IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **36** (2025) 095402 (16pp) https://doi.org/10.1088/1361-6501/ae005c

MSAFNet: multi-scale self-adaptive feature fusion network for AI-generated image detection

Liwei Yao[1], Sen Niu[1], Kaili Liao[1], Guobing Zou[2], Kefeng Li[1,*] and Tengbo Zhao[1]

[1] The School of Computer and Information Engineering, Institute for Artificial Intelligence, Shanghai Polytechnic University, Shanghai 201209, People's Republic of China
[2] The school of Computer Engineering and Science, Shanghai University, Shanghai 200444, People's Republic of China

E-mail: y_20010830@163.com

Received 30 April 2025, revised 21 July 2025
Accepted for publication 28 August 2025
Published 4 September 2025

Abstract

With the development of advanced generative technologies such as generative adversarial networks and diffusion models, AI-generated images have become extremely realistic. As a result, there is an urgent need for efficient detection tools to combat misinformation and protect societal trust and personal privacy. Current detection methods perform well when identifying content from known generative models, but their performance drops when facing unknown or emerging technologies. Moreover, most detection methods focus on detecting local artifacts, overlooking the extraction and effective fusion of global relationships and multi-scale features, which makes it difficult to capture the complex patterns in AI-generated images. To address these challenges, this paper proposes a multi-scale self-adaptive feature fusion network (MSAFNet). Specifically, the proposed MSAFNet utilizes a global noise feature extraction module to capture global noise patterns in images, compensating for the shortcomings of existing methods that focus solely on local noise features. Meanwhile, it adopts a texture feature extraction module to capture subtle texture anomalies and a color feature extraction module to enrich color feature information, enhancing the expression capability of color features and improving the recognition of complex features in AI-generated images. Lastly, we introduce a self-adaptive feature fusion module to capture complementary information and the importance of multi-scale features, ensuring efficient fusion of feature information across different dimensions. Experimental results on the large-scale DFFD and CIFAKE datasets demonstrate that the proposed MSAFNet achieves higher classification accuracy compared to other AI-generated image detection methods.

Keywords: imaging, AI-generated image detection, IoT security, deep learning

1. Introduction

The development of efficient detection methods for AI-generated images is not only a critical technological safeguard to curb the spread of misinformation but also contributes to building a more transparent and trustworthy information environment. However, the urgency of this need stems from the exponential development of generative models. From the initial proposal of the generative adversarial networks (GANs [1]) to the subsequent emergence of the variational autoencoders (VAEs [2]), the GLOW [3], the stable diffusion [4], and numerous derivative models [5–9], the rapid iteration of generative models has significantly improved the quality and realism of synthetic images. At the same time, this technological

* Author to whom any correspondence should be addressed.

advancement has also brought about the risk of a decline in societal trust. For example, the proliferation of fake or misleading information, tampered facial images, and the spread of other deceptive and libelous content is eroding public trust in information. Therefore, developing efficient AI-generated image detection methods is crucial to address technological challenges and preserve information authenticity.

The task of detecting AI-generated images is the first introduced by study [10], which explores the ability of traditional detectors and deep learning-based detectors to identify GAN-generated fake images in a social network environment. It points out that only deep learning detectors maintain high accuracy (ACC) even when facing compressed data [10]. However, conventional detection methods heavily depend on the distribution of training data, making it difficult to reveal the fundamental differences between AI-generated images and real images. To further explore the unique characteristics of GAN-generated images, researchers propose a series of methods [11–15] that analyze artifact features in the frequency domain or spectrum to identify synthetic images. These methods leverage specific frequency features inevitably introduced during the image synthesis process by generative models, improving the robustness and generalization ability of detection. Meanwhile, researchers focus on the differences between AI-generated and real images in multiple dimensions. For instance, a global texture enhancement method is introduced to improve detection performance [16]. Additionally, a study highlights that detection methods based on frequency-domain analysis or global texture features primarily rely on specific informational traces [17]. In the search for more effective detection criteria, researchers discover that color information also plays an important role in the decision-making process. However, existing methods primarily rely on simplified, single-type random grayscale data augmentation strategies, such as the color-robust universal detector. This results in an overly simplistic representation of color feature information, making it difficult to capture the interactions and correlations between different color spaces, thereby limiting the generalization ability and adaptability of detection methods.

With the emergence of diffusion generative models, Ojha *et al* [18]. are the first to identify the significant limitations of GAN-based image detection methods in recognizing images generated by diffusion models. To address this issue, some detectors [19, 20] specifically utilize the artifacts inevitably left by diffusion models to identify their generated images. Meanwhile, to improve the generalization ability of detection methods, researchers propose various approaches capable of recognizing data from different sources, including those for GAN-based generative models [5–8, 21, 22] and diffusion model-based generative methods [23–26]. Building on this, subsequent research shifts detection strategies toward noise pattern analysis. Methods [27–32] apply the steganalysis rich model (SRM) filters [33] or other local feature extraction techniques to extract local noise artifacts, which serve as key cues for detection. However, these methods primarily focus on local noise analysis, often relying on fixed-size patches or selecting only a subset of patches for detection [34]. This localized approach tends to overlook the unique noise patterns that span the entire image in generative images, resulting in an incomplete representation of noise features and thereby affecting the comprehensiveness and ACC of detection.

Based on the above analysis, a multi-scale self-adaptive feature fusion network (MSAFNet) is proposed for the efficient detection of AI-generated images. This method comprehensively utilizes global noise, color features, and texture information to enhance detection generalization and robustness. Firstly, a global noise feature extraction module (GNFEM) is designed to capture the global noise patterns in the input image. Within this module, the noise net further extracts deep-level global noise features to uncover the noise artifacts left by the generative model in the image. Next, a color feature extraction module (CFEM) is introduced, which converts the original RGB image into five different color spaces to enrich the representation of color features. Meanwhile, the color net explores the potential correlations between different color spaces, enhancing the model's ability to perceive color features. Additionally, the texture feature extraction module (TFEM) uses the local binary pattern (LBP) algorithm to compute the global texture features of the input image, and the texture net further extracts subtle texture differences from the local regions of the image to enhance the detection of texture artifacts in generated images. Finally, a self-adaptive feature fusion module (SAFM) is proposed to integrate complementary information from different sub-features and adaptively adjust the weight distribution of each sub-feature, ensuring the ACC and robustness of feature fusion [35, 36]. This improves the detection capability of AI-generated images. Our main contributions can be summarized as follows:

(1) To address the issue that existing methods based on local block analysis often overlook the global noise patterns in images, the GNFEM is proposed. This module utilizes the SRM filter to directly extract the global noise patterns from the entire input image, and through the noise net, deep global noise features are mined, thereby improving the sensitivity of the detection method to artifacts in AI-generated images.
(2) Existing methods for color feature extraction are often overly simplistic and overlook the dependencies between different color spaces. To address this, the CFEM is proposed. The input image is first converted from the RGB color space into five different color spaces ('gray', 'lab', 'ycbcr', 'cmyk', and 'hsv'). The channels from these color spaces are then concatenated to enrich the color feature information. Next, the color net extracts more diverse and discriminative color features while learning the complex dependencies between different color spaces, thereby enhancing the expression capability of color features.
(3) To efficiently integrate different sub-features, the SAFM is proposed. This module is capable of learning the complementary information between different sub-features

and adaptively optimizing the fusion ratio of each feature through a dynamic weight adjustment mechanism. This ensures the optimal fusion of multi-source features within a unified framework. This strategy significantly enhances the model's ability to learn the complex relationships between multi-source features, thereby improving the ACC and robustness of AI-generated image detection.

2. Related work

This section briefly introduces the mainstream image generation technologies and detection methods for AI-generated images.

2.1. Image generation

In recent years, with breakthroughs in deep learning technology, image generation techniques make significant progress, enabling the synthesis of visual content that closely resembles real images. The current mainstream image generation methods include the GANs [1], the VAEs [2], and the stable diffusion [4], each offering advantages and being suitable for different application scenarios. The GANs continuously optimize the realism of generated images through adversarial training between a generator and a discriminator. However, the training process tends to be unstable and often encounters mode collapse issues. In contrast, the VAEs use an encoder-decoder structure for probabilistic modeling, demonstrating a balanced performance in stability and diversity of generated samples. However, generated images often suffer from blurred details due to probabilistic approximation bias. In recent years, the stable diffusion rapidly develops, achieving image generation through a mechanism of gradually adding and removing noise. They outperform the GANs and the VAEs in terms of training stability and generation quality, making them one of the most advanced image generation methods. The stable diffusion shows broad application potential in areas such as text-to-image generation and medical image synthesis.

2.2. AI-generated image detection

Early research on the detection of AI-generated images primarily focuses on frequency-domain artifact features in GAN-generated images [37]. A study [12] discovers that GANs rely on convolution-based up-sampling methods, making it difficult to replicate the frequency spectrum distribution of real data. Researchers leverage this flaw and focus on extracting frequency spectrum features for detection. Furthermore, Dzanic *et al* [13]. propose a high-frequency signal analysis method based on the discrete Fourier transform, differentiating generated images through spectral features under high resolution and low compression conditions. In addition, some researchers find that global texture features [16] and color features [17, 38] are also important distinguishing cues. However, method [17] primarily employs random grayscale transformations as a means of color information augmentation, which limits its representational richness. Similarly, the strategy [38] is based on color imaging forensics but is restricted to analysis within the standard RGB color space. This limitation prevents a comprehensive exploration of potential discriminative information across multiple color spaces. Consequently, the utilization of color features may be inadequate. With the rise of diffusion models, Ricker *et al* [39]. point out that images generated by diffusion models do not exhibit obvious artifacts in the frequency domain. Chen *et al* [27]. further discover structural differences between images generated by diffusion models and those produced by GANs. These differences lead to a significant decline in the performance of traditional frequency-domain detection methods when identifying diffusion-generated images.

To address the above issues, methods are proposed to analyze specific artifacts left by various generative models in the high-frequency components of synthetic images [7, 8]. As a result, subsequent researchers commonly use the SRM [33] to extract noise patterns from images when faced with images generated by unknown generative models. For example, in the SSP method [27], researchers focus on extracting hidden noise from the simplest texture regions of an image (i.e. areas with the lowest texture complexity), while the AIDE method [28] uses discrete cosine transform (DCT) scores, selects two patch blocks from the highest and lowest frequencies based on the scores, and uses the SRM to extract the noise patterns of the four patches, followed by feature fusion with global semantic information extracted by the contrastive language-image pre-training[40]. Similarly, Li *et al* [30]. proposed a collaborative modeling strategy that integrates noise features from the spatial domain with energy distributions in the frequency domain. They utilized SRM to extract both salient and subtle noise patterns from strong-texture and weak-texture regions, respectively, and employed an enhanced DCT-based channel attention mechanism for frequency-domain guidance. This approach improves the model's ability to perceive artifacts under varying noise intensities. Meanwhile, Cavia *et al* [29]. focused solely on fixed-size patch-level modeling. They applied convolutional networks with extremely small receptive fields to independently assign a forgery score to each patch and then aggregated these scores via pooling to obtain a global decision, thereby forming a detection framework based on local information. However, these methods are mainly based on local patch blocks for noise feature extraction, which may overlook the global noise distribution characteristics of the image. Therefore, Zhong *et al* [41]. propose a new strategy that divides the entire image into multiple patch blocks and then reorganizes them based on texture richness, disrupting the original semantic information of the image to extract noise patterns. Their research focuses on the correlation of noise patterns between rich and barren regions rather than only on the features of the noise patterns themselves. However, this method disrupts the semantic structure of the image, which may weaken the long-range correlations of the image's noise patterns, thus affecting detection performance.

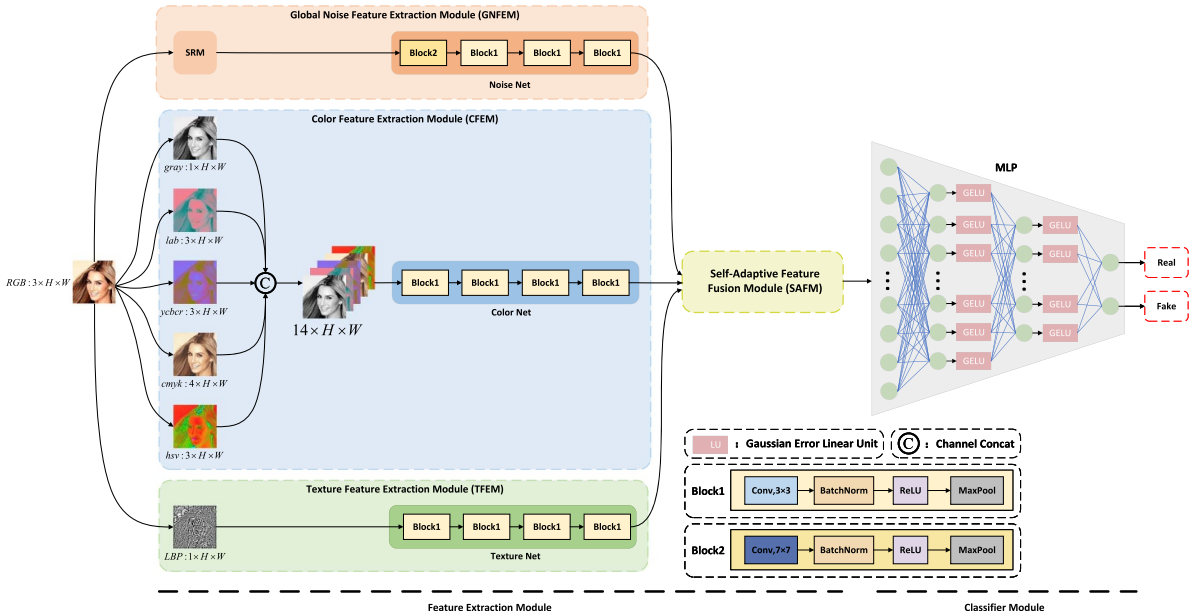


Figure 1. The architecture of the proposed MSAFNet.

3. Proposed method

The architecture of the proposed MSAFNet is shown in figure 1. This method mainly consists of two core modules: the feature extraction module and the classifier module. The feature extraction module includes four sub-modules: the GNFEM, the CFEM, the TFEM and the SAFM. The GNFEM is designed to extract the global noise patterns, the CFEM captures the color features, and the TFEM extracts the texture features. Additionally, the SAFM is proposed to analyze image features from multiple dimensions, learn complementary information between sub-features, and adaptively allocate weights to enhance system robustness. The classifier module consists of a multi-layer perceptron (MLP), which processes the fused features and outputs the final classification decision.

3.1. Feature extraction module

3.1.1. GNFEM. The previous AIDE [28] method primarily focuses on analyzing local noise patterns in images. This method divides the image into multiple patch blocks and uses a DCT scoring module to select certain patch blocks, then extracts the noise patterns from these patches for subsequent analysis. However, this strategy disrupts the long-range correlated noise patterns in AI-generated images and overlooks global noise information, thus limiting the feature expression capability of the noise patterns. To address this issue, our module abandons the local patch block strategy and instead uses the entire image to preserve the global noise patterns. Specifically, the SRM [33, 37, 42–44] method is applied to the input image $X \in \mathbb{R}^{B\times C\times H\times W}$ (where B, C, H, and W represent the batch size, the channel, the height, and the width , respectively.) to extract the global noise pattern of the entire image, which can be represented as

$$X_1^{(1)} = \gamma(X). \tag{1}$$

The global noise feature $X_1^{(1)} \in \mathbb{R}^{B\times C_1\times H_1\times W_1}$ is extracted through the function $\gamma(\cdot)$, which employs the SRM method based on predefined high-pass filters to enhance subtle residual noise features in the image, thereby facilitating forensic analysis.

Next, the deep global noise pattern $X_1 \in \mathbb{R}^{B\times C'\times H'\times W'}$ (where $H' = \frac{H_1}{32}$ and $W' = \frac{W_1}{32}$) can be expressed as

$$X_1 = \psi\left(X_1^{(1)}\right) \tag{2}$$

where $\psi(\cdot)$ represents the noise net, which consists of one block2 and three block1. The specific structure of the noise net is shown in figure 1.

3.1.2. CFEM. Although previous method [17, 45] recognizes the importance of color features, a random grayscale processing approach is used. This approach results in a singular color feature representation. Additionally, the interrelationships between different color spaces are ignored. To enhance the expressive power of color information, the CFEM is proposed. This module converts the input image $X \in \mathbb{R}^{B\times C\times H\times W}$ from the default RGB color space to five different color spaces ('lab', 'cmyk', 'ycbcr', 'hsv', and 'gray'), thus constructing a richer multi-dimensional color representation system. It can be formulated as

$$X_j = \mathcal{T}_j(X), \quad j = \{\text{lab},\text{cmyk},\text{ycbcr},\text{hsv},\text{gray}\}. \tag{3}$$

The converted images $X_{\text{gray}} \in \mathbb{R}^{B\times 1\times H\times W}$, $X_{\text{ycbcr}}, X_{\text{hsv}}, X_{\text{lab}} \in \mathbb{R}^{B\times 3\times H\times W}$, and $X_{\text{cmyk}} \in \mathbb{R}^{B\times 4\times H\times W}$ can be obtained through

$\mathcal{T}_j(X)$, which converts the RGB image X into the $j-$th color space.

Subsequently, the images from different color spaces are concatenated along the channel dimension, producing the fused output $X_2^{(1)} \in \mathbb{R}^{B\times 14\times H\times W}$. This enables the collaborative exploration of multiple color spaces, thereby enhancing the expressive power of color features. This strategy effectively addresses the issue of overly simplified color information caused by random grayscale processing [17], allowing the model to comprehensively utilize feature information from different color spaces. The process is specifically represented as

$$X_2^{(1)} = \mathrm{Ch}\left(X_{\mathrm{hsv}}, X_{\mathrm{ycbcr}}, X_{\mathrm{lab}}, X_{\mathrm{cmyk}}, X_{\mathrm{gray}}\right) \tag{4}$$

where $\mathrm{Ch}(\cdot)$ represents the channel concatenation.

Finally, the deep color feature $X_2 \in \mathbb{R}^{B\times C'\times H'\times W'}$ can be described as

$$X_2 = \varphi\left(X_2^{(1)}\right) \tag{5}$$

where $\varphi(\cdot)$ represents the color net, which consists of four block1. The specific structure of the color net is shown in figure 1. Specifically, Block1 performs nonlinear combinations through convolution operations to reveal the hidden correlations between different color spaces. The design goal of the color net is to facilitate cross-channel information interaction across multiple color domains, fully exploring the correlated features within the color spaces. This enables the model to more effectively capture local inconsistencies and global distortions in the image, thereby enhancing the expressive power of color information.

3.1.3. TFEM. Previous studies [16] show that embedding texture features into a network helps improve the model's generalization ability and has confirmed that texture features are key criteria for detecting AI-generated images. Inspired by this, the TFEM is introduced, specifically designed to extract texture information from the input image. Given the input image $X \in \mathbb{R}^{B\times C\times H\times W}$, the texture feature map can be processed using the LBP algorithm and can be described as

$$X_3^{(1)} = \eta(X). \tag{6}$$

The texture feature map $X_3^{(1)} \in \mathbb{R}^{B\times 1\times H\times W}$ can be obtained via $\eta(\cdot)$, where $\eta(\cdot)$ denotes the application of the LBP method to enhance texture representation by encoding texture variations in the image.

Subsequently, the subtle texture anomaly feature $X_3 \in \mathbb{R}^{B\times C'\times H'\times W'}$ can be computed as

$$X_3 = \phi\left(X_3^{(1)}\right) \tag{7}$$

where $\phi(\cdot)$ represents the texture net, which consists of four block1. The specific structure of the texture net is shown in figure 1. Through progressive feature extraction, the texture net enhances the model's ability to recognize local texture anomalies, further improving its detection of subtle texture irregularities.

3.1.4. SAFM. Previous research methods (such as [28] and [46]) used traditional fixed-weight fusion strategies, such as average fusion or channel concatenation. However, these methods are difficult to adapt to the dominant artifact features of different generation models and struggle to effectively learn the complementary information between individual sub-features. The SAFM is designed to address this issue, drawing inspiration from the multi-head self-attention mechanism in Transformer architectures [47–49]. This module models cross-feature dependencies, enabling information complementarity between sub-features and adaptive weight allocation, thereby enhancing the model's flexibility and generalization ability. The specific structure of the SAFM is shown in figure 2.

The input includes three feature maps, denoted as $X_1, X_2, X_3 \in \mathbb{R}^{B\times C'\times H'\times W'}$, representing the deep global noise pattern, the deep color feature, and the subtle texture anomaly feature, respectively. To obtain a global description of each feature map, adaptive average pooling (AAP) is applied to each X_i, reducing its spatial dimensions to 1×1. This process can be expressed as

$$X_i^{\mathrm{pool}} = \alpha(X_i), \quad i = \{1,2,3\}. \tag{8}$$

Each spatially compressed feature map can be obtained through $\alpha(\cdot)$, where $\alpha(\cdot)$ denotes the AAP.

Next, the three global description vectors are concatenated along the feature dimension to construct a comprehensive feature representation, ensuring that the information from each sub-feature is preserved and utilized in subsequent processing, which can be computed as

$$Fe = \delta\left(X_1^{\mathrm{pool}}, X_2^{\mathrm{pool}}, X_3^{\mathrm{pool}}\right). \tag{9}$$

The fused feature $Fe \in \mathbb{R}^{B\times C'\times 3}$ (where 3 represents the number of input features) can be obtained through $\delta(\cdot)$, where $\delta(\cdot)$ represents concatenation along the feature dimension. This fused feature ensures the integrity of each sub-feature's information, providing a richer representation for subsequent processing.

To leverage the attention mechanism for adaptive weighting of features and capture complementary information between sub-features, the Query (Q), the Key (K), and the Value (V) representations need to be constructed. To achieve this, three separate 1D convolutional layers are applied to Fe, as represented by

$$\begin{aligned} Q &= W^Q * Fe \in \mathbb{R}^{B\times C'\times 3} \\ K &= W^K * Fe \in \mathbb{R}^{B\times C'\times 3} \\ V &= W^V * Fe \in \mathbb{R}^{B\times C'\times 3} \end{aligned} \tag{10}$$

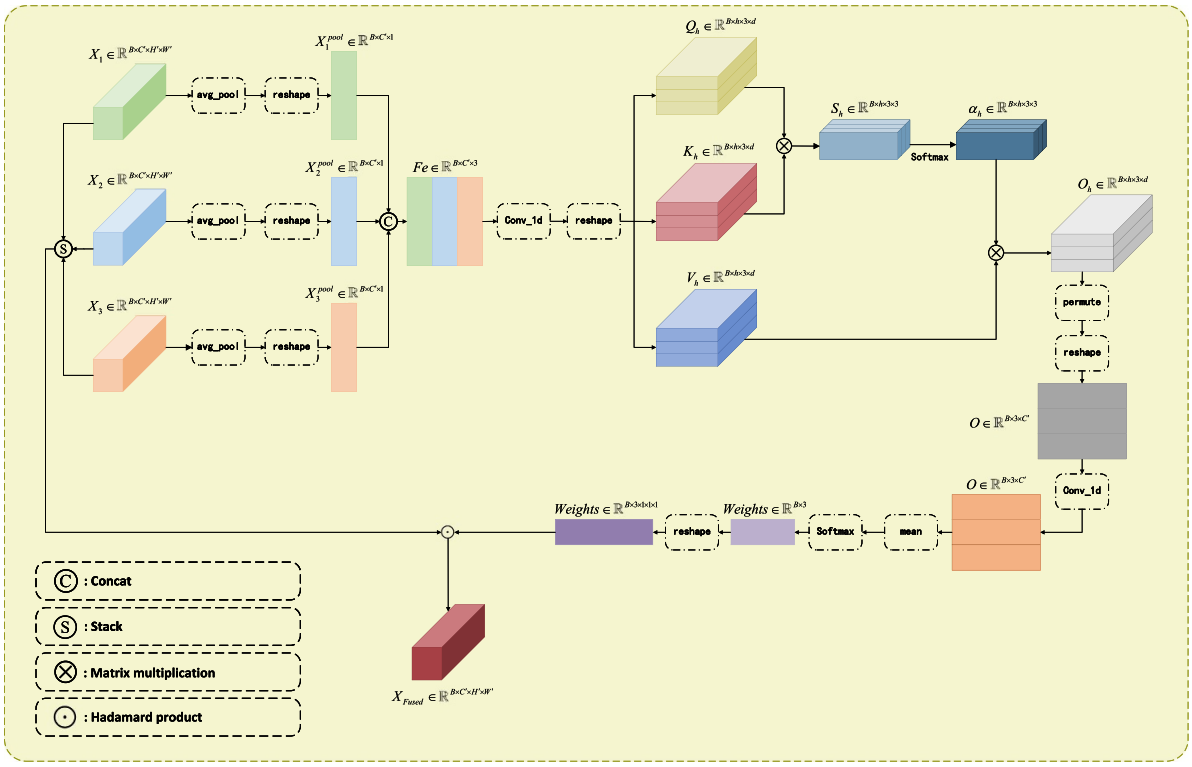


Figure 2. The architecture of the SAFM.

where W^Q, W^K, $W^V \in \mathbb{R}^{C' \times C' \times 1}$ represent the corresponding convolution kernels with a size of 1×1, and $*$ denotes the convolution. This process constructs the Q, K, and V representations. These serve as the foundation for the attention mechanism.

To enhance the model's expressive power, the channel dimension C' is divided into h attention heads. Each attention head processes a subset of the channel information independently. The dimension of each head can be defined as

$$d = \frac{C'}{h}. \tag{11}$$

Additionally, to ensure the computational validity of the multi-head attention mechanism, C' must be divisible by h. By reshaping the Q, the K, and the V to achieve attention head segmentation, the shapes can be transformed as

$$Q_h, K_h, V_h \in \mathbb{R}^{B \times h \times 3 \times d}. \tag{12}$$

This operation allows subsequent attention calculations to occur independently within each subspace, enabling different attention heads to focus on the feature interaction patterns within their respective subspaces, thereby improving the model's adaptability to various generated artifacts.

For each attention head, the attention scores S_h between the Q and the K are first calculated through matrix multiplication, and scaled by $\sqrt{d}$ to prevent numerical instability caused by excessively large values. The computation process can be presented as

$$S_h = \frac{Q_h K_h^T}{\sqrt{d}} \in \mathbb{R}^{B \times h \times 3 \times 3}. \tag{13}$$

The attention scores are then normalized through the operation, resulting in the generation of the attention weight matrix α_h, which can be represented as

$$\alpha_h = \mathrm{Softmax}\left(S_h\right) \in \mathbb{R}^{B \times h \times 3 \times 3} \tag{14}$$

where $\mathrm{Softmax}(\cdot)$ denotes the normalization.

Then, the attention weight matrix α_h is applied to the values V_h to enhance important features and suppress redundant information, while ensuring that each attention head can optimize feature interactions within independent subspaces. The final output of the attention head O_h can be generated as

$$O_h = \alpha_h V_h \in \mathbb{R}^{B \times h \times 3 \times d}. \tag{15}$$

To fuse the diverse feature patterns learned by multiple heads, dimension permutation is first performed on the output of each head O_h, after which the results from all heads are merged and restored to the channel dimension. This ensures that the fused feature information is both complete and diverse. This process can be formed as

$$O = \mathrm{PR}\left(O_h\right). \tag{16}$$

The final feature pattern $O \in \mathbb{R}^{B \times 3 \times C'}$ can be obtained through $\mathrm{PR}(\cdot)$, where $\mathrm{PR}(\cdot)$ represents the process of fusing

the multi-head outputs. This fusion ensures that each output feature (a total of 3 features) retains complete channel information, enhancing the consistency and integrity of the feature representation.

To further extract more refined cross-channel complementary information, a 1D convolution layer is applied to the previously fused output O for information extraction, generating the fused feature description $\tilde{O} \in \mathbb{R}^{B\times 3\times C'}$ can be expressed as

$$\tilde{O} = W^{\text{out}} * O \in \mathbb{R}^{B\times 3\times C'} \tag{17}$$

where W^{out} represents the convolution kernel of size 1×1.

Next, average pooling is applied along the channel dimension to each feature description (corresponding to the ith input feature), and its mean is computed, thereby obtaining the scalar response ω_i for each feature

$$\omega_i = \frac{1}{C'}\sum_{j=1}^{C'} \tilde{O}_{ij} \in \mathbb{R}^{B\times 3}, \quad i = 1,2,3. \tag{18}$$

Moreover, normalized weights are generated to ensure that the weight distribution between different features is interpretable and stable. The specific process can be formulated as

$$w_i = \frac{\exp(\omega_i)}{\sum_{k=1}^{3}\exp(\omega_k)} \in \mathbb{R}^{B\times 3}, \quad i = 1,2,3. \tag{19}$$

These weights (w_i) represent the contribution of each input feature in the final fusion process, enabling adaptive weighted fusion.

Subsequently, the original three input feature maps X_1, X_2, X_3 are stacked along a new dimension to form a unified tensor $X \in \mathbb{R}^{B\times 3\times C'\times H'\times W'}$, for subsequent feature processing and the application of the attention mechanism a.s. which can be outlined as

$$X = S(X_1, X_2, X_3) \tag{20}$$

where $S(\cdot)$ represents the process of stacking the three original features.

Finally, using the previously calculated weights $\{w_i\}_{i=1}^{3}$, a weighted operation is performed on each component of X, and the results are summed along the feature dimension, The final adaptive fused features $X_{\text{Fused}} \in \mathbb{R}^{B\times C'\times H'\times W'}$ can be denoted as

$$X_{\text{Fused}} = \sum_{i=1}^{3} w_i \odot X_i \tag{21}$$

where $\odot$ represents the weighted operation.

The final fused feature map X_{Fused} integrates global information from the three input features across different spatial locations, while capturing the complementary relationships between them. The adaptive weight mechanism further enhances key information, providing a richer and more effective feature representation for the subsequent classifier module, thereby improving the model's discriminative ability.

3.2. Classifier module

Finally, the fused features are flattened and input into the MLP. The decision scores are generated after passing through three linear layers and two GELU [50] activation functions. This transformation can be expressed as

$$\hat{y}_i = \text{MLP}(X_{\text{Fused}}). \tag{22}$$

The learnable parameters in the feature extraction module and classifier module are optimized to minimize the cross-entropy loss function. Its mathematical expression is as follows

$$l = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log \hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \tag{23}$$

where y_i is the ground truth label, and $\hat{y}_i$ is the predicted probability. During the inference phase, the model can directly extract features from input images of any size and output the detection results, without requiring additional preprocessing or postprocessing steps.

4. Experiments

4.1. Datasets

To comprehensively and effectively evaluate the proposed MSAFNet, we select two representative datasets: the DFFD [51] and the CIFAKE [52]. The DFFD is a comprehensive dataset specifically designed for deepfake detection, covering four types of digital face manipulations, and includes various manipulation methods and different degrees of tampering. The CIFAKE dataset consists of two parts: one part includes real images directly sourced from the CIFAR-10 [53], and the other part contains an equal number of synthetic images generated using the stable diffusion model [4]. These two datasets each offer unique challenges and characteristics, providing a more comprehensive evaluation of the performance of our detection method.

4.1.1. DFFD datasets. The DFFD is a comprehensive dataset specifically designed for deepfake detection and localization research, renowned for its rich diversity and broad challenges. The dataset covers various types of forged faces, including identity swaps, expression changes, attribute manipulations, and completely synthetic face images. Specifically, the DFFD dataset constructs forged images from four main dimensions: first, complex identity and expression swaps based on video clips from FaceForensics++; second, fine-grained attribute editing of images from the FFHQ [8] and CelebA [54] datasets using FaceAPP and StarGAN [55];

Table 1. Detailed split of training and testing sets in the DFFD dataset.

Subdatasets	img_align_celeba	ffhq	pggan_v1	pggan_v2	stylegan_celeba	stylegan_ffhq	stargan	faceapp
REAL/FAKE	REAL	REAL	FAKE	FAKE	FAKE	FAKE	FAKE	FAKE
Train	162 770	10 000	9975	9982	10 000	9999	10 000	6309
Test	19 867	999	998	1000	1000	1000	1000	999

Table 2. Detailed split of training and testing sets in the CIFAKE dataset.

Class	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
Train	10 000	10 000	10 000	10 000	10 000	10 000	10 000	10 000	10 000	10 000
Test	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000

additionally, high-quality synthetic face images are generated using pre-trained models from PGGAN [7] and StyleGAN [8]. This multi-layered application of techniques makes DFFD not only cover a wide range of forgery types but also possess significant variability within each category, greatly enriching the dataset. In this study, we primarily used eight subsets from the DFFD dataset, including img_align_celeba, ffhq (real images), pggan_v1, pggan_v2, stylegan_ffhq, stylegan_celeba, faceapp, and stargan (fake images), totaling 229 035 training images and 26 863 testing images. These subsets are balanced in terms of gender, age, and face size, ensuring the data is unbiased and diverse. Detailed information about the subsets can be shown in table 1.

4.1.2. CIFAKE datasets. The CIFAKE dataset is a unique dataset specifically designed for recognizing AI-generated synthetic images. Its structure is consistent with the classic CIFAR-10 [53] dataset, containing real and synthetic images of size 32×32 pixels across ten categories (e.g. airplanes, cars, birds, cats, etc), providing rich diversity for evaluating the model's robustness and generalization ability. The CIFAKE dataset consists of 60 000 real images directly from CIFAR-10 and 60 000 high-quality synthetic images generated using the stable diffusion model, with 6000 images per category, totaling 120 000 images. The dataset is randomly divided into 50 000 training images and 10 000 testing images. Table 2 provides a breakdown of the number of images in each category for both the training (Train) and testing (Test) sets.

4.2. Experimental setup

In our method, we use the Adam optimizer with no weight decay (i.e. the weight decay coefficient is set to 0), and the learning rate is set to 0.0001. Additionally, we construct a learning rate scheduler that combines linear warm-up and cosine annealing strategies. The linear warm-up is used to gradually increase the learning rate in the initial stages of training to stabilize the training process, while cosine annealing smoothly decreases the learning rate throughout the training cycle to promote better model convergence. The batch size is set to 64. For the DFFD dataset, the training epoch is limited to 15, while for the smaller CIFAKE dataset, the number of epochs is increased to 25. All experiments are conducted on a computer equipped with an NVIDIA GeForce RTX 3090 GPU, using the PyTorch framework, Python version 3.9, and CUDA version 11.3.

This study addresses a binary classification problem and uses two key evaluation metrics—the ACC and the area under the curve (AUC)—to assess the model's performance on the validation set. These metrics have been widely used in previous research [46, 56–58] and can evaluate the model's classification capability and robustness from different perspectives.

The ACC is a fundamental metric for evaluating the performance of classification models, especially in binary classification tasks. The ACC reflects the overall classification ACC of the model by calculating the ratio of correctly predicted samples to the total number of samples. The AUC, on the other hand, provides an evaluation method that is unaffected by class distribution. Specifically, the AUC represents the area under the ROC curve, which illustrates the relationship between the true positive rate and the false positive rate at different classification thresholds. Therefore, the AUC not only measures the model's ability to distinguish between positive and negative samples but also provides a more stable evaluation in the case of imbalanced class distributions.

By using both the ACC and the AUC, a more comprehensive evaluation of the model's classification ability and robustness can be achieved. This ensures that the selected model maintains high ACC overall, while still being able to distinguish effectively in imbalanced datasets.

4.3. Comparative experiments

To validate the superiority of the proposed framework, comparative experiments are conducted with methods (Vision Transformer [59], Swin Transformer [60], AIDE [28], ConvNeXt [61], DeiT [62], FreDect [15], GramNet [16], and LGrad [63]). To ensure fairness, all models, including ours, are trained and tested on the same dataset with identical batch size and epoch settings. The experimental results are shown in tables 3 and 4, where bold and underlined values indicate the best and second-best results in each column, respectively. On the DFFD dataset, our method achieved 99.24%

Table 3. Comparison results of our method and other methods on the DFFD dataset.

Method	Pre-training	ACC (%)	AUC (%)
Vision Transform-T (2020)	FALSE	95.60	98.79
Swin Transform-T (2021)	FALSE	98.33	99.79
CovNexT-B (2022)	FALSE	97.51	99.73
DeiT-B (2020)	FALSE	96.22	99.30
AIDE (2024)	TRUE	98.56	99.90
FreDect (2020)	FALSE	98.54	99.88
GramNet (2020)	FALSE	97.63	99.77
LGrad (2023)	TRUE	98.04	99.78
Ours	FALSE	**99.24**	**99.95**

Table 4. Comparison results of our method and other methods on the CIFAKE dataset.

Method	Pre-training	ACC (%)	AUC (%)
Vision Transform-T (2020)	FALSE	90.82	96.85
Swin Transform-T (2021)	FALSE	94.57	98.84
CovNexT-B (2022)	FALSE	96.95	99.61
DeiT-B (2020)	FALSE	95.50	99.29
AIDE (2024)	TRUE	96.78	99.51
FreDect (2020)	FALSE	89.56	96.28
GramNet (2020)	FALSE	97.01	99.61
LGrad (2023)	TRUE	93.65	98.29
Ours	FALSE	**97.60**	**99.66**

ACC and 99.95% AUC without pre-training, outperforming other state-of-the-art methods with an average ACC improvement of 1.69%. On the more challenging CIFAKE dataset, our method achieves an average ACC improvement of 3.25% compared to the other methods. These results strongly demonstrate the significant advantages of our framework in handling AI-generated images from various generators.

Our method achieves strong detection performance on both DFFD and CIFAKE, validating the effectiveness of the MSAFNet in addressing fake images generated by various generators. The experimental results show that our method outperforms traditional generic image classification methods, such as the Vision Transformer, the Swin Transformer, the ConvNeXt, and the DeiT, in terms of both the ACC and the AUC. This result proves that our method successfully captures the complex internal artifact relationships in AI-generated images, showcasing its specialized advantage in the field of AI-generated image detection. Furthermore, compared to detection frameworks like the AIDE, the FreDect, the Gram-Net, and the LGrad, which are designed for the same task, our method achieves optimal detection ACC even without pretraining. This advantage further highlights the importance of the three types of sub-features (global noise patterns, color features, and texture features) extracted by our method, which form the key basis for AI-generated image detection.

It is worth noting that the FreDect [15] achieves the second-highest ACC and AUC on the DFFD dataset, but its ACC drops significantly on the CIFAKE dataset. The main reason for this is that the FreDect relies on frequency-domain anomaly detection, which identifies based on the frequency-domain features of the image. On the DFFD dataset, since most of the generated images are produced by GANs, these images exhibit obvious artifacts in the frequency domain, which explains the method's good performance on this dataset. However, on the CIFAKE dataset, images generated by Stable Diffusion lack distinct frequency-domain artifacts, resulting in reduced detection performance for the FreDect.

To address this limitation, our method introduces improvements in two aspects. Firstly, the GNFEM, CFEM, and TFEM modules are incorporated to extract global noise patterns, color features, and texture features, respectively. These modules enhance the representation of image features from multiple perspectives, compensating for potential information loss caused by relying on a single feature. Secondly, the SAFM module is adopted for multi-scale adaptive feature fusion, enabling the model to learn complementary information between different sub-features, thereby improving the overall detection ability and generalization performance. Through these improvements, our method overcomes the limitations of existing techniques on specific types of datasets and enhances the robustness of the model.

Additionally, the GramNet [16] improves detection generalization by embedding global texture information in the ResNet-50 backbone network. It performs well on both the DFFD and CIFAKE datasets, further validating the importance of global texture information in AI-generated image detection, and supporting the effectiveness of the TFEM module in extracting texture features in our approach.

4.4. Ablation study

4.4.1. Validating the effectiveness of global noise patterns. Ablation experiments are conducted on both datasets to validate the effectiveness of global noise patterns in comparison to local noise patterns. The specific experimental setup is as follows: To ensure a fair comparison with the local patch-based noise pattern strategy, the method proposed by the AIDE [28] is referenced. First, the DCT scoring is used to select two patches with the highest and lowest frequencies based on the score, and then the SRM is applied to extract the noise patterns from these four patches. Subsequently, the proposed noise net is employed to mine deep fake artifact features, and decision outputs are made through the MLP. In contrast, the global noise analysis method discards the image segmentation strategy and directly applies the SRM filtering to the entire input image. The subsequent processing steps remain identical to those in the local patch-based approach.

In the experimental process, noise patterns are first extracted through the SRM, and then processed using the noise net to obtain feature maps. To ensure fairness and consistency, the number of channels in the feature maps is set to 64, and the height and width of the feature maps are controlled by adjusting the number of block1 layers in the noise net. This ensures that the feature map sizes for both global and local noise patterns are consistent during decision making. In the experimental results tables, the dimensions of the feature maps are shown only in terms of height and width. The experimental results are presented in the tables 5 and 6, where the bold and

Table 5. Comparison of noise pattern extraction strategies on the DFFD dataset.

Strategy	Feature size	ACC (%)	AUC (%)
Global noise	16×16	98.81	<u>99.91</u>
	8×8	**98.97**	**99.93**
	4×4	<u>98.87</u>	99.85
Patch noise	16×16	96.07	99.07
	8×8	96.18	99.17
	4×4	96.19	99.18

Table 6. Comparison of noise pattern extraction strategies on the CIFAKE dataset.

Strategy	Feature size	ACC (%)	AUC (%)
Global noise	16×16	95.62	98.99
	8×8	**96.06**	**99.10**
	4×4	<u>95.97</u>	<u>99.16</u>
Patch noise	16×16	84.76	92.38
	8×8	84.62	92.08
	4×4	85.08	92.57

underlined values indicate the first and second best results in each column, respectively.

The experimental results show that, when the feature map size is the same, the global noise pattern analysis outperforms the local noise pattern analysis in terms of both ACC and AUC on both datasets. On the DFFD dataset, the global noise analysis method achieved an ACC of 98.97% and an AUC of 99.93% at the optimal resolution ($64 \times 8 \times 8$), which is an improvement of 2.78 and 0.75 percentage points over the best local noise method (the ACC = 96.19%, the AUC = 99.18%).

It is worth noting that the advantage of global noise pattern analysis is more pronounced in the CIFAKE dataset. At the same optimal resolution ($64 \times 8 \times 8$), this method achieved an ACC of 96.06% and an AUC of 99.10%, far surpassing the local method with the same feature map size (the ACC = 84.62%, the AUC = 92.08%). Even compared to the best local noise method (the ACC = 85.08%, the AUC = 92.57%), the ACC still improved by 10.98 percentage points, demonstrating a significant performance advantage. Further analysis of the impact of feature map size on detection performance revealed that the global noise analysis method achieves the best performance at the resolution of ($64 \times 8 \times 8$), which is why all subsequent ablation experiments are conducted using this feature map size. This result may be attributed to the ability of the global method to capture the long-range dependencies of image noise patterns, whereas the local method, due to information isolation between image patches, struggles to fully leverage the global noise information. These results provide quantitative evidence that global noise analysis offers a clear performance edge in AI-generated image detection.

Table 7. Ablation experiment results of feature extraction modules on the DFFD dataset.

GNFEM	CFEM	TFEM	ACC (%)	AUC (%)
✓	×	×	98.97	99.93
×	✓	×	98.35	99.82
×	×	✓	96.89	99.54
✓	✓	×	<u>99.17</u>	<u>99.94</u>
✓	×	✓	99.02	99.89
×	✓	✓	98.31	99.87
✓	✓	✓	**99.24**	**99.95**

Table 8. Ablation experiment results of feature extraction modules on the CIFAKE dataset.

GNFEM	CFEM	TFEM	ACC (%)	AUC (%)
✓	×	×	96.06	99.10
×	✓	×	96.99	99.47
×	×	✓	92.93	97.92
✓	✓	×	<u>97.57</u>	<u>99.64</u>
✓	×	✓	96.20	99.27
×	✓	✓	97.30	99.59
✓	✓	✓	**97.60**	**99.66**

4.4.2. Verifying the effectiveness of each feature extraction module. To systematically evaluate the independent contributions and synergistic effects of the GNFEM, the CFEM, and the TFEM feature extraction modules, multiple ablation experiments are conducted on the DFFD and CIFAKE datasets. The experimental results are shown in tables 7 and 8. In all experiments, the feature maps generated by each feature extraction module are kept consistent, and other network parameters are kept unchanged to eliminate the influence of confounding factors.

The experimental results show that on both the DFFD and CIFAKE datasets, the model achieves the best ACC and AUC values when all three sub-feature extraction modules (the GNFEM, the CFEM,the TFEM) are combined. Among them, the global noise pattern extracted by the GNFEM demonstrates excellent detection capabilities on both datasets. When using the GNFEM alone, its ACC value is typically the highest or second highest, proving the critical role of global noise features in high-precision generated image detection. For example, on the DFFD dataset, using the GNFEM alone yields an ACC of 98.97% and an AUC of 99.93%; on the CIFAKE dataset, the ACC reaches 96.06% and the AUC is 99.10%.

Furthermore, on the CIFAKE dataset, the CFEM demonstrates the strongest independent detection capability (ACC = 96.99%, AUC = 99.47%), which fully validates its effectiveness in enriching color feature information and enhancing the expression of color features. The CFEM module is able to capture local inconsistencies within the color features more comprehensively, providing strong support for the detection of AI-generated images. Meanwhile, although the independent detection performance of the TFEM is relatively lower

Table 9. Ablation experiment results of the SAFM's number of heads on the DFFD dataset.

num_heads	ACC (%)	AUC (%)
4	99.07	99.96
8	99.14	**99.96**
16	**99.24**	99.95
32	99.20	99.95

Table 10. Ablation experiment results of the SAFM's number of heads on the CIFAKE dataset.

num_heads	ACC (%)	AUC (%)
4	97.46	99.65
8	97.41	99.63
16	97.47	**99.67**
32	**97.60**	99.66

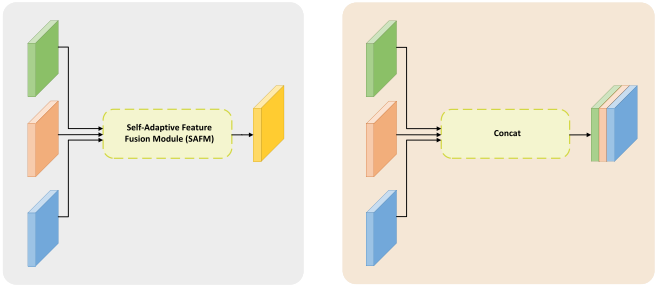
(CIFAKE: ACC = 92.93%, AUC = 97.92%), when combined with the TFEM, the model achieves the highest ACC and AUC in the three-feature fusion detection. This result indicates that the TFEM plays an irreplaceable and important role in capturing micro-texture anomalies in generated images.

Ultimately, the combination of the three feature extraction modules further enhances the overall detection capability of the model, allowing it to achieve the best performance. Specifically, on the DFFD dataset, when all features are combined, the ACC reaches 99.24% and the AUC is 99.95%; on the CIFAKE dataset, the ACC increases to 97.60% and the AUC reaches 99.66%. This indicates that there is a significant complementary effect between the global noise pattern, color features, and texture features, which together form the key basis for detecting fake images, thereby further improving the robustness and ACC of the detection system.

4.4.3. Validating the optimal number of heads in SAFM. To determine the optimal number of heads (num_heads) in the SAFM, ablation experiments are conducted on the DFFD and CIFAKE datasets. The specific experimental results are shown in tables 9 and 10.

On the DFFD dataset, the ACC and AUC values for different numbers of attention heads are compared. The experimental results show that when num_heads is set to 16, the highest ACC (99.24%) and a high AUC (99.95%) are achieved. Although the AUC value at num_heads is 32 is the same as that at 16, the ACC is slightly lower (99.20%). Therefore, on the DFFD dataset, the best performance is achieved with 16 attention heads.

On the CIFAKE dataset, the model performs best with num_heads is 32, achieving an ACC of 97.60% and an AUC of 99.66%. Although the AUC is slightly higher (99.67%) when num_heads is 16, the ACC is lower (97.47%). Considering both ACC and AUC, the best choice on the CIFAKE dataset is 32 attention heads.



(a) (b)

Figure 3. Schematic diagram of the SAFM and Concat fusion. (a) The SAFM method. (b) The channel concatenation method.

Table 11. Comparison of SAFM and concat fusion methods on DFFD and CIFAKE datasets.

Dataset	Fusion method	ACC (%)	AUC (%)
DFFD	Concat	99.15	99.89
	SAFM	**99.24**	**99.95**
CIFAKE	Concat	97.56	99.64
	SAFM	**97.60**	**99.66**

The experimental results indicate that the number of attention heads in SAFM significantly impacts model performance. A reasonable number of attention heads enhances the model's ability to capture fine-grained interactions between sub-features, enabling more effective feature fusion and improving overall detection performance.

4.4.4. Verification of SAFM's effectiveness. To verify the advantages of the SAFM over traditional fusion strategies, comparative experiments are conducted on the DFFD and CIFAKE datasets. The baseline method uses conventional channel concatenation (Concat), where global noise, color, and texture features are merged along the channel axis and then passed to the classifier. The schematic diagrams of the two fusion methods are shown in figure 3, with the left side (a) illustrating fusion through the SAFM module and the right side (b) showing fusion through traditional channel concatenation. The specific experimental results are presented in table 11.

The experimental results show that, compared to the traditional channel concatenation fusion method, SAFM achieves improvements in both the ACC and the AUC. On the DFFD dataset, even when the model performance is close to saturation (ACC $>$ 99%), SAFM still manages to further enhance both the ACC and the AUC, fully demonstrating its effectiveness in capturing and fusing complementary information from different sub-features. It also has the ability to adaptively adjust the weight distribution of sub-features, optimizing overall detection performance. On the CIFAKE dataset, due to the inclusion of high-fidelity images generated by diffusion models, the original resolution is relatively low, and the differences between real and fake samples are very subtle, making

Table 12. Comparison of our method and other approaches in cross-dataset experiments.

Method	DFFD to CIFAKE		CIFAKE to DFFD		AVG (%)	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Vision Transform-T (2020)	47.89	43.98	44.13	50.42	46.01	47.20
Swin Transform-T (2021)	46.05	40.50	67.69	50.58	56.87	45.54
CovNexT-B (2022)	43.11	38.11	71.57	47.08	57.34	42.60
DeiT-B (2020)	45.52	39.92	73.48	41.45	59.50	40.69
AIDE (2024)	46.97	37.06	39.59	46.90	43.28	41.98
FreDect (2020)	42.14	42.19	42.13	**66.19**	42.14	54.19
GramNet (2020)	50.02	43.11	75.35	45.10	62.69	44.11
LGrad (2023)	37.43	32.81	47.95	38.72	42.69	35.77
Ours	**52.72**	**51.32**	**76.52**	59.54	**64.62**	**55.43**

the detection task more challenging. Although the performance improvement is relatively small in this case, the SAFM still outperforms the traditional channel concatenation fusion method, further validating its ability to capture and fuse subtle features when handling high-difficulty image data.

4.5. Cross-dataset experiments

To evaluate the model's generalization ability to unseen generative models, we conducted cross-dataset experiments by training on the DFFD dataset and testing on the CIFAKE dataset, as well as training on CIFAKE and testing on DFFD. The detailed results are presented in table 12. In the table, the first column ('DFFD to CIFAKE') represents the performance when the model is trained on DFFD and tested on CIFAKE, while the second column ('CIFAKE to DFFD') shows the results when trained on CIFAKE and tested on DFFD. The third column ('Avg') indicates the average performance across both training-testing combinations, serving as an overall measure of cross-dataset detection capability. We primarily focus on this average metric to assess the model's generalization performance in cross-dataset scenarios.

It is worth noting that, compared to the within-dataset experimental results, the performance in cross-dataset experiments shows a noticeable decline. This can be attributed to several key factors. First, there is a significant difference in image resolution: images in the DFFD dataset are generally high-resolution, whereas those in the CIFAKE dataset are extremely low-resolution (32×32 pixels), which severely limits the amount of available information and greatly constrains the model's ability to perceive fine-grained artifact features, thereby impacting its classification performance. Second, there is a fundamental difference in the generation methods between the two datasets. The DFFD primarily contains images generated by GAN-based models, while the CIFAKE mainly consists of images produced by diffusion models such as the Stable Diffusion. These different generative mechanisms lead to domain-specific differences in artifact types and texture patterns, making it difficult for the discriminative patterns learned during training to transfer effectively across datasets. Lastly, the semantic categories of the images differ entirely. The DFFD focuses on human face images, whereas the CIFAKE includes a wide variety of common object categories. This semantic-level inconsistency further increases the difficulty of cross-domain generalization.

Nevertheless, under such a challenging cross-distribution scenario, our proposed model still achieves the highest scores in both Avg ACC and Avg AUC, clearly demonstrating its strong generalization ability and robustness in the face of multiple challenges, including unseen generative models, significant resolution gaps, and semantic domain discrepancies.

4.6. Robustness analyzes

In real-world applications, images often undergo unforeseen perturbations during dissemination and interaction, posing significant challenges for the detection of AI-generated images. To evaluate the robustness of different detection methods under such potential distortions, this study adopts three common and impactful types of image degradation: JPEG compression (with quality factors $Q = 70$ and 90), Gaussian blur ($\sigma = 2.0$), and image downsampling (scaling the image to one-quarter of its original width and height, $r = 0.25$). Robustness experiments were conducted on both the DFFD and CIFAKE datasets. Tables 13 and 14 present the robustness results under intra-dataset settings (training and testing on the same dataset), while tables 15 and 16 report the results of cross-dataset robustness evaluations (training on one dataset and testing on the other). To comprehensively assess the overall robustness of each method across various perturbation scenarios, the average performance (Avg ACC and Avg AUC) across all distortion types is used as the primary evaluation metric.

In this set of experiments, our method continues to demonstrate superior performance, achieving the highest average ACC and AUC scores among all methods. Among these metrics, ACC serves as a core indicator of the model's classification capability and more directly reflects its effectiveness and reliability in practical applications. Particularly in scenarios with balanced positive and negative samples, ACC provides a more faithful reflection of the model's decision-making ability near classification boundaries, making it a more critical performance metric for real-world deployment.

Table 13. Robustness comparison of our method and other approaches trained and tested on the DFFD dataset.

	JPEG($Q = 70$)		JPEG($Q = 90$)		Blur($\sigma = 2.0$)		Downsampling($r = 0.25$)		Avg	
Method	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Vision Transform-T (2020)	94.88	98.36	95.49	98.68	94.48	98.49	93.97	98.41	94.71	98.49
Swin Transform-T (2021)	93.21	97.92	95.97	98.77	90.02	98.25	90.79	98.40	92.50	98.34
CovNexT-B (2022)	96.40	99.44	96.78	99.59	95.60	99.41	95.31	98.29	96.02	99.18
DeiT-B (2020)	96.28	99.29	96.28	99.30	96.31	99.30	96.30	99.30	96.29	99.30
AIDE (2024)	94.79	**99.97**	**97.86**	99.79	91.16	98.35	94.56	98.41	94.59	99.13
FreDect (2020)	95.96	98.08	96.26	98.72	95.99	93.81	95.24	97.81	95.86	97.11
LGrad (2023)	96.16	99.01	96.30	99.31	93.98	98.51	87.61	98.15	93.51	98.75
Ours	**97.42**	99.73	97.21	**99.81**	**97.22**	**99.65**	**96.42**	**99.32**	**97.07**	**99.63**

Table 14. Robustness comparison of our method and other approaches trained and tested on the CIFAKE dataset.

	JPEG($Q = 70$)		JPEG($Q = 90$)		Blur($\sigma = 2.0$)		Downsampling($r = 0.25$)		Avg	
Method	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Vision Transform-T (2020)	90.77	96.81	90.72	96.75	71.70	**82.70**	**71.26**	**81.31**	81.11	89.39
Swin Transform-T (2021)	94.57	98.83	94.52	98.86	64.49	78.22	65.05	78.56	79.66	88.62
CovNexT-B (2022)	96.52	99.55	96.84	99.56	62.41	77.77	66.79	80.76	80.64	89.41
AIDE (2024)	96.74	99.48	96.73	99.49	72.54	82.50	68.82	76.22	83.71	89.42
FreDect (2020)	89.13	95.91	88.65	95.74	50.21	54.75	48.94	49.89	69.23	74.07
GramNet (2020)	96.62	99.57	96.99	99.59	63.26	72.67	63.41	72.72	80.07	86.14
LGrad (2023)	93.65	98.36	93.73	98.42	50.93	50.99	51.17	52.73	72.37	75.13
Ours	**97.28**	**99.65**	**97.50**	**99.64**	**73.77**	82.55	69.46	77.98	**84.50**	**89.96**

Table 15. Cross-dataset robustness comparison of our method and other approaches trained on the DFFD dataset and tested on the CIFAKE dataset.

	JPEG($Q = 70$)		JPEG($Q = 90$)		Blur($\sigma = 2.0$)		Downsampling($r = 0.25$)		Avg	
Method	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Vision Transform-T (2020)	47.40	43.24	48.54	44.51	50.57	51.46	**51.03**	51.97	49.39	47.80
Swin Transform-T (2021)	46.06	40.99	46.85	41.27	34.51	29.63	36.13	28.98	40.89	35.22
CovNexT-B (2022)	43.63	35.53	44.46	39.22	39.91	36.02	37.61	33.55	41.40	36.08
DeiT-B (2020)	45.50	39.77	45.90	40.46	45.27	43.11	44.91	42.59	45.40	41.48
AIDE (2024)	47.13	36.00	47.12	37.00	41.31	37.15	41.78	38.55	44.34	37.18
FreDect (2020)	42.29	42.49	41.55	41.24	50.00	53.31	50.00	52.73	45.96	47.44
GramNet (2020)	49.87	43.03	50.14	43.20	50.01	43.55	50.00	43.99	50.01	43.44
LGrad (2023)	37.25	32.18	50.01	50.64	50.01	50.31	50.01	50.63	46.82	45.94
Ours	**52.74**	**51.33**	**52.61**	**51.13**	**51.36**	**54.47**	50.13	**53.72**	**51.71**	**52.66**

Table 16. Cross-dataset robustness comparison of our method and other approaches trained on the CIFAKE dataset and tested on the DFFD dataset.

	JPEG($Q = 70$)		JPEG($Q = 90$)		Blur($\sigma = 2.0$)		Downsampling($r = 0.25$)		Avg	
Method	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Vision Transform-T (2020)	44.31	50.59	44.15	50.35	63.97	**63.38**	68.11	65.11	55.14	57.36
Swin Transform-T (2021)	67.69	50.50	67.76	50.64	71.78	54.72	72.78	55.66	70.00	52.88
CovNexT-B (2022)	71.54	47.06	71.61	47.17	73.63	50.16	74.02	50.84	72.70	48.81
DeiT-B (2020)	73.48	41.40	73.50	41.54	74.85	43.15	75.14	43.48	74.24	42.39
AIDE (2024)	39.01	41.34	36.00	38.75	37.76	52.57	45.16	59.51	39.48	48.04
FreDect (2020)	39.32	**67.21**	43.28	**68.72**	28.38	31.49	25.22	**71.96**	34.05	59.85
GramNet (2020)	75.24	44.15	72.56	44.12	76.90	42.77	77.05	41.14	75.44	43.05
LGrad (2023)	33.47	52.26	51.99	48.34	33.45	52.25	61.56	54.01	45.12	51.72
Ours	**76.55**	59.88	**76.56**	59.42	**77.52**	59.91	**77.42**	60.24	**77.01**	**59.86**

Meanwhile, we introduce AUC as a supplementary metric to evaluate the overall ranking capability of the model under varying decision thresholds. Specifically, AUC measures the extent to which the model can rank positive samples ahead of negative ones. Therefore, in scenarios with imbalanced sample distributions or ambiguous class boundaries, AUC provides a useful reference for assessing the model's discriminative ability.

However, it is important to note that a high AUC does not necessarily indicate strong classification performance. Since AUC is threshold-independent, a model may still achieve a high AUC even when many samples near the decision boundary are misclassified, as long as the overall ranking is roughly correct. In particular, under the balanced positive and negative sample setting in this study, frequent misclassifications around the boundary often suggest that the learned decision function is not sufficiently sharp, making it difficult to achieve stable and accurate predictions—ultimately compromising performance in real-world applications. This issue is not reflected by the AUC metric but is clearly captured by ACC. For instance, although FreDect achieves a relatively high AUC in table 15, its ACC is significantly lower, indicating that it performs poorly near decision boundaries. This suggests that the model fails to capture fine-grained features of borderline samples and is vulnerable to noise perturbations. Poor performance on such samples directly leads to a drop in overall ACC, thereby undermining the model's reliability and applicability in practical tasks. Therefore, ACC provides a more faithful reflection of the model's real-world classification performance and should be regarded as the primary metric for evaluating model effectiveness.

In contrast, our method not only maintains strong ranking capability in terms of average AUC, but also achieves the highest performance in average ACC, indicating that the model makes more stable and reliable classification decisions. This is particularly valuable in real-world applications where tolerance for misclassification is low. Therefore, while AUC offers a comprehensive view of a model's discriminative ability, ACC remains the key metric for evaluating practical effectiveness. This highlights the core advantage of our approach.

In the robustness experiments conducted on the DFFD and CIFAKE datasets, a more noticeable performance drop was observed on the CIFAKE dataset. This is primarily due to the extremely low image resolution in CIFAKE, with each image being only 32×32 pixels, which severely limits the amount of useful information. When such low-resolution images are further degraded by operations like Gaussian blur and downsampling, critical visual details and features are significantly diminished and blurred, substantially increasing the difficulty of the detection task and consequently degrading detection performance.

In contrast, the DFFD dataset consists mostly of high-quality, high-resolution images that preserve richer texture, structural, and artifact information, making the detection task relatively easier to perform. On this dataset, our method achieves the highest scores in both average ACC and AUC, clearly demonstrating the model's superior detection capability in high-quality image scenarios.

Table 17. Comparison of our method and other approaches in terms of efficiency and performance.

Method	FLOPS (GFLOPS)	Params (M)	Latency (ms)
Vision Transform-T(2020)	16.86	85.65	<u>10.76</u>
Swin Transform-T(2021)	7.11	27.50	18.25
CovNexT-B(2022)	20.05	85.65	15.94
DeiT-B(2020)	16.86	85.68	13.82
AIDE(2024)	225.69	893.54	57.42
FreDect(2020)	5.40	23.51	17.55
LGrad(2023)	<u>5.39</u>	<u>23.51</u>	20.66
Ours	**4.10**	**12.26**	**10.50**

More importantly, under the extreme conditions of the CIFAKE dataset—characterized by low resolution and strong perturbations—our method still demonstrates a significant advantage, achieving the highest average ACC and AUC among all methods and outperforming all baseline models by a notable margin. This result indicates that our proposed approach not only performs well under ideal conditions but also exhibits strong robustness and generalization capabilities in challenging scenarios involving severe image distortion and information loss.

4.7. Analysis of model efficiency and deployability

To evaluate the model's efficiency in practical applications, we compared our model with several baseline methods in terms of the number of parameters, FLOPs (floating-point operations), and inference latency. The detailed experimental results are presented in table 17, which shows the relevant metrics for each model.

The experimental results show that our model exhibits excellent computational efficiency, with a relatively low number of parameters (12.26 M) and FLOPs (4.10 GFLOPS). In particular, when compared to baseline methods such as AIDE (2024) and ConvNeXt-T-B (2022), it demonstrates a clear computational advantage. This indicates that despite having fewer parameters, our model is still capable of maintaining strong detection performance while offering promising potential for real-world deployment.

5. Conclusion

In this paper, the MSAFNet method is proposed. The GNFEM captures global noise patterns, addressing limitations in local block analysis. The TFEM extracts subtle texture anomalies, enriching feature representation. The CFEM improves color feature expression by converting images into multiple color spaces and concatenating channels, revealing cross-channel dependencies and providing richer discriminative cues. The SAFM enables efficient sub-feature integration by optimizing fusion ratios and capturing complementary relationships among sub-features. Experimental results on the DFFD and CIFAKE datasets demonstrate that the proposed MSAFNet

achieves comparable ACC in contrast to other state-of-the-art methods.

In the future, we will focus on developing theoretically grounded modeling mechanisms for AI-generated image forgery fingerprints, and designing more efficient multi-scale feature fusion strategies, thereby further enhancing the model's generalization ability and interpretability in complex scenarios.

Data availability statement

The data cannot be made publicly available upon publication because they are not available in a format that is sufficiently accessible or reusable by other researchers. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272290, and in part by the Xinjiang Natural Science Foundation under Grant 2022D01A236. The authors would like to express their sincere gratitude for this support, which has significantly contributed to the completion of this research.

ORCID iDs

Liwei Yao 0009-0009-4170-5107
Sen Niu 0000-0002-6259-5463

References

[1] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Proc. 28th Int. Conf. Neural Information Processing Systems - Volume 2 (NIPS'14)* (MIT Press) pp 2672–80
[2] Kingma D P and Welling M 2022 Auto-encoding variational bayes (arXiv:1312.6114)
[3] Kingma D P and Dhariwal P 2018 arXiv:1807.03039
[4] Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B 2022 High-resolution image synthesis with latent diffusion models *2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 10674–85
[5] Brock A, Donahue J and Simonyan K 2019 Large scale GAN training for high fidelity natural image synthesis (arXiv:1809.11096)
[6] Ho J, Jain A and Abbeel P 2020 Denoising diffusion probabilistic models (arXiv:2006.11239)
[7] Karras T, Aila T, Laine S and Lehtinen J 2018 Progressive growing of gans for improved quality, stability, and variation (arXiv:1710.10196)
[8] Karras T, Laine S and Aila T 2019 A style-based generator architecture for generative adversarial networks *2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 4396–405
[9] Song Y and Ermon S 2020 Generative modeling by estimating gradients of the data distribution (arXiv:1907.05600)
[10] Marra F, Gragnaniello D, Cozzolino D and Verdoliva L 2018 Detection of GAN-generated fake images over social networks *2018 IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)* pp 384–9
[11] Zhang X, Karaman S and Chang S F 2019 Detecting and simulating artifacts in gan fake images *2019 IEEE Int. Workshop on Information Forensics and Security (WIFS)* pp 1–6
[12] Durall R, Keuper M and Keuper J 2020 Watch your up-convolution: cnn based generative deep neural networks are failing to reproduce spectral distributions *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*
[13] Dzanic T, Shah K and Witherden F 2020 Fourier spectrum discrepancies in deep network generated images *Advances in Neural Information Processing Systems* vol 33, ed H Larochelle, M Ranzato, R Hadsell, M Balcan and H Lin (Curran Associates, Inc.) pp 3022–32 (available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/1f8d87e1161af68b81bace188a1ec624-paper.pdf)
[14] Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D and Holz T 2020 Leveraging frequency analysis for deep fake image recognition *Proc. 37th Int. Conf. Machine Learning (ICML'20)* (JMLR.org)
[15] Frank J C, Eisenhofer T, Schönherr L, Fischer A, Kolossa D and Holz T 2020 arXiv:2003.08685
[16] Liu Z, Qi X and Torr P H 2020 Global texture enhancement for fake face detection in the wild *2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 8057–66
[17] Chandrasegaran K, Tran N T, Binder A and Cheung N M 2022 Discovering transferable forensic features for cnn-generated images detection *Computer Vision — ECCV 2022: 17th European Conf., (Tel Aviv, Israel, 23 October–27 October 2022, Proc., Part XV* (Springer) pp 671–89 (available at: https://doi.org/10.1007/978-3-031-19784-0_39)
[18] Ojha U, Li Y and Lee Y J 2023 Towards universal fake image detectors that generalize across generative models *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 24480–9
[19] 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 *Proc. 2023 ACM SIGSAC Conf. Computer and Communications Security (CCS'23)* (Association for Computing Machinery) pp 3418–32 (available at: https://doi.org/10.1145/3576915.3616588)
[20] Wang Z, Bao J, Zhou W, Wang W, Hu H, Chen H and Li H 2023 Dire for diffusion-generated image detection *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)* pp 22445–55
[21] Park T, Liu M Y, Wang T C and Zhu J Y 2019 Semantic image synthesis with spatially-adaptive normalization *2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 2332–41
[22] Zhu J Y, Park T, Isola P and Efros A A 2017 Unpaired image-to-image translation using cycle-consistent adversarial networks *2017 IEEE Int. Conf. Computer Vision (ICCV)* pp 2242–51
[23] Dhariwal P and Nichol A 2021 Diffusion models beat gans on image synthesis *Proc. 35th Int. Conf. Neural Information Processing Systems (NIPS'21)* (Curran Associates Inc.)
[24] 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 *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 10696–706
[25] Nichol A Q, Dhariwal P, Ramesh A, Shyam P, Mishkin P, Mcgrew B, Sutskever I and Chen M 2022 GLIDE: towards photorealistic image generation and editing with text-guided diffusion models *Proc. 39th Int. Conf. Machine Learning* (*Proc. Machine Learning Research*) vol 162, ed K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and

S Sabato (PMLR) pp 16784–804 (available at: https://proceedings.mlr.press/v162/nichol22a.html)
[26] Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B 2022 High-resolution image synthesis with latent diffusion models *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 10684–95
[27] Chen J, Yao J and Niu L 2024 A single simple patch is all you need for AI-generated image detection (arXiv:2402.01123)
[28] Yan S, Li O, Cai J, Hao Y, Jiang X, Hu Y and Xie W 2025 A sanity check for AI-generated image detection (arXiv:2406.19435)
[29] Cavia B, Horwitz E, Reiss T and Hoshen Y 2024 Real-time deepfake detection in the real-world (arXiv:2406.09398)
[30] Li J, Jiang W, Shen L and Ren Y 2025 Optimized Frequency Collaborative Strategy Drives AI Image Detection *IEEE Internet Things J.* **12** 16192–203
[31] 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 *Advances in Neural Information Processing Systems* vol 37, ed A Globerson, L Mackey, D Belgrave, A Fan, U Paquet, J Tomczak and C Zhang (Curran Associates, Inc.) pp 59570–96 (available at: https://proceedings.neurips.cc/paper_files/paper/2024/file/6dddcff5b115b40c998a08fbd1cea4d7-paper-Conference.pdf)
[32] Gye S, Ko J, Shon H, Kwon M and Kim J K 2025 Reducing the content bias for AI-generated image detection *IEEE/CVF Winter Conf. Applications of Computer Vision, WACV 2025, (Tucson, AZ, USA, 26 February–6 March 2025)* (IEEE) pp 399–408 (available at: https://doi.org/10.1109/WACV61041.2025.00049)
[33] Fridrich J and Kodovsky J 2012 Rich Models for Steganalysis of Digital Images *IEEE Trans. Inf. Forensics Security* **7** 868–82
[34] Cai Y, Li L, Wang D, Li X and Liu X 2023 HTMatch: an efficient hybrid transformer based graph neural network for local feature matching *Signal Process.* **204** 108859
[35] Zhang S and Zhang Y 2024 Multi-layer feature fusion and attention enhancement for fine-grained vehicle recognition research *Meas. Sci. Technol.* **36** 015012
[36] Lai Y and Liu B 2024 Research on road crack segmentation based on deep convolution and transformer with multi-branch feature fusion *Meas. Sci. Technol.* **35** 115017
[37] Xie H, Ni J, Zhang J, Zhang W and Huang J 2022 Evading generated-image detectors: a deep dithering approach *Signal Process.* **197** 108558
[38] McCloskey S and Albright M 2019 Detecting GAN-generated imagery using saturation cues *2019 IEEE Int. Conf. Image Processing (ICIP)* pp 4584–8
[39] Ricker J, Damm S, Holz T and Fischer A 2022 arXiv:2210.14571
[40] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I 2021 Learning transferable visual models from natural language supervision (arXiv:2103.00020)
[41] Zhong N, Xu Y, Li S, Qian Z and Zhang X 2024 PatchCraft: Exploring texture patch for efficient AI-generated image detection (arXiv:2311.12397)
[42] Wang K, Zhu Y, Chang Q, Wang J and Yao Y 2025 High-accuracy image steganography with invertible neural network and generative adversarial network *Signal Process.* **234** 109988
[43] Li Q, Tan S, Li B and Huang J 2025 Elastic supernet with dynamic training for JPEG steganalysis *Signal Process.* **236** 110038
[44] Chen Z, Zhao Y and Ni R 2017 Detection of operation chain: JPEG-resampling-JPEG *Signal Process., Image Commun.* **57** 8–20
[45] Li H, Li B, Tan S and Huang J 2020 Identification of deep network generated images using disparities in color components *Signal Process.* **174** 107616
[46] Lanzino R, Fontana F, Diko A, Marini M R and Cinque L 2024 Faster than lies: real-time deepfake detection using binary neural networks *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) Workshops* pp 3771–80
[47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2023 Attention is all you need (arXiv:1706.03762)
[48] Zhang Z, Yang Y and Jian X 2025 MASNet: a novel deep learning approach for enhanced detection of small targets in complex scenarios *Meas. Sci. Technol.* **36** 045402
[49] Tao H, Huang Z, Wang Y, Qiu J and Vladimir S 2025 Efficient feature fusion network for small objects detection of traffic signs based on cross-dimensional and dual-domain information *Meas. Sci. Technol.* **36** 035004
[50] Hendrycks D and Gimpel K 2023 Gaussian error linear units (GELUs) (arXiv:1606.08415)
[51] Dang H, Liu F, Stehouwer J, Liu X and Jain A K 2020 On the detection of digital face manipulation *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*
[52] Bird J J and Lotfi A 2024 CIFAKE: image Classification and Explainable Identification of AI-Generated Synthetic Images *IEEE Access* **12** 15642–50
[53] Krizhevsky A and Hinton G 2009 Learning multiple layers of features from tiny images *Technical Report* (available at: www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf)
[54] Liu Z, Luo P, Wang X and Tang X 2015 Deep learning face attributes the wild *2015 IEEE Int. Conf. Computer Vision (ICCV)* pp 3730–8
[55] 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 (arXiv:1711.09020)
[56] Yang Z, Liang J, Xu Y, Zhang X-Y and He R 2023 Masked Relation Learning for DeepFake Detection *IEEE Trans. on Information Forensics and Security* **18** 1696–708
[57] Yang S, Guo H, Hu S, Zhu B, Fu Y, Lyu S, Wu X and Wang X 2024 CrossDF: improving cross-domain deepfake detection with deep information decomposition (arXiv:2310.00359)
[58] Yu Y, Zhao X, Ni R, Yang S, Zhao Y and Kot A C 2023 Augmented Multi-Scale Spatiotemporal Inconsistency Magnifier for Generalized DeepFake Detection *IEEE Trans. on Multimedia* **25** 8487–98
[59] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N 2021 An image is worth 16×16 words: transformers for image recognition at scale (arXiv:2010.11929)
[60] 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 (arXiv:2103.14030)
[61] Liu Z, Mao H, Wu C Y, Feichtenhofer C, Darrell T and Xie S 2022 A convnet for the 2020s *2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 11966–76
[62] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A and Jégou H 2021 Training data-efficient image transformers and distillation through attention (arXiv:2012.12877)
[63] Tan C, Zhao Y, Wei S, Gu G and Wei Y 2023 Learning on gradients: generalized artifacts representation for GAN-generated images detection *2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* pp 12105–14