Beyond Semantic Features: Pixel-level Mapping for Generalized AI-Generated Image Detection

Chenming Zhou[1,2], Jiaan Wang[1,2], Yu Li[1,2*], Lei Li[1,2], Juan Cao[1,2], Sheng Tang[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{zhouchenming21b,wangjiaan24s,liyu,lilei,caojuan,ts}@ict.ac.cn

Abstract

The rapid evolution of generative technologies necessitates reliable methods for detecting AI-generated images. A critical limitation of current detectors is their failure to generalize to images from unseen generative models, as they often overfit to source-specific semantic cues rather than learning universal generative artifacts. To overcome this, we introduce a simple yet remarkably effective *pixel-level mapping* pre-processing step to disrupt the pixel value distribution of images and break the fragile, non-essential semantic patterns that detectors commonly exploit as shortcuts. This forces the detector to focus on more fundamental and generalizable high-frequency traces inherent to the image generation process. Through comprehensive experiments on GAN and diffusion-based generators, we show that our approach significantly boosts the cross-generator performance of state-of-the-art detectors. Extensive analysis further verifies our hypothesis that the disruption of semantic cues is the key to generalization.

Introduction

The rapid advancement of generative models, from Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) to contemporary Diffusion Models (Ho, Jain, and Abbeel 2020), has propelled AI-generated images with remarkable fidelity and diversity that are now virtually indistinguishable from authentic photographs. While these generative techniques have catalyzed innovation in creative arts and industrial applications, they have simultaneously precipitated a crisis of visual authenticity. The proliferation of AI-generated forgeries poses significant societal risks, particularly through the dissemination of misinformation and the erosion of evidentiary reliability in critical domains. This emerging threat landscape underscores the urgent need for robust detection methodologies in AI security.

In the task of in-distribution detection, classifiers often achieve remarkably high accuracy, yet their detection performance significantly deteriorates when faced with unknown generative models. To address the generalization challenge in detection, existing techniques typically fall into two categories: data-centric techniques that pre-process images to highlight forensic traces, and model-centric techniques that aim to learn more generalized features. Wang et al. (2020) demonstrate that aggressive data augmentation can enable convolutional classifiers to achieve cross-model generalization, while Durall, Keuper, and Keuper (2020) analyze frequency-domain anomalies in GAN-generated images, identifying high-frequency artifacts caused by upsampling operations for generalized detection. Recent methods leverage the powerful feature learning capabilities of pre-trained large-scale models (Ojha, Li, and Lee 2023; Tan et al. 2025; Cozzolino et al. 2024; Khan and Dang-Nguyen 2024), assuming that their pre-training on vast real-image datasets allows them to detect synthetic content through deviations in feature distributions.

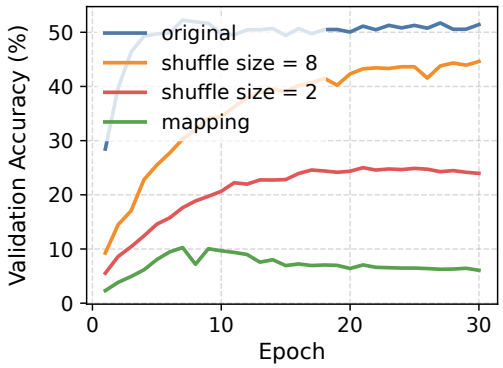


Figure 1: The impact of different image processing methods on ImageNet classification results.

Although existing methods generalize well when test and training distributions align, their accuracy declines significantly under substantial semantic shifts. This issue primarily stems from *semantic bias* discrepancies caused by imperfect fitting to training data in generative models (Yan et al. 2025; Guillaro et al. 2025), manifesting as visual artifacts like blurring and texture anomalies. However, different models exhibit specific semantic biases. As generative models continue to evolve with improved architectures and sampling techniques, these semantic biases diminish, leading to performance degradation in detectors that rely on them. Therefore, reducing the influence of semantic bias during classifier

*Corresponding author.

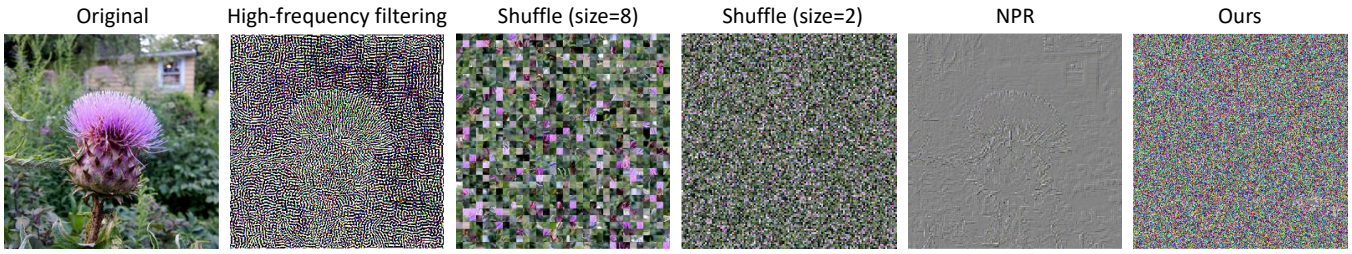


Figure 2: Visualization results of various semantic-reduction methods.

training is crucial for detection generalization.

To mitigate the impact of *semantic bias*, some approaches leverage the Fourier transform to convert images into the frequency domain, where low-frequency components are removed via masking before reconstructing the images (Tan et al. 2024b; Bammey 2023; Chu et al. 2024). Although such spectral pruning can partially suppress low-frequency information, it fails to completely eliminate its interference and inevitably incurs information loss that degrades generative artifact detection. Alternative methods employ patch shuffling (Fu et al. 2025; Liu et al. 2022; Zheng et al. 2024) to reduce the classifier's receptive field, thereby preventing overfitting to abstract semantic patterns. However, as shown in Figure 2, high-pass filtering and NPR (Tan et al. 2024b) methods can still preserve noticeable semantic information in images. Moreover, although the shuffle method progressively destroys semantics as the patch size decreases, the ImageNet classification experiments in Figure 1 demonstrate that, despite slower convergence and lower validation accuracy, models still extract sufficient semantic information from shuffled patches even with the minimal patch size of 2.

To address the limitations of existing semantic-reduction methods, we propose a pixel-level mapping approach that reduces semantic bias through pixel value transformations. Since semantic bias mainly resides in low-frequency components, our method amplifies inter-pixel disparities to suppress low-frequency patterns while preserving detectable high-frequency artifacts. As shown in Figure 1 and Figure 2, our mapping achieves greater semantic reduction and lower validation accuracy than both baseline and shuffling methods, confirming the effective reduction of semantic information. By transforming low-frequency information, the impact of semantic bias on classifiers is significantly reduced, thereby amplifying the influence of high-frequency generative artifacts during training. While manipulating pixel values, our method preserves pixels' local correlations, offering two distinct advantages over existing bias mitigation strategies: 1) minimal information loss during image processing, and 2) enhanced emphasis on high-frequency forensic features during classifier training. Comprehensive experiments across diverse datasets and generative models (GANs and Diffusion models) demonstrate consistent improvements in cross-model generalization. Our key contributions are:

- We empirically demonstrate that simple high-pass filtering and image patch shuffling fail to effectively eliminate semantic information, making them inadequate for reducing the impact of semantic bias on generalization in detection tasks. Our findings reveal critical limitations in existing approaches and highlight the need for more sophisticated semantic suppression methods.
- We introduce a novel pixel-level mapping method that attenuates semantic bias during classifier training. By transforming low-frequency information while amplifying high-frequency artifacts, our method significantly reduces classifier reliance on biased semantic features, addressing the core generalization challenge in synthetic image detection.
- Extensive validation across multiple benchmarks confirms the effectiveness of our method. The results show consistent performance gains when detecting images from unseen generative architectures, demonstrating the generalization capability of the proposed method.

Related Work

Generative Models

Generative models differ fundamentally from traditional autoencoders (Masci et al. 2011; Vincent et al. 2008; Salah et al. 2011) by their ability to sample novel samples that conform to the distribution of existing data. GANs initially led synthetic image generation, with key variants addressing specific constraints: ProGAN (Karras et al. 2017) for progressive training, StyleGAN (Karras, Laine, and Aila 2019) for disentangled representations, BigGAN (Brock, Donahue, and Simonyan 2018) for large-scale synthesis, and StarGAN (Choi et al. 2018) for multi-domain translation. Although GANs offer fast inference and modular extensibility, they produce noticeable semantic artifacts due to limited generation quality. Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Song et al. 2020) later emerged as a mathematically grounded framework, overcoming GANs' training instability while ensuring better sample diversity. Recent large-scale implementations (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022) trained on extensive datasets achieve photorealistic high-resolution generation. These rapid advances in quality and diversity have markedly reduced semantic artifacts, thereby weakening detection methods dependent on semantic features.

AI-Generated Image Detection

The generalization challenge in synthetic image detection was first addressed by Wang et al. (2020), demonstrating that

aggressive data augmentation enables cross-GAN detection. Durall, Keuper, and Keuper (2020) later revealed that GAN upsampling introduces distinct high-frequency artifacts, detectable via spectral analysis. However, such GAN-based observations limit classifier generalization. The emergence of diffusion models further challenged these approaches, as higher image fidelity eliminated many detectable artifacts. This prompted methods (Ojha, Li, and Lee 2023; Cozzolino et al. 2024; Tan et al. 2025) leveraging large-scale pretrained models to detect synthetic images by their deviations from natural distributions. Yet these models primarily rely on semantic features and overlook high-frequency traces correlated with generation artifacts. When test data diverges from fine-tuning distributions, their detection performance degrades significantly and shows inconsistency across domains.

To mitigate classifiers' over-reliance on semantic bias, existing methods disrupt global semantics via frequency manipulation or patch shuffling. Some studies (Tan et al. 2024b; Bammey 2023; Chu et al. 2024) mitigate semantic interference by extracting high-frequency components through residual operations or frequency filtering. However, removing specific frequency bands cannot fully eliminate semantic influence, and the inherent coupling between generative artifacts and semantic content causes collateral damage under direct spectral suppression. While the patch shuffling strategies (Fu et al. 2025; Liu et al. 2022; Zheng et al. 2024) aim to limit the classifier's receptive field by randomly permuting image patches, these semantic-suppression techniques force classifiers to prioritize forensic traces that generalize better across generator architectures. Yet, our experiments reveal that classifiers can still capture semantic information from shuffled patches through powerful fitting capacity. Meanwhile, the shuffling operation may disrupt the global structural patterns of generative artifacts.

Methodology

Preliminaries

The detection generalization of generated images constitutes a classification task, aiming to perform binary classification on input images to determine whether they are synthesized by generative models (Wang et al. 2020). Specifically, the classifier is trained on images generated by a *limited set* of known generative models, yet must maintain effectiveness when identifying images from *unseen* generative architectures. Let $\mathcal{X} \subseteq \mathbb{R}^{H\times W\times C}$ denote the image space and $\mathcal{G} = \{G_1, ..., G_k\}$ represent a set of known generative models. We define:

- **Training Data:** $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ where:

$$y_i = \begin{cases} 1, & \text{if } x_i \sim P_{\text{gen}}(x|G_j), G_j \in \mathcal{G} \\ 0, & \text{if } x_i \sim P_{\text{real}}(x) \end{cases} \tag{1}$$

- **Objective:** Learn a classifier $f_\theta : \mathcal{X} \to [0, 1]$ that minimizes the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log f_\theta(x_i) + (1 - y_i)\log(1 - f_\theta(x_i))\right], \tag{2}$$

while generalizing to unseen models $\mathcal{G}' = \{G_{k+1}, ..., G_{k+m}\}$.

Most existing studies on detection generalization focus on training data containing samples from only *one category* of generative models (Wang et al. 2020; Frank et al. 2020; Luo et al. 2024). The trained classifier is required to:

1. Achieve *intra-class generalization*: Maintain detection accuracy across:
   - Different architectures of the same model family, e.g., StyleGAN2 $\rightarrow$ BigGAN.
   - Same model architecture trained with different data or hyperparameters, e.g., Stable Diffusion-V1.4 $\rightarrow$ Stable Diffusion-V1.5.
2. Demonstrate *cross-class generalization*: Transfer effectively to:

$$\mathcal{G}_{\text{test}} \in \{\text{Diffusion Models}\} \times \{\text{GANs}\} \setminus \mathcal{G}_{\text{train}}, \tag{3}$$

where $\mathcal{G}_{\text{train}}$ contains only one model category.

Training classifiers directly on RGB images can achieve high accuracy on generated images from the same source model, but performance significantly degrades on out-of-distribution data. This occurs because the classifier tends to overfit to semantic distribution patterns in the training set, limiting its generalization capability. Two primary factors contribute to this phenomenon: 1) Different generative models introduce distinct semantic artifacts (e.g., blurring, texture anomalies) due to variations in model architectures and training procedures - even identical models trained on the same data produce subtle differences from random initialization. 2) As generative models improve, their outputs increasingly approximate real data distributions, making semantic-level artifacts harder to detect in high-fidelity samples. To generalize better on unseen models, detectors must avoid relying on semantic bias in detection.

Pixel-level Mapping for Semantic Bias Reduction

Generative artifacts in synthetic images exhibit tight coupling with semantic content, where perturbations to semantics inevitably distort artifact distributions. Existing works generally associate semantic bias with the low-frequency components of the smooth regions of the image, while generative traces are related to high-frequency details (Tan et al. 2024a; Bammey 2023). The inductive bias of convolutional classifiers towards low-frequency features (Tang et al. 2022) exacerbates semantic dominance during training. Therefore, weakening low-frequency semantic bias while enhancing high-frequency trace is a promising pathway to improve detection generalization.

To this end, we propose a pixel-level mapping approach that converts monotonically ordered pixel values (0-255) into a new set of pixel values, thereby altering the tight spatial arrangement between neighboring pixels. This transformation converts the image's low-frequency information into high-frequency while preserving the correlations between pixels. The overall process is illustrated in Figure 3(a). The input image undergoes semantic transformation through a

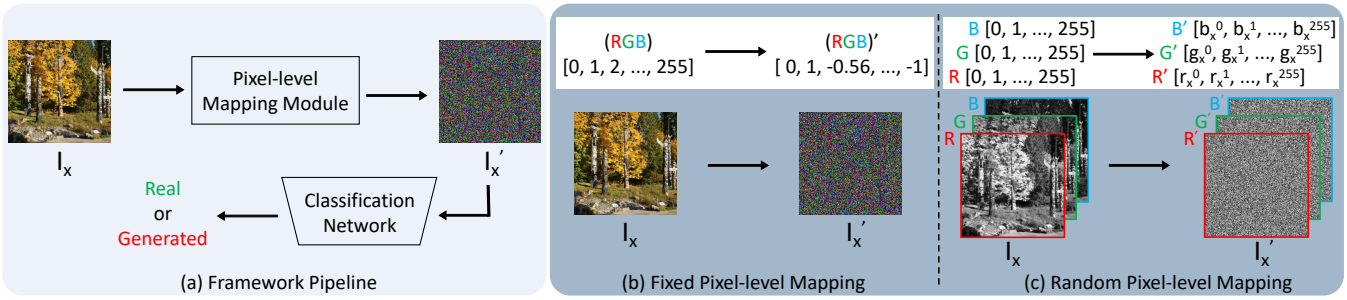


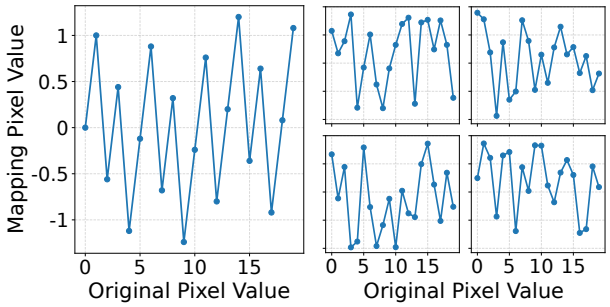
Figure 3: (a) The framework pipeline of the proposed method. The input image first passes through the pixel-level mapping module before being sent into the classification head. (b) The fixed pixel-level mapping module applies the same fixed mapping to all three channels of images. (c) The random pixel-level mapping module applies a different random mapping to the three channels of each image.

pixel-level mapping module prior to being fed into the classification head. To effectively convert low-frequency information into high-frequency components by amplifying inter-pixel value differences, we propose a computationally efficient fixed mapping approach within our pixel-level mapping module in Figure 3(b). Our extended experimental results reveal an important insight: the specific pixel mapping relationship itself is not the critical factor. Rather, the key lies in disrupting the original monotonic pixel arrangement. Even when applying completely randomized mapping relationships that vary per sample as in Figure 3(c), the detection accuracy remains statistically comparable to deterministic mappings. Therefore, the fixed mapping proposed in this work can be viewed as one specific instantiation of this broader mapping paradigm. The implementation of the proposed pixel-level mapping module operates as follows:

Fixed pixel-level mapping module. Given an input RGB image $I \in \mathbb{R}^{H\times W\times 3}$ with pixel values ranging from 0 to 255, we propose a mapping function that amplifies the differences between adjacent pixel values while normalizing the pixel intensities to facilitate classifier training. Formally, for each pixel value $v \in [0, 256)$, the mapping function ϕ_f can be expressed as:

$$\phi_f(v) = v - \text{round}\left(\frac{v}{256}, 2\right) \times 256, \tag{4}$$

where v denotes the original pixel intensity and round represents the round operation (equivalent to NumPy's `np.round` with `decimals=2`). The selection of `decimals` aligns with our monotonicity disruption principle. Setting `decimals=1` preserves pixel linearity due to coarse quantization (1/256≈0.0039), while `decimals>1` effectively disrupt monotonic arrangements and `decimals=2` can simultaneously normalize the transformed pixel values to the approximate range of [-1.28, 1.28]. The correspondence between original and mapping pixel values in the fixed mapping module is illustrated in Figure 4(a). To demonstrate the relation clearly, we present only the mapping results for pixel values in the range [0, 20]. As evident from the figure, the disparities between adjacent pixel values are significantly accentuated compared to the regularly normalized pixel values, thereby



(a) (b)

Figure 4: (a) Fixed pixel-level mapping table, which remains the same for each channel of each sample. (b) Four examples of random pixel-level mapping tables, which maintain randomness for each channel of each sample.

transforming originally smooth low-frequency regions in the image into high-frequency components.

Random pixel-level mapping module. The fixed mapping method applies identical mapping rules to all input samples. Through further experiments, we observe that even when each sample (and every channel within each sample) undergoes transformations with randomly generated mapping tables, the classifier can still extract detection-generalizable features from the varying high-frequency information, achieving classification performance comparable to that of the fixed mapping approach. Concretely, for each input image $I \in \mathbb{R}^{H\times W\times 3}$, we construct a per-channel mapping table $T_c \in \mathbb{R}^{256}$ where each entry follows an independent and identically distributed (i.i.d.) uniform distribution over 256 dimensions as follows:

$$T_c \sim \mathcal{U}(-1, 1)^{256}, \quad c \in \{0, 1, 2\}. \tag{5}$$

The transformed image $I' \in \mathbb{R}^{H\times W\times 3}$ is then generated by:

$$I'_c[x, y] = T_c\big[I_c[x, y]\big], \quad \forall\, x \in [0, H),\ y \in [0, W), \tag{6}$$

where $I_c[x, y] \in \{0, ..., 255\}$ denotes the original pixel value at location (x, y) in channel c.

In Figure 4(b), we visualize four randomly generated mapping tables. For clarity of presentation, only the mapped results for pixel values in the range of [0, 20] are displayed. The figure demonstrates that the randomized mappings can similarly amplify the disparities between adjacent pixels. Although distinct samples and image channels undergo different transformations, our subsequent experiments reveal that employing varied mappings still preserves the effectiveness of detection-relevant features.

Experiments

Setup

Training datasets. We follow the settings from NPR (Tan et al. 2024b) and C2P-CLIP (Tan et al. 2025), which use the ForenSynths (Wang et al. 2020) and GenImage datasets (Zhu et al. 2023). The ForenSynths dataset contains 20 semantic classes, of which we exclusively employ 4 classes (i.e., car, cat, chair, horse) during training to maintain consistency with previous works. For the GenImage dataset, we adopt SDv1.4 (Rombach et al. 2022) as the generative model. The real images in both ForenSynths and GenImage originate from the LSUN (Yu et al. 2015) and ImageNet datasets (Russakovsky et al. 2015).

Test datasets. To evaluate the generalization capability of the proposed method in real-world scenarios, we incorporate diverse real images alongside GAN and Diffusion models. The evaluation benchmark comprises two datasets of Self-Synthesis (Tan et al. 2024a) and GenImage (Zhu et al. 2023). The detailed descriptions are available in the Supplementary Materials.

Implementation details. Our method is implemented using PyTorch (Paszke et al. 2019) with 8 Nvidia 3090 GPUs. We implement the detector network with a ResNet-50 (He et al. 2016) architecture. During training, we randomly crop images to size 128×128 to avoid resizing bias. During testing, we center-crop the images. We utilize the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 2×10^{-4}. The first-order moment decay rate and the second-order moment decay rate are set to 0.9 and 0.999, respectively, and weight decay is set to 2×10^{-4}. We train the detector network for 200 epochs with a batch size of 128.

Metrics. We follow existing works (Ojha, Li, and Lee 2023; Liu et al. 2024; Tan et al. 2025), to compare and report classification accuracy (Acc) and average precision (AP). A uniform classification threshold of 0.5 is maintained on all evaluation benchmarks to ensure an equitable comparison of detection performance.

Quantitative Analysis

We present a comprehensive evaluation of our proposed method against competing approaches through cross-dataset and cross-model testing.

Evaluation on Self-Synthesis GAN dataset. The results of accuracy are shown in Table 1. The results of comparative methods are from NPR (Tan et al. 2024b). The training setting is the same as NPR which uses ProGAN (4 classes). This benchmark contains data from 9 state-of-the-art GAN models to evaluate generalization ability across different GAN models. Our method surpasses the baseline UniFD (Ojha, Li, and Lee 2023) by 20.3% in classification accuracy and outperforms the current state-of-the-art NPR (Tan et al. 2024b) by 4.7%, demonstrating remarkable generalization over different GAN models.

Evaluation on GenImage dataset. The results of accuracy are shown in Table 2. The results of the compared methods are from GenImage (Zhu et al. 2023), C2P-CLIP (Tan et al. 2025) and DRCT (Chen et al. 2024). All the detection models are trained on the GenImage training set with SDv1.4 as the generative model. The GenImage dataset incorporates synthetic images generated by state-of-the-art diffusion models, including commercial closed-source models such as MidJourney and WuKong. Notably, a subset of generated images employs substantially higher resolutions. For instance, MidJourney outputs 1024×1024 pixel images, whose resolution discrepancy from conventional datasets introduces resolution bias that poses significant challenges to detection robustness. Our pixel-level mapping method achieves a new state-of-the-art result with an average accuracy of 98.4%, exceeding baseline UniFD and state-of-the-art C2P-CLIP by 9.6% and 2.6% respectively. We additionally conduct experiments on the UniversalFakeDetect dataset (Ojha, Li, and Lee 2023), with detailed results provided in the Supplementary Materials.

Comparison with different semantic reduction methods. To evaluate the impact of different semantic reduction methods on detection performance, we conduct comparative experiments following the protocol in Table 2. Our analysis includes: 1) baseline approaches (high-pass filtering and patch shuffling with sizes 8×8 and 2×2) implemented on ResNet-50, and 2) state-of-the-art variants - NPR (spectrum-based) (Tan et al. 2024b) and BSA (shuffle-based) (Zheng et al. 2024) - representing advanced improvements over these core ideas. For NPR, we directly report results from the original paper, while for BSA (which lacked GenImage benchmarks in its publication), we faithfully reproduced the method using the author's official codebase and report new evaluation results in Table 3. The results in Table 3 reveal several key insights: 1) High-pass filtering underperforms the baseline due to excessive information loss from discarding low-frequency components, while residual semantic biases persist in mid-high frequencies; 2) While patch shuffling with size 8×8 demonstrates marginal performance gains, the extreme case of 2×2 patches proves counterproductive - the excessive fragmentation prevents meaningful feature learning, as evidenced by the classifier's failure in detection. 3) Both BSA and NPR demonstrate stronger performance through their specialized designs (receptive field restriction and residual operations respectively); 4) Our proposed pixel mapping approach achieves significant performance gains, validating its dual advantage in effectively suppressing semantic bias while preserving discriminative generative artifacts. Notably, the random mapping variant exhibits slightly inferior performance compared to fixed mapping, suggesting that consistent transformation patterns benefit training

Method	AttGAN	BEGAN	CramerGAN	InfoMaxGAN	MMDGAN	RelGAN	S3GAN	SNGAN	STGAN	**Mean**
CNNDetection (Wang et al. 2020)	51.1 / 83.7	50.2 / 44.9	81.5 / 97.5	71.1 / 94.7	72.9 / 94.4	53.3 / 82.1	55.2 / 66.1	62.7 / 90.4	63.0 / 92.7	62.3 / 82.9
Frank (Frank et al. 2020)	65.0 / 74.4	39.4 / 39.9	31.0 / 36.0	41.1 / 41.0	38.4 / 40.5	69.2 / 96.2	69.7 / 81.9	48.4 / 47.9	25.4 / 34.0	47.5 / 54.7
Durall(Durall, Keuper, and Keuper 2020)	39.9 / 38.2	48.2 / 30.9	60.9 / 67.2	50.1 / 51.7	59.5/ 65.5	80.0 / 88.2	**87.3** / 97.0	54.8 / 58.9	62.1 / 72.5	60.3 / 63.3
Patchfor (Chai et al. 2020)	68.0 / 92.9	97.1 / **100.0**	97.8 / 99.9	93.6 / 98.2	97.9 / **100.0**	99.6 / **100.0**	66.8 / 68.1	97.6 / 99.8	92.7 / 99.8	90.1 / 95.4
F3Net (Qian et al. 2020)	85.2 / 94.8	87.1 / 97.5	89.5 / 99.8	67.1 / 83.1	73.7 / 99.6	98.8 / **100.0**	65.4 / 70.0	51.6 / 93.6	60.3 / 99.9	75.4 / 93.1
SelfBlend (Shiohara and Yamasaki 2022)	63.1 / 66.1	56.4 / 59.0	75.1 / 82.4	79.0 / 82.5	68.6 / 74.0	73.6 / 77.8	53.2 / 53.9	61.6 / 65.0	61.2 / 66.7	65.8 / 69.7
GANDetection (Mandelli et al. 2022)	57.4 / 75.1	67.9 / **100.0**	67.8 / 99.7	67.6 / 92.4	67.7 / 99.3	60.9 / 86.2	69.6 / 83.5	66.7 / 90.6	69.6 / 97.2	66.1 / 91.6
LGrad (Tan et al. 2023)	68.6 / 93.8	69.9 / 89.2	50.3 / 54.0	71.1 / 82.0	57.5 / 67.3	89.1 / 99.1	78.5 / 86.0	78.0 / 87.4	54.8 / 68.0	68.6 / 80.8
UnivFD (Ojha, Li, and Lee 2023)	78.5 / 98.3	72.0 / 98.9	77.6 / 99.8	77.6 / 98.9	77.6 / 99.7	78.2 / 98.7	85.2 / **98.1**	77.6 / 98.7	74.2 / 97.8	77.6 / **98.8**
NPR (Tan et al. 2024b)	83.0 / 96.2	99.0 / 99.8	98.7 / 99.0	94.5 / 98.3	98.6 / 99.0	99.6 / **100.0**	79.0 / 80.0	88.8 / 97.4	98.0 / **100.0**	93.2 / 96.6
Fixed-mapping	**99.6** / 99.9	99.1 / **100.0**	98.9 / **100.0**	**99.8** / **100.0**	**99.4** / **100.0**	99.7 / **100.0**	85.2 / 91.7	**99.1** / **100.0**	**99.9** / **100.0**	**97.9** / **98.8**
Random-mapping	99.4 / **100.0**	**99.9** / **100.0**	**99.4** / 99.7	99.0 / 99.9	99.2 / 99.8	**99.9** / **100.0**	77.4 / 83.1	98.3 / 99.8	**99.9** / **100.0**	96.9 / 98.0

Table 1: Cross-GAN performance (ACC./A.P.) analysis using the **Self-Synthesis** 9 GANs dataset, where **Bold** and underline values denote the top and runner-up results.

Method	Ref	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	**mAcc**
ResNet-50 (He et al. 2016)	CVPR2016	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S (Touvron et al. 2021)	ICML2021	55.6	99.9	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T (Liu et al. 2021)	ICCV2021	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot (Wang et al. 2020)	CVPR2020	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec (Zhang, Karaman, and Chang 2019)	WIFS2019	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net (Qian et al. 2020)	ECCV2020	50.1	99.9	**99.9**	49.9	50.0	99.9	49.9	49.9	68.7
GramNet (Liu, Qi, and Torr 2020)	CVPR2020	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
UnivFD (Ojha, Li, and Lee 2023)	CVPR2023	93.9	96.4	96.2	71.9	85.4	94.3	81.6	90.5	88.8
DIRE (Wang et al. 2023)	ICCV2023	50.4	**100.0**	**99.9**	52.3	67.2	**100.0**	50.1	50.0	71.2
FreqNet (Tan et al. 2024a)	AAAI2024	89.6	98.8	98.6	66.8	86.5	97.3	75.8	81.4	86.8
NPR (Tan et al. 2024b)	CVPR2024	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
FatFormer (Liu et al. 2024)	CVPR2024	92.7	**100.0**	**99.9**	75.9	88.0	99.9	**98.8**	55.8	88.9
DRCT (Chen et al. 2024)	ICML2024	91.5	95.0	94.4	79.4	89.2	94.7	90.0	81.7	89.5
C2P-CLIP (Tan et al. 2025)	AAAI2025	88.2	90.9	97.9	96.4	**99.0**	98.8	96.5	98.7	95.8
VIB-Net (Zhang et al. 2025)	CVPR2025	88.1	99.6	99.2	73.9	74.3	98.3	89.4	91.2	89.3
B-Free (Guillaro et al. 2025)	CVPR2025	89.0	93.5	93.5	92.7	83.3	94.1	90.8	94.6	91.4
Effort (Yan et al. 2024)	ICML2025	82.4	99.8	99.8	78.7	93.3	97.4	91.7	77.6	91.1
Fixed-mapping	Ours	**96.8**	98.9	98.8	**98.7**	98.4	98.2	**98.8**	**98.8**	**98.4**
Random-mapping	Ours	95.3	98.3	97.6	98.0	97.2	97.5	98.2	98.0	97.5

Table 2: Accuracy (Acc) comparison across models tested on **GenImage**, with training conducted on SDv1.4. **Bold** and underline indicate top and runner-up performance respectively.

Method	mAcc	mAP
ResNet-50	67.0	76.9
High-frequency filtering	64.4	70.2
Patch shuffling (size=8)	70.7	80.6
Patch shuffling (size=2)	50.5	51.0
BSA (Zheng et al. 2024)	74.7	85.6
NPR (Tan et al. 2024b)	88.6	93.7
Fixed-mapping	**98.4**	**99.8**
Random-mapping	97.5	99.6

Table 3: Comparison with different semantic reduction methods tested on **GenImage**, with training conducted on SDv1.4. **Bold** and underline indicate top and runner-up performance respectively.

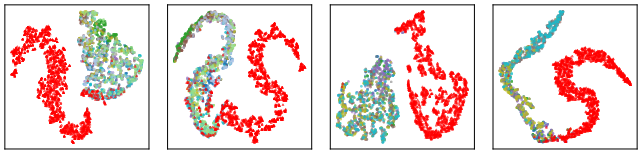
convergence despite equivalent semantic suppression.

Qualitative Analysis

We present visual comparisons of different semantic reduction approaches.

Visualization of t-SNE Results. We present the t-SNE results of the fixed pixel-level mapping method trained using the aforementioned settings (ProGAN with 4 classes and SDv1.4) on both the GAN-generated and Diffusion-generated image sets in Fig. 5. We also reproduce the state-of-the-art method NPR using the same training and test setting. (a) and (b) utilize the Self-Synthesis dataset, which includes all GAN models, whereas (c) and (d) employ the GenImage dataset, comprising seven Diffusion models. The red triangles represent the features of real images, while the others correspond to the features of GAN-generated and Diffusion-generated images from the two datasets, respectively. The results demonstrate that the features extracted by mapping method effectively separate generated images from real images for both GAN and Diffusion models. This indicates the superior generalization capability of our approach.

Analysis of Anomalies in Mapped Images. Our pixel-level mapping analysis reveals distinctive artifacts in generated images, as demonstrated in Fig. 6. For clearer visualization, we display results from the low-frequency components of high-resolution images. Using high-resolution datasets (Midjourney and RAISE (Dang-Nguyen et al. 2015)), we randomly crop low-frequency smooth regions for examina-



(a) (b) (c) (d)

Figure 5: (a) Pixel-level mapping GAN model t-SNE results. (b) NPR GAN model t-SNE results. (c) Pixel-level mapping Diffusion model t-SNE results. (d) NPR Diffusion model t-SNE results.

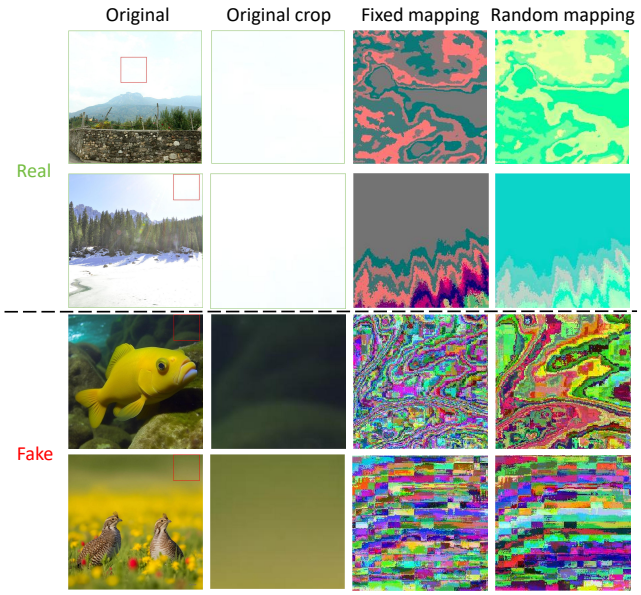
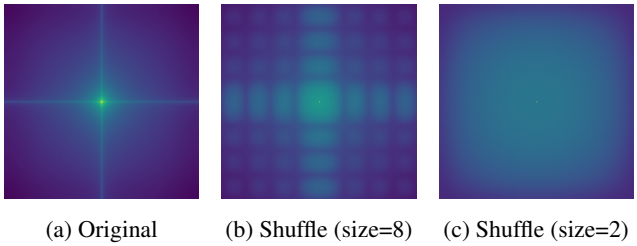


Figure 6: Visualization results of anomalies in mapped images.



(a) Original (b) Shuffle (size=8) (c) Shuffle (size=2)

(d) NPR (e) Fixed mapping (f) Random mapping

Figure 7: Frequency spectra of different semantic-reduction methods.

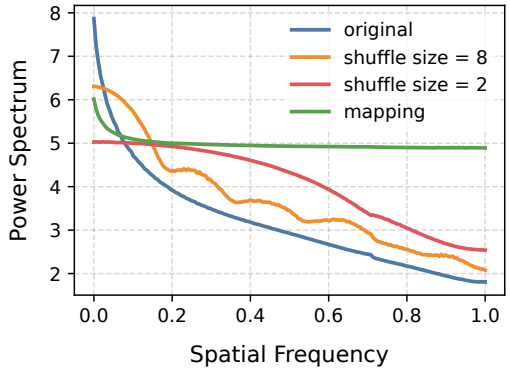


Figure 8: Power spectral representation of images under different processing methods.

tion. The pixel-level mapping exposes abnormal checkerboard noise patterns which may originate from the upsampling process, while natural images maintain smooth transitions. Compared to original images, the mapped versions exhibit significantly reduced semantic features while amplifying anomalous traces, validating our method's capability to isolate generation-specific artifacts.

Frequency Spectra Comparisons. We compare the spectral performance of different semantic-reduction methods by averaging 1,000 spectral images as shown in Figure 7, while the shuffle method narrows the low/high-frequency gap (particularly with patch size=2), it still exhibits clear energy drops at spectral boundaries. The NPR approach fails to effectively reduce the frequency gap through residual operation, whereas our mapping method significantly equalizes the energy distribution, enhancing high-frequency features for classifier training. We convert the 2D spectral data into 1D azimuthal integral spectrum (Durall, Keuper, and Keuper 2020), which illustrates the spectral energy distribution across frequencies, with the horizontal axis ranging from low to high frequencies and the vertical axis representing spectral energy, as shown in Figure 8. It reveals that although shuffling attenuates low-frequency power, it fails to enhance high-frequency components. Instead, the proposed method effectively narrows the spectral power gap between low- and high-frequency components and forces classifiers to prioritize high-frequency artifacts during training.

Conclusion

This paper presents a pixel-level mapping method that addresses classifier reliance on semantic bias by suppressing low-frequency features while enhancing high-frequency artifacts, significantly improving cross-model and cross-distribution detection generalization. The computationally efficient preprocessing step applies pixel-value transformations before classification to amplify inter-pixel differences and disrupt low-frequency biases. Extensive experiments on multiple benchmarks demonstrate the remarkable performance of the proposed method in generalization scenarios, offering a new direction for semantic bias reduction in synthetic image detection.

Acknowledgments

This work was supported by the Innovation Funding of Institute of Computing Technology, Chinese Academy of Sciences under Grant No. E561090 and E561160.

References

Bammey, Q. 2023. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5: 1–9.

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

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

Chen, B.; Zeng, J.; Yang, J.; and Yang, R. 2024. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*.

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 *CVPR*, 8789–8797.

Chu, B.; Xu, X.; Wang, X.; Zhang, Y.; You, W.; and Zhou, L. 2024. FIRE: Robust Detection of Diffusion-Generated Images via Frequency-Guided Reconstruction Error. *arXiv preprint arXiv:2412.07140*.

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.

Dang-Nguyen, D.-T.; Pasquini, C.; Conotter, V.; and Boato, G. 2015. Raise: A raw images dataset for digital image forensics. In *MMSys*, 219–224.

Durall, R.; Keuper, M.; and Keuper, J. 2020. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 7890–7899.

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. PMLR.

Fu, X.; Yan, Z.; Yao, T.; Chen, S.; and Li, X. 2025. Exploring Unbiased Deepfake Detection via Token-Level Shuffling and Mixing. *arXiv preprint arXiv:2501.04376*.

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

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.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

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*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*, 4401–4410.

Khan, S. A.; and Dang-Nguyen, D.-T. 2024. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *ICMR*, 1006–1015.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, 10770–10780.

Liu, S.; Lian, Z.; Gu, S.; and Xiao, L. 2022. Block shuffling learning for deepfake detection. *arXiv preprint arXiv:2202.02819*.

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 *ICCV*, 10012–10022.

Liu, Z.; Qi, X.; and Torr, P. H. 2020. Global texture enhancement for fake face detection in the wild. In *CVPR*, 8060–8069.

Luo, Y.; Du, J.; Yan, K.; and Ding, S. 2024. LaREˆ 2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection. In *CVPR*, 17006–17015.

Mandelli, S.; Bonettini, N.; Bestagini, P.; and Tubaro, S. 2022. Detecting gan-generated images by orthogonal training of multiple cnns. In *ICIP*, 3091–3095. IEEE.

Masci, J.; Meier, U.; Cireşan, D.; and Schmidhuber, J. 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In *ICANN*, 52–59. Springer.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 24480–24489.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. An imperative style, high-performance deep learning library. *NeurIPS*, 32: 8026.

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

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

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

Salah, R.; Vincent, P.; Muller, X.; Gloro, X.; and Bengio, Y. 2011. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*, 833–840.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In *CVPR*, 18720–18729.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

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 *AAAI*, volume 39, 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 *AAAI*, volume 38, 5052–5060.

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 *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, L.; Shen, W.; Zhou, Z.; Chen, Y.; and Zhang, Q. 2022. Defects of convolutional decoder networks in frequency representation. *arXiv preprint arXiv:2210.09020*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, 10347–10357. PMLR.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*, 1096–1103.

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. 2023. Dire for diffusion-generated image detection. In *ICCV*, 22445–22455.

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 *ICML*.

Yan, Z.; Wang, J.; Wang, Z.; Jin, P.; Zhang, K.-Y.; Chen, S.; Yao, T.; Ding, S.; Wu, B.; and Yuan, L. 2024. Effort: Efficient Orthogonal Modeling for Generalizable AI-Generated Image Detection. *arXiv preprint arXiv:2411.15633*.

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

Zhang, H.; He, Q.; Bi, X.; Li, W.; Liu, B.; and Xiao, B. 2025. Towards Universal AI-Generated Image Detection by Variational Information Bottleneck Network. In *CVPR*, 23828–23837.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019. Detecting and simulating artifacts in gan fake images. In *WIFS*, 1–6. IEEE.

Zheng, C.; Lin, C.; Zhao, Z.; Wang, H.; Guo, X.; Liu, S.; and Shen, C. 2024. Breaking Semantic Artifacts for Generalized AI-generated Image Detection. *NeurIPS*, 37: 59570–59596.

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. *NeurIPS*, 36: 77771–77782.