DySy-Det: A Synergistic Framework with Dynamic Reconstruction-Path Consistency for AI-Generated Image Detection

Fanli Jin[1,2], Feng Lin[1,2]*, Gaojian Wang[1,2], Tong Wu[1,2], Zhisheng Yan[3]

[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3]Department of Information Sciences and Technology, George Mason University

Abstract

Advanced image generative models have led to concerns about malicious use, underscoring the necessity for generalizable detection methods. However, existing approaches tend to overfit to domain-specific forgery patterns, while overlooking complementary cues from different domains. Therefore, we introduce **DySy-Det** (**Dy**namic **Sy**nergy **Det**ector), a novel framework that mines collaborative and robust forgery artifacts from multiple evidence sources. First, DySy-Det fine-tunes a CLIP vision transformer to extract high-level semantics for identifying conceptual inconsistencies, while generating attention maps that pinpoint key discriminative regions. Then, this semantic guidance, in the form of a mask, directs a targeted reconstruction process. By focusing on these salient areas, our approach effectively extracts localized reconstruction errors, thereby filtering out irrelevant background noise. Furthermore, inspired by the intrinsic generative mechanics of diffusion models, we introduce the concept of Reconstruction-Path Consistency (RPC), which quantifies the temporal stability of the denoising trajectory to expose dynamic generative artifacts. We capture this by computing noise alignment scores across multiple timesteps and encode them via a lightweight network. Extensive evaluations on GenImage and UniversalFakeDetect benchmarks demonstrate that DySy-Det outperforms the state-of-the-art detector by **6.14%** and **1.57%** in mean accuracy, respectively.

Code — https://github.com/Vanleya/DySy-Det

Introduction

The rapid development of generative techniques, such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2014; Karras, Laine, and Aila 2019) and Diffusion Models (DMs) (Ho, Jain, and Abbeel 2020; Rombach et al. 2022), has enabled the synthesis of highly realistic images. Despite the creative advantages, they also raise critical concerns, including misinformation dissemination (Bontridder and Poullet 2021), copyright infringement (Gaffar and Albarashdi 2025), and privacy breaches (Golda et al. 2024). As these threats escalate, verifying the authenticity of visual content becomes crucial. Consequently, there is an urgent need to develop generalizable detectors.

*Corresponding author

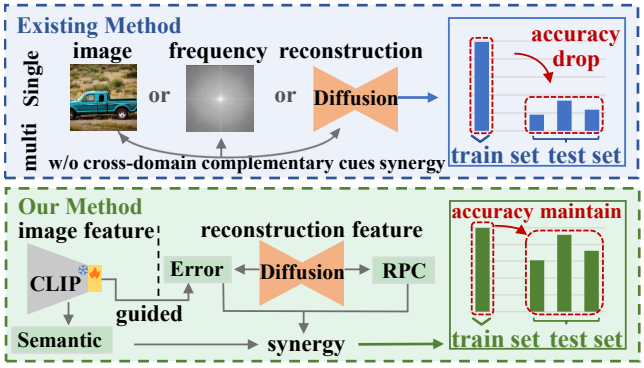


Figure 1: Existing methods often utilize independent cross-domain features, performing poorly on unseen test sets. In contrast, our method synergizes semantic inconsistencies, guided reconstruction error and RPC to maintain high accuracy.

However, traditional detection methods often struggle with this challenge due to their reliance on specific forgery artifacts such as pixel-level inconsistencies (Wang et al. 2020; Ojha, Li, and Lee 2023), frequency anomalies (Jeong et al. 2022; Dong, Kumar, and Liu 2022), and reconstruction errors (Wang et al. 2023; Cazenavette et al. 2024). This dependency leads to significant degradation in detection performance on images produced by unseen models, as shown in Figure 1.

In response to this limitation, a recent line of work has sought to leverage multi-perspective signals. For instance, FatFormer (Liu et al. 2024a) and AIDE (Yan et al. 2024) integrate spatial and frequency cues, while LaRE^2 (Luo et al. 2024) combines CLIP embeddings with reconstruction error to benefit from both high-level semantics and low-level residual information. Despite their efforts, generalization on challenging datasets remains insufficient due to two shortcomings. First, they lack effective synergy among different feature streams. In particular, the reconstruction error is typically computed over the entire image, which may introduce redundant noise from background regions, hindering detection performance. Second, most methods remain confined to static image analysis, ignoring the dynamic inconsistencies introduced by the generative process.

To address these shortcomings, we introduce DySy-Det (Dynamic Synergy Detector), a unified detection framework designed to achieve broader cross-domain synergy by capturing dynamic generative artifacts alongside other forensic cues. First, we propose the Semantic-Driven Anomaly Extractor (SDAE), which applies Low-Rank Adaptation (LoRA) to efficiently fine-tune a CLIP transformer. It produces abstract semantic representations for capturing conceptual inconsistencies, as well as spatial attention maps that localize discriminative regions. These maps are fused into a single binary guidance mask, which then steers our Attention-Guided Residual Enhancer (AGRE). By focusing the reconstruction process exclusively on these salient areas, AGRE effectively suppresses background noise and enhances the discriminative quality of the residual signal.

Beyond static cues, we explore the dynamics of the generative process. Previous work has shown that real images, which lie outside the learned generative manifold, often yield larger reconstruction errors. Building on this observation, we shift focus from static endpoints to dynamic denoising trajectories. Since generated images reside on the manifold, their denoising paths tend to align closely with the forward noising process, while real images exhibit misaligned trajectories. To characterize this temporal behavior, we introduce a feature called Reconstruction-Path Consistency (RPC) and design Reconstruction-Path Consistency Analyzer (RPCA) to extract it. Specifically, we compute a sequence of similarity scores across several timesteps along the denoising path, then feed it into a lightweight convolutional network. This enables the module to learn to detect subtle temporal inconsistencies while preserving the interpretability of the consistency scores.

We conduct extensive experiments on two widely-used benchmarks, GenImage (Zhu et al. 2023) and UniversalFakeDetect (Ojha, Li, and Lee 2023). Our model is trained solely on images generated by Stable Diffusion V1.4 for GenImage and ProGAN for UniversalFakeDetect, and evaluated on samples from both the training generator and over 20 unseen models. The overall mean accuracy across all models is 93.02% and 91.56% respectively, surpassing previous state-of-the-art methods by 6.14% and 1.57%.

Our contributions are summarized as follows:

- We design a synergistic mechanism that leverages CLIP attention maps to localize and emphasize meaningful residual signals during the reconstruction process.
- We propose the RPC as a dynamic feature quantifying the correspondence between noise injection and denoising trajectories in diffusion models, exposing temporal artifacts missed by static analysis.
- We present a unified detection framework that captures more comprehensive and complementary traces. Experiments show that we achieve the state-of-the-art detection performance, enabling strong generalization across unseen generators under single-source training.

Related Work

Detection Based on Learned Artifacts

Early detectors focus on hand–crafted features, such as color disparities (McCloskey and Albright 2018), flawed saturation (McCloskey and Albright 2019), or illogical illumination (Farid 2022; Matern, Riess, and Stamminger 2019), which prove too brittle against rapidly evolving generators. This limitation prompts a shift towards learning forgery artifacts directly from data. For example, Convolutional Neural Networks are widely used to autonomously learn discriminative features directly from image pixels (Wang et al. 2020; Liu et al. 2022; Tan et al. 2023). These methods capture subtle low-level artifacts, such as unnatural textures and statistical irregularities introduced by the generative process. More recently, this learning-based paradigm has expanded to leverage powerful representations from pre-trained Vision-Language Models such as CLIP to improve generalization (Ojha, Li, and Lee 2023; Sha et al. 2023; Tan et al. 2025). In addition to image-domain methods, frequency-based approaches (Frank et al. 2020; Tan et al. 2024a) posit that generative models leave behind periodic spectral artifacts. Methods like BiHPF (Jeong et al. 2022) apply high-pass filtering to amplify these subtle patterns, making them more distinguishable for detection networks.

Detection Based on Reconstruction Error

The emergence of powerful diffusion models spurs the development of specialized detection techniques that leverage the unique characteristics of these models (Wang et al. 2023; Cazenavette et al. 2024; Ricker, Lukovnikov, and Fischer 2024). DIRE (Wang et al. 2023) pioneers this field by inverting an image to Gaussian noise using DDIM Inversion (Song, Meng, and Ermon 2020) and then reconstructing it with a diffusion model, using the resulting residual error to authenticate the image. This method is based on the principle that synthetic images align with the model's manifold and are easier to reconstruct, whereas real images tend to yield larger reconstruction errors. AEROBLADE (Ricker, Lukovnikov, and Fischer 2024) introduces a training-free detection method for Latent Diffusion Models, relying on autoencoder reconstruction errors for rapid detection.

Detection Based on Multi-Perspective Cues

To build more generalizable detectors, recent work begins to jointly analyze signals from different perspectives (Yan et al. 2024; Liu et al. 2024a; Luo et al. 2024; Liu et al. 2024b). For instance, AIDE (Yan et al. 2024) utilizes the OpenCLIP model to extract semantic embeddings and combines them with low- and high-frequency features to train a classifier. LaRE^2 (Luo et al. 2024) proposes to optimize CLIP-extracted image features with latent reconstruction error, thereby improving detection accuracy. However, these approaches are constrained by two key issues: ineffective synergy between features and an oversight of dynamic generative artifacts. Our framework is designed to address these two limitations.

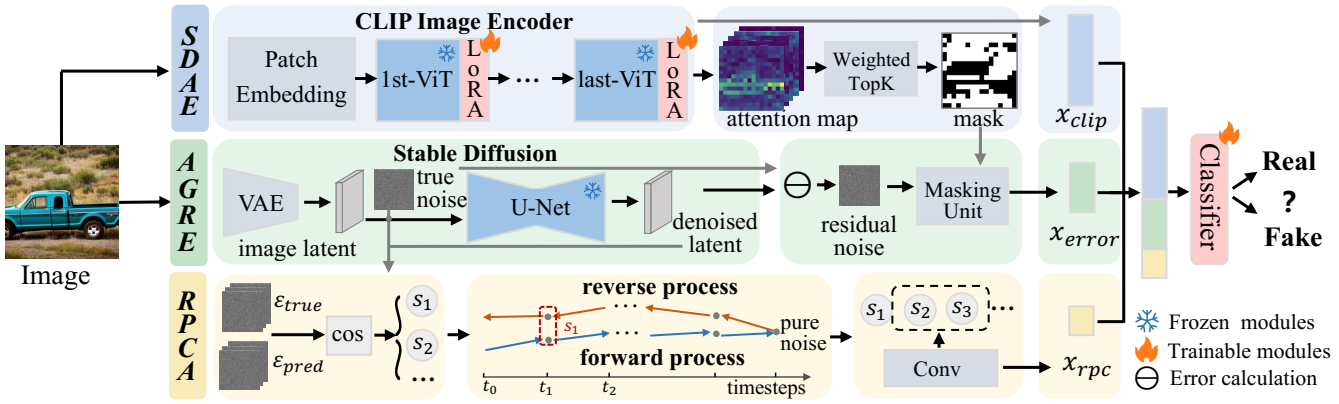


Figure 2: Architecture of DySy-Det for Generalizable Deepfake Detection. The model integrates three complementary cues: the top module extracts high-level semantic inconsistencies; the middle module computes localized reconstruction errors guided by attention mask; and the bottom module analyzes dynamic generative artifacts using the proposed RPC feature.

Methodology

In this section, we detail our proposed framework, DySy-Det, with its architecture illustrated in Figure 2.

Problem Formulation

The primary objective of AI-generated image detection is to develop a generalizable model that can accurately identify synthetic images by any unknown generators. However, due to the rapid and ongoing evolution of image generation techniques, it is impractical to cover all potential generators during training. To address this, we adopt the constraint introduced in (Wang et al. 2020), where the detection model is allowed to train on data from a single generator.

During training, we construct a dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}$, where pair (x_i, y_i) consists of synthetic samples from a designated generator G_{seen} ($x_i \sim P_{\text{data}}(G_{\text{seen}})$, $y_i = 1$) and real samples from the natural distribution ($x_i \sim P_{\text{data}}(\text{real})$, $y_i = 0$). A neural network model f_θ is trained as the detector by minimizing the loss over $\mathcal{D}_{\text{train}}$:

$$\theta^* = \arg\min_\theta \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \mathcal{L}(f_\theta(x), y). \tag{1}$$

During testing, we evaluate model performance under two scenarios: *in-distribution* (ID), using held-out samples from G_{seen} to form $\mathcal{D}^{\text{ID}}_{\text{test}}$; and *out-of-distribution* (OOD), using samples generated by unseen generators in $\mathcal{G}_{\text{OOD}} = \mathcal{G} \setminus G_{\text{seen}}$ to construct $\mathcal{D}^{\text{OOD}}_{\text{test}}$.

This strict setting reflects a realistic threat model where detectors must operate with minimal prior knowledge of unseen generators. Building generalizable detectors under such conditions remains challenging in AI-generated content forensics.

Semantic-Driven Anomaly Extractor (SDAE)

This module extracts semantic embeddings and generates a spatial guidance mask using the corresponding attention maps. To this end, we adapt a CLIP vision encoder for the forensic detection task by incorporating the parameter-efficient LoRA mechanism.

Specifically, we inject learnable low-rank matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$ ($r \ll d$) into the projection matrices of the Query, Key, and Value layers in each Transformer block. This low-rank adaptation enables efficient fine-tuning with minimal additional parameters.

The CLIP image encoder then processes the input image x_0 to generate semantic embeddings, formulated as:

$$x_{clip} = \text{CLIP}^{\text{lora}}_{\text{img}}(x_0). \tag{2}$$

Concurrently, we obtain the attention matrix from each Transformer layer and extract the attention weights from the [CLS] token to all patch tokens. We average the multi-head attention maps to compute attention scores $A_i \in \mathbb{R}^{1\times N}$, where N is the number of image patches. These scores reflect how the model allocates its focus across spatial regions, providing informative cues for identifying forgery cues.

Inspired by (Chen et al. 2023), we adopt an Exponential Moving Average (EMA) strategy to recursively integrate attention scores across layers. This helps balance shallow and deep representations for highlighting salient image regions. The cumulative attention $\tilde{A}_i$ is updated at each layer i as:

$$\tilde{A}_i = \beta \cdot \tilde{A}_{i-1} + (1-\beta)\cdot A_i, \tag{3}$$

where β denotes the momentum term.

After aggregating attention across all layers, we reshape the resulting attention vector into a 2D spatial map. We then identify the top-k locations with the highest attention values and denote their 2D coordinates as $\mathcal{I}$. These selected positions are used to construct a binary attention mask M, defined as:

$$M[p] = \begin{cases} 1, & \text{if } p \in \mathcal{I}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $p = (i, j)$ indexes a spatial location in the 2D map.

This binary mask serves as spatial guidance in the subsequent module, where it selectively filters feature representations to suppress irrelevant background signals and enhance detection accuracy.

Attention-Guided Residual Enhancer (AGRE)

This module leverages the attention mask generated by the previous stage to guide a single-step noising and denoising process, producing localized reconstruction errors that serve as strong signals for distinguishing real from AI-generated content.

The diffusion forward process involves gradually adding noise to the original image x_0 until it becomes pure noise x_T. Since the forward process of diffusion models has a closed-form solution, given a time step t, we can directly obtain x_t from x_0. In the reverse process, the noisy sample x_T is progressively denoised to reconstruct the original image. During this procedure, the model is trained to predict the injected noise. Given a noisy sample x_t and its corresponding time step t, the network is optimized to recover the ground-truth noise by minimizing:

$$L_\theta(x_0, t) = \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2. \quad (5)$$

Therefore, diffusion models inherently have the ability to obtain x_t from x_0 in a single step during the forward process and to recover x_0 from x_t during the reverse process.

Based on this property, we extract the reconstruction error as follows: For an image x_0, we first obtain its latent variable z_0 through a Variational Auto Encoder (Kingma, Welling et al. 2013). Then, we sample noise ϵ from the noise distribution $\mathcal{N}(0, I)$, and add the noise to z_0 to obtain the latent variable z_t at the time step t with predefined noise schedule α_t:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad (6)$$

Subsequently, we use pre-trained denoising U-Net ϵ_θ to estimate the noise , and compute the element-wise squared error between the prediction and ground truth noise as a spatial residual error tensor:

$$\text{LRL}_{i,j} = \left(\epsilon_{i,j} - \epsilon_\theta(z_t, t)_{i,j}\right)^2, \quad (7)$$

where (i, j) indexes the spatial location in the latent space.

Then, we perform a mean reduction over both the temporal and spatial dimensions to obtain LRL_{avg}.

To enhance the error information in critical regions while suppressing noise in other areas, we process the extracted reconstruction features using a mask from the previous module. Specifically, we upsample the mask to match the dimensions of the latent error and then multiply them element-wise to obtain x_{error}. This process can be represented by the following formula:

$$x_{\text{err}} = U(M) \odot \text{LRL}, \quad (8)$$

where M denotes the attention mask, and $U(\cdot)$ represents an upsampling operation to match the dimensions of the latent error LRL. In this way, we can not only significantly improve the efficiency of feature extraction but also retains the key information required to distinguish between real and generated images.

In summary, this module generates localized reconstruction cues focusing on salient areas, which is necessary for robust discrimination.

Reconstruction-Path Consistency Analyzer (RPCA)

To characterize temporal artifacts arising from the diffusion process, this module explicitly analyzes the denoising trajectory by quantifying noise alignment at multiple points in time.

Since evaluating all timesteps is computationally intensive, we perform sparse sampling by selecting a subset of representative timesteps $\mathcal{T} = \{t^{(1)}, t^{(2)}, \ldots, t^{(n)}\}$ from the full trajectory. This strategy retains major dynamic features while ensuring computational efficiency.

At each selected timestep $t^{(i)}$, we compute the cosine similarity between the predicted noise $\epsilon_{\text{pred}} = \epsilon_\theta(z_{t^{(i)}}, t^{(i)})$ and the ground-truth noise ϵ_{true}:

$$s^{(i)} = \frac{\epsilon_{\text{true}} \cdot \epsilon_{\text{pred}}}{\|\epsilon_{\text{true}}\| \cdot \|\epsilon_{\text{pred}}\|}, \quad i = 1, \ldots, n. \quad (9)$$

These scores are then concatenated into a temporal sequence vector:

$$\mathbf{s} = [s^{(1)}, s^{(2)}, \ldots, s^{(n)}]. \quad (10)$$

Although this hand-crafted feature provides interpretable information about denoising consistency, directly using the sequence only captures the overall similarity level in a statistical sense. It overlooks subtle yet discriminative temporal dynamics that may arise between individual time steps. To enhance the expressiveness of this feature, we further model the temporal evolution of denoising consistency using a lightweight convolutional neural network:

$$\mathbf{x}_{\text{rpc}} = \text{Conv1D}(\mathbf{s}). \quad (11)$$

The network captures local transitions along the reconstruction path, producing an interpretable feature sensitive to temporal inconsistencies.

Ultimately, the RPCA produces a robust feature that encapsulates the dynamic signature of the generative process, serving as a powerful cue for detection.

Classifier and Loss Function To leverage the complementary strengths of all three modules, we fuse their outputs into a unified feature representation. Specifically, we concatenate the semantic embedding $\mathbf{x}_{\text{clip}}$ from SDAE, the localized reconstruction error $\mathbf{x}_{\text{error}}$ from AGRE, and the dynamic signature $\mathbf{x}_{\text{rpc}}$ from RPCA:

$$\mathbf{x}_{\text{unified}} = \text{Concat}(\mathbf{x}_{\text{clip}}, \mathbf{x}_{\text{error}}, \mathbf{x}_{\text{rpc}}). \quad (12)$$

This feature is then fed into a final classifier f, to produce the prediction probability $\hat{y}$. The entire model is optimized using the standard cross-entropy loss:

$$\mathcal{L} = -\sum_i y_i \log(\hat{y}_i). \quad (13)$$

Experiments

Datasets

We conduct experiments on two widely-used datasets including GenImage (Zhu et al. 2023) and UniversalFakeDetect (Ojha, Li, and Lee 2023).

Methods	Pub	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	mAcc
ResNet-50	CVPR2016	54.90	**99.90**	99.70	53.50	61.90	98.20	56.60	52.00	72.10
Spec	WIFS2019	52.00	99.40	99.20	49.70	49.80	94.80	55.60	49.80	68.80
F3Net	ECCV2020	50.10	**99.90**	**99.90**	49.90	50.00	**99.90**	49.90	49.90	68.70
CNNSpot	CVPR2020	52.80	96.30	95.90	50.10	39.80	78.60	62.30	46.80	64.20
GramNet	CVPR2020	54.20	99.20	99.10	50.30	54.60	98.90	50.80	51.70	69.90
DeiT-S	ICML2021	55.60	**99.90**	99.80	49.80	58.10	98.90	56.90	53.50	71.60
Swin-T	ICCV2021	62.10	**99.90**	99.80	49.80	67.60	99.10	62.30	57.60	74.80
UnivFD	CVPR2023	83.70	86.05	86.25	59.00	83.30	77.55	61.55	86.15	77.95
FreqNet	AAAI2024	**87.32**	95.73	95.52	51.96	90.81	90.30	51.34	93.71	82.08
NPR	CVPR2024	87.05	99.07	98.92	64.15	94.40	97.44	56.06	88.96	85.76
LaRE^2	CVPR2024	69.85	98.24	97.93	71.73	80.26	98.89	**87.48**	80.47	85.61
AIDE	ICLR2025	79.38	99.75	99.75	78.49	91.77	98.85	80.24	66.83	86.88
Ours		86.51	99.88	99.79	**80.54**	**95.78**	99.76	86.24	**95.68**	**93.02**

Table 1: Accuracy (%) Comparison on the GenImage Dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Methods	Pub	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	mAP
UnivFD	CVPR2023	95.30	96.51	96.11	71.93	94.04	92.15	80.86	96.50	90.43
FreqNet	AAAI2024	94.80	99.51	99.57	48.14	96.76	97.15	49.33	98.13	85.42
NPR	CVPR2024	96.22	99.99	99.94	70.88	98.97	99.70	63.98	97.33	90.88
LaRE^2	CVPR2024	98.08	99.88	99.80	94.87	95.14	99.86	97.55	97.40	97.82
AIDE	ICLR2025	97.95	99.95	99.98	94.58	99.08	99.97	97.07	93.48	97.76
Ours		**99.20**	**100.00**	**100.00**	**99.14**	**99.79**	**99.99**	**99.26**	**99.79**	**99.65**

Table 2: Average Precision (%) Comparison on the GenImage Dataset.

GenImage This dataset primarily consists of images from diffusion-based generative models. We use images from Stable Diffusion V1.4 (Rombach et al. 2022) for training and evaluate over eight generators: Midjourney (Midjourney Team 2022), Stable Diffusion V1.4, Stable Diffusion V1.5 (Rombach et al. 2022), ADM (Dhariwal and Nichol 2021), GLIDE (Nichol et al. 2021), Wukong (Wukong Model Zoo 2022), VQDM (Gu et al. 2022) and BigGAN (Brock, Donahue, and Simonyan 2018).

UniversalFakeDetect Following prior work (Wang et al. 2020), we train on four classes (horses, chairs, cats, cars) from ProGAN (Karras et al. 2018). Our evaluation is conducted on a test set composed of 19 subsets, including: ProGAN, CycleGAN (Zhu et al. 2017), BigGAN (Brock, Donahue, and Simonyan 2018), StyleGAN (Karras, Laine, and Aila 2019), GauGAN (Park et al. 2019), StarGAN (Choi et al. 2018), SITD (Chen et al. 2018), CRN (Chen and Koltun 2017), DeepFake (Rossler et al. 2019), IMLE (Li, Zhang, and Malik 2019), SAN (Dai et al. 2019), Guided Diffusion (Dhariwal and Nichol 2021), DALLE (Ramesh et al. 2021), LDM (Rombach et al. 2022) and GLIDE (Nichol et al. 2021).

Implementation Details

Our model is implemented in PyTorch (Paszke et al. 2017) and builds on two pre-trained backbones. For semantic inconsistencies, we adapt CLIP ViT-L/14 (Radford et al. 2021) using LoRA ($r = 8$, $\alpha = 16$, dropout=0.1) and generate a spatial mask from its top-k ($k = 100$) attention tokens using EMA with a momentum of $\beta = 0.99$. For reconstruction-based features, we employ Stable Diffusion v1.5 (Rombach et al. 2022) with the prompt "a photo" at time step $t \in \{200, 250, 300\}$. A final MLP then fuses the feature streams from both backbones for classification. Input images are resized to 224×224. The model is trained for a single epoch using the Adam optimizer with a learning rate of $1e{-}4$ and a batch size of 48. All experiments are conducted on two NVIDIA A6000 GPUs with a fixed random seed of 32 to ensure reproducibility. Further details on the selection of these hyperparameters are provided in the Appendix.

We report mean Accuracy (mAcc) and mean Average Precision (mAP) as evaluation metrics.

Comparison to State-of-the-Art Models

Evaluation On GenImage Table 1 and Table 2 present a comprehensive performance comparison on the GenImage dataset, reporting detection Accuracy (Acc) and Average Precision (AP), respectively. The accuracy results for several foundational methods (from ResNet-50 (He et al. 2016) to Swin-T (Liu et al. 2021)) are cited from the original GenImage paper (Zhu et al. 2023). As the source paper did not report AP scores for these methods, they are consequently excluded from the comparison in Table 2. For all other methods, including UnivFD (Ojha, Li, and Lee 2023), FreqNet (Tan et al. 2024a), NPR (Tan et al. 2024b), LaRE^2 (Luo et al. 2024), and AIDE(Yan et al. 2024), we obtained the results using their official pre-trained models or our re-implementations.

Subsets	Freq-spec WIFS2019	CNNSpot CVPR2020	Patchfor ECCV2020	UnivFD CVPR2023	LGrad CVPR2023	FreqNet AAAI2024	NPR CVPR2024	FatFormer CVPR2024	Ours
ProGAN	49.90	**99.99**	75.03	99.85	99.85	99.58	99.94	99.89	99.98
CycleGAN	**99.90**	85.20	68.97	98.55	85.28	95.84	90.27	99.36	97.35
BigGAN	50.50	70.20	68.47	94.90	82.93	90.45	87.28	**99.50**	**99.50**
StyleGAN	49.90	85.70	79.16	95.60	94.74	90.22	96.27	97.13	**99.44**
GauGAN	50.30	78.95	64.23	99.35	72.46	93.41	85.36	**99.43**	99.39
StarGAN	99.70	91.70	63.94	95.60	99.57	85.67	99.65	**99.75**	97.90
SITD	50.00	66.67	75.14	62.22	56.39	65.56	62.22	81.39	**91.11**
CRN	50.60	**86.31**	72.33	56.45	50.57	59.04	50.01	69.46	84.10
DeepFake	50.10	53.47	75.54	58.85	57.95	88.92	78.83	**93.27**	89.77
IMLE	50.10	86.26	55.30	68.70	50.60	59.06	50.01	69.46	**88.68**
SAN	48.00	48.69	**75.28**	56.62	55.47	71.92	68.26	68.04	59.59
Guided	50.90	60.07	67.41	70.00	76.60	67.25	**76.70**	76.00	70.15
DALLE	50.00	55.58	67.91	87.45	89.40	97.25	93.85	**98.10**	97.85
LDM_200	50.40	54.03	76.50	74.15	95.27	97.25	98.30	90.40	**99.20**
LDM_200_CFG	50.40	54.96	76.10	95.15	95.35	97.75	**98.65**	97.90	94.25
LDM_100	50.30	54.14	75.77	94.55	94.75	97.04	98.45	97.80	**99.20**
Glide_100_27	51.70	60.78	74.81	79.20	91.25	86.55	**97.55**	94.65	89.80
Glide_50_27	51.40	63.80	73.28	78.05	90.65	87.80	**97.60**	89.00	90.80
Glide_100_10	50.40	65.66	68.52	78.65	88.05	84.40	**97.10**	89.30	91.60
mAcc	55.45	69.58	71.24	81.26	80.38	85.00	85.60	89.99	**91.56**

Table 3: Accuracy (%) Comparison on the UniversalFakeDetect Dataset. Each row corresponds to one generative-model subset, and each column corresponds to one detection method.

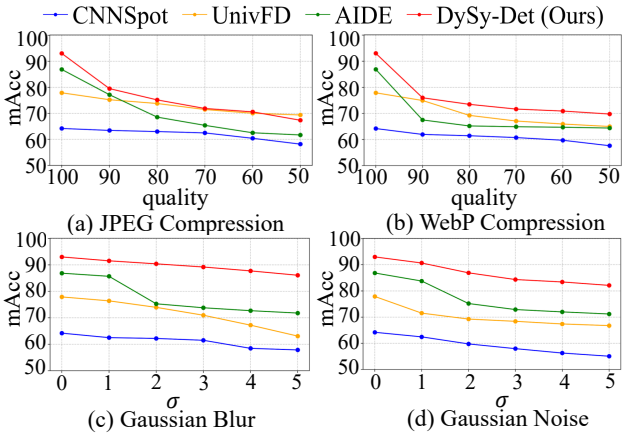


Figure 3: Detection performance of DySy-Det and other methods when handling perturbed images, measured in mAcc.

As highlighted in (Zhu et al. 2023), generators like SD V1.4, SD V1.5, and Wukong share similar architectures, making them relatively easier to detect. In contrast, subsets such as Midjourney, ADM, GLIDE, VQDM, and BigGAN exhibit higher diversity in their generative mechanisms, thus better reflecting generalization ability. On these challenging subsets, our method achieves competitive or leading performance, demonstrating strong generalization against unseen generation patterns.

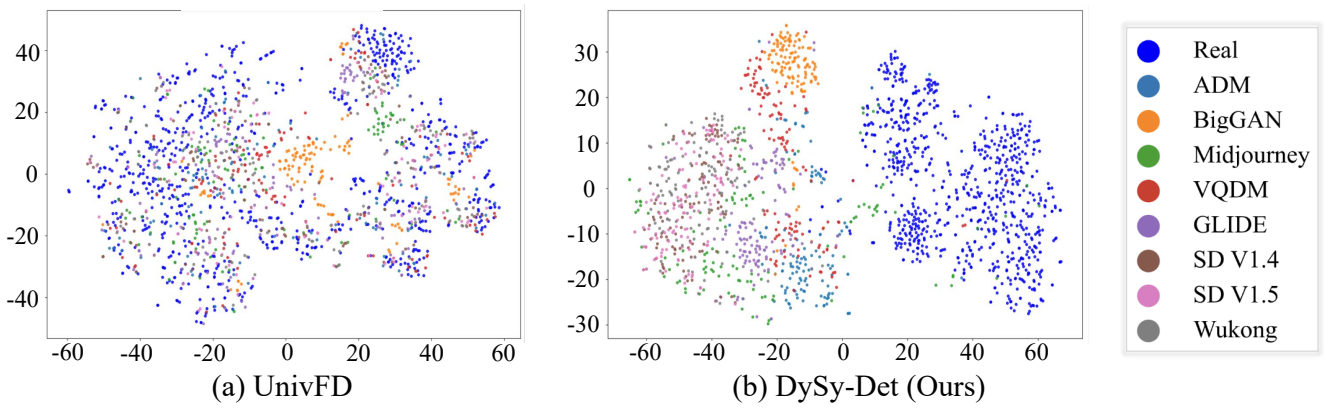
Quantitatively, our proposed method achieves the state-of-the-art performance with a mAcc of 93.02% and a mAP of 99.65% on the GenImage dataset. This represents a substantial improvement over the baseline UnivFD (77.95% mAcc, 90.43% mAP) and the prior state-of-the-art AIDE (86.88% mAcc, 97.76% mAP). These significant gains in both accuracy and average precision validate the effectiveness of our synergy-driven design, which integrates multi-level features that span semantic understanding, structural focus, and generative behavior.

Evaluation On UniversalFakeDetect The performance of Acc on the UniversalFakeDetect dataset is presented in Table 3. Results for Freq-spec, CNNSpot, and Patchfor (Chai et al. 2020) are cited from the paper (Ojha, Li, and Lee 2023). We obtained the results for UnivFD, LGrad (Tan et al. 2023), FreqNet, NPR, and FatFormer (Liu et al. 2024a) by using their official pre-trained models or our re-implementations.

The experimental results demonstrate that DySy-Det establishes a new state-of-the-art performance. Our model achieves a mean accuracy of 91.56%, outperforming the baseline UnivFD (81.26%) by 10.30% and the previous state-of-the-art, FatFormer (89.99%), by 1.57%. For a more comprehensive comparison, the results measured by AP are provided in Appendix.

Robustness to Unseen Perturbations In real-world scenarios, images undergo various unforeseen perturbations during acquisition, transmission, post-processing, and uploads, which can degrade the performance of detection methods. To evaluate our method under such non-ideal conditions, we tested four types of perturbations: JPEG compression (quality = 100-50), WebP compression (quality = 100-50), Gaussian blur (σ = 0-5), and Gaussian noise (σ = 0-5).

As illustrated in Figure 3, all methods experience a decline in performance as the severity of perturbations increases. However, DySy-Det consistently achieves the high-



(a) UnivFD (b) DySy-Det (Ours)

Figure 4: The t-SNE visualization of spatial representation from UnivFD (a) and our method (b).

est accuracy across nearly all distortion levels. The only exception arises under extreme JPEG compression (quality = 50), where its performance marginally falls behind that of UnivFD. Despite this, the overall results substantiate the robustness of our approach, further affirming its practical suitability for real-world scenarios involving noisy or degraded inputs.

Visualization Analysis To further assess the discriminative ability of our method, we visualize the feature embeddings extracted from the final layer of UnivFD and our DySy-Det using t-SNE (Van der Maaten and Hinton 2008), as shown in Figure 4.

Compared to UnivFD (Figure 4a), Our Method (Figure 4b) produces embeddings with much clearer class-wise separation. Real images form a more compact and distinct cluster, while samples from different generative models are better disentangled.

These results demonstrate that DySy-Det learns more discriminative and generalizable representations, which is crucial for reliable detection across diverse and unseen synthetic sources. Furthermore, visualizations of the logit distributions of extracted forgery features are provided in the Appendix.

Ablation Studies

We conduct ablation experiments on the GenImage dataset to evaluate the individual effectiveness of each feature type and the impact of their interaction, as shown in Table 4.

Guided-Reconstruction Error Comparing Setting 5 (full model) with Setting 4 (w/o reconstruction error) shows that removing error features reduces accuracy from 93.02% to 91.92%. This confirms that low-level residual cues from the denoising process contribute meaningful complementary information to CLIP semantics by revealing subtle texture inconsistencies.

RPC The contribution of the RPC is evaluated by comparing Setting 5 (full model) and Setting 3 (w/o RPC). Removing RPC results in a drop from 93.02% to 90.45%, indicating that RPC captures valuable temporal trajectory information during denoising and enhances the model's ability to distinguish real from generated content.

Setting	CLIP	Error	Mask	RPC	mAcc
1	✓				84.85
2	✓	✓		✓	91.52
3	✓	✓	✓		90.45
4	✓			✓	91.92
5	✓	✓	✓	✓	**93.02**

Table 4: Ablation study of our model's components. We report the mAcc (%) on the GenImage dataset.

Attention Mask To evaluate the role of the attention mask in guiding residual feature extraction, we compare Setting 5 (full model) and Setting 2 (w/o mask). Accuracy drops from 93.02% to 91.52%, showing that spatial guidance improves the relevance of reconstruction features.

Summary Each component meaningfully improves detection accuracy. Their synergy yields the best performance (93.02%), validating the effectiveness of our complementary multi-cues framework.

Conclusion

In this paper, we first analyze the core limitations of existing AI-generated image detectors, particularly their lack of effective synergy of different forensic cues and overlooking dynamic inconsistencies. Then we introduced DySy-Det, a novel framework designed to capture more holistic forgery signatures by synergizing three complementary sources of evidence: high-level semantic inconsistencies, localized reconstruction errors guided by semantic attention, and dynamic generative artifacts called RPC feature. Extensive experiments on GenImage and UniversalFakeDetect demonstrate that DySy-Det achieves state-of-the-art performance and strong robustness under common image perturbations. We believe that DySy-Det offers a promising direction for improving the robustness and generalization for AI-generated image detection.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2023YFB2904000 and 2023YFB2904001, in part by the National Natural Science Foundation of China under Grant U2436206, 62032021, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LZ25F020005, and in part by Ant Group through CCF-Ant Research Fund.

References

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

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Cazenavette, G.; Sud, A.; Leung, T.; and Usman, B. 2024. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10759–10769.

Chai, L.; Bau, D.; Lim, S.-N.; and Isola, P. 2020. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, 103–120. Springer.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3291–3300.

Chen, M.; Lin, M.; Li, K.; Shen, Y.; Wu, Y.; Chao, F.; and Ji, R. 2023. Cf-vit: A general coarse-to-fine method for vision transformer. In *Proceedings of the AAAI conference on artificial intelligence*, 7042–7052.

Chen, Q.; and Koltun, V. 2017. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, 1511–1520.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11065–11074.

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

Dong, C.; Kumar, A.; and Liu, E. 2022. Think Twice Before Detecting GAN-Generated Fake Images From Their Spectral Domain Imprints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7865–7874.

Farid, H. 2022. Lighting (in) consistency of paint by text.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the 37th International Conference on Machine Learning*, 3247–3258.

Gaffar, H.; and Albarashdi, S. 2025. Copyright protection for AI-generated works: Exploring originality and ownership in a digital landscape. *Asian Journal of International Law*, 15(1): 23–46.

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*, 12: 48126–48144.

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.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10696–10706.

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.

Jeong, Y.; Kim, D.; Min, S.; Joe, S.; Gwon, Y.; and Choi, J. 2022. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 48–57.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

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

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Li, K.; Zhang, T.; and Malik, J. 2019. Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4220–4229.

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

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024a. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10770–10780.

Liu, R.; Zhang, S.; Xu, Y.; Xu, W.; and He, X. 2024b. High-resolution network-based multi-feature fusion for generalized forgery detection. *Multimedia Systems*, 35.

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

Luo, Y.; Du, J.; Yan, K.; and Ding, S. 2024. LaREˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17006–17015.

Matern, F.; Riess, C.; and Stamminger, M. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83–92. IEEE.

McCloskey, S.; and Albright, M. 2018. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*.

McCloskey, S.; and Albright, M. 2019. Detecting GAN-generated imagery using saturation cues. In *2019 IEEE international conference on image processing (ICIP)*, 4584–4588. IEEE.

Midjourney Team. 2022. Midjourney. https://www.midjourney.com/home/. Accessed 2022.

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.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

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.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. 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.

Sha, Z.; Li, Z.; Yu, N.; and Zhang, Y. 2023. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, 3418–3432.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502*.

Tan, C.; Tao, R.; Liu, H.; Gu, G.; Wu, B.; Zhao, Y.; and Wei, Y. 2025. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7184–7192.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024a. Frequency-aware deepfake detection: improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024b. 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.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

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.

Wukong Model Zoo. 2022. Wukong. https://xihe.mindspore.cn/modelzoo/wukong. Accessed May 2022.

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*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

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. In *Advances in Neural Information Processing Systems*, 77771–77782.