Aggregating Diverse Cue Experts for AI-Generated Image Detection

Lei Tan[1], **Shuwei Li**[1], **Mohan Kankanhalli**[1], **Robby T. Tan**[1, 2]

[1]National University of Singapore,
[2]ASUS Intelligent Cloud Services (AICS)
lei.tan@nus.edu.sg, shuwei@u.nus.edu, mohan@comp.nus.edu.sg, robby.tan@nus.edu.sg

Abstract

The rapid emergence of image synthesis models poses challenges to the generalization of AI-generated image detectors. However, existing methods often rely on model-specific features, leading to overfitting and poor generalization. In this paper, we introduce the Multi-Cue Aggregation Network (MCAN), a novel framework that integrates different yet complementary cues in a unified network. MCAN employs a mixture-of-encoders adapter to dynamically process these cues, enabling more adaptive and robust feature representation. Our cues include the input image itself, which represents the overall content, and high-frequency components that emphasize edge details. Additionally, we introduce a Chromatic Inconsistency (CI) cue, which normalizes intensity values and captures noise information introduced during the image acquisition process in real images, making these noise patterns more distinguishable from those in AI-generated content. Unlike prior methods, MCAN's novelty lies in its unified multi-cue aggregation framework, which integrates spatial, frequency-domain, and chromaticity-based information for enhanced representation learning. These cues are intrinsically more indicative of real images, enhancing cross-model generalization. Extensive experiments on the **GenImage**, **Chameleon**, and **UniversalFakeDetect** benchmark validate the state-of-the-art performance of MCAN. In the GenImage dataset, MCAN outperforms the best state-of-the-art method by up to **7.4%** in average **ACC** across eight different image generators.

Introduction

Distinguishing synthetic images from real ones is becoming increasingly challenging due to the rapid advancements in generative models. As new architectures emerge, they produce highly realistic images with fewer detectable artifacts, making traditional detection methods less effective. While many methods have been proposed (e.g., (Tan et al. 2024b; Sarkar et al. 2024; Cozzolino et al. 2025; Wang et al. 2023; Luo et al. 2024; Ricker, Lukovnikov, and Fischer 2024)), most rely on model-specific features, leading to overfitting and poor generalization.

A common approach in existing methods is to enhance the distinction between real and generated images using reconstruction error, which quantifies the discrepancy between an input image and its reconstructed version produced by a diffusion model (Wang et al. 2023; Luo et al. 2024; Ricker, Lukovnikov, and Fischer 2024). While effective for many generative models, reconstruction error is highly dependent on the specific model used for reconstruction, limiting its generalizability. Another approach involves leveraging high-frequency information (Liu et al. 2024; Tan et al. 2024a), which has demonstrated broader applicability across both GANs and diffusion models. However, high-frequency representations discard most semantic information, leading to failure cases where the lack of contextual details results in incorrect classifications.

The rise of large-scale pre-trained models has greatly ad-

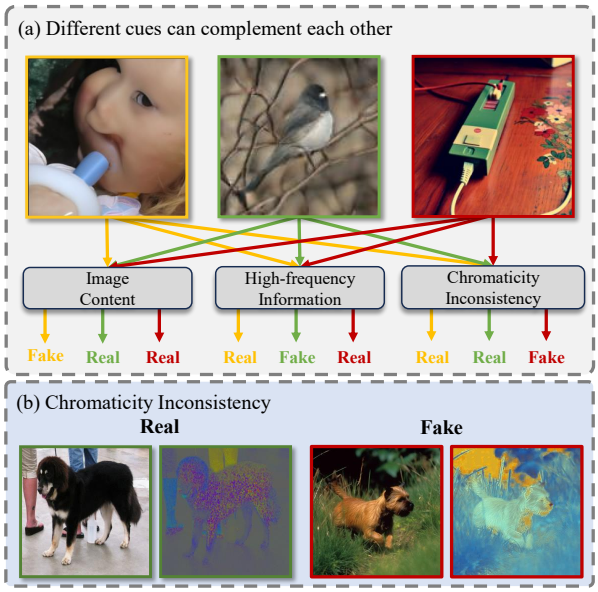


Figure 1: (a) Motivation for MCAN: In the top panel, yellow, green, and red represent the three generated images; Using multiple cues: image, frequency, and chromaticity enhances AI-generated image detection by leveraging the cues' complementary strengths. (b) Motivation for Chromaticity Inconsistency: In the bottom panel, real images show chromaticity inconsistencies due to noise, while fake images appear smoother with uniform chromaticity.

vanced AI-generated image detection. UniFD (Ojha, Li, and Lee 2023) pioneered using image features from a frozen CLIP model to train a classifier, achieving strong detection performance. RBAD (Cozzolino et al. 2024) demonstrated that CLIP can reliably detect images, even with limited generated samples. Fatformer (Liu et al. 2024) enhanced detection by fine-tuning CLIP with frequency-sensitive adapters. However, these methods rely on a single feature type, limiting their generalization ability.

To address the problem, as illustrated in Figure 1(a), we analyze three AI-generated images across classifiers, each using a different cue, and observe that these cues exhibit complementary properties in detecting synthetic content. For instance, AI-generated images missed by high-frequency representations often appear simple from an image content perspective. Hence, in this paper, we propose integrating multiple cues into a unified network. Inspired by multi-modal learning, we treat these cues as multi-modal inputs and introduce the Multi-Cue Aggregation Network (MCAN). MCAN employs a Mixture-of-Encoder Adapter (MoEA), incorporating the mixture-of-experts concept to dynamically adapt diverse cues within a unified feature space. Specifically, MoEA adopts a dynamic strategy that allows each image token to flexibly integrate different adapter encoders while utilizing a shared adapter decoder to project the fused cue representations.

Beyond the widely used original images, which primarily capture content, and high-frequency representations, which emphasize edge details, we introduce Chromaticity Inconsistency (CI), a novel representation designed to highlight noise differences between real and AI-generated images. Noise components are challenging to extract directly from pixel differences due to variations in lighting intensity. To mitigate this, we apply a chromaticity transformation to minimize the influence of illumination. Theoretically, chromaticity should remain consistent across surfaces with uniform material and color temperature. However, real-world factors such as camera noise introduce inconsistencies at the pixel level in real images. As shown in Figure 1(b), AI-generated images often appear smoother than real images due to the absence of such noise. Leveraging this distinction, we integrate CI with original and high-frequency representations to enhance the effectiveness of our MCAN.

To sum up, our key contributions are as follows:

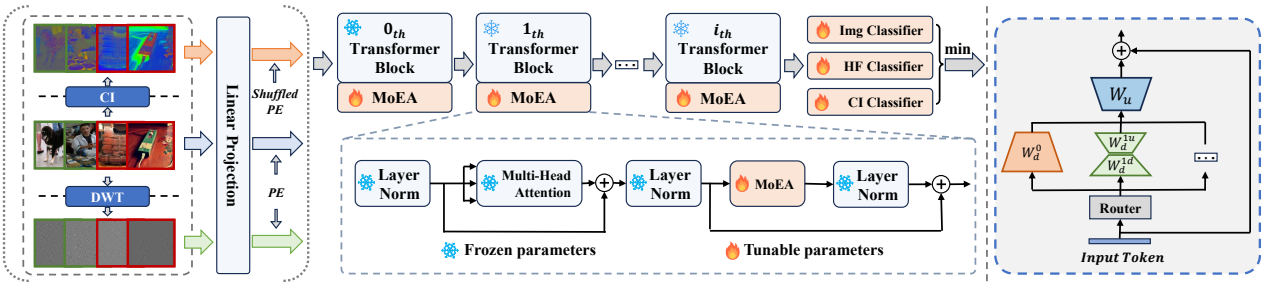
- **Chromaticity Inconsistency (CI)**: A new representation that mitigates lighting intensity effects through chromaticity-based transformation, highlighting noise differences between real and generated images.
- **MCAN**: A novel framework that enhances AI-generated image detection by dynamically integrating spatial, frequency, and chromaticity through a unified multi-cue aggregation strategy, setting it apart from existing methods.
- **Extensive Experiments**: In-depth evaluation on the GenImage, Chameleon, and UniversalFakeDetect benchmark shows MCAN superior performance. In the GenImage dataset, MCAN outperforms the state-of-the-art method by up to 10.6% in average Accuracy.

Related Work

AI-generated image detection has long been a focus in the computer vision community, with increasing attention as generative models advance. Early approaches relied on hand-crafted cues such as chromatic aberration (Mayer and Stamm 2018), color (McCloskey and Albright 2019), saturation (McCloskey and Albright 2019), blending (Li et al. 2020), and reflections (O'brien and Farid 2012). These methods struggle with generalization as generation techniques evolve. Leveraging deep learning, neural networks have been applied to detect AI-generated images (Liu, Qi, and Torr 2020; Wang et al. 2020; Marra et al. 2018). CNNSpot (Wang et al. 2020) uses a standard CNN-based image classifier to achieve strong performance in detecting GAN-generated images. FreDect (Frank et al. 2020) reveals that GAN images contain unique artifacts identifiable in the frequency domain and enhances detection by leveraging frequency-based representations. MoE-FFD (Kong et al. 2025a) is an SOTA face forgery detection method that substantially improves model generalizability and reduces the parameter overhead. PIM (Kong et al. 2025b) effectively extracts inherent pixel-inconsistency forgery fingerprints and achieves SOTA performance in both generalization and robustness. With the rise of diffusion models, detecting diffusion-generated images has gained traction. Ojha et al. (Ojha, Li, and Lee 2023) introduced a universal fake image detector using pre-trained vision-language models. Fatformer (Liu et al. 2024) improves detection by fine-tuning CLIP with frequency-sensitive adapters and language-guided training. DIRE (Wang et al. 2023) observed that diffusion-generated images are more easily reconstructed by diffusion models than real images. Lare^2 (Luo et al. 2024) improved DIRE in both efficiency and effectiveness. NPR (Tan et al. 2024b) refined upsampling in generative models, amplifying artifacts for better detection. ZED (Cozzolino et al. 2025) identified discrepancies between real and generated images through lossless encoder coding differences, capturing them via a multi-resolution structure. While these methods show promise by targeting specific features, the rapid emergence of diverse generative networks continues to challenge their generalizability.

Proposed Method

Figure 2 illustrates the architecture of our proposed Multi-Cues Aggregating Network (MCAN). MCAN extends beyond image and high-frequency information (Tan et al. 2024a; Liu et al. 2024) by incorporating chromaticity inconsistency as an additional cue. It first applies a Discrete Wavelet Transform (DWT) to extract high-frequency features and uses a chromaticity inconsistency transformation to highlight noise discrepancies. A pre-trained CLIP ViT-B/16 model, with frozen parameters, serves as the backbone, while the Mixture-of-Encoder Adapter (MOEA) enables fine-tuning. To enhance detection, separate classifiers process multiple cues, and the final classification is determined by aggregating their minimal outputs.



(a) The overall structure of Multi-cue Aggregation Network

(b) Mixture-of-Encoder Adapter

Figure 2: Overall architecture: MCAN combines image representation, high-frequency representation, and the novel chromaticity inconsistency as three distinct cues. To effectively integrate these cues, MCAN uses a mixture of encoder adapters that adapt efficiently to each cue's representation.

Chromaticity Inconsistency

Noise introduced during image acquisition has been explored for detecting image manipulation and synthesis (Mahdian and Saic 2009; Lyu, Pan, and Zhang 2014; Zhou et al. 2018). However, due to the inherently low magnitude of noise signals, generative models struggle to capture them using conventional losses, like L_1 or L_2 loss. A key obstacle is variations in lighting intensity, which often dominate local discrepancies, making it difficult to identify noise from the original image. To address this, we leverage chromaticity information instead of raw images, reducing the impact of light intensity.

Following the common image formation model for Lambertian surfaces (e.g. (Finlayson et al. 2005)), we express the response of each pixel (x, y) in the camera sensor under a light source across different wavelengths λ as:

$$\rho_j(x,y) = \sigma(x,y) \int_{\lambda_j} E_j(\lambda,x,y) S(\lambda,x,y) Q(\lambda) d\lambda, \quad (1)$$

where λ represents the wavelength, and $E(\lambda)$ and $S(\lambda)$ denote the spectral power distribution of the incident light and the surface spectral reflectance, respectively. $Q(\lambda)$ corresponds to the spectral sensitivity of the camera sensor. The subscript $j \in \{r, g, b\}$ indicates the channel in the spectral domain. $\sigma(x, y)$ represents the Lambertian reflection term, computed as the dot product between the surface normal and the illumination direction.

Following the Eq. (1), we leverage Wien's approximation to Planck's law (Wyszecki and Stiles 2000) to parameterize the illuminant SPD by its light source color temperature T:

$$E(\lambda, T) = I c_1 \lambda^{-5} e^{\frac{-c_2}{T\lambda}}, \quad (2)$$

where c_1 and c_2 are constants, and I is a variable controlling the overall intensity of the light. With Eq. (2), the Eq. (1) can be rewrite as:

$$\rho_j(x,y) = c_1 \sigma(x,y) I(x,y) \int_{\lambda_j} \lambda^{-5} e^{\frac{-c_2}{T(x,y)\lambda}} F(\lambda,x,y) d\lambda, \quad (3)$$

with $F(\lambda, x, y) = S(\lambda, x, y)Q(\lambda)$.

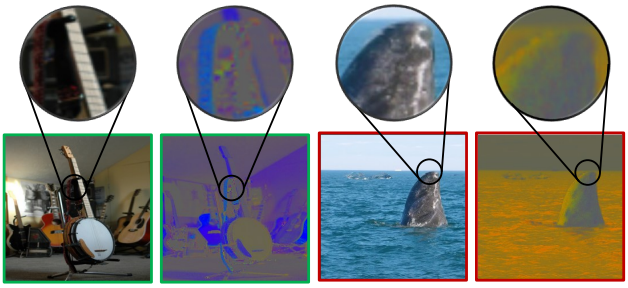


Figure 3: Visualization of Chromaticity Inconsistency (CI): Real images, affected by noise, show weaker consistency in CI images compared to generated images.

We can then eliminate the light intensity I by the chromaticity definition as:

$$\frac{\rho_j(x,y)}{\rho_k(x,y)} = \frac{\int_{\lambda_j} \lambda^{-5} e^{\frac{-c_2}{T(x,y)\lambda}} S(\lambda,x,y) Q_j(\lambda) d\lambda}{\int_{\lambda_k} \lambda^{-5} e^{\frac{-c_2}{T(x,y)\lambda}} S(\lambda,x,y) Q_k(\lambda) d\lambda}, \quad (4)$$

where j and k refer to two different color channels of the image.

Since $Q(\lambda)$ and $S(\lambda)$ are intrinsic functions defined by the spectral sensitivity of the camera sensor and the surface material's reflectance, respectively, we make a mild assumption that local regions share uniform illumination conditions. This implies that the color temperature T remains consistent within these regions, allowing the chromaticity information to be treated as constant locally. Consequently, for pixels corresponding to surfaces of the same material, denoted as (x_a, y_a) and (x_b, y_b), this relationship can be expressed as:

$$\frac{\rho_j(x_a,y_a)}{\rho_k(x_a,y_a)} = \frac{\rho_j(x_b,y_b)}{\rho_k(x_b,y_b)}. \quad (5)$$

Eq. (5) defines the ideal pixel relationship in a real image. However, when accounting for noise introduced during the acquisition process, the formulation becomes:

$$\frac{\rho_j(x_a,y_a) + N_j(x_a,y_a)}{\rho_k(x_a,y_a) + N_k(x_a,y_a)} \neq \frac{\rho_j(x_b,y_b) + N_j(x_b,y_b)}{\rho_k(x_b,y_b) + N_k(x_b,y_b)}, \quad (6)$$

where, N_j and N_k represent the random noise components in different color channels. From Eq. (6), the noise component can be extracted by computing the difference between pixels on the same surface. Leveraging this insight, we define the Chromaticity Inconsistency (CI) transformation as:

$$I_{ci} = [e^{-\frac{\rho_r}{\rho_g}}, e^{-\frac{\rho_g}{\rho_b}}, e^{-\frac{\rho_b}{\rho_r}}]. \tag{7}$$

To assess whether Eq.(7) effectively highlights noise differences between real and generated images, we visualize samples from the GenImage (Zhu et al. 2024) dataset in Figure 3. Due to the inherent noise introduced during image acquisition, real images exhibit lower consistency on the same material surface in CI images compared to generated ones.

Multi-cue Aggregation Network

Existing AI-generated image detection methods primarily focus on specific visual cues, such as image representations (Ojha, Li, and Lee 2023), reconstruction errors (Wang et al. 2023), and others (Sarkar et al. 2024). While some approaches integrate multiple cues (Liu et al. 2024; Luo et al. 2024), they often assign unequal importance to each, leading to suboptimal optimization and failing to fully leverage their complementary properties. To overcome this limitation, we take a multi-modal approach, incorporating image representation, high-frequency representation, and our proposed chromaticity inconsistency (Liu et al. 2024; Luo et al. 2024) to build a robust detection model.

A crucial yet often overlooked factor is the difference in domain shifts between real and generated images across datasets. Real images exhibit relatively minor quality variations, whereas generated images undergo more significant shifts due to discrepancies in the training data and the structure of generative models. This discrepancy increases the risk of false positives, where generated images are misclassified as real. To improve generalization, we aggregate multiple complementary cues. To this end, we introduce the Multi-Cue Aggregation Network (MCAN), which treats all cues as multi-modal inputs and encodes them within a unified framework. The following sections provide a detailed overview of MCAN's architectural design.

Position Embedding Shuffle As shown in Figure 3, while CI amplifies noise, it also preserves some original image content, allowing the network to extract content information to some extent. Since MCAN already leverages original images for learning representations, we aim to minimize the influence of image content in CI representations.

In Vision Transformers (ViTs), positional embeddings determine the spatial locations of tokens. By perturbing the order of these embeddings, we disrupt the spatial structure of CI images, reducing content integrity while preserving noise characteristics. Specifically, for CI images, the shuffled positional embeddings are defined as:

$$p_{ci} = [p_{\text{cls}}; \underbrace{p_{s1}; p_{s2}; \ldots; p_{\text{sL}}}_{\text{Shuffle Term}}], \tag{8}$$

where L is the length of image tokens, p_{si} refers to the shuffled position embedding in i_{th} position.

Mixture-of-encoder Adapter In MCAN, different cues serve as distinct modalities. Extracting features through a shared network across these diverse modalities can lead to suboptimal performance (Gou et al. 2023). To address this, we adopt the mixture-of-experts paradigm (Riquelme et al. 2021; Chi et al. 2022) and introduce the Mixture-of-Encoder Adapter (MoEA) to effectively capture and integrate the unique characteristics of each cue. This design enhances feature discrimination and representation, ultimately improving detection performance.

Specifically, as illustrated in Figure 2 (b), given an input token $z_i \in \mathbb{R}^d$ and a set of N encoder experts, we use a cosine-based router to normalize the token feature, ensuring stable routing scores. The routing function is defined as:

$$g(z_i) = \sigma(\frac{W_1 z_i W_e}{\tau \left\| W_1 z_i \right\| \left\| W_e \right\|}), \tag{9}$$

where $\sigma(\cdot)$ denotes the softmax function, $\|\cdot\|$ represents the L_2 normalization, and τ is the learnable temperature scalar. $W_1 \in \mathbb{R}^{d \times d_e}$ is the weight matrix of a linear projection, which maps the input token into a lower-dimensional subspace, while $W_e \in \mathbb{R}^{d_e \times N}$ serves as the expert embedding, transforming the feature into the final scoring distribution.

Leveraging the additive properties of different encoders, we integrate them through the scoring distribution, enabling the construction of a mixed encoder without separately computing results for each encoder. This significantly reduces computational cost during MoEA's inference stage. Consequently, the weight matrix of the mixed encoder, denoted as W_d, which encodes the input token into a c-dimensional subspace, is given by:

$$\begin{aligned} W_d = (g(z_i)_0 W_d^0 + g(z_i)_1 W_d^1 + \cdots + g(z_i)_N W_d^N), \\ \text{if i} > 0 : W_d^i = W_d^{id} W_d^{iu}, \end{aligned} \tag{10}$$

where, $W_d^0 \in \mathbb{R}^{d \times c}$, $W_d^{id} \in \mathbb{R}^{d \times \frac{c}{i}}$, and $W_d^{iu} \in \mathbb{R}^{\frac{c}{i} \times c}$.

Using an identical structure across all experts can lead to homogenized representations, limiting the diversity of learned features, especially when faced with heterogeneous inputs. To address this, we introduce W_d^{id} and W_d^{iu} to enhance expert diversity, ensuring that each expert contributes unique information to the model. This design increases the variability of expert outputs, leading to more effective feature extraction. Furthermore, since this transformation remains re-parameterizable, only a single mixed encoder needs to be maintained during inference, significantly reducing computational overhead. The overall processing of MoEA is formulated as:

$$M(z_i) = z_i W_u W_d + z_i. \tag{11}$$

Optimization

To optimize the training of MCAN, alongside the binary cross entropy losses $\mathcal{L}_{\text{img}}$, $\mathcal{L}_{ci}$, $\mathcal{L}_{hf}$ for the classification, we introduce two additional loss functions according to the (Riquelme et al. 2021; Li et al. 2023): importance loss ($\mathcal{L}_{\text{imp}}$) and entropy loss ($\mathcal{L}_{\text{ent}}$) to ensure both balanced expert assignment and effective specialization in MoEA. $\mathcal{L}_{\text{imp}}$ promotes a balanced usage of different experts, while the $\mathcal{L}_{ent}$ encourages each token to select a specific encoder expert in each layer. Therefore, the overall loss can be given as:

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \mathcal{L}_{ci} + \mathcal{L}_{hf} + \mathcal{L}_{\text{imp}} + \mathcal{L}_{\text{ent}}. \tag{12}$$

Method	Venue	Testing Subset								Avg
		Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Accuracy(%)
ResNet-50 (He et al. 2016)	CVPR'16	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S (Touvron et al. 2021)	ICML'21	55.6	99.9	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T (Liu et al. 2021)	ICCV'21	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot (Wang et al. 2020)	CVPR'20	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec (Zhang, Karaman, and Chang 2019)	WIFS'19	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net (Qian et al. 2020)	ECCV'20	50.1	99.9	99.9	49.9	50.0	99.9	49.9	49.9	68.7
GramNet (Liu, Qi, and Torr 2020)	CVPR'20	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
DIRE (Wang et al. 2023)	ICCV'23	65.8	99.7	99.7	54.5	58.1	99.4	54.3	49.8	72.7
LaRE^2 (Luo et al. 2024)	CVPR'24	74.0	**100.0**	**99.9**	61.7	88.5	**100.0**	**97.2**	68.7	86.2
LaRE^2-ViT† (Luo et al. 2024)	CVPR'24	72.5	76.1	76.0	67.9	80.6	73.7	69.0	70.5	73.3
FatFormer† (Liu et al. 2024)	CVPR'24	93.1	97.0	97.2	82.0	95.0	95.8	88.8	49.9	87.4
NPR (Tan et al. 2024b)	CVPR'24	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
DRCT (Chen et al. 2024)	ICML'24	91.5	95.0	94.4	79.4	89.2	94.7	90.0	81.7	89.5
VIB-Net (Zhang et al. 2025)	CVPR'25	88.1	99.6	99.2	73.9	74.3	98.3	89.4	-	88.9
AIDE (Yan et al. 2025)	ICLR'25	79.4	99.7	99.8	78.5	91.8	98.7	80.3	66.9	86.9
MCAN	-	**94.3**	98.8	98.5	**90.2**	**98.6**	**98.8**	**96.8**	**98.8**	**96.9**

Table 1: Comparisons on between the proposed MCAN and state-of-the-art methods on the GenImage testing set. The symbol †indicates our reproduction using the source code on a CLIP-ViT-B/16 version.

Training Setting	Methods									
	CNNSpot	FreDect	Fusing	LNP	UniFD	DIRE	Patchcraft	NPR	AIDE	MCAN
ProGAN	56.94	55.62	56.98	57.11	57.22	58.19	53.76	57.29	58.37	**60.81**
SDV1.4	60.11	56.86	57.07	55.63	55.62	59.71	56.32	58.13	62.60	**69.61**

Table 2: Comparisons between the proposed MCAN and state-of-the-art methods on the Chameleon dataset.

In summary, MCAN adapts to diverse cues, enhancing AI-generated image detection and generalization. Chromaticity Inconsistency (CI) highlights noise discrepancies between real and generated images, critically complementing both image and high-frequency features.

Experimental Results

Datasets and Evaluation Metrics We evaluate our method on three challenging benchmarks that comprehensively cover various real and synthetic image domains: GenImage (Zhu et al. 2024), Chameleon (Yan et al. 2025), and UniversalFakeDetect (Ojha, Li, and Lee 2023). For training and evaluation, following the official protocol and previous works (Zhu et al. 2024; Luo et al. 2024; Yan et al. 2025; Tan et al. 2024b), we use Accuracy (ACC) as the evaluation metric.

Implementation Details We implement our MCAN using PyTorch and conduct all experiments on an RTX H100 GPU. The mini-batch size is set to 64, with each batch containing an equal number of randomly sampled real and generated images. The learning rate is fixed at 1×10^{-4} during the training. We employ the CLIP-based ViT-B/16 (Radford et al. 2021) as the backbone. All images, in both training and testing phases, are resized to 224×224 without applying additional data augmentation strategies.

Comparison with State-of-the-Arts

We compare the proposed MCAN with a broad set of state-of-the-art (SOTA) methods across three challenging benchmarks, including GenImage (Zhu et al. 2024), Chameleon (Yan et al. 2025), and UniversalFakeDetect (Ojha, Li, and Lee 2023).

Table 1 summarizes the average performance of different methods in the GenImage dataset. To ensure a fair comparison, we re-implement LaRE^2 with the same backbone as MCAN (CLIP ViT-B/16), as the original model uses CLIP-RN50. As shown in Table 1, MCAN achieves a leading average accuracy of 96.9%, surpassing DRCT (Chen et al. 2024) by a large margin of 7.4%. Moreover, Table 1 reveals that most existing methods suffer notable performance drops under cross-model protocols due to overfitting to single cues. In contrast, MCAN integrates multiple complementary visual cues, leading to consistently strong performance across all subsets and superior generalization.

We further evaluate MCAN on the challenging Chameleon dataset by comparing it against a comprehensive set of SOTA detectors, including CNNSpot (Wang et al. 2020), FreDect (Frank et al. 2020), Fusing (Ju et al. 2022), LNP (Liu et al. 2022), UniFD (Ojha, Li, and Lee 2023), DIRE (Wang et al. 2023), Patchcraft (Zhong et al. 2023), NPR (Tan et al. 2024b), AIDE (Yan et al. 2025). As shown in Table 2, MCAN outperforms all SOTAs. On the ProGAN training protocol, it achieves 60.81% accuracy, surpassing AIDE (58.37%) by 2.44%. The performance gap widens on the SDV1.4 training protocol, where MCAN reaches 69.61%, outperforming AIDE (62.60%) by 7.01%. These results underscore MCAN's capability in detecting high-quality synthetic content and its robustness across different generative paradigms.

On the UniversalFakeDetect dataset, MCAN is evaluated against methods specifically designed for generaliza-

Methods	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAcc
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
CNNSpot	100.0	85.2	70.2	85.7	79.0	91.7	53.5	66.7	48.7	86.3	86.3	60.1	54.0	55.0	54.1	60.8	63.8	65.7	55.6	69.6
Patchfor	75.0	69.0	68.5	79.2	64.2	63.9	75.5	75.1	75.3	72.3	55.3	67.4	76.5	76.1	75.8	74.8	73.3	68.5	67.9	71.2
F3Net	99.4	76.4	65.3	92.6	58.1	100.0	63.5	54.2	47.3	51.5	51.5	69.2	68.2	75.4	68.8	81.7	83.3	83.1	66.3	71.3
UniFD	100.0	98.5	94.5	82.0	99.5	97.0	66.6	63.0	57.5	59.5	72.0	70.0	94.2	73.8	94.4	79.1	79.9	78.1	86.8	81.4
LGrad	99.8	85.4	82.9	94.8	72.5	99.6	58.0	62.5	50.0	50.7	50.8	77.5	94.2	95.9	94.8	87.4	90.7	89.6	88.4	80.3
FreqNet	97.9	95.8	90.5	97.6	90.2	93.4	97.4	88.9	59.0	71.9	67.4	86.7	84.6	99.6	65.6	85.7	97.4	88.2	59.1	85.1
NPR	99.8	95.0	87.6	96.2	86.6	99.8	76.9	66.9	98.6	50.0	50.0	84.6	97.7	98.0	98.2	96.3	97.2	97.4	87.2	87.6
FatFormer	99.9	99.3	99.5	97.2	99.4	99.8	93.2	81.1	68.0	69.5	69.5	76.0	98.6	94.9	98.7	94.4	94.7	94.2	98.8	90.9
MCAN	100.0	99.6	98.8	97.0	99.3	100.0	94.0	86.7	68.9	87.3	87.3	70.9	98.8	94.2	98.5	97.4	97.2	97.1	98.8	93.3

Table 3: Comparisons between the proposed MCAN and state-of-the-art methods on the UniversalFakeDetect dataset.

Setting	Testing Subset								Avg ACC.(%)
	Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	
Img only	88.1	**99.6**	**99.5**	75.6	93.2	98.7	89.8	51.4	87.0
HF only	90.2	96.5	96.5	83.5	95.6	96.2	93.6	96.7	93.6
CI only	87.5	96.9	96.7	81.7	85.2	96.7	94.1	51.7	86.3
CI-Shuffled Only	89.1	97.2	97.0	86.7	86.5	96.9	95.0	70.0	89.8
Naïve Combination(Img, HF, CI-Shuffled)	92.8	97.6	97.4	85.6	96.9	97.2	94.5	97.5	95.9
MCAN-(Img, HF)	93.7	97.4	96.9	88.6	96.4	97.0	94.8	98.2	95.4
MCAN(Img, CI-Shuffled)	93.3	98.6	98.5	88.8	97.5	98.4	96.5	69.5	92.6
MCAN(HF, CI-Shuffled)	94.1	98.2	98.3	89.8	95.5	97.3	96.3	98.8	96.0
MCAN-(Img, HF, CI-Shuffled)	**94.3**	98.8	98.5	**90.2**	**98.6**	**98.8**	**96.8**	**98.8**	**96.9**

Table 4: Ablation study of the different components in MCAN: 'Img', 'CI', and 'HF' refer to the image cue, chromaticity inconsistency cue, and high-frequency cue, respectively.

tion across multiple generative models. As shown in Tables 3, MCAN consistently outperforms prior approaches including CNNSpot (Wang et al. 2020), Patchfor (Chai et al. 2020), F3Net (Qian et al. 2020), UniFD (Ojha, Li, and Lee 2023), LGrad (Tan et al. 2023), FreqNet (Tan et al. 2024a), NPR (Tan et al. 2024b), Fatformer (Liu et al. 2024) in mean Accuracy (mAcc). Notably, MCAN improves upon UniFD (Ojha, Li, and Lee 2023), which freezes the CLIP ViT-L/14 encoder and trains only a classifier on image features, by 11.9% in mAcc. Compared to Fatformer (Liu et al. 2024), which employs an adapter-based architecture, MCAN achieves 2.4% higher mAcc. These findings further validate MCAN's strong generalization capability and robustness to distribution shifts across generative techniques.

Ablation Studies

we conduct comprehensive ablation studies on MCAN by incrementally integrating different modules into the baseline model to evaluate their contributions. Following LaRE^2(Luo et al. 2024), we train the model on Stable Diffusion V1.4 and assess its performance across all eight subsets of GenImage. The results are summarized in Table 4.

The configurations in Table 4 are defined as follows: '**Img only**' trains the baseline model solely on raw images, while '**HF only**' uses only high-frequency representations. '**CI only**' is trained exclusively on Chromaticity Inconsistency (CI) representations, whereas '**CI-shuffled**' applies a positional embedding shuffle strategy within CI training. '**Naïve Combination**' aggregates predictions from three independently trained models on these cues. Finally, '**MCAN**' represents our proposed model, which effectively integrates multiple cues using MoEAs.

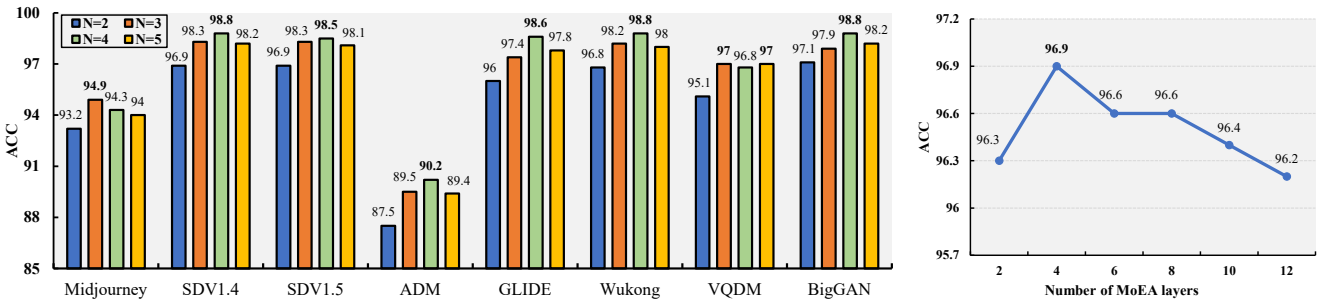
From Table 4, incorporating positional embedding shuffling into CI training improves average accuracy by 3.5% compared to the 'CI only' setting. This suggests that disrupting the spatial structure of CI images helps the model focus on noise patterns rather than image content, enhancing feature learning.

Since raw images, high-frequency components, and CI representations offer complementary information, simply aggregating predictions from separately trained networks already enhances performance. However, MCAN further optimizes this process by seamlessly integrating multiple cues within a unified network, achieving superior results.

To assess the impact of CI, we conducted an ablation study by combining CI with 'HF', 'Img', and 'HF+Img'. These combinations yielded performance improvements of 5.6%, 2.4%, and 1.5%, respectively, highlighting the crucial role of CI in enhancing MCAN's effectiveness.

Discussion

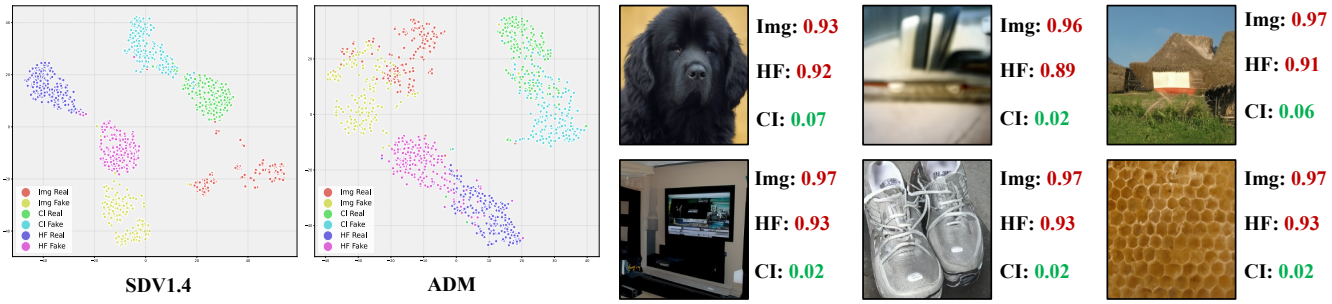
Number of Encoder Experts In the MoE framework, the number of encoder experts plays a crucial role in model performance. To assess its impact, we vary the number of encoder experts from 2 to 5 and conduct a detailed performance comparison, as shown in Figure 4 (a). The results show that when the number of encoder experts is fewer than the number of cues, performance drops significantly, indicating that effective feature extraction requires sufficient expert capacity. However, increasing the number of encoder experts does not always yield further improvements. In most cases,



(a) Performance under different number of encoder experts

(b) Performance under different numbers of MoEA

Figure 4: Performance of MCAN under structures: Optimal reaches when MoEA contains 4 experts in the last 4 blocks.



(a) Visualization of learned feature space on SDV1.4 subset.

(b) Classification results on AI-generated images.

Figure 5: Visualization of learned feature space of MCAN (a) and classification results for generated images under different cues (b). (a) While all cues face generalization challenges, MCAN improves generalization by leveraging complementary features across cues. (b) To demonstrate the complementary nature of CI to 'Img' and 'HF', we specifically select failure cases for 'Img' and 'HF'. The values shown indicate the confidence scores from each cue, representing the likelihood that the corresponding image is classified as real.

the optimal performance is achieved with four encoder experts, reaching an average accuracy of 96.9%.

Number of MoEA Layers The placement of MoEA layers within the network is a critical design choice. To evaluate its impact, we conduct empirical experiments by inserting MoEA layers at different positions within the CLIP-B/16 model while keeping the remaining layers equipped with a single expert adapter following the MoEA structure. The results, shown in Figure 4 (b), provide key insights. Our findings reveal that integrating MoEA into every layer is unnecessary. In MCAN, the best performance is achieved when MoEA layers are inserted in the last four layers. This suggests that shallower layers primarily capture general features, while deeper layers focus on more specialized information. Based on this, MCAN integrates MoEA only in the last four layers of CLIP-B/16, optimizing both efficiency and effectiveness.

Visualization Results To further demonstrate the effectiveness of MCAN, we train the model on the SDV1.4 subset and visualize the learned feature spaces on both SDV1.4 and the unseen ADM subset in Figure 5 (a). The results highlight a key challenge: generalization remains difficult when relying on individual cues alone. However, the inherent diversity among different cues enables each to provide a unique perspective for AI-generated image detection. In Figure 5 (b), we also visualize selected examples where 'Img' and 'HF' classifiers fail but 'CI' provides correct predictions. These images exhibit high visual quality and realistic content, leading to high-confidence predictions by 'Img' and 'HF' classifiers. However, subtle noise artifacts present in these images enable 'CI' to effectively identify them as fake. This complementary nature helps mitigate generalization errors, enhancing MCAN's robustness. These findings suggest that integrating multiple cues reduces the risk of overfitting associated with any single cue, ultimately improving both reliability and adaptability.

Conclusion

In this paper, we introduce the Multi-Cue Aggregation Network (MCAN), a novel framework for AI-generated image detection. MCAN effectively leverages the complementary properties of multiple cues and integrates them dynamically using a Mixture-of-Encoder Adapter (MoEA) within a unified network. In addition to the image and high-frequency cues, we introduce Chromaticity Inconsistency (CI), a new cue based on chromaticity that amplifies noise discrepancies between real and generated images. Given the complementary nature of these cues, integrating CI into MCAN further enhances detection performance.

References

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

Chi, Z.; Dong, L.; Huang, S.; Dai, D.; Ma, S.; Patra, B.; Singhal, S.; Bajaj, P.; Song, X.; Mao, X.-L.; et al. 2022. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35: 34600–34613.

Cozzolino, D.; Poggi, G.; Corvi, R.; Nießner, M.; and Verdoliva, L. 2024. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4356–4366.

Cozzolino, D.; Poggi, G.; Nießner, M.; and Verdoliva, L. 2025. Zero-Shot Detection of AI-Generated Images. In *European Conference on Computer Vision*, 54–72. Springer.

Finlayson, G. D.; Hordley, S. D.; Lu, C.; and Drew, M. S. 2005. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1): 59–68.

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.

Gou, Y.; Liu, Z.; Chen, K.; Hong, L.; Xu, H.; Li, A.; Yeung, D.-Y.; Kwok, J. T.; and Zhang, Y. 2023. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*.

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

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.

Kong, C.; Luo, A.; Bao, P.; Yu, Y.; Li, H.; Zheng, Z.; Wang, S.; and Kot, A. C. 2025a. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *IEEE Transactions on Dependable and Secure Computing*.

Kong, C.; Luo, A.; Wang, S.; Li, H.; Rocha, A.; and Kot, A. C. 2025b. Pixel-inconsistency modeling for image manipulation localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, B.; Shen, Y.; Yang, J.; Wang, Y.; Ren, J.; Che, T.; Zhang, J.; and Liu, Z. 2023. Sparse Mixture-of-Experts are Domain Generalizable Learners. In *International Conference on Learning Representations*.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020. Face x-ray for more general face forgery detection. In *CVPR*, 5001–5010.

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.

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024. 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, 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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17006–17015.

Lyu, S.; Pan, X.; and Zhang, X. 2014. Exposing region splicing forgeries with blind local noise estimation. *International journal of computer vision*, 110: 202–221.

Mahdian, B.; and Saic, S. 2009. Using noise inconsistencies for blind image forensics. *Image and vision computing*, 27(10): 1497–1503.

Marra, F.; Gragnaniello, D.; Cozzolino, D.; and Verdoliva, L. 2018. Detection of gan-generated fake images over social networks. In *MIPR*, 384–389. IEEE.

Mayer, O.; and Stamm, M. C. 2018. Accurate and efficient image forgery detection using lateral chromatic aberration. *IEEE Transactions on information forensics and security*, 13(7): 1762–1777.

McCloskey, S.; and Albright, M. 2019. Detecting GAN-generated imagery using saturation cues. In *ICIP*, 4584–4588. IEEE.

O'brien, J. F.; and Farid, H. 2012. Exposing photo manipulation with inconsistent reflections. *ACM Trans. Graph.*, 31(1): 4–1.

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.

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.

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 *ICML*, 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.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts.

Advances in Neural Information Processing Systems, 34: 8583–8595.

Sarkar, A.; Mai, H.; Mahapatra, A.; Lazebnik, S.; Forsyth, D. A.; and Bhattad, A. 2024. Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28140–28149.

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*, 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 *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.

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

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. *arXiv preprint arXiv:2303.09295*.

Wyszecki, G.; and Stiles, W. S. 2000. *Color science: concepts and methods, quantitative data and formulae*, volume 40. John wiley & sons.

Yan, S.; Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Xie, W. 2025. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23828–23837.

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

Zhong, N.; Xu, Y.; Qian, Z.; and Zhang, X. 2023. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *CoRR*.

Zhou, P.; Han, X.; Morariu, V. I.; and Davis, L. S. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1053–1061.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2024. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36.