributional shift, the optimal classifier must adapt its decision boundary to the posterior induced by the new test distribution. Since retraining the full model is often impractical, we introduce a lightweight, theoretically grounded post-hoc calibration method: a learnable scalar bias α applied to the model's output logits. This scalar globally adjusts the decision threshold, correcting for both label and input shift. Notably, α can be efficiently optimized using only a few unlabeled samples from the target distribution, while keeping the model backbone fixed. Under mild assumptions, this calibration approximates the Bayes-optimal classifier for the shifted distribution. As our approach requires no access to training data, loss functions, or model internals, as well as avoids retraining or additional supervision, it can be effortlessly applied to any AI-generated image detector, regardless of architecture.

In summary, our method provides a principled yet efficient way to restore model robustness under realistic distribution shift. Without modifying the original detector, our scalar calibration significantly improves performance across a wide range of generative scenarios. These results demonstrate that test-time decision bias, when properly diagnosed and corrected, can be mitigated with minimal computational and data overhead, offering a practical solution to the brittleness of modern AI-generated image detectors.

Our contributions are as follows:

- We identify a systematic bias of most existing AI-generated image detectors during test time, which frequently misclassify fake images as real.
- Using Bayesian theory, we show that this systematic bias stems from the class-conditional input shift and label shift between mismatched train-test distributions.
- Based on mild assumptions, we propose a post-hoc calibration method that optimizes a learnable scalar correction to the model's logits (while keeping the backbone frozen), largely improving most existing AI-generated image detectors' accuracy during test.

2 Related Work

AI-Generated Image Detection. The rapid advancement of generative models, including GANs, VAEs, and diffusion models (Ho, Jain, and Abbeel 2020; Xu et al. 2024), has enabled the creation of highly photorealistic synthetic images, raising pressing concerns around misinformation and visual authenticity. This has led to an increasing demand for reliable methods to distinguish real from generated content (Chen, Yao, and Niu 2024; Chen et al. 2024; Nie et al. 2024; Zhong et al. 2025; Nguyen, Azizpour, and Stamm 2025; Xiao et al. 2025b; Jia et al. 2025; Guillaro et al. 2025).

A wide range of detection approaches has been proposed, leveraging pixel- or patch-level artifacts (Nataraj et al. 2019; Chai et al. 2020; Wang et al. 2020; Ju et al. 2022; Zhong et al. 2023; Lorenz, Durall, and Keuper 2023; Tan et al. 2024; Fu et al. 2025), and modeling generation-specific fingerprints through gradient features or network signatures (Marra et al. 2019; Yu, Davis, and Fritz 2019; Liu et al. 2022; Jeong et al. 2022; Tan et al. 2023; Wang et al. 2023a; Li et al. 2025), as well as utilizing reconstruction inconsistencies from pretrained generative models (Wang et al. 2023c; Ricker et al. 2024). On the other hand, frequency-based analyses reveal spectral discrepancies among generated images (Frank et al. 2020; Dzanic, Shah, and Witherden 2020), and AIDE (Yan et al. 2025a) further integrates spectral and semantic features for improved robustness. Due to the prevalence of vision-language models (VLMs) and their wide applications (Wang et al. 2023b; Yang et al. 2025b; Min et al. 2025), cross-modal inconsistencies have also been explored using CLIP or other VLMs, enabling zero-shot or lightweight fake image detectors (Ojha, Li, and Lee 2023; Liu et al. 2024; Cozzolino et al. 2024; Koutlis and Papadopoulos 2024).

While these approaches have demonstrated promising results, most assume static inference distributions and are vulnerable to domain shift, limiting their generalization to unseen generative models, which is the key challenge for real-world and practical deployment.

Post-Hoc Calibration. Post-hoc calibration is an effective strategy for adapting model outputs to the test data distribution, which often differs from the training distribution. It has been extensively explored in contexts such as class-imbalanced learning (Buda, Maki, and Mazurowski 2018; Tang, Huang, and Zhang 2020; Wu et al. 2021; Hong et al. 2021), where models tend to overfit to majority classes, and continual learning (Wu et al. 2019; Hou et al. 2019; Zhao et al. 2020), where newer data is often favored. Notably, Menon et al. (2021) introduced *logit adjustment*, a post-processing method that adjusts biased classifier outputs using theoretically optimal shifts based on class frequencies, offering a unifying framework for various heuristic bias-correction approaches (Kang et al. 2019; Ye et al. 2020; Islam et al. 2021; Kim and Kim 2020). Similar ideas have also been extended to mitigate prediction bias in foundation models (Zhu et al. 2023; Mai et al. 2024).

In this paper, we demonstrate that AI-generated image detectors exhibit significant prediction bias when the target data distribution shifts, even under class-balanced settings, which contrasts with previously studied scenarios. We provide a theoretical explanation grounded in Bayesian theories and propose a principled yet efficient post-hoc calibration approach that corrects for this bias using only a small number of labeled (or even unlabeled) target examples.

3 Train-Test Misalignment in AI-Generated Image Detection

3.1 A Probabilistic Problem Formulation

We consider a binary classification problem for AI-generated image detection, where each input $x \in \mathcal{X}$ is assigned a binary label $y \in \{0, 1\}$, with $y = 0$ denoting a real image and $y = 1$ a fake image. Let the joint distributions over data and labels in the training and test domains be denoted by $P_{\mathsf{tr}}(x, y)$ and $P_{\mathsf{te}}(x, y)$, respectively.

We assume that the classifier is trained to minimize cross-entropy under the training distribution P_{tr}, and outputs a logit $f(x) \in \mathbb{R}$, whose sigmoid $\sigma(f(x)) = \frac{1}{1+\exp(-f(x))}$ estimates the probability of the input being fake. The final decision is made by thresholding the probability at τ, *i.e.*, predicting $\hat{y} = 1$ if $\sigma(f(x)) > \tau$, where τ denotes the decision threshold, typically set to $\tau = 0.5$ (*i.e.*, when $f(x) = 0$).

However, in realistic open-world AI-generated image detection settings, the test distribution P_{te} often differs from the training distribution P_{tr}. In particular, the fake images in the test set may be synthesized by different generative models not seen during training (*e.g.*, training on StarGAN fakes, testing on Stable Diffusion fakes). This introduces a covariate mismatch in $P(x|y = 1)$, while $P(x|y = 0)$ (real images) remains approximately unchanged. Additionally, the class priors $P_{\mathsf{tr}}(y)$ and $P_{\mathsf{te}}(y)$ may also differ. Hence, we adopt a more expressive modeling assumption that jointly considers the following shifts:

- *Class-Conditional Input Shift:* $P_{\mathsf{tr}}(x|y) \neq P_{\mathsf{te}}(x|y)$, especially for $y = 1$.
- *Label Prior Shift:* $P_{\mathsf{tr}}(y) \neq P_{\mathsf{te}}(y)$,

In practice, both shifts may occur simultaneously, leading to $P_{\mathsf{tr}}(x, y) \neq P_{\mathsf{te}}(x, y)$, and thus $P_{\mathsf{te}}(y|x) \neq P_{\mathsf{tr}}(y|x)$. In such cases, a model trained on P_{tr} may yield biased decision boundaries when evaluated on P_{te}, as illustrated in Fig. 1.

3.2 Default Threshold Is Not Bayes-Optimal

We interpret the phenomenon shown in Fig. 1 through the following propositions grounded in Bayesian decision theory (Devroye, Györfi, and Lugosi 1997).

Proposition 1 (Bayes Non-optimality)**.** *The default threshold $f(x) = 0$ (or $\tau = 0.5$) is not Bayes-optimal under class-conditional input shift and label prior shift.*

Proof. By Bayes' theorem, the posterior under the training distribution is $P_{\mathsf{tr}}(y = 1|x) = \frac{P_{\mathsf{tr}}(x|y=1)P_{\mathsf{tr}}(y=1)}{P_{\mathsf{tr}}(x)}$, which gives the model output logit as

$$f(x) = \log\frac{P_{\mathsf{tr}}(y=1|x)}{P_{\mathsf{tr}}(y=0|x)} = \log\frac{P_{\mathsf{tr}}(x|y=1)P_{\mathsf{tr}}(y=1)}{P_{\mathsf{tr}}(x|y=0)P_{\mathsf{tr}}(y=0)}\,. \quad (1)$$

Under the testing distribution P_{te}, the Bayes-optimal classifier predicts $y = 1$ if

$$P_{\mathsf{te}}(y = 1|x) > P_{\mathsf{te}}(y = 0|x) \iff \log\frac{P_{\mathsf{te}}(x|1)}{P_{\mathsf{te}}(x|0)} + \log\frac{P_{\mathsf{te}}(1)}{P_{\mathsf{te}}(0)} > 0\,. \quad (2)$$

The Bayes-optimal decision boundary satisfies

$$\log\frac{P_{\mathsf{te}}(x|1)}{P_{\mathsf{te}}(x|0)} + \log\frac{P_{\mathsf{te}}(1)}{P_{\mathsf{te}}(0)} = f(x) + \underbrace{\log\frac{P_{\mathsf{te}}(x|1)/P_{\mathsf{tr}}(x|1)}{P_{\mathsf{te}}(x|0)/P_{\mathsf{tr}}(x|0)} + \log\frac{P_{\mathsf{te}}(1)/P_{\mathsf{tr}}(1)}{P_{\mathsf{te}}(0)/P_{\mathsf{tr}}(0)}}_{\Delta(x)} = 0\,. \quad (3)$$

Eq. (3) implies that the mismatch between P_{tr} and P_{te} causes the decision boundaries to differ by $\Delta(x)$:

$$\Delta(x) \doteq \log\frac{P_{\mathsf{te}}(x|1)}{P_{\mathsf{tr}}(x|1)} + \log\frac{P_{\mathsf{te}}(1)(1 - P_{\mathsf{tr}}(1))}{P_{\mathsf{tr}}(1)(1 - P_{\mathsf{te}}(1))}\,, \quad (4)$$

where we assume that the real image distribution remains stable between training and testing, *i.e.*, $P_{\mathsf{te}}(x|0) \doteq P_{\mathsf{tr}}(x|0)$, and that the shift mainly occurs in the fake class.

In Eq. (4), unless $\Delta(x) = 0$ holds for all x, which would require both covariate and label distributions to align, the Bayes-optimal decision boundary does not correspond to $f(x) = 0$ or $\tau(f(x)) = 0.5$. As a result, using $f(x) = 0$ as the default threshold leads to suboptimal decisions under distribution shift. □

3.3 A Scalar Value Can Correct for the Threshold

Let us analyze the behavior of $\Delta(x)$ over $x \sim P_{\mathsf{te}}(x|1)$, *i.e.*, test samples that are truly fake.

Assumption 1 (Systematic Conditional Shift)**.** *Let the log-likelihood ratio between test and train fake distributions be approximately constant:*

$$\log\frac{P_{\mathsf{te}}(x|1)}{P_{\mathsf{tr}}(x|1)} \doteq c, \quad \forall x \sim P_{\mathsf{te}}(x|1)\,. \quad (5)$$

This assumption is justified by the fact that fake images generated by different GANs or diffusion models tend to exhibit coherent and systematic deviations in visual features (*e.g.*, frequency artifacts, textures) (Ricker et al. 2022; You et al. 2025). As a result, the likelihood under the training fake distribution is consistently misaligned, causing the classifier to under- or over-estimate $f(x)$ in a fixed direction.

Assumption 2 (Consistent Prior Shift)**.** *The prior over fake samples in the test set is shifted by a constant factor:*

$$\log \frac{P_{\mathsf{te}}(1)}{P_{\mathsf{tr}}(1)} = \delta\,. \tag{6}$$

This assumption is well justified by Menon et al. (2021), and the rightmost term in Eq. (4) can be written as δ', which is another constant that can be easily derived from δ.

Proposition 2 (Scalar Correction)**.** *The suboptimal threshold can be corrected using a global additive scalar $\tilde{\alpha}$.*

Proof. Under Assumptions 1 and 2, the model output bias $\Delta(x)$ in Eq. (4) becomes

$$\Delta(x) \doteq c + \delta' = \text{const}\,, \quad \forall x \sim P_{\mathsf{te}}(x|1)\,, \tag{7}$$

which justifies the use of a global additive logit correction scalar, $\tilde{\alpha} := -\Delta(x)$, for post-hoc calibration to correct for the biased decision boundary shown in Proposition 1. The calibrated output logit is then written as

$$\tilde{f}(x) := f(x) - \tilde{\alpha}\,, \quad \tilde{\alpha} = -(c + \delta')\,, \tag{8}$$

where $\tilde{\alpha} \in \mathbb{R}$ manifests the optimal threshold adjustment value, which can be estimated using a small subset of data sampled from P_{te}. □

Therefore, post-hoc calibration—namely, adapting τ to minimize expected test error—is not only justified but necessary for robust performance under distribution shift.

4 Post-Hoc Calibration as a Remedy

Building upon the probabilistic analysis in the previous section, we have established that two types of distribution shift between training and testing domains induce a systematic bias in the classifier's output logits. In particular, the logit function $f(x)$, learned under the training distribution $P_{\mathsf{tr}}(x, y)$, tends to systematically underestimate or overestimate the posterior probability $P_{\mathsf{te}}(y = 1|x)$ when evaluated on the test distribution $P_{\mathsf{te}}(x, y)$. To address this issue, we propose a simple yet effective method: calibrating the logit output at test-time using a learnable additive bias $\alpha \in \mathbb{R}$.

We explore two variants of post-hoc calibration, depending on whether ground-truth labels from the test distribution are available. Both variants leverage kernel density estimation (KDE) (Rosenblatt 1956; Parzen 1962) as their underlying mechanism for its simplicity and effectiveness.

Supervised Calibration. In scenarios where a small amount of labeled target data is available, we propose a supervised logit calibration method based on KDE and accuracy maximization. The core idea is to model the distribution of classifier logits for each class using KDE, and to select a calibration value that maximizes the classification accuracy on the target distribution.

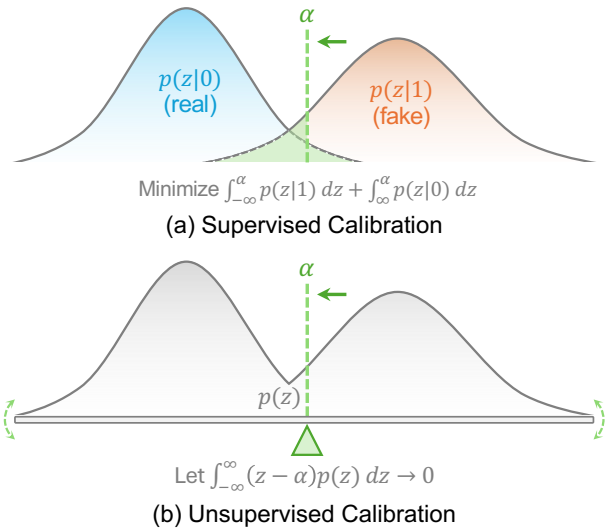

Figure 2: Conceptual illustration of our proposed (a) supervised and (b) unsupervised calibration methods, both designed to identify an optimal scalar α that achieves an ideal separation between real and fake distributions, with or without access to ground-truth labels.

Let $z = f(x) \in \mathbb{R}$ denote the real-valued logits produced by a binary classifier. Given a small set of labeled samples from the target domain, we estimate two class-conditional densities using Gaussian kernel density estimation:

$$p_0(z) := p(z|y = 0)\,, \quad p_1(z) := p(z|y = 1)\,. \tag{9}$$

As shown in Fig. 2, the expected classification error for a given calibration value α is then defined as

$$\begin{aligned}\mathcal{R}(\alpha) &= P(z > \alpha \mid y = 0) + P(z \leqslant \alpha \mid y = 1) \\ &= \int_{-\infty}^{\alpha} p_1(z)\, dz + \int_{\infty}^{\alpha} p_0(z)\, dz\,.\end{aligned} \tag{10}$$

The optimal calibration is thus found by minimizing $\mathcal{R}(\alpha)$:

$$\alpha^\star = \arg\min_{\alpha} \mathcal{R}(\alpha)\,, \tag{11}$$

where we leave the detailed method description and optimization procedure to Appendix.

This method is both simple and effective, requiring only a handful of labeled examples and no assumptions on the parametric form of the underlying logit distributions. Its ability to adaptively model asymmetric or multimodal distributions makes it especially useful in real-world transfer and domain adaptation scenarios. In Sec. 5.2 we also compare it with several other supervised threshold selection methods.

Unsupervised Calibration. In fully unsupervised settings where no labeled target data is available, we propose a calibration method that leverages the intrinsic structure of the logit distribution to recover an effective decision boundary. Our approach builds upon a fundamental assumption in binary classification: when the underlying data distribution is

Method ↓	ProGAN	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN	StyleGAN2	WFIR	ADM	Glide	Midjourney	SD v1.4	SD v1.5	VQDM	Wukong	DALLE2	**Average**	Δ
CNNSpot (2020)	100.00	90.17	71.17	87.62	94.60	81.43	86.91	91.65	60.40	58.06	51.39	50.57	50.53	56.45	51.02	51.25	70.83	
+ *Ours (Sup.)*	100.00	97.95	83.30	88.64	95.02	86.83	96.15	98.30	67.94	65.63	67.48	64.64	65.00	63.12	59.75	51.70	78.22	**+7.39**
+ *Ours (Unsup.)*	100.00	97.59	83.25	88.80	95.15	86.70	95.41	98.10	67.42	65.56	67.61	63.98	64.42	62.68	57.90	51.90	77.90	**+7.08**
FreDect (2020)	99.36	78.03	81.97	78.77	94.62	80.56	66.19	50.75	63.40	54.12	45.87	38.77	39.21	77.80	40.29	34.70	64.03	
+ *Ours (Sup.)*	99.70	77.78	86.48	80.77	97.40	80.21	74.92	58.00	68.19	63.19	45.99	49.52	42.43	78.05	49.62	49.75	68.88	**+4.85**
+ *Ours (Unsup.)*	99.74	73.49	86.10	81.19	96.85	80.45	74.04	57.70	66.55	61.44	43.93	34.66	34.29	77.98	38.32	34.65	65.09	**+1.06**
Fusing (2022)	99.99	85.21	77.35	87.02	97.02	76.92	83.27	66.85	56.53	57.16	52.17	51.05	51.36	55.13	51.72	52.85	68.85	
+ *Ours (Sup.)*	99.99	96.28	84.62	89.02	98.05	84.36	96.44	87.05	67.53	69.92	65.20	62.42	63.21	68.83	60.11	61.75	78.42	**+9.57**
+ *Ours (Unsup.)*	99.99	89.58	81.97	88.64	97.52	80.49	88.26	82.80	65.43	67.44	61.88	59.86	60.23	65.97	58.25	61.30	75.60	**+6.75**
LNP (2022)	99.78	92.15	83.05	84.60	99.90	75.17	93.87	55.15	78.98	79.67	57.83	79.37	79.24	69.94	75.74	93.05	81.09	
+ *Ours (Sup.)*	99.81	94.48	84.75	88.00	100.00	77.21	97.25	91.40	85.38	81.89	79.77	87.55	86.86	80.53	85.56	94.30	88.42	**+7.33**
+ *Ours (Unsup.)*	99.52	94.20	83.65	86.64	99.90	76.88	96.82	91.00	84.14	78.21	78.33	83.41	83.86	79.89	83.86	94.35	87.17	**+6.07**
LGrad (2023)	99.98	90.47	88.92	85.69	99.62	82.81	87.77	58.50	66.13	71.67	70.47	65.24	65.91	74.51	60.30	71.25	77.45	
+ *Ours (Sup.)*	99.99	94.23	90.18	86.11	99.85	86.96	95.55	59.70	66.07	75.08	70.23	66.29	67.26	74.31	62.81	78.40	79.56	**+2.11**
+ *Ours (Unsup.)*	99.95	94.16	90.05	86.11	99.65	86.74	95.61	58.20	66.00	75.24	70.41	66.07	66.83	74.17	62.07	78.35	79.35	**+1.90**
UnivFD (2023)	99.81	84.93	95.08	98.33	95.75	99.47	74.96	86.90	66.87	62.46	56.13	63.66	63.49	85.31	70.93	50.75	78.43	
+ *Ours (Sup.)*	99.72	90.09	95.70	98.33	96.62	99.48	92.60	90.15	78.38	76.96	68.66	79.83	79.17	89.81	83.08	63.00	86.35	**+7.92**
+ *Ours (Unsup.)*	99.74	89.53	95.33	98.30	96.45	99.50	92.22	90.30	78.38	76.89	68.77	79.59	79.09	89.16	83.11	62.80	86.20	**+7.77**
RINE (2024)	100.00	88.86	99.60	99.32	99.55	99.77	94.50	97.35	74.61	80.72	57.12	83.96	83.35	89.79	84.95	54.85	86.77	
+ *Ours (Sup.)*	99.96	95.73	99.62	99.74	99.82	99.87	99.31	97.15	89.40	92.95	80.48	93.90	93.96	95.66	93.97	81.50	94.56	**+7.80**
+ *Ours (Unsup.)*	99.99	93.57	98.47	99.74	97.20	99.82	99.01	95.90	88.92	91.35	80.16	92.88	92.75	95.88	93.40	81.60	93.79	**+7.02**
AIDE (2025a)	99.99	99.65	83.95	98.49	99.90	73.24	98.01	94.75	93.47	95.09	77.20	92.95	92.84	95.12	93.49	96.60	92.80	
+ *Ours (Sup.)*	99.99	99.75	85.55	98.60	99.92	91.01	99.52	95.65	94.20	95.08	79.05	93.14	93.19	95.62	93.80	96.60	94.42	**+1.62**
+ *Ours (Unsup.)*	99.98	99.75	85.30	97.31	99.77	82.17	99.30	95.60	94.08	94.95	78.64	93.37	93.16	95.54	93.64	96.50	93.69	**+0.90**
Effort (2025b)	99.92	89.17	99.12	99.96	100.00	99.94	90.21	70.30	59.18	70.43	50.58	71.62	71.42	75.13	69.34	53.25	79.35	
+ *Ours (Sup.)*	99.89	94.02	99.45	100.00	100.00	99.94	96.02	93.05	76.78	86.58	68.63	88.88	88.55	89.91	87.17	62.75	89.48	**+10.13**
+ *Ours (Unsup.)*	99.92	90.22	99.12	99.96	100.00	99.94	91.68	82.20	71.62	78.25	68.62	79.86	79.79	79.10	78.64	62.35	85.08	**+5.73**

Table 1: Results on AIGCDetectBenchmark (Zhong et al. 2023). Accuracies (%) of different detectors (rows) in detecting real and fake images from different generators (columns) are reported. These methods are trained on real images from LSUN and fake images generated by ProGAN and evaluated over 16 generators.

separable, the classifier logits tend to form a bimodal distribution, implicitly reflecting the presence of two latent modes corresponding to the two classes. An ideal decision threshold should hence lie near the valley between these modes and induce a symmetric partitioning of the distribution.

Formally, given a set of unlabelled target logits $\{z_i\}_{i=1}^n$, we begin by estimating the continuous probability density function $p(z)$ via Gaussian kernel density estimation:

$$p(z) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{z - z_i}{h}\right), \tag{12}$$

where $K(\cdot)$ is the Gaussian kernel and h is the bandwidth, whose details are introduced in Appendix.

Our goal is to find a calibration value α that aligns with the distributional symmetry of $p(z)$. To do this, we introduce a symmetry surrogate objective by treating α as a reference point and minimizing the first-order weighted shift of the distribution with respect to α:

$$\Phi(\alpha) = \int_{-\infty}^{\infty} (z - \alpha) \cdot p(z)\, dz\,. \tag{13}$$

This expression quantifies the imbalance of the logit distribution with respect to α as a reflective center. Intuitively, when $\Phi(\alpha) = 0$, the distribution $p(z)$ is symmetric around α, which serves as a natural partition boundary. Notably, we propose a moment-balancing optimization method to ensure robust unsupervised performance across distributions by adaptively penalizing low-confident logits, whose details are deferred to Appendix.

This method requires no label supervision and is robust to mild distributional shift, as it operates directly on the empirical logit geometry. It is particularly effective when the logits exhibit a latent bimodal structure, enabling recovery of semantically meaningful decision thresholds even under unsupervised conditions.

5 Experiment

5.1 Experimental Setup

Baselines. We evaluate nine representative off-the-shelf AI-generated image (AIGI) detectors, including CNNSpot (Wang et al. 2020), FreDect (Frank et al. 2020), Fusing (Ju et al. 2022), LNP (Liu et al. 2022), LGrad (Tan et al. 2023), UnivFD (Ojha, Li, and Lee 2023), RINE (Koutlis and Papadopoulos 2024), AIDE (Yan et al. 2025a), and Effort (Yan et al. 2025b). Rather than directly comparing with these methods, we apply our calibration strategy on top of them to assess its applicability and effectiveness in enhancing detection performance.

Benchmarks. We use two popular benchmarks in our main experiments. (1) *AIGCDetectBenchmark* (Zhong et al. 2023) covers 16 generative models, including GANs and text-to-image models such as Midjourney and DALL·E 2. The training set contains real images from LSUN and fake

Method ↓	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	**Average**	Δ
CNNSpot (2020)	55.13	99.96	99.85	50.11	50.47	99.28	50.18	49.98	69.37	
+ *Ours (Sup.)*	76.12±0.11	99.95±0.05	99.85±0.13	54.39±0.35	64.98±0.52	99.72±0.22	51.20±0.87	49.08±0.49	74.41±0.11	**+5.04**
+ *Ours (Unsup.)*	75.92±0.52	99.94±0.05	99.82±0.18	49.06±1.02	64.51±0.38	99.51±0.19	40.08±1.22	38.07±0.12	70.86±0.21	**+1.49**
FreDect (2020)	61.34	99.60	99.46	50.40	54.89	97.13	49.87	50.65	70.42	
+ *Ours (Sup.)*	81.67±0.27	99.62±0.13	99.48±0.20	50.57±0.31	75.29±0.59	98.01±0.51	49.40±0.64	55.90±1.37	76.24±0.21	**+5.83**
+ *Ours (Unsup.)*	81.67±0.33	99.60±0.02	99.45±0.04	39.21±0.78	75.24±0.48	97.97±0.12	34.86±0.23	55.53±0.44	72.94±0.11	**+2.52**
Fusing (2022)	58.72	99.98	99.91	57.05	73.50	99.96	64.75	55.43	76.16	
+ *Ours (Sup.)*	82.56±1.50	99.99±0.15	99.89±0.06	86.31±2.67	96.29±0.66	99.90±0.61	94.44±0.97	62.18±0.76	90.20±0.51	**+14.03**
+ *Ours (Unsup.)*	66.09±1.83	99.98±0.00	99.91±0.00	67.18±1.23	80.42±0.85	99.96±0.00	73.80±0.89	60.17±0.78	80.94±0.45	**+4.78**
LNP (2022)	50.34	99.93	99.86	60.88	50.10	99.78	75.44	62.02	74.79	
+ *Ours (Sup.)*	58.34±0.93	99.91±0.05	99.85±0.06	94.75±0.18	85.69±0.12	99.77±0.11	96.38±0.13	61.95±0.32	87.08±0.09	**+12.29**
+ *Ours (Unsup.)*	54.44±1.56	99.97±0.03	99.83±0.03	94.35±0.57	85.83±0.06	99.78±0.06	96.37±0.22	62.00±0.64	86.57±0.25	**+11.78**
LGrad (2023)	68.70	99.40	99.41	51.90	61.57	96.97	51.04	49.73	72.34	
+ *Ours (Sup.)*	81.33±0.22	99.27±0.06	99.41±0.07	54.07±1.75	76.64±0.33	97.82±0.14	56.56±0.58	52.38±1.86	77.19±0.43	**+4.85**
+ *Ours (Unsup.)*	80.33±0.64	99.24±0.07	99.39±0.08	51.76±0.61	76.78±0.19	97.72±0.19	56.53±0.13	46.65±0.35	76.05±0.13	**+3.71**
UnivFD (2023)	85.16	96.89	96.74	53.86	73.61	92.17	56.56	61.40	77.05	
+ *Ours (Sup.)*	87.82±0.37	96.89±0.38	96.67±0.20	57.67±0.92	81.23±0.07	92.67±0.26	66.94±0.66	79.05±0.43	82.37±0.18	**+5.32**
+ *Ours (Unsup.)*	87.71±0.24	96.77±0.38	96.49±0.52	56.62±0.12	81.16±0.17	92.52±0.37	66.85±0.06	78.97±0.20	82.14±0.08	**+5.09**
RINE (2024)	69.38	99.97	99.90	56.73	53.24	99.95	88.95	86.15	81.78	
+ *Ours (Sup.)*	96.33±0.13	99.98±0.02	99.91±0.10	92.86±0.31	97.34±0.26	99.95±0.03	98.70±0.11	98.45±0.29	97.94±0.08	**+16.16**
+ *Ours (Unsup.)*	96.06±0.29	99.97±0.00	99.86±0.02	92.73±0.15	95.67±0.59	99.88±0.07	98.63±0.20	97.90±0.51	97.59±0.12	**+15.80**
AIDE (2025a)	79.38	99.74	99.75	78.54	91.80	98.67	80.27	77.20	88.17	
+ *Ours (Sup.)*	91.66±0.48	99.74±0.14	99.75±0.17	89.81±0.25	96.97±0.29	99.22±0.41	91.63±0.61	77.22±0.48	93.25±0.13	**+5.08**
+ *Ours (Unsup.)*	86.29±0.84	99.78±0.02	99.74±0.02	85.62±0.43	94.88±0.39	98.98±0.15	86.94±0.94	76.60±0.13	91.10±0.20	**+2.94**
Effort (2025b)	82.40	99.83	99.81	78.78	93.31	97.42	91.70	88.05	91.41	
+ *Ours (Sup.)*	94.09±0.43	99.82±0.15	99.83±0.08	96.57±0.16	98.38±0.03	98.47±0.12	97.10±0.18	88.85±0.14	96.64±0.09	**+5.23**
+ *Ours (Unsup.)*	93.40±0.36	99.71±0.07	99.70±0.06	96.57±0.08	98.31±0.13	98.47±0.08	96.88±0.20	87.58±0.84	96.33±0.16	**+4.92**

Table 2: Results on GenImage (Zhu et al. 2024b). Accuracies (%) of different detectors (rows) in detecting real and fake images from different generators (columns) are reported. These methods are trained on real images from ImageNet and fake images generated by SD v1.4 and evaluated over 8 generators.

images generated by ProGAN, while the test set includes diverse fake images from the 16 generative models. Likewise, (2) *GenImage* (Zhu et al. 2024b) comprises ImageNet's 1,000 classes generated using 8 SOTA generators such as Stable Diffusion (SD) and Midjourney, and its training set contains real images from ImageNet and fake images generated by SD v1.4. Benchmark details are in Appendix, where we also conduct experiments on other recent benchmarks.

Implementation Details. We randomly sample a small subset of the test data (referred to as *validation set*) to optimize the scalar calibration parameter α. By default, we use 100 images, which is approximately 1% of the entire test set. The optimized α is then applied to the test data for evaluation. All experiments are conducted on a single NVIDIA RTX A6000 GPU. For each method, we use the official pre-trained checkpoints without any fine-tuning or modification. If no checkpoint is available, we train the model using the official codebase. To ensure statistical robustness, all results are averaged over 10 independent runs, and we report both the mean accuracies and standard deviations.

5.2 Results and Analysis

Results on Standard Benchmarks. Tabs. 1 and 2 illustrate the performance of representative AIGI detectors on the AIGCDetectBenchmark and GenImage datasets, respectively. Our proposed post-hoc calibration strategy consistently improves detection accuracies across different baselines. In particular, especially when the baseline models are equipped with strong pre-trained feature extractors such as CLIP (*e.g.*, RINE (Koutlis and Papadopoulos 2024) and Effort (Yan et al. 2025b)), our method is able to further enhance their performance by larger margins, indicating its complementary benefit. Overall, these results demonstrate that our calibration approach can effectively unlock the latent potential of existing detectors, regardless of the backbone representation quality, and is especially beneficial in handling domain shift and subtle generation artifacts commonly present in AIGI datasets.

Robustness to Image Perturbations. Tab. 3 shows the results under different types of image perturbations, evaluating the robustness of AIGI detectors in real-world applications under potential unseen perturbations. Notably, our method shows strong resistance to image perturbations, especially for JPEG compression, significantly improving the performance by large margins, *e.g.*, +15.39% accuracy for AIDE under JPEG compression (QF $= 90$). These results further validate the robustness and applicability of our method in real-world scenarios.

Method ↓	Original	JPEG Compression		Gaussian Blur	
	Average	QF=95	QF=90	σ=1.0	σ=2.0
CNNSpot (2020)	70.83	64.43	63.18	70.67	69.64
+ *Ours (Sup.)*	78.22	77.61	77.40	76.79	74.46
+ *Ours (Unsup.)*	77.90	77.21	76.98	76.32	73.83
FreDect (2020)	64.03	69.16	68.10	65.75	66.53
+ *Ours (Sup.)*	68.88	77.92	74.37	68.42	70.52
+ *Ours (Unsup.)*	65.09	76.01	72.93	65.49	69.62
Fusing (2022)	68.85	61.82	61.04	68.08	66.66
+ *Ours (Sup.)*	78.42	77.91	77.22	73.15	70.27
+ *Ours (Unsup.)*	75.60	73.79	73.53	71.82	69.02
LNP (2022)	81.09	79.37	79.42	70.21	69.23
+ *Ours (Sup.)*	88.42	83.40	83.47	71.77	69.88
+ *Ours (Unsup.)*	87.17	82.82	82.90	71.43	69.79
LGrad (2023)	77.45	51.79	50.00	54.20	50.03
+ *Ours (Sup.)*	79.56	61.47	56.32	61.43	49.98
+ *Ours (Unsup.)*	79.35	61.13	55.46	61.24	49.53
UnivFD (2023)	78.43	74.10	71.65	70.31	65.66
+ *Ours (Sup.)*	86.35	82.17	80.78	77.38	70.04
+ *Ours (Unsup.)*	86.20	82.08	80.59	77.23	69.59
RINE (2024)	86.77	78.94	76.19	81.33	73.81
+ *Ours (Sup.)*	94.56	90.93	89.19	86.17	76.53
+ *Ours (Unsup.)*	93.79	90.34	88.63	85.61	75.54
AIDE (2025a)	92.80	65.91	60.22	79.48	68.43
+ *Ours (Sup.)*	94.42	79.41	75.61	82.49	73.04
+ *Ours (Unsup.)*	93.69	77.90	73.84	81.83	70.30
Effort (2025b)	79.35	67.37	63.56	75.30	62.52
+ *Ours (Sup.)*	89.48	77.59	74.72	83.47	70.04
+ *Ours (Unsup.)*	85.08	75.30	71.67	81.11	67.89

Table 3: Robustness on JPEG compression and Gaussian blur. We report the accuracies (%) averaged over 16 test sets in AIGCDetectBenchmark (Zhong et al. 2023) with different quality factor (QF) and variance (σ).

Effect of Validation Set Size. Fig. 3 shows the performance using different validation set sizes. We can observe that both our supervised and unsupervised calibration methods perform stably with varying amounts of validation data, demonstrating strong performance with as few as 10 samples, which is less than 0.1% of the data in each test set. These results further confirm the effectiveness of our method, demonstrating its utility as a lightweight and practical test-time enhancement for AIGI detection.

Comparison with Different Estimation Methods for α. Fig. 4 further showcases the performance using different supervised searching methods to estimate α in place of our proposed KDE-based method, such as binary search and training with binary cross-entropy (BCE) loss, whose details are deferred to Appendix. We can observe that our proposed method shows strong performance even with 4 target samples in the validation set, and consistently outperforms the other two counterparts by large margins with more target samples. These results demonstrate the data efficiency of our proposed calibration method, highlighting its practicality and suitability for real-world deployment.

6 Conclusion

In this work, we investigated a critical yet underexplored challenge in AI-generated image detection: the systematic misclassification of fake images under distribution shift, even in class-balanced settings. Through empirical and theoretical analyses, we identified that this phenomenon arises from a mismatch between the model's learned decision threshold and the shifted test-time distribution, particularly due to label prior and class-conditional input shifts in the fake class. To address this, we introduced a lightweight and principled post-hoc calibration method that applies a learnable scalar correction to the model's logits. Without modifying the detector's backbone, our method significantly improves detection robustness across diverse generators. Our findings underscore the importance of aligning the decision boundary with the test distribution, and demonstrate that substantial performance gains can be achieved through minimal yet targeted calibration. This highlights the untapped potential of existing detectors and opens new directions for reliable and adaptive AI-generated image detection under realistic deployment conditions.

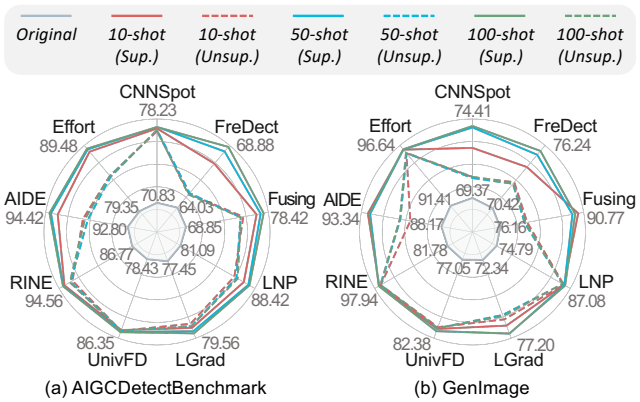

Figure 3: Effect of validation set size on the proposed supervised and unsupervised calibration methods. We report the average accuracies on the two benchmarks.

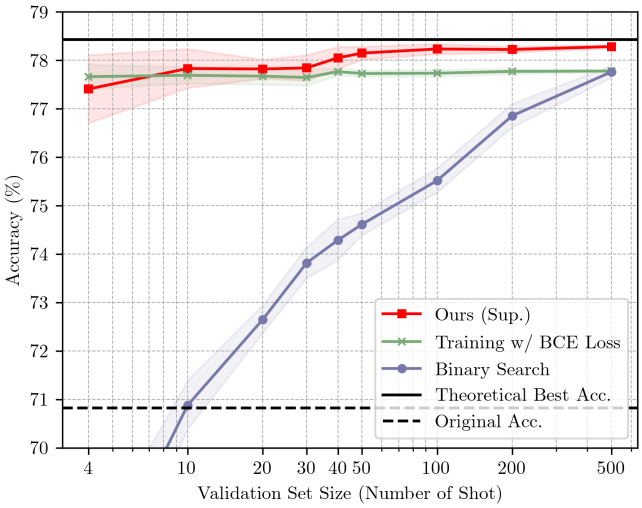

Figure 4: Performance comparison of different supervised calibration methods for estimating α. Average accuracies of CNNSpot (Wang et al. 2020) on AIGCDetectBenchmark (Zhong et al. 2023) are reported.

Acknowledgments

This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021), and the National Natural Science Foundation of China under Grant 62571412. The authors sincerely thank the anonymous reviewers for their valuable comments.

References

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Netw.*, 106: 249–259.

Chai, L.; Bau, D.; Lim, S.-N.; and Isola, P. 2020. What Makes Fake Images Detectable? Understanding Properties That Generalize. In *ECCV*, 103–120.

Chen, J.; Yao, J.; and Niu, L. 2024. A Single Simple Patch Is All You Need for AI-Generated Image Detection. *arXiv:2402.01123*.

Chen, Y.; Zhang, L.; Niu, Y.; Tan, L.; and Chen, P. 2024. Learning on Less: Constraining Pre-trained Model Learning for Generalizable Diffusion-Generated Image Detection. *arXiv:2412.00665*.

Cozzolino, D.; Poggi, G.; Corvi, R.; Nießner, M.; and Verdoliva, L. 2024. Raising the Bar of AI-Generated Image Detection with CLIP. In *CVPR*, 4356–4366.

Devroye, L.; Györfi, L.; and Lugosi, G. 1997. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media.

Dzanic, T.; Shah, K.; and Witherden, F. 2020. Fourier Spectrum Discrepancies in Deep Network Generated Images. *NeurIPS*, 3022–3032.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *ICML*, 3247–3258.

Fu, X.; Yan, Z.; Yang, Z.; Yao, T.; Zhao, Y.; Ding, S.; and Li, X. 2025. PiD: Generalized AI-Generated Images Detection with Pixelwise Decomposition Residuals. In *ICML*.

Guillaro, F.; Zingarini, G.; Usman, B.; Sud, A.; Cozzolino, D.; and Verdoliva, L. 2025. A Bias-Free Training Paradigm for More General AI-Generated Image Detection. In *CVPR*, 18685–18694.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 6840–6851.

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling Label Distribution for Long-Tailed Visual Recognition. In *CVPR*, 6626–6636.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a Unified Classifier Incrementally via Rebalancing. In *CVPR*, 831–839.

Islam, M.; Seenivasan, L.; Ren, H.; and Glocker, B. 2021. Class-Distribution-Aware Calibration for Long-Tailed Visual Recognition. *arXiv:2109.05263*.

Jeong, Y.; Kim, D.; Ro, Y.; Kim, P.; and Choi, J. 2022. FingerprintNet: Synthesized Fingerprints for Generated Image Detection. In *ECCV*, 76–94.

Jia, Z.; Huang, C.; Zhu, Y.; Fei, H.; Duan, X.; Yuan, Z.; Deng, Y.; Zhang, J.; Zhang, J.; and Zhou, J. 2025. Secret Lies in Color: Enhancing AI-Generated Images Detection with Color Distribution Analysis. In *CVPR*, 13445–13454.

Ju, Y.; Jia, S.; Ke, L.; Xue, H.; Nagano, K.; and Lyu, S. 2022. Fusing Global and Local Features for Generalized AI-Synthesized Image Detection. In *ICIP*, 3465–3469.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. *arXiv:1910.09217*.

Kim, B.; and Kim, J. 2020. Adjusting Decision Boundary for Class Imbalanced Learning. *IEEE Access*, 8: 81674–81685.

Koutlis, C.; and Papadopoulos, S. 2024. Leveraging Representations From Intermediate Encoder-Blocks for Synthetic Image Detection. In *ECCV*, 394–411.

Li, X.; Chen, J.; Yu, Y.; Song, X.; Shan, H.; and Jiang, Y.-G. 2025. Revealing the Implicit Noise-based Imprint of Generative Models. *arXiv:2503.09314*.

Liu, B.; Yang, F.; Bi, X.; Xiao, B.; Li, W.; and Gao, X. 2022. Detecting Generated Images by Real Images. In *ECCV*, 95–110.

Liu, H.; et al. 2024. Forgery-Aware Adaptive Transformer for Generalizable Synthetic Image Detection. In *CVPR*.

Lorenz, P.; Durall, R. L.; and Keuper, J. 2023. Detecting Images Generated by Deep Diffusion Models Using Their Local Intrinsic Dimensionality. In *ICCV*, 448–459.

Mahara, A.; and Rishe, N. 2025. Methods and Trends in Detecting Generated Images: A Comprehensive Review. *arXiv:2502.15176*.

Mai, Z.; Chowdhury, A.; Zhang, P.; Tu, C.-H.; Chen, H.-Y.; Pahuja, V.; Berger-Wolf, T.; Gao, S.; Stewart, C.; Su, Y.; et al. 2024. Fine-Tuning Is Fine, If Calibrated. In *NeurIPS*, 136084–136119.

Marra, F.; Gragnaniello, D.; Verdoliva, L.; and Poggi, G. 2019. Do GANs Leave Artificial Fingerprints? In *MIPR*, 506–511.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-Tail Learning via Logit Adjustment. In *ICLR*.

Min, Y.; Yang, M.; Zhang, J.; Wang, Y.; Wu, A.; and Deng, C. 2025. Vision-Language Interactive Relation Mining for Open-Vocabulary Scene Graph Generation. In *ICCV*, 16755–16764.

Nadimpalli, A. V.; and Rattani, A. 2022. On Improving Cross-Dataset Generalization of Deepfake Detectors. In *CVPR*, 91–99.

Nataraj, L.; Mohammed, T. M.; Chandrasekaran, S.; Flenner, A.; Bappy, J. H.; Roy-Chowdhury, A. K.; and Manjunath, B. 2019. Detecting GAN generated Fake Images using Co-occurrence Matrices. *arXiv:1903.06836*.

Nguyen, T. D.; Azizpour, A.; and Stamm, M. C. 2025. Forensic Self-Descriptions Are All You Need for Zero-Shot Detection, Open-Set Source Attribution, and Clustering of AI-Generated Images. In *CVPR*, 3040–3050.

Nie, J.; Zhang, Y.; Liu, T.; Cheung, Y.-m.; Han, B.; and Tian, X. 2024. Detecting Discrepancies Between AI-Generated and Natural Images Using Uncertainty. *arXiv:2412.05897*.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards Universal Fake Image Detectors That Generalize Across Generative Models. In *CVPR*, 24480–24489.

Parzen, E. 1962. On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist.*, 33(3): 1065–1076.

Rajan, A. S.; and Lee, Y. J. 2025. Stay-Positive: A Case for Ignoring Real Image Features in Fake Image Detection. In *ICML*.

Ricker, J.; Damm, S.; Holz, T.; and Fischer, A. 2022. Towards the Detection of Diffusion Model Deepfakes. *arXiv:2210.14571*.

Ricker, J.; et al. 2024. AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error. In *CVPR*.

Rosenblatt, M. 1956. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Statist.*, 27(3): 832–837.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. Rethinking the Up-Sampling Operations in CNN-Based Generative Network for Generalizable Deepfake Detection. In *CVPR*, 28130–28139.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *CVPR*, 12105–12114.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. In *NeurIPS*, 1513–1524.

Wang, H.; Fei, J.; Dai, Y.; Leng, L.; and Xia, Z. 2023a. General GAN-Generated Image Detection by Data Augmentation in Fingerprint Domain. In *ICME*, 1187–1192.

Wang, H.; Yang, M.; Wei, K.; and Deng, C. 2023b. Hierarchical Prompt Learning for Compositional Zero-Shot Recognition. In *IJCAI*, 1470–1478.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-Generated Images Are Surprisingly Easy To Spot... for Now. In *CVPR*, 8695–8704.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023c. DIRE for Diffusion-Generated Image Detection. In *ICCV*, 22445–22455.

Wu, T.; Liu, Z.; Huang, Q.; Wang, Y.; and Lin, D. 2021. Adversarial Robustness Under Long-Tailed Distribution. In *CVPR*, 8659–8668.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large Scale Incremental Learning. In *CVPR*, 374–382.

Xiao, S.; Guo, Y.; Peng, H.; Liu, Z.; Yang, L.; and Wang, Y. 2025a. Generalizable AI-Generated Image Detection Based on Fractal Self-Similarity in the Spectrum. *arXiv:2503.08484*.

Xiao, Y.; Yang, B.; Chen, W.; Chen, J.; Cao, Z.; Dong, Z.; Ji, X.; Lin, L.; Ke, W.; and Wei, P. 2025b. Are High-Quality AI-Generated Images More Difficult for Models to Detect? In *ICML*.

Xu, C.; Yan, J.; Yang, M.; and Deng, C. 2024. Rethinking Noise Sampling in Class-Imbalanced Diffusion Models. *IEEE Trans. Image Process.*, 33: 6298–6308.

Yan, S.; Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Xie, W. 2025a. A Sanity Check for AI-Generated Image Detection. In *ICLR*.

Yan, Z.; Wang, J.; Wang, Z.; Jin, P.; Zhang, K.-Y.; Chen, S.; Yao, T.; Ding, S.; Wu, B.; and Yuan, L. 2025b. Effort: Efficient Orthogonal Modeling for Generalizable AI-Generated Image Detection. In *ICML*.

Yang, M.; Goenawan, G. J.; Qin, H.; Han, K.; Peng, X.; Yang, Y.; and Zhu, H. 2025a. Detecting Open World Objects via Partial Attribute Assignment. In *CVPR*, 20318–20328.

Yang, M.; Yin, J.; Gu, Y.; Deng, C.; Zhang, H.; and Zhu, H. 2025b. Consistent Prompt Tuning for Generalized Category Discovery. *Int. J. Comput. Vis.*, 133: 4014–4041.

Ye, H.-J.; Chen, H.-Y.; Zhan, D.-C.; and Chao, W.-L. 2020. Identifying and Compensating for Feature Deviation in Imbalanced Deep Learning. *arXiv:2001.01385*.

You, Z.; Zhang, X.; Guo, H.; Wang, J.; and Li, C. 2025. Are Images Indistinguishable to Humans Also Indistinguishable to Classifiers? In *CVPR*, 28790–28800.

Yu, N.; Davis, L. S.; and Fritz, M. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *ICCV*, 7556–7566.

Zhao, B.; Xiao, X.; Gan, G.; Zhang, B.; and Xia, S.-T. 2020. Maintaining Discrimination and Fairness in Class Incremental Learning. In *CVPR*, 13208–13217.

Zhong, N.; Chen, H.; Xu, Y.; Qian, Z.; and Zhang, X. 2025. Beyond Generation: A Diffusion-Based Low-Level Feature Extractor for Detecting AI-Generated Images. In *CVPR*, 8258–8268.

Zhong, N.; Xu, Y.; Li, S.; Qian, Z.; and Zhang, X. 2023. PatchCraft: Exploring Texture Patch for Efficient AI-Generated Image Detection. *arXiv:2311.12397*.

Zhu, B.; Tang, K.; Sun, Q.; and Zhang, H. 2023. Generalized Logit Adjustment: Calibrating Fine-Tuned Models by Removing Label Bias in Foundation Models. In *NeurIPS*, 64663–64680.

Zhu, F.; Ma, S.; Cheng, Z.; Zhang, X.-Y.; Zhang, Z.; and Liu, C.-L. 2024a. Open-World Machine Learning: A Review and New Outlooks. *arXiv:2403.01759*.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2024b. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. In *NeurIPS*, 77771–77782.