Dual-Branch Asymmetric Discrepancy Learning Based on Fake Image Pattern-Coexistence for AI-Generated Image Detection

Chunli Song, Jie Liu*, Peiyang Wang, Ying Huang, Guixuan Zhang, Zhi Zeng, Shuwu Zhang

School of Artificial Intelligence, Beijing University of Posts and Telecommunications

Abstract

With the rapid advancement of generative models, high-fidelity AI-generated images have become increasingly indistinguishable from real images, posing significant challenges to traditional detection methods that rely on explicit artifacts or uniform feature learning. We hypothesize that detection ambiguity originates from pattern coexistence: synthetic images simultaneously embed (a) authentic patterns inherited from real-image distributions and (b) synthetic patterns induced by generative architectures, whereas real images maintain consistent patterns. We validate this hypothesis through SHAP-based quantitative analysis, demonstrating that synthetic images inherently exhibit a dual distribution—simultaneously containing authentic patterns and synthetic traces—while real images show a unimodal distribution. Building on this insight, this paper proposes a Dual-Branch Asymmetric Discrepancy Learning (DADL) framework. The DADL leverages multi-scale feature extraction and Asymmetric Feature Discrepancy Loss to capture and amplify such pattern differences across multiple scales. Extensive experiments on three benchmarks (AIGCDetectBenchmark, GenImage, and Chameleon) show that DADL achieves state-of-the-art performance, with particular strengths in detecting high-fidelity synthetic images from diffusion models (e.g., Midjourney, SDv1.4, SDv1.5) and enhancing generalization across diverse generative paradigms. This study not only offers an effective approach for AIGI detection but also sheds light on the intrinsic properties of synthetic images, providing a new perspective for advancing AIGI forensics.

Code — https://github.com/songchunli1999/DADL

Introduction

The rapid advancement and widespread adoption of Artificial Intelligence-Generated Image (AIGI) technologies, such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2014; Park et al. 2019) and diffusion models (Ho, Jain, and Abbeel 2020; Nichol et al. 2021; Dhariwal and Nichol 2021; Rombach et al. 2022), demonstrate significant potential for innovation and efficiency gains across various domains, including artistic creation, advertising design, and virtual simulation. However, their potential misuse poses significant risks, such as the spread of disinformation (Bontridder and Poullet 2021), the proliferation of deepfakes that erode personal reputation (Rossler et al. 2019), and violations of privacy (Golda et al. 2024). Consequently, effective AIGI detection techniques are urgently required. These techniques are critical for preserving information authenticity, protecting intellectual property, and mitigating societal harms caused by deceptive synthetic media.

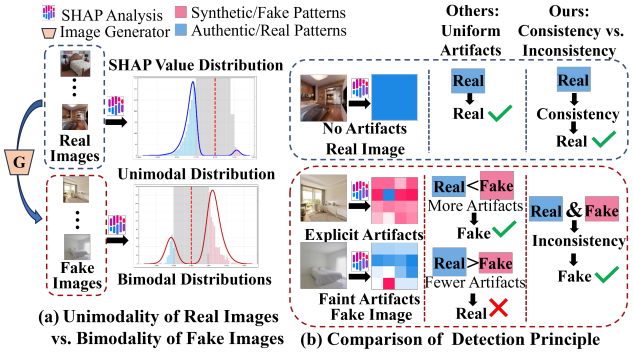


Figure 1: Illustration of (a) pattern coexistence in fake images via bimodal SHAP value distributions, and (b) our detection principle leveraging internal inconsistency.

Current detection paradigms predominantly rely on supervised training, where binary classifiers distinguish synthetic from real images using discriminative artifacts. However, modern AIGI achieves unprecedented realism by replicating intricate textures, lighting, and scene nuances. This evolution erodes traditional authenticity markers, exposing a critical limitation: conventional detectors depend on explicit synthetic artifacts. When these artifacts are minimized by realistic features, discriminative cues become faint or indistinguishable. The core issue lies in homogenized training objectives. Conventional approaches treat synthetic images as uniform entities, ignoring their inherent duality—the coexistence of authentic patterns (inherited from real images) and synthetic traces (induced by generators).This oversight causes ambiguous representations for high-fidelity forgeries. To empirically validate this duality, we performed SHAP-based (Lundberg and Lee 2017) feature attribution (Fig. 1) on 1,000 real LSUN images and 1,000 fake images (from

*Corresponding author: AILJ@bupt.edu.cn

ADM and VQDM generators). The visualization reveals divergent patterns: synthetic images show bimodal distributions (blue/red regions indicating conflicting authentic/synthetic evidence), while real images show unimodal consistency. This distributional divergence motivates a new detection principle: amplify the internal inconsistency within forgeries while reinforcing the consistency of real images.

By leveraging this insight, we propose a novel **D**ual-Branch **A**symmetric **D**iscrepancy **L**earning (**DADL**) framework. This framework is specifically designed to capture and amplify the inherent duality within fake images, thereby enhancing generalization across diverse and evolving AIGI forgeries. DADL utilizes a lightweight ResNet to construct a dual-branch structure, with two parallel feature extraction branches that adopt distinct perspectives on the input image. The key innovation lies in the Asymmetric Feature Discrepancy (AD) Loss, which drives asymmetric learning between the branches: for real images, the AD Loss minimizes the $\mathbf{L}_2$ distance between corresponding features from the two branches, enforcing consistency in their representational space to reflect the uniformity of natural patterns; for fake images, it deliberately maximizes this $\mathbf{L}_2$ distance, amplifying the inherent divergence between authentic and synthetic patterns embedded within the same image. Since such pattern-coexistence arises at multiple scales in fake images, the AD Loss is computed on both shallow-texture and deep-semantic layers, exposing subtle forgery traces otherwise buried in high-fidelity textures. This multi-scale strategy more effectively amplifies these elusive traces of forgery inherent in fake images, consequently improving generalization. Furthermore, to effectively leverage the differences between the two branch outputs, DADL computes the inter-branch feature difference as the final representation, which is then optimized via a Cross-Entropy (CE) Loss for binary classification (real vs. fake). The AD Loss and CE Loss operate synergistically: CE Loss handles the core discrimination task, while the AD Loss explicitly disentangles inherent duality in fake images, providing critical evidence for authenticity determination. Experiments confirm that our approach achieves state-of-the-art (SOTA) performance across multiple benchmark datasets. Our main contributions are:

- **Pattern-Coexistence Hypothesis & Empirical Validation.** We formalize the Pattern-Coexistence Hypothesis and empirically validate that fake images are not "pure" artifacts—they simultaneously contain authentic textures inherited from real training data and generation-specific traces. This inherent duality underpins the failure of conventional detectors, which treat synthetic images as homogeneous entities.
- **DADL Architecture with AD Loss.** We design Dual-Branch Asymmetric Discrepancy Learning, a lightweight dual-branch framework built on ResNet that exploits the coexistence phenomenon through asymmetric learning. The AD Loss enforces feature agreement on real images (minimizing inter-branch distances) and feature disagreement on fake images (maximizing inter-branch distances), thereby amplifying latent inconsistencies only when they exist.
- **Multi-Scale Discrepancy Mining.** The AD Loss is applied hierarchically across shallow texture layers and deep semantic layers, enabling synergistic mining of fine-grained forgery traces at every scale. This design addresses the problem of subtle synthetic cues being buried in high-fidelity textures.
- **Superior Cross-Domain Generalization.** Our method achieves SOTA performance across representative datasets (AIGCDetectBenchmark, GenImage, Chameleon), with notable gains in detecting unseen generators. This validates its effectiveness and robustness in handling diverse and evolving AIGI technologies.

Related Work

Driven by sophisticated generative models (GANs, diffusion models), AIGI detection has significantly advanced recently. This has led to various techniques distinguishing real from fake AI-generated images, briefly introduced below.

Artifact-Based Detection

A prominent line of work focuses on identifying explicit synthetic artifacts introduced during image generation:
Frequency-domain analysis: FreDect (Frank et al. 2020) maps images to the frequency domain, detecting GAN-generated content by identifying upsampling-induced artifacts in the frequency spectrum.
Model-specific traces for diffusion models: DIRE (Wang et al. 2023) introduces the Diffusion Reconstruction Error, distinguishing real and generated images by measuring reconstruction discrepancies. Building on this, AEROBLADE (Ricker, Lukovnikov, and Fischer 2024) enhances detection by analyzing noise pattern handling during the diffusion model's reverse denoising process.
Texture feature analysis: PatchCraft (Zhong et al. 2023) compares rich and poor texture patches, extracting inter-pixel correlation differences as a general fingerprint. Similarly, NPR (Tan et al. 2024) models adjacent pixel relationships to identify upsampling-induced local distribution bias.

Feature Learning via Augmentation & Pre-training

To enhance generalization, recent methods focus on robust feature learning through augmentation and pre-trained model integration:
Data augmentation strategies: CNNSpot (Wang et al. 2020) demonstrates data augmentation enables a classifier trained on ProGAN (Karras et al. 2017) to generalize to unseen GANs. SAFE (Li et al. 2025) further improves cross-generator detection by leveraging comprehensive augmentations (ColorJitter, RandomRotation, RandomMask), forcing the model to focus on invariant rather than spurious artifacts.
Pre-trained model utilization: UnivFD (Ojha, Li, and Lee 2023) leverages pre-trained CLIP (Radford et al. 2021) features with nearest-neighbor and linear probing for cross-model generalization. AIDE (Yan et al. 2024) integrates high-level semantic information from CLIP with low-level frequency artifacts, aiming for comprehensive judgment. Effort (Yan et al. 2025) enhances detection by projecting features into orthogonal subspaces using pre-trained extractors.

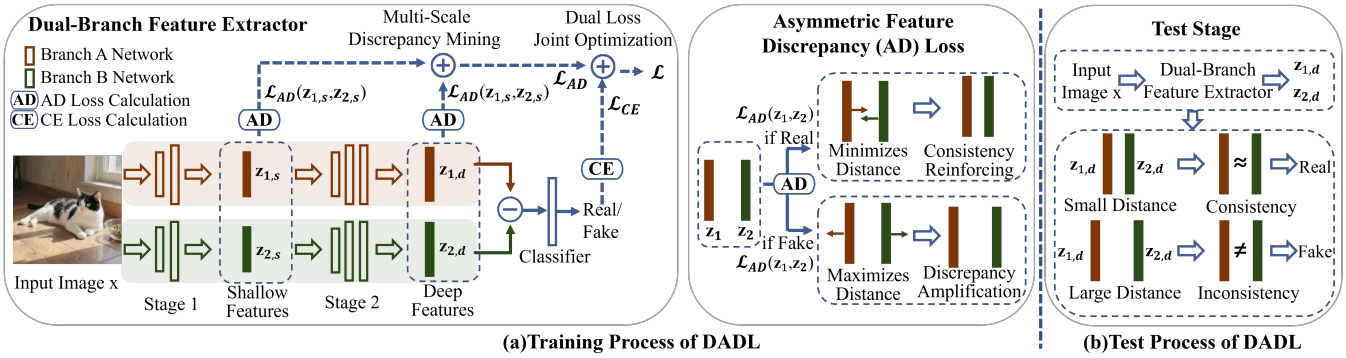


Figure 2: Overall Framework of DADL. (a) DADL training processes consists of dual-branch feature extractor, asymmetric discrepancy loss, multi-scale discrepancy mining, dual loss joint optimization. (b) During the testing phase, a classification decision is made by determining whether the distance between the dual-branch features indicates Inconsistency or Consistency.

Explainability-Enhanced Approaches

Emerging work explores explainability boosts robustness:
Multimodal reasoning: AIGI-Holmes (Zhou et al. 2025) utilizes multimodal large language models to enhance explainability in AIGI detection, generating natural language justifications for decisions.

While the aforementioned detection methods have achieved specific generalization performance and can successfully combat low-fidelity fakes, they exhibit significant vulnerability to high-quality fake images, mainly due to the presence of regions that mimic real image textures. Most approaches treat fake images as uniform, ignoring the coexistence of authentic patterns and synthetic traces. These limitations highlight the need for a framework that explicitly models the intrinsic bimodality of fake images.

Methodology

This section details the technical implementation of the proposed DADL framework, including its network architecture, loss functions, and multi-scale discrepancy mining strategy. The design of DADL is explicitly aimed at leveraging the inherent duality of fake images and operationalize the detection principle of "amplifying internal inconsistency in forgeries while reinforcing consistency in real images."The following subsections detail each component.

Overall Framework of DADL

DADL is a dual-branch network designed to capture and amplify the internal inconsistency within fake images while reinforcing the consistency of real images. The overall architecture is shown in Fig. 2 (a). Given an input image x, DADL training processes it through the following stages:
Dual-Branch Feature Extraction: The image is fed into two parallel branches, Branch-A and Branch-B. This dual-branch design enables the network to capture both authentic patterns and synthetic traces within the image.
AD Loss: The AD Loss drives asymmetric learning between the two branches. For real images, the AD Loss enforces feature consistency; for fake images, it amplifies latent discrepancies.
Multi-Scale Discrepancy Mining: The AD Loss is computed across multiple layers, capturing discrepancies at different scales. This multi-scale strategy exposes and amplifies subtle forgery traces while addressing subtle synthetic cues buried in high-fidelity textures.
Dual Loss Joint Optimization: The AD Loss and CE Loss are combined to optimize the network. The CE Loss handles the core discrimination task, while the AD Loss provides critical evidence for authenticity determination by disentangling inherent duality in fake images.

Through the described training process, our model minimizes the distance between the two branches' features for real samples, enforcing consistency. Conversely, for fake samples, it maximizes this inter-branch feature distance, amplifying inconsistency. The testing process as shown in Fig.2(b), this learned discrepancy serves as a robust detection criterion. When the feature distance of an incoming image is small, it indicates consistency, leading to a classification as real. Conversely, a large distance signifies inconsistency, resulting in a classification as fake.

Dual-Branch Feature Extractor

The Dual-Branch Feature Extractor is the core component of DADL, responsible for generating complementary feature representations from the input image. It consists of two parallel branches, Branch-A and Branch-B, each with identical yet weight-independent architectures. To balance robust feature extraction with computational efficiency, both branches are built upon an identical lightweight ResNet backbone (He et al. 2016), similar to that used in SAFE (Li et al. 2025), comprising two stages that progressively extract hierarchical features from raw pixels to high-level semantics. This design ensures full structural symmetry while enabling the emergence of distinct inductive biases: despite identical architectures, Branch-A and Branch-B are updated independently during training. This parameter independence is the sole source of their divergent feature learning, without structural specialization, each branch naturally develops unique focuses based on gradient flows in its parameter space. Over time, this leads to complementary pattern capture. Such a

setup is critical for capturing both the authentic patterns and synthetic traces coexisting in fake images, generating distinct feature representations that lay the foundation for subsequent discrepancy analysis.

Asymmetric Feature Discrepancy (AD) Loss

The AD Loss operationalizes DADL's core principle: amplifying inconsistency in fake images while reinforcing consistency in real ones. Built upon the complementary feature representations generated by the dual-branch extractor, this loss function introduces an asymmetric regulatory strategy that explicitly models the inherent duality of synthetic content. By dynamically adjusting the relationship between features from Branch-A and Branch-B based on image authenticity, the AD Loss ensures that the network learns to distinguish real and fake images through their inherent distributional differences rather than relying on explicit synthetic artifacts.

Therefore, for real images, the AD loss minimizes the normalized L_2 distance between features from two branches to enhance representation consistency. Conversely, for fake images, it maximizes this distance, using a margin m_t to amplify their intrinsic inconsistencies. We extract $\mathbf{z}_1$ and $\mathbf{z}_2$, one-dimensional feature vectors, from an arbitrary network layer. These vectors correspond to two branches for an input image x. We define the AD loss as:

$$\mathcal{L}_{AD}(\mathbf{z}_1, \mathbf{z}_2) = \begin{cases} \text{dist}(\mathbf{z}_1, \mathbf{z}_2), & \text{if } x \text{ is real} \\ \max(0, m_t - \text{dist}(\mathbf{z}_1, \mathbf{z}_2)), & \text{if } x \text{ is fake.} \end{cases} \tag{1}$$

Here, m_t is a boundary threshold. We define the function $\text{dist}(\mathbf{z}_1, \mathbf{z}_2)$ as the normalized L_2 distance:

$$\text{dist}(\mathbf{z}_1, \mathbf{z}_2) = \left\| \frac{\mathbf{z}_1}{\|\mathbf{z}_1\|_2} - \frac{\mathbf{z}_2}{\|\mathbf{z}_2\|_2} \right\|_2^2. \tag{2}$$

This approach pulls the $\text{dist}(\mathbf{z}_1, \mathbf{z}_2)$ of real samples closer, which ensures their distributional consistency. For fake samples, it pushes the distance between the two branch features further away to a boundary value, thereby amplifying the inherent distributional discrepancies within the fake data.

We use an Exponential Moving Average (EMA) mechanism to dynamically adjust the boundary value m_t at each step. We obtain the threshold m_t at step t using:

$$m_t = \beta \cdot m_{t-1} + (1 - \beta) \cdot \mu_{real,t}. \tag{3}$$

Here, β is the EMA momentum factor, which we set to 0.99 to ensure a smooth and stable update trajectory. We set the initial value m_0 to 0.5. m_{t-1} is the threshold from the previous training step ($t-1$ for $t > 1$). We calculate $\mu_{t,real}$ as the average distance from the subset of real samples $\mathcal{B}_{t,real}$ within the current batch $\mathcal{B}_t$:

$$\mu_{t,real} = \frac{1}{|\mathcal{B}_{t,real}|} \sum_{x \in \mathcal{B}_{t,real}} \text{dist}(\mathbf{z}_1, \mathbf{z}_2). \tag{4}$$

We design m_t to push the feature distance of fake images to be at least as large as the average feature distance of real images. This effectively separates real and fake samples in the feature space. This dynamic threshold also ensures that the discriminative requirements adaptively increase as the model's capabilities improve. This entire process completes our calculation of the AD loss for the features $(\mathbf{z}_1, \mathbf{z}_2)$ of a given sample x.

Multi-Scale Discrepancy Mining

To capture multi-scale forgery traces, DADL incorporates a Multi-Scale Discrepancy Mining strategy, which extends the AD Loss across hierarchical layers of the dual-branch feature extractor. As shown in Fig. 2(a), we extract:

Shallow features: From Stage 1, sensitive to high-frequency artifacts

Deep features: From Stage 2, capturing structural distortions

This design ensures that subtle forgery traces, often buried in high-fidelity textures at specific scales, are exposed and amplified, thereby enhancing the model's ability to generalize to diverse and unseen generative models. The total multi-scale AD Loss is then calculated as follows:

$$\mathcal{L}_{AD} = \alpha_1 \cdot \mathcal{L}_{AD}(\mathbf{z}_{1,s}, \mathbf{z}_{2,s}) + \alpha_2 \cdot \mathcal{L}_{AD}(\mathbf{z}_{1,d}, \mathbf{z}_{2,d}). \tag{5}$$

Here, $\mathcal{L}_{AD}(\mathbf{z}_{1,s}, \mathbf{z}_{2,s})$ and $\mathcal{L}_{AD}(\mathbf{z}_{1,d}, \mathbf{z}_{2,d})$ are calculated by substituting the shallow features $(\mathbf{z}_{1,s}, \mathbf{z}_{2,s})$ and deep features $(\mathbf{z}_{1,d}, \mathbf{z}_{2,d})$ into Eq.(1), respectively. α_1 and α_2 are learnable weights that quantify the contributions of the shallow and deep AD Loss components to the total $\mathcal{L}_{AD}$. The model optimizes these weights alongside other parameters during training, subject to the constraint $\alpha_1 + \alpha_2 = 1$.

By mining discrepancies across hierarchical layers, this strategy addresses a key limitation of single-scale methods: it ensures that no matter where the synthetic traces are embedded (whether in local textures or global semantics), they are explicitly amplified by the AD Loss. This comprehensive capture of multi-scale cues significantly improves the model's robustness to diverse generative strategies.

Dual Loss Joint Optimization

To optimize the network for robust forgery detection, DADL employs a joint optimization strategy that combines the AD Loss with the CE Loss. This dual loss framework ensures that the network not only amplifies internal inconsistencies within fake images but also maintains strong discriminative power between real and fake images.

Feature Representation and Cross-Entropy Loss To effectively leverage the discrepancies captured by the dual-branch network, we use the difference between the deep features extracted from the two branches as our ultimate representation. Specifically, we compute:

$$\mathbf{z}_{\text{diff}} = \mathbf{z}_{1,d} - \mathbf{z}_{2,d}, \tag{6}$$

where $(\mathbf{z}_{1,d})$ and $(\mathbf{z}_{2,d})$ are the deep features extracted from Branch-A and Branch-B, respectively. This difference vector $(\mathbf{z}_{\text{diff}})$ captures the essential discrepancies between the two branches, which are crucial for forgery detection.

We feed this representation $(\mathbf{z}_{\text{diff}})$ into a classifier to obtain the predicted logits. The CE Loss $(\mathcal{L}_{CE})$ is then computed between these logits and the ground-truth label. The

CE Loss inherently constrains the model, allowing it to develop robust discriminative capabilities for the core classification task.

Integrating AD Loss The AD Loss amplifies the internal inconsistencies within fake images while reinforcing the consistency in real images. This loss function provides crucial evidence for authenticity discrimination by revealing internal pattern discrepancies within fake samples.

To integrate the AD Loss into the primary classification objective, we add it to the CE Loss with a fixed weighting factor (λ). This forms our final optimization objective:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{AD}, \tag{7}$$

where (λ) is set to 0.5. This combined objective aims to enhance the model's overall robustness and discriminative power by balancing the contributions of the CE Loss and the AD Loss.

Experiments

Empirical Validation of the Pattern-Coexistence Hypothesis in Fake Images

Our SHAP-based analysis empirically validates the Pattern-Coexistence Hypothesis, which is clearly manifested in the bimodal distribution of SHAP values for fake images versus the unimodal distribution of SHAP (Lundberg and Lee 2017) values for real images. We analyzed image patches from 1,000 real LSUN images and 1,000 fake images (from ADM and VQDM generators) on AIGCDetectBenchmark using two SOTA detectors, NPR and SAFE. As shown in Fig. 3, we analyzed the influence of each patch on the prediction result by dividing an image into 4x4 patches. For these SHAP values, those less than 0 (blue bars) indicate a 'real' prediction, while values greater than 0 (red bars) suggest a 'fake' prediction. To more accurately visualize the SHAP values of patches influencing the model's decision, we discarded SHAP values with an absolute magnitude below a certain threshold (0.001 for NPR and 1e-5 for SAFE). Real images (Fig. 3 (a) and (d)) consistently exhibit a unimodal SHAP distribution, reflecting cohesive internal patterns. In contrast, fake images from multiple generators (ADM in Fig. 3 (b) and (e), VQDM in Fig. 3 (c) and (f)) consistently display a distinct bimodal distribution, confirming the simultaneous presence of both authentic (negative SHAP values) and synthetic (positive SHAP values) patterns within them. This consistent finding across both NPR and SAFE detectors and various fake data sources robustly validates the universality of our pattern-coexistence hypothesis for fake images.

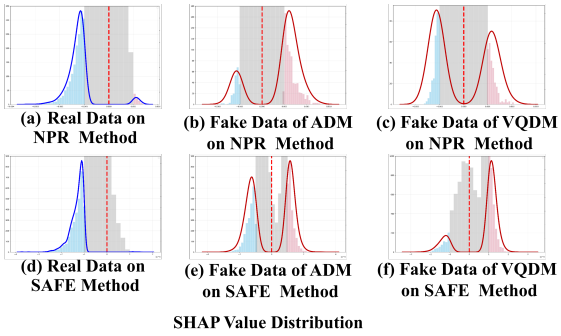


Figure 3: Empirical Validation of the Pattern-Coexistence Hypothesis in Fake Images.

Experimental Settings

Datasets. To comprehensively evaluate the generalization ability of existing approaches, we evaluate the detectors on three general and comprehensive benchmarks of AIGCDetectBenchmark (Zhong et al. 2023), GenImage (Zhu et al. 2023) and Chameleon (Yan et al. 2024).

Evaluation Metrics. Following existing AI-generated detection approaches (Wang et al. 2020) and (Yan et al. 2024), we report both classification accuracy (Acc) and average precision (AP) in our experiments. To intuitively evaluate the detection performance, we also report the averaged metrics for each test set, termed ACC_M and AP_M.

Implementation Details. Our dual-branch feature extractor is built upon a lightweight ResNet architecture adapted from (Li et al. 2025). For image preprocessing, we first transform the input into the frequency domain using the Discrete Wavelet Transform (DWT). We then apply standard data augmentation techniques, including RandomCrop, RandomHorizontalFlip, RandomRotation, and RandomMask. We employ the Discrete Wavelet Transform (DWT). We employ the AdamW optimizer for 100 epochs with a batch size of 32, a learning rate of 5×10^{-3}, and a weight decay of 0.01.

Comparing Methods. We compare our DADL with 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), DIRE (Wang et al. 2023), PatchCraft (Zhong et al. 2023), NPR (Tan et al. 2024), AIDE (Yan et al. 2024), SAFE (Li et al. 2025), Effort (Yan et al. 2025), AIGI-Holmes (Zhou et al. 2025).

Performance on AIGCDetectBenchmark

We follow the experimental setup of the AIGCDetectBenchmark (Zhong et al. 2023), utilizing four categories from the ProGAN dataset (car, cat, chair, and horse) for training. As shown in Table 1, our DADL method demonstrates superior performance, achieving over 90% accuracy on the vast majority of datasets. With an ACC_M that surpasses the current SOTA result by 1.55%, our approach is effectively validated as a highly effective solution. DADL's advantage is particularly significant when dealing with high-quality generative models based on diffusion models, such as Midjourney, SDv1.4, SDv1.5, VQDM, and Wukong. This demonstrates that DADL can effectively capture and utilize the internal pattern discrepancies of fake images, a crucial capability that addresses a key challenge in the field. Despite these strengths, DADL's results on the WFIR dataset are relatively weaker. This may stem from the unique artifact patterns of WFIR not fully aligning with our multi-scale feature extraction mechanism. This finding offers valuable insights for future work, suggesting that we could explore more re-

Method	Pro-GAN	Style-GAN	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	Style-GAN2	WFIR	ADM	Gl-ide	Midj-ourney	SDv-1.4	SDv-1.5	VQ-DM	Wu-kong	DAL-LE2	ACC_M
CNNSpot	**100.0**	90.17	71.17	87.62	94.60	81.42	86.91	91.65	60.39	58.07	51.39	50.57	50.53	56.46	51.03	50.45	70.78
FreDect	99.36	78.02	81.97	78.77	94.62	80.57	66.19	50.75	63.42	54.13	45.87	38.79	39.21	77.80	40.30	34.70	64.03
Fusing	**100.0**	85.20	77.40	87.00	97.00	77.00	83.30	66.80	49.00	57.20	52.20	51.00	51.40	55.10	51.70	52.80	68.38
GramNet	99.99	87.05	67.33	86.07	95.05	69.35	87.28	86.80	58.61	54.50	50.02	51.70	52.16	52.86	50.76	49.25	68.67
LNP	99.67	91.75	77.75	84.10	99.92	75.39	94.64	70.85	84.73	80.52	65.55	85.55	85.67	74.46	82.06	88.75	83.84
LGrad	99.83	91.08	85.62	86.94	99.27	78.46	85.32	55.70	67.15	66.11	65.35	63.02	63.67	72.99	59.55	65.45	75.34
DIRE-G	95.19	83.03	70.12	74.19	95.47	67.79	75.31	58.05	75.78	71.75	58.01	49.74	49.83	53.68	54.46	66.48	68.68
DIRE-D	52.75	51.31	49.70	49.58	46.72	51.23	51.72	53.30	**98.25**	92.42	89.45	91.24	91.63	91.90	90.90	92.45	71.53
UnivFD	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
PatchCraft	**100.0**	92.77	**95.80**	70.17	99.97	71.58	89.55	85.80	82.17	83.79	90.12	95.38	95.30	88.91	91.07	96.60	89.31
NPR	99.90	96.10	87.30	90.30	99.60	85.40	98.10	60.70	84.90	**96.70**	92.60	97.40	97.50	90.10	91.70	**99.60**	91.70
AIDE	99.99	**99.64**	83.95	98.48	99.91	73.25	98.00	94.20	93.43	95.09	77.20	93.00	92.85	95.16	93.55	96.60	92.77
SAFE	99.86	98.03	89.72	98.86	99.89	91.52	98.56	51.95	82.05	96.29	95.26	99.40	99.26	96.29	98.20	95.30	93.15
Effort	98.51	86.51	91.68	97.95	96.27	97.73	86.95	88.65	70.39	73.28	66.32	74.77	74.55	81.55	77.33	77.29	83.73
AIGI-Holmes	**100.0**	98.35	94.51	97.03	**100.0**	95.19	98.88	**95.71**	88.43	91.53	81.56	91.28	91.38	90.94	89.46	85.32	93.16
DADL	99.97	99.38	94.17	**99.12**	**100.0**	90.55	**99.79**	52.82	95.60	94.62	**95.70**	**99.63**	**99.43**	**97.71**	**99.56**	97.30	**94.71**

Table 1: Comparison on the AIGCDetectBenchmark (Zhong et al. 2023). Accuracy (%) of different detectors (rows) in detecting real and fake images from different generators (columns). DIRE-D indicates this result comes from DIRE detector trained over fake images generated by ADM following its official setup (Wang et al. 2023). DIRE-G indicates this baseline is trained on the same ProGAN training data as others. The best result and the second-best result are marked in **bold** and underline, respectively.

fined analysis methods to further enhance the model's generalization ability.

Performance on GenImage

Based on the results presented in Table 2, we evaluate our DADL method using the experimental setup of the GenImage benchmark. The model is trained on real images from ImageNet and fake images generated by SD v1.4. DADL demonstrates superior generalization capabilities, achieving an ACC_M of 92.52%, which surpasses the next best method by 1.42%. This confirms the overall effectiveness of our approach in detecting a diverse range of AI-generated images. Notably, our method DADL exhibits exceptional generalization on the high-quality commercial dataset, Midjourney. It achieves an accuracy of 99.15%, outperforming the second-best method by a significant 7.65%. This highlights our method's strong ability to capture and utilize the internal pattern discrepancies present in the latest, high-quality generated content. Furthermore, on the challenging ADM dataset, our model still outperforms most methods, thus demonstrating the robustness of our approach.

Performance on Chameleon

To thoroughly evaluate our method, we conduct experiments on the challenging Chameleon dataset, which is known for its high-quality generated content. Following the benchmark's protocol, we test three distinct training scenarios. The first two scenarios, using the ProGAN and SD v1.4 datasets respectively, directly correspond to the training settings of the AIGCDetectBenchmark and GenImage benchmark that we discussed earlier. The final scenario utilizes the comprehensive All GenImage dataset for training. As shown in Table 3, our method demonstrates highly competitive performance even when trained on a single generator. Specifically, when trained on the ProGAN dataset, our DADL framework achieves the best among all baselines, which highlights its ability to learn robust features from a single GAN generator. More significantly, our method exhibits remarkable generalization capabilities when trained on the comprehensive All GenImage dataset, which contains a variety of GANs and diffusion models. In this scenario, our method achieves an accuracy of 67.77%, notably surpassing the second-best baseline by 2.0%. This result underscores the superior ability of our DADL framework to capture and amplify intrinsic discrepancies, thereby enhancing the generalization performance in detecting AI-generated images from a wide range of unseen generators.

Ablation Study

Ablation Study on Multi-scale AD Loss. To validate our multi-scale AD Loss, we ablated it on the AIGCDetectBenchmark. As shown in Fig. 4, adding either the shallow-feature AD Loss ($\mathcal{L}_{CE} + \mathcal{L}_{AD}(\mathbf{z}_{1,s}, \mathbf{z}_{2,s})$) or the deep-feature AD Loss ($\mathcal{L}_{CE} + \mathcal{L}_{AD}(\mathbf{z}_{1,d}, \mathbf{z}_{2,d})$) to the baseline ($\mathcal{L}_{CE}$) significantly improves performance in both Acc_M and AP_M. This demonstrates the effectiveness of the AD Loss in capturing feature discrepancies. The best performance is achieved by fusing both shallow and deep AD Losses ($\mathcal{L}_{CE} + \mathcal{L}_{AD}$), with an Acc_M of 94.71% and an AP_M of 97.51%. This result highlights the necessity of a multi-scale Strategy.

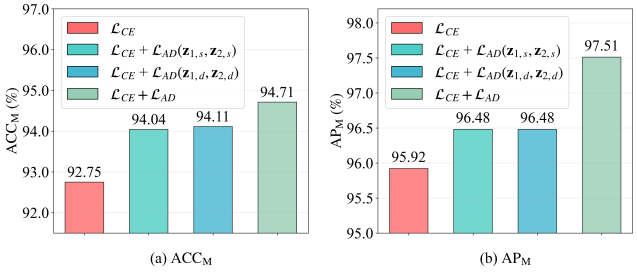
Sensitivity Analysis of Hyperparameter λ. To study the impact of the hyperparameter λ on DADL's performance, we conducted a sensitivity analysis by varying λ from 0.25 to 1.5. As defined in Eq.7, λ controls the weight between $\mathcal{L}_{CE}$

Method	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	ACC_M
ResNet-50	54.90	99.90	99.70	53.50	61.90	98.20	56.60	52.00	72.09
DeiT-S	55.60	99.90	99.80	49.80	58.10	98.90	56.90	53.50	71.56
Swin-T	62.10	99.90	99.80	49.80	67.60	99.10	62.30	57.60	74.78
CNNSpot	52.80	96.30	95.90	50.10	39.80	78.60	53.40	46.80	64.21
Spec	52.00	99.40	99.20	49.70	49.80	94.80	55.60	49.80	68.79
F3Net	50.10	99.90	99.90	49.90	50.00	99.90	49.90	49.90	68.69
GramNet	54.20	99.20	99.10	50.30	54.60	98.90	50.80	51.70	69.85
DIRE	60.20	99.90	99.80	50.90	55.00	99.20	50.10	50.20	70.66
UnivFD	73.20	84.20	84.00	55.20	76.90	75.60	56.90	80.30	73.29
GenDet	89.60	96.10	96.10	58.00	78.40	92.80	66.50	75.00	81.56
PatchCraft	79.00	89.50	89.30	77.30	78.40	89.30	83.70	72.40	82.30
DRCT	91.50	95.01	94.41	**79.42**	89.18	94.67	90.03	**81.67**	89.49
AIDE	79.38	99.74	**99.76**	78.54	91.82	98.65	80.26	66.89	86.88
Effort	82.40	**99.80**	**99.80**	78.70	93.30	97.40	91.70	77.60	91.10
DADL	**99.15**	99.50	99.43	67.76	**96.33**	**99.42**	**98.08**	80.50	**92.52**

Table 2: Comparison on the GenImage (Zhu et al. 2023). Accuracy (%) of different detectors (rows) in detecting real and fake images from different generators (columns). These methods are trained on real images from ImageNet and fake images generated by SD v1.4. The best result and the second-best result are marked in **bold** and underline, respectively.

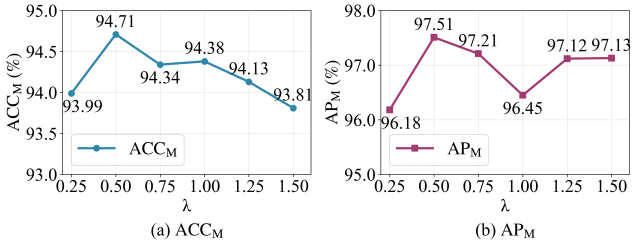
Training Dataset	CNNSpot	FreDect	Fusing	GramNet	LNP	UnivFD	DIRE	PatchCraft	NPR	AIDE	**DADL**
ProGAN	56.94	55.62	56.98	58.94	57.11	57.22	58.19	53.76	57.29	58.37	**59.24**
SD v1.4	60.11	56.86	57.07	60.95	55.63	55.62	59.71	56.32	58.13	**62.60**	59.35
All GenImage	60.89	57.22	57.09	59.81	58.52	60.42	57.83	55.70	57.81	65.77	**67.77**

Table 3: Comparison on the Chameleon (Yan et al. 2024). Accuracy (%) of different detectors (columns) when trained on various datasets (rows) in detecting real and fake images. The best result and the second-best result are marked in **bold** and underline, respectively.



(a) ACC_M (b) AP_M

Figure 4: Ablation study of multi-scale AD Loss on the AIGCDetectBenchmark.



(a) ACC_M (b) AP_M

Figure 5: Sensitivity analysis of hyperparameter λ on the AIGCDetectBenchmark.

and our proposed $\mathcal{L}_{AD}$. As shown in the Fig. 5, the model achieves its best performance when $\lambda = 0.5$, with an ACC_M and an AP_M. This suggests that a weight of 0.5 effectively balances the combined effect of the two loss functions. Furthermore, the values for both ACC_M and AP_M remain consistently high across the entire tested range of λ. This observation demonstrates that DADL is robust to the choice of the λ hyperparameter, maintaining strong performance even when the weight between the two losses is varied.

Conclusion

In this paper, we propose the Pattern-Coexistence Hypothesis: fake images contain both authentic patterns and synthetic traces, which we then empirically validate. Based on this, we propose DADL with the AD Loss to amplify the inconsistency of fake images and reinforce the consistency of real images. Furthermore, the AD Loss is applied hierarchically across shallow texture layers and deep semantic layers, enabling synergistic mining of fine-grained forgery traces at every scale. Extensive experiments demonstrate that DADL effectively captures internal pattern discrepancies in synthetic images, achieving SOTA performance on multiple benchmarks and enhancing generalization across diverse generative models and fidelity levels. Overall, this study offers an effective AIGI detection approach and insights into synthetic images' intrinsic properties, providing a new perspective for advancing AIGI forensics.

Acknowledgements

This work has been supported by the National Key R&D Program of China under Grant NO. 2024YFF0907200 and the BUPT innovation and entrepreneurship support program 2025-YC-A060.

References

Bontridder, N.; and Poullet, Y. 2021. The role of artificial intelligence in disinformation. *Data & Policy*, 3: e32.

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

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, 3247–3258. PMLR.

Golda, A.; Mekonen, K.; Pandey, A.; Singh, A.; Hassija, V.; Chamola, V.; and Sikdar, B. 2024. Privacy and security concerns in generative AI: a comprehensive survey. *IEEE Access*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

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 *2022 IEEE International Conference on Image Processing (ICIP)*, 3465–3469. IEEE.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Feng, F. 2025. Improving Synthetic Image Detection Towards Generalization: An Image Transformation Perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2405–2414.

Liu, B.; Yang, F.; Bi, X.; Xiao, B.; Li, W.; and Gao, X. 2022. Detecting generated images by real images. In *European Conference on Computer Vision*, 95–110. Springer.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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

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

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ricker, J.; Lukovnikov, D.; and Fischer, A. 2024. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9130–9140.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12105–12114.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695–8704.

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

Yan, S.; Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Xie, W. 2024. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*.

Yan, Z.; Wang, J.; Jin, P.; Zhang, K.-Y.; Liu, C.; Chen, S.; Yao, T.; Ding, S.; Wu, B.; and Yuan, L. 2025. Orthogonal Subspace Decomposition for Generalizable AI-Generated Image Detection. In *Forty-second International Conference on Machine Learning*.

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

Zhou, Z.; Luo, Y.; Wu, Y.; Sun, K.; Ji, J.; Yan, K.; Ding, S.; Sun, X.; Wu, Y.; and Ji, R. 2025. AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models. *arXiv preprint arXiv:2507.02664*.

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