Computer Science Review 60 (2026) 100908

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Methods and trends in detecting AI-generated images: A comprehensive review☆

Arpan Mahara*, Naphtali Rishe

Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, Florida, USA

ARTICLE INFO

Keywords:
Forensic image detection
Fingerprint analysis
Generative adversarial networks
Diffusion models
Frequency domain analysis
Vision-language models (VLMs)
Patch-based detection
Cross-model generalization
Cross-domain detection
Cross-scene analysis
Synthetic image identification

ABSTRACT

The proliferation of generative models, such as Generative Adversarial Networks (GANs), Diffusion Models, and Variational Autoencoders (VAEs), has enabled the synthesis of high-quality multimedia data. However, these advancements have also raised significant concerns regarding adversarial attacks, unethical usage, and societal harm. Recognizing these challenges, researchers have increasingly focused on developing methodologies to detect synthesized data effectively, aiming to mitigate potential risks. Prior reviews have predominantly focused on deepfake detection and often overlook recent advancements in synthetic image forensics, particularly approaches that incorporate multimodal frameworks, reasoning-based detection, and training-free methodologies. To bridge this gap, this survey provides a comprehensive and up-to-date review of state-of-the-art techniques for detecting and classifying synthetic images generated by advanced generative AI models. The review systematically examines core detection paradigms, categorizes them into spatial-domain, frequency-domain, fingerprint-based, patch-based, training-free, and multimodal reasoning-based frameworks, and offers concise descriptions of their underlying principles. We further provide detailed comparative analyses of these methods on publicly available datasets to assess their generalizability, robustness, and interpretability. Finally, the survey highlights open challenges and future directions, emphasizing the potential of hybrid frameworks that combine the efficiency of training-free approaches with the semantic reasoning of multimodal models to advance trustworthy and explainable synthetic image forensics.

1. Introduction

The advent of advanced generative models has enabled the creation of highly realistic synthetic images. These images are synthesized through various approaches, including conditional methods such as image-to-image translation and text-to-image translation, as well as unconditional generation. Generative models based on Generative Adversarial Networks (GANs) [31], Diffusion Models [36,79], Variational Autoencoders (VAEs) [43], and Autoregressive Models [83,84] dominate the current literature on image generation. Fig. 1 provides a simple illustration of the architectures of these model families. While these four generative model families—GANs, Diffusion Models, VAEs, and Autoregressive—have dominated the landscape of image generation, other methods such as Normalizing Flows [72] have also made significant contributions and deserve attention. With these advancements, several state-of-the-art generative models have been made publicly available alongside commercially accessible tools such as Adobe Firefly [1], MidJourney [60], DALL.E 3 [64], and Imagen 3 [32].

Although generative models have enabled advancements in image synthesis, their accessibility introduces critical concerns related to misinformation, privacy, and security. The ability to generate highly realistic AI-synthesized images raises ethical and security risks, including deceptive media [76], identity fraud [96], and geopolitical manipulation [61]. Consequently, forensic detection has become an essential area of research, aiming to reliably distinguish AI-generated imagery from real-world photographs. Despite significant progress, existing detection techniques face challenges in generalizability, robustness, scalability, and reasoning across evolving generative architectures. The detection of images generated by both conditional and unconditional generative methods has emerged as a rapidly growing research area. Despite this increasing interest, a comprehensive review that consolidates and

☆ This work is based in part upon work supported by the National Science Foundation under Grant Nos. MRI20 CNS-2018611 and MRI CNS-1920182.
* Corresponding author.
Email addresses: amaha038@fiu.edu (A. Mahara), rishen@cs.fiu.edu (N. Rishe).

https://doi.org/10.1016/j.cosrev.2026.100908
Received 2 April 2025; Received in revised form 25 November 2025; Accepted 13 January 2026
Available online 21 January 2026

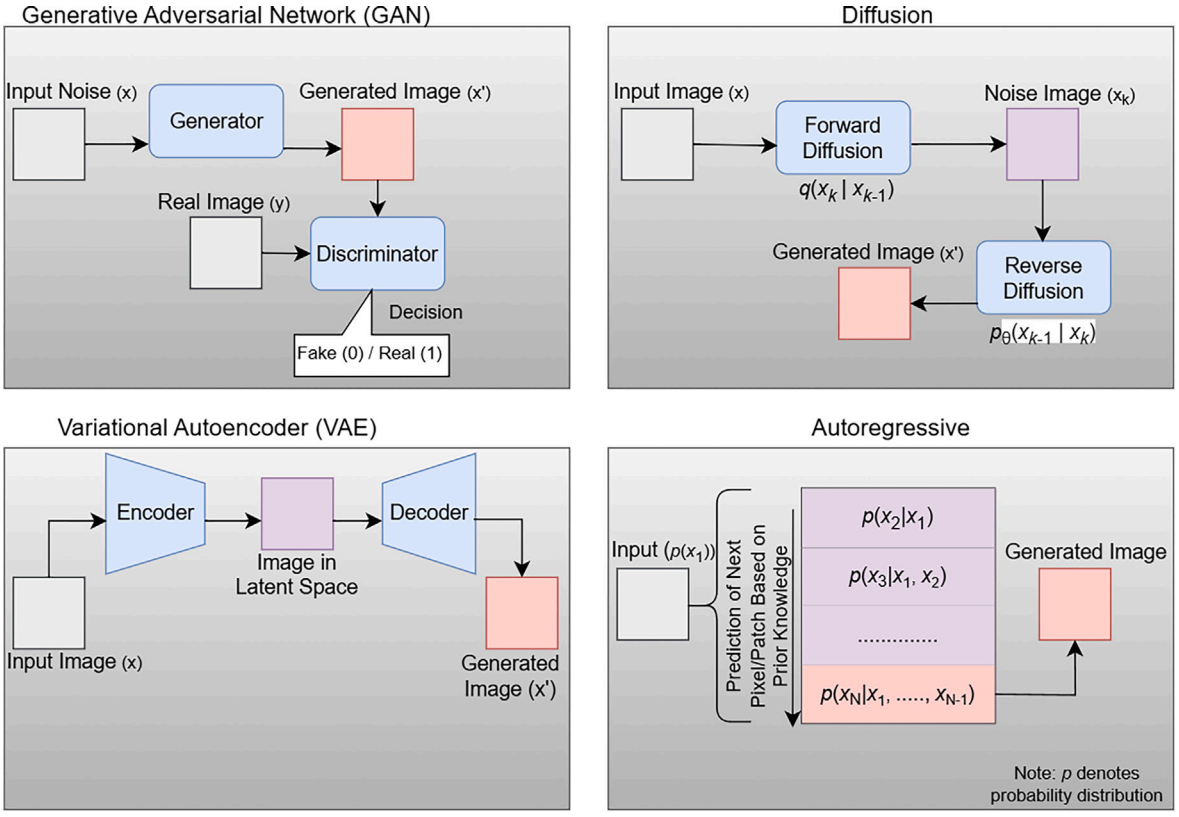

Fig. 1. Illustration of simplified architecture of GAN, diffusion, VAE and autoregressive generative models.

systematically categorizes existing detection methodologies is still lacking. While prior surveys have presented reviews on the detection of image manipulations achieved through traditional computer-based techniques [78], image editing tools [98], and deepfake methodologies [88], our focus diverges from these areas and addresses the limitations in this domain. This survey fills this gap by providing a structured analysis of detection approaches, classifying them based on their underlying techniques and contributions, and focusing on recent advances that incorporate multimodal techniques, reasoning with large language models, and training-free detectors for enhanced forensic accuracy. Given the increasing relevance of such methods in forensic image analysis, it is crucial to explore their long-term applicability and effectiveness in generative image detection.

To establish a unified framework, this survey introduces a core taxonomy that organizes artificial intelligence-generated image (AIGI) detection methodologies into seven categories: spatial-domain analysis, frequency-domain analysis, fingerprint analysis, patch-based analysis, training-free methods, multimodal and reasoning-based models, and commercial detection frameworks. Each category represents a distinct perspective on identifying AI-generated imagery and serves as the foundation for the subsequent sections of this paper. By examining the evolving landscape of adversarial threats in generative AI, this study highlights key challenges, methodological trends, and emerging opportunities, reinforcing the need for continued research in forensic detection of AIGI.

2. Reviews

In the present survey of detection methods for AI-generated images, we categorize them into seven distinct groups, with the final category comprising commercial methods. For each group, we discuss the core proposal of each method in ascending order based on publication date. At the end of each section, we present a table summarizing whether the experimental evaluations of the methods satisfy three key criteria, described as follows:

1. Cross-Family Generators: This criterion evaluates whether a detection method, trained on images from one type of generative model (e.g., GAN), is tested on and demonstrates effectiveness in detecting images generated by a different type of generative model (e.g., Diffusion models). A method that evaluates images from multiple generative model types satisfies the cross-family criterion.
2. Cross-Category: This criterion examines whether a method was trained and tested on images belonging to different classes. For example, a detector trained on human face images and evaluated on its ability to detect generated images of animals satisfies the cross-category requirement.
3. Cross-Scene: This criterion assesses whether a method's performance was tested across datasets with distinct data distributions. For instance, a detector trained on images from the ImageNet [75] dataset and evaluated on images from the LSUN-Bedroom [95] dataset satisfies the cross-scene requirement. Importantly, methods that meet the cross-scene criterion typically also satisfy the cross-category requirement, although the converse is not necessarily true.

It is worth mentioning that these criteria are not extreme cases and do not imply that such evaluations are impossible for any given method. Instead, they are based on the descriptions and experimental setups reported in the original papers. The goal is to assist future work in considering these criteria to demonstrate generalizability and provide real-world evidence (Fig. 2)

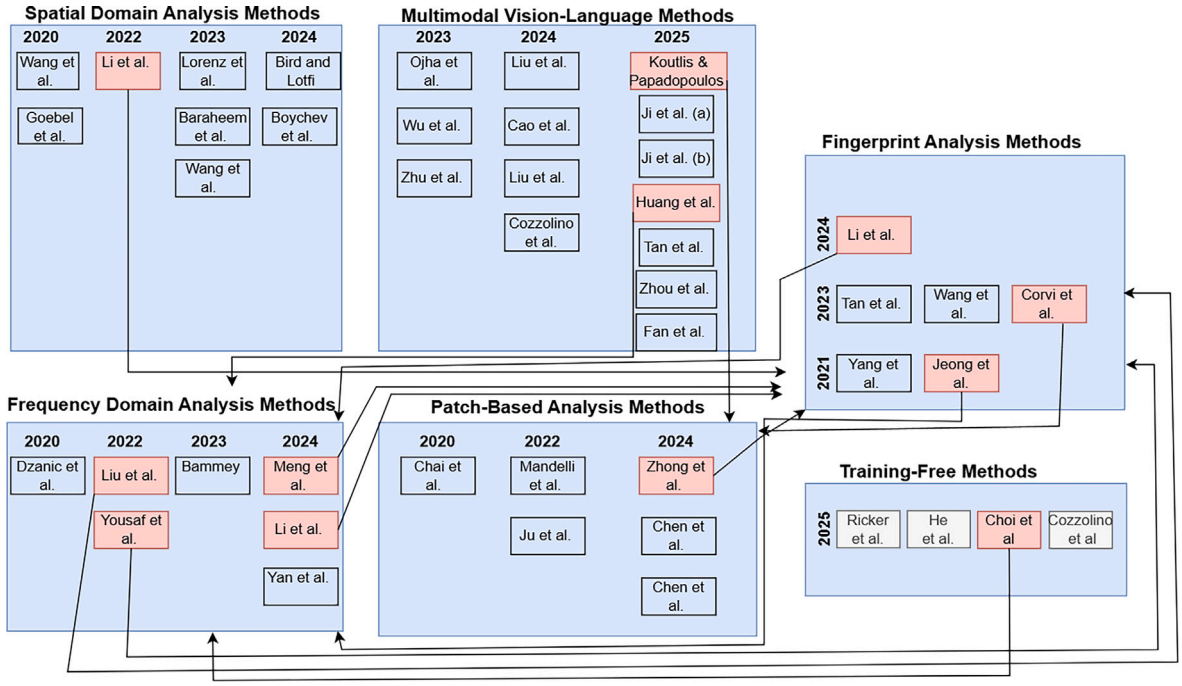

Fig. 2. Categorization of detection methods based on core architecture and methodological proposals. The figure illustrates the division of detection methods into categories. Each category is highlighted using blue-colored rectangles. Some methods are connected to multiple categories, shown using light-red highlights and arrows pointing to the respective sub-categories. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.1. Spatial domain analysis / spatial feature analysis methods

Spatial domain analysis methods focus on detecting generated images by analyzing intrinsic features within the pixel-level spatial representation. These methods leverage spatial patterns such as texture irregularities, unnatural edge formations, and color inconsistencies, which are common artifacts introduced during the image synthesis process. By examining spatial features like intensity gradients, local pixel dependencies, and edge sharpness, these approaches can effectively uncover subtle distortions. Techniques often employed include convolutional neural networks (CNNs) for automated feature extraction, statistical analysis of pixel intensity distributions, and handcrafted feature-based classifiers. Spatial domain methods excel in capturing localized anomalies and abrupt visual transitions, providing a robust and interpretable approach for image forgery detection.

2.1.1. CNN-generated images are surprisingly easy to spot: wang et al. (2020)

Wang et al. [87] conducted a pivotal study demonstrating that a detection method trained on images generated by a single generative model, particularly a GAN, can generalize to detect synthetic images from a variety of unseen CNN-based generative models. This finding challenges the previously established view that cross-model generalization is inherently difficult for forensic classifiers. To evaluate this, the authors trained ProGAN [40] on the LSUN dataset [95], producing a dataset of 720K images for training and 4K for validation, with an equal split between real and generated images. They employed a ResNet-50 [35], pretrained on ImageNet, as a binary classifier. Robust feature learning was facilitated through various data augmentation strategies, including: (a) no augmentation, (b) Gaussian blur, (c) JPEG compression, and (d) combinations of both with varying probabilities (50% and 10%). The trained classifier demonstrated strong generalization capabilities, successfully detecting images synthesized by other prominent generative models such as StyleGAN [41], BigGAN [8], CycleGAN [100], StarGAN [17], and GauGAN [66]. This work highlights the existence of common artifacts across CNN-generated images, suggesting that classifiers can leverage these shared patterns for generalizable detection across different architectures and tasks.

2.1.2. Detection and attribution of GAN images: goebel et al. (2020)

Goebel et al. [30] introduced a comprehensive framework to detect, attribute, and localize GAN-generated images through co-occurrence matrix analysis and deep learning. This method builds on insights from steganalysis, leveraging pixel-level co-occurrence features to identify artifacts introduced during image synthesis. The approach begins by computing co-occurrence matrices from the RGB channels of the input image in four orientations: horizontal, vertical, diagonal, and anti-diagonal. Each co-occurrence matrix captures a 256×256 histogram of pixel value pairs, normalized and stacked into a tensor of size $256 \times 256 \times 12$. The matrices are defined as:

$$C_{i,j} = \sum_{m,n} \begin{cases} 1, & \text{if } I[m,n] = i \text{ and } I[m+1,n] = j \\ 0, & \text{otherwise.} \end{cases}$$

These features are then processed using a modified XceptionNet [18], designed for three key tasks: 1. Binary Detection: Classifying images as real or GAN-generated. 2. Multi-Class Attribution: Identifying the GAN architecture (e.g., ProGAN, CycleGAN). 3. Localization: Generating heatmaps to identify manipulated regions through patch-based analysis.

Extensive experiments conducted on over 2.76 million images demonstrated the effectiveness of this method across various GAN models, including ProGAN, StyleGAN, CycleGAN, StarGAN, and SPADE/GauGAN. Additionally, t-SNE [84] visualizations showed a clear separation between real and GAN-generated images, reinforcing the model's interpretability and robustness against varying JPEG compression levels and patch sizes. This framework advances GAN forensics by integrating detection, attribution, and localization into a unified pipeline.

2.1.3. Estimating artifact similarity with representation learning: Li et al. (2022), GASE-Net

Li et al. [48] introduced GASE-Net, a framework designed to detect GAN-generated images by estimating artifact similarity. This method addresses challenges in cross-domain generalization and robustness against post-processing, using a two-stage approach: representation learning and representation comparison. In the representation learning stage, a ResNet-50 backbone is modified with instance normalization (IN) applied to the shallow layers to enhance feature extraction by filtering out instance-specific biases. This ensures that the learned representations remain invariant across different domains while retaining category-level distinctions. Feature maps from reference images are aggregated element-wise to form domain prototypes, which serve as robust representations of GAN or pristine image domains.

In the representation comparison stage, the feature map of a consult (suspicious) image is concatenated with the domain prototypes along the channel dimension. A shallow CNN processes the concatenated tensor to output similarity scores. The network is optimized using a Category and Domain-Aware (CDA) loss, which maximizes inter-class separation and minimizes intra-class variation by leveraging both domain and category information. The ground truth similarity scores v_{true} for optimization are defined as:

$$\hat{s}_n = \begin{cases} 1, & \text{if } y^* = y_n, \\ 0, & \text{if } y^* \neq y_n, \end{cases}$$

where y^* and y_n denote the category labels of the consult image and the n-th domain prototype, respectively. The predicted similarity scores v_{pred} are optimized against v_{true} using Mean Square Error (MSE) loss.

During inference, similarity scores are averaged across GAN-generated and pristine domains. An image is classified as GAN-generated if the average fake score exceeds the pristine score. Extensive experiments demonstrate that GASE-Net outperforms state-of-the-art methods in cross-domain scenarios, preserving resilience against various post-processing techniques, including JPEG compression, Gaussian blur, and resizing.

2.1.4. Local intrinsic dimensionality analysis: Lorenz et al. (2023), AdaptedMultiLID

Lorenz et al. [54] proposed a framework utilizing the Multi-Local Intrinsic Dimensionality (multiLID) method for detecting diffusion-generated images. This approach builds upon the earlier work [55] and demonstrates strong performance in distinguishing both diffusion and GAN-generated images. The method starts by extracting low-dimensional feature maps using an untrained ResNet18 model. MultiLID is then computed to capture local growth rates of feature densities within the latent space. The multiLID feature vector for each sample x is defined as:

$$\text{multiLID}_d(x)[i] = -\log\left(\frac{d_i(x)}{d_k(x)}\right),$$

where $d_i(x)$ and $d_k(x)$ represent the Euclidean distances to the i^{th} and k^{th} nearest neighbors, respectively. A random forest classifier is trained on the labeled multiLID scores to perform image classification. Extensive experiments validate the effectiveness of this method, with high detection accuracy achieved for both diffusion and GAN-generated images across multiple datasets. The framework is also resilient to post-processing operations such as JPEG compression and Gaussian blur, making it suitable for real-world detection scenarios.

2.1.5. AI-generated image detection: Baraheem et al. (2023)

Baraheem et al. [4] proposed a framework to detect GAN-generated images using transfer learning on pretrained classifiers. The authors compiled a diverse dataset called Real and Synthetic Images (RSI), consisting of 48,000 images synthesized by 12 GAN architectures across tasks such as image-to-image, sketch-to-image, and text-to-image generation. EfficientNetB4 [82] achieved the best detection performance after fine-tuning on the RSI dataset. The model's architecture was modified by replacing the classifier head with layers for global average pooling, batch normalization, ReLU activation, dropout, and a sigmoid output. The training utilized the Adam optimizer with a batch size of 64, an initial learning rate of 0.001, and data augmentation techniques such as horizontal flips. To facilitate model explainability, the authors incorporated four Class Activation Map (CAM) methods, GradCAM [77], AblationCAM [69], LayerCAM [38], and Faster ScoreCAM [86], to visualize the discriminative regions influencing classification decisions.

2.1.6. Diffusion reconstruction error (DIRE): Wang et al. (2023)

Wang et al. [89] proposed Diffusion Reconstruction Error (DIRE), a novel method to detect diffusion-generated images by leveraging reconstruction errors from pre-trained diffusion models. The method addresses the limitations of previous detectors, which struggled to generalize across different diffusion models. The key idea is that diffusion-generated images can be reconstructed more accurately by the pre-trained DDIM model [80] than real images. DIRE is defined as the L_1-norm of the difference between the input image x_0 and its reconstructed counterpart x_0':

$$\text{DIRE}(x_0) = \|x_0 - x_0'\|_1.$$

The process involves applying forward noise to the input image and then performing a reverse denoising process to generate the reconstruction. A ResNet-50 classifier is trained using binary cross-entropy loss on these DIRE representations to distinguish between real and generated images. The authors demonstrated that DIRE achieves state-of-the-art performance on their proposed DiffusionForensics dataset, which includes images generated by various diffusion models across multiple domains (e.g., LSUN-Bedroom [95], ImageNet [75], and CelebA-HQ [40]). Extensive experiments showed that DIRE not only excels in detecting diffusion-generated images but also generalizes well to unseen models and maintains robustness under perturbations like Gaussian blur and JPEG compression.

2.1.7. Classification and explainable identification: Bird and Lotfi (2024)

Bird and Lotfi [6] introduced a framework for detecting AI-generated images using a large-scale dataset, CIFAKE, which is detailed in the datasets section. Their classification approach employs a Convolutional Neural Network (CNN) that processes images through stacked convolutional and pooling layers, followed by fully connected layers with a final sigmoid activation for binary classification. As seen in [4], the study emphasizes explainability by implementing Gradient Class Activation Mapping (Grad-CAM) [77], which generates heatmaps to highlight regions influencing the model's decisions. Grad-CAM computes importance weights α_k for each feature map A_k, producing a visual explanation as:

$$L_c^{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k A_k\right), \quad \alpha_k = \frac{1}{Z}\sum_{i,j}\frac{\partial y_c}{\partial A_k^{i,j}},$$

where Z is the spatial size of the feature map. The heatmaps reveal that the model focuses on subtle imperfections, often in the image background, to differentiate between real and synthetic images. This approach improves trust in AI-generated image detection by combining high classification performance with visual interpretability, making it a valuable contribution to computer vision and data authenticity research.

2.1.8. Self-contrastive learning on ImagiNet: Boychev et al. (2024)

Boychev et al. [7] introduce a large-scale dataset, ImagiNet, consisting of 200,000 real and synthetic images, as detailed in the datasets section. The detection framework consists of two stages: pretraining with Self-Contrastive Learning (SelfCon) and a calibration step. During the pretraining phase, the ResNet-50 [35] backbone, initialized with ImageNet weights, is paired with a sub-network that projects intermediate feature maps into a shared latent space. This setup produces two

output embeddings per input image, facilitating contrastive learning. The SelfCon loss is defined as:

$$L_{\text{SelfCon}} = \sum_{i\in A,\omega\in\Omega} -\frac{1}{|P(i)||\Omega|} \sum_{p\in P(i),\omega'\in\Omega} \log \frac{\exp(\omega(x_i)\cdot\omega'(x_p)/\tau)}{\sum_{l\in Q(i)}\exp(\omega(x_i)\cdot\omega'(x_l)/\tau)},$$

where A represents the set of anchor images in a batch, $P(i)$ denotes positive samples for anchor x_i, $Q(i)$ contains negative samples, and $\omega(x)$ is the normalized embedding. The method balances feature similarities and differences using a temperature parameter τ.

In the calibration step, the sub-network and projection heads are removed, and a multilayer perceptron (MLP) classifier is trained using cross-entropy loss on an equal number of real and synthetic images. This stage enhances robustness by fine-tuning the learned features for both detection and model identification tasks. Experimental results indicate that the framework achieves up to 0.99 AUC and 95% balanced accuracy, demonstrating robust zero-shot performance on various synthetic image benchmarks.

Comparative analysis of spatial domain analysis methods. Spatial-domain forensic methods share the unified goal of identifying AI-generated imagery through pixel-level and localized artifact analysis, yet they differ in their architectural principles and scope of generalization. Early CNN-based detectors, such as Wang et al. [87], revealed that discriminative spatial artifacts are shared across different CNN generators, demonstrating cross-model generalization using a standard ResNet-50 backbone trained solely on ProGAN data. However, these early detectors were limited in interpretability and reliance on dataset-specific features. Goebel et al. [30] addressed these issues by introducing co-occurrence matrix–based representations that capture pixel dependencies across multiple orientations, enabling both detection and attribution within a unified framework. To overcome the poor cross-domain transferability of such handcrafted representations, Li et al. [48] proposed GASE-Net, emphasizing representation learning and artifact similarity estimation between domain prototypes and consult images, which improved robustness to post-processing and domain shifts. Lorenz et al. [54] further expanded generalization beyond GANs by introducing AdaptedMultiLID, which models intrinsic feature-space dimensionality rather than explicit pixel statistics, addressing the prior limitation of model specificity. Baraheem et al. [4] contributed the RSI benchmark and demonstrated that transfer learning on large-scale diverse GAN datasets with EfficientNet and CAM visualizations enhances explainability and model generalization. To address the gap in detecting diffusion-generated content, Wang et al. [89] introduced DIRE, which integrates spatial-domain analysis with diffusion-based reconstruction, achieving higher performance on unseen diffusion architectures. More recently, Bird and Lotfi [6] and Boychev et al. [7] tackled the remaining limitations of explainability and training efficiency by incorporating Grad-CAM interpretability and self-contrastive learning (SelfCon), respectively, offering improved transparency and zero-shot generalization. Collectively, these studies illustrate a clear methodological evolution—from handcrafted co-occurrence features and CNN-based spatial detection, to representation and reconstruction learning, and finally to contrastive and explainable frameworks. The trajectory underscores the field's movement from model-specific and dataset-dependent detection toward unified, interpretable, and generalizable spatial-domain analysis for AI-generated image forensics (Table 1).

Table 1
Evaluation of cross-family generators, cross-category, and cross-scene generalization in detection models from spatial domain analysis category.

Models	Cross-family generators	Cross-category	Cross-scene
Wang et al. [87]	✗	✓	✗
Goebel et al. [30]	✗	✓	✗
Li et al. [48]	✗	✗	✗
Lorenz et al. [54]	✓	✓	✗
Baraheem et al. [4]	✗	✓	✗
Wang et al. [89]	✓	✓	✓
Bird & Lotfi [6]	✗	✓	✗
Boychev et al. [7]	✓	✓	✗

2.2. Multi-modal vision-language methods (and/or multi-modal large language models)

This section encompasses two closely related but distinct subcategories of forensic detection methods: Multimodal Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs). The first category leverages vision-language models (VLMs) (see Fig. 3 for representative architectures) trained on large-scale image–text datasets to detect generated images by aligning visual and textual features. These approaches primarily depend on a visual encoder and typically lack explicit reasoning capabilities. They detect inconsistencies in synthetic images through cross-modal embeddings, enabling robust detection across a broad range of generative models. Representative examples include CLIP [67], which belongs to the contrastive learning paradigm [15] (see Fig. 3 for categorization), and its numerous transfer-learning adaptations that distinguish real images from generated ones. The second category, developed more recently, predominantly employs Multimodal Large Language Models (MLLMs) that jointly utilize visual and textual inputs to provide not only detection but also interpretive reasoning explaining the model's decision. These methods commonly adopt recent VLM or MLLM backbones such as Qwen [104], LLaMA, LLaVA [113], and GPT [103], achieving enhanced explainability and alignment with human-understandable reasoning.

2.2.1. Universal fake image detector by Ojha (2023)

Ojha et al. [63] identified that existing fake image detectors struggle with generalization, often misclassifying images from unseen generative models as real. This limitation arises from classifiers being asymmetrically tuned to detect artifacts specific to training data. To address this, the authors propose constructing a meaningful feature space using CLIP:ViT [27,67], trained on 400 million image–text pairs. They employ two approaches: (a) mapping training images to CLIP's final layer to create feature representations, then during inference, classifying images based on cosine distance to the nearest neighbor in real and fake feature spaces, and (b) augmenting CLIP with a linear layer for binary classification. Both methods demonstrated generalization, effectively detecting synthetic images from state-of-the-art generative models.

2.2.2. Language-guided synthetic image detection by Wu (2023)

Wu et al. [90] propose a language-guided approach for detecting synthetic images by integrating image–text contrastive learning in a VLM. They reformulate the detection task as an identification problem, determining whether a query image aligns with an anchor set of text-labeled images. This method enhances generalization by aligning visual features with carefully designed textual labels such as "Real Photo," "Real Painting," "Synthetic Photo," and "Synthetic Painting." The authors found these labels more effective than simpler "real" or "fake" categories, as they account for differences in image types, such as camera-captured versus digitally created content. The proposed LASTED framework encodes images using a ResNet-50x64 vision encoder and text labels using a transformer-based text encoder. During training, both encoders generate visual and textual representations, e_v and e_t, respectively. A contrastive loss aligns these features, ensuring that matched pairs have higher similarity scores than unmatched ones:

$$L_I = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(e_{v,i}\cdot e_{t,i}/\tau)}{\sum_{j=1}^{N}\exp(e_{v,i}\cdot e_{t,j}/\tau)},$$

where τ is a temperature parameter.

During testing, the text encoder is discarded. A mean representation vector is computed from an anchor set of known images. The

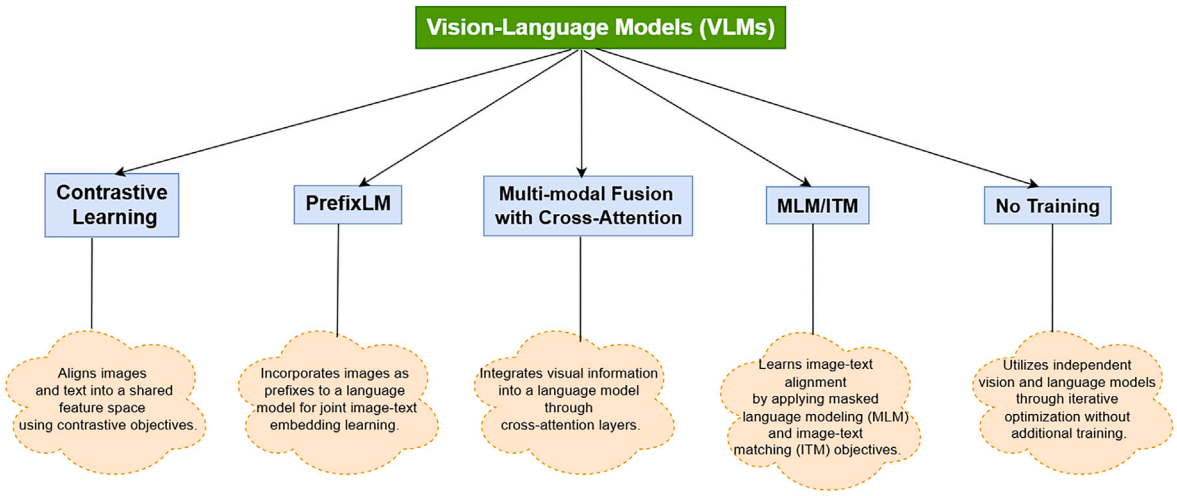

Fig. 3. Taxonomy of Vision-Language Models (VLMs) categorized by learning paradigms.

input image's representation is compared to this anchor vector using cosine similarity, which determines its category based on a predefined threshold. This approach allows the model to generalize across diverse generative models and contexts. While LASTED model has achieved better performance on the selected dataset, as seen in the paper text-labels are influential factors, selecting only four labels with photos and paintings might limit towards generalizability when the images are from very different domain such as medical images and the satellite imagery. The researchers should design appropriate labels for their specific tasks related images.

2.2.3. GenDet: good generalizations by Zhu (2023)

Zhu et al. [101] address the challenge of detecting synthetic images from unseen generators, which existing methods struggle to classify due to limited generalization. The authors propose **GenDet**, a detection model composed of two key components: *Teacher-Student Discrepancy-Aware Learning* and *Generalized Feature Augmentation*. These components are trained through an adversarial framework to improve generalization to both seen and unseen generators. The model employs a feature extractor E, based on CLIP:ViT [67], to extract features from input images. The *Teacher-Student Discrepancy-Aware Learning* is designed to: a) Reduce the difference in output between a teacher network (N_t) and a student network (N_s) for real images, and b) Amplify this difference for fake images to enhance detection. The discrepancy losses are defined as:

$$\mathcal{L}_{\min} = \frac{1}{b}\sum_{i=1}^{b} \| N_t(f_i) - N_s(f_i) \|_2^2, \quad \mathcal{L}_{\max} = -\mathcal{L}_{\min},$$

where f_i represents features from real or fake images, and b is the batch size.

To further enhance generalization, the *Generalized Feature Augmenter* adversarially generates augmented features. This augmenter facilitates the reduction of the difference between the teacher and student networks for augmented fake features, encouraging the model to detect unseen synthetic images by maintaining a large discrepancy during inference. Finally, a binary classifier N_c is trained on the variation between the teacher and student outputs to classify images as real or fake. Experiments demonstrate that GenDet achieves state-of-the-art performance on the UniversalFakeDetect and GenImage datasets, surpassing prior methods in both accuracy and mean average precision (mAP).

2.2.4. Mixture of low-rank experts: Liu 2024

Liu et al. [53] propose a transferable AI-generated image detection model utilizing CLIP-ViT as the backbone with parameter-efficient fine-tuning. The method modifies the MLP layers of the last three ViT-B/32 blocks through a Mixture of Low-Rank Experts (MoLE), integrating both shared and separate Low-Rank Adapters (LoRAs). Shared LoRAs capture common feature representations across datasets, while separate LoRAs specialize in diverse generative patterns. A trainable gating mechanism dynamically assigns input tokens to appropriate experts, with a load-balancing loss ensuring uniform expert utilization. The model freezes most CLIP parameters, adapting only LoRA modules and a new MLP classification head with a sigmoid activation. The forward operation in each MLP block is expressed as:

$$\Delta W x = \frac{\alpha}{r} BAx + \sum_{i=1}^{N} G_i(x) \frac{\alpha_i}{r_i} B_i A_i x,$$

where $G_i(x)$ is the gating function, and A, B, A_i, and B_i are low-rank matrices.

The approach achieves state-of-the-art generalization across unseen diffusion and autoregressive models, with superior robustness to post-processing perturbations like Gaussian blur and JPEG compression. Experimental results on benchmarks such as UnivFD and GenImage show that the proposed method surpasses existing detectors, including UnivFD and DIRE, by up to +12.72% in classification accuracy.

2.2.5. Merging a Mixture of hyper LORAS: Hyperdet by Cao 2024

Cao et al. [9] propose a generalizable synthetic image detection framework, HyperDet, leveraging the large pretrained multimodal model, CLIP: ViT-L/14, as the backbone, similar to Liu et al. [53], but with few notable innovations. Unlike Liu et al., the authors introduce a novel grouping of Spatial Rich Model (SRM) filters into five distinct groups to generate multiple filtered views of input images, capturing varying levels of pixel artifacts and features. Along this, HyperDet employs Hyper LoRAs, a hypernetwork-based approach that generates Low-Rank Adaptation (LoRA) weights for fine-tuning the CLIP model. These LoRA weights are computed using three types of embeddings: task embeddings, layer embeddings, and position embeddings. The outputs of these LoRA experts are merged to form a unified representation for classification, effectively integrating shared and specific knowledge for generalizable feature extraction. During training, HyperDet fine-tunes the last eight fully connected layers of the CLIP: ViT-L/14 model, along with the newly introduced Hyper LoRAs modules. To address imbalanced optimization, the framework employs a composite binary cross-entropy loss function, incorporating both original and filtered views of the images. This design achieves robust performance in detecting synthetic images across diverse generative models and datasets.

2.2.6. Forgery-aware adaptive transformer (FatFormer) by Liu: 2024

Liu et al. [51] introduce FatFormer, a generalizable synthetic image detection framework utilizing the pretrained CLIP model inspired by the work of Ojha et al. [63]. The authors address the limitations of freezing CLIP's layers, which hinder the generalization of forgery detection. FatFormer integrates two modules: the Forgery-Aware Adapter (FAA) and Language-Guided Alignment (LGA), for effective adaptation of CLIP's features. The FAA module extracts forgery artifacts from both image and frequency domains. The Image Forgery Extractor applies lightweight convolution layers to capture low-level artifacts, while the Frequency Forgery Extractor employs Discrete Wavelet Transform (DWT) and grouped attention mechanisms to dynamically aggregate multi-band frequency clues. The final adapted feature representation at each ViT stage is defined as:

$$g^{(j)} = g^{(j)}_{\text{img}} + \lambda \cdot g^{(j)}_{\text{freq}},$$

where λ balances image and frequency contributions.

The LGA module enhances text prompts using a Patch-Based Enhancer (PBE) and aligns image patch tokens with text embeddings through the Text-Guided Interactor (TGI). Contrastive loss is applied to the cosine similarities between image and text embeddings:

$$S^{(i)} = \cos(f^{(0)}_{\text{img}}, f^{(i)}_{\text{text}}), \quad S'^{(i)} = \frac{1}{N}\sum_{t=1}^{N} \cos(f^{(t)}_{\text{img}}, f^{(i)}_{\text{text}}),$$

where N is the number of patches.

2.2.7. Raising the bar with CLIP: by Cozzolino 2024

Cozzolino et al. [21] implement the CLIP: ViT-L/14 pretrained VLM, similar to the approaches in [9,44,53], for detecting synthetic images with a straightforward yet impactful adjustment. The authors propose generating synthetic images by feeding real-image captions to text-to-image models and then extracting feature vectors for both real and synthetic images using CLIP's image encoder. Specifically, feature vectors are obtained from the second-to-last layer of the ViT module:

$$r_i = \text{CLIP}(R_i), \quad f_i = \text{CLIP}(F_i),$$

where r_i and f_i represent feature vectors for real image R_i and synthetic image F_i, respectively.

For classification, the authors employ a simple linear Support Vector Machine (SVM). Notably, the analysis presented suggests that the CLIP-based detector does not rely on the same low-level traces exploited by most existing detectors, making it potentially more robust against adversarial attacks that target low-level features. The authors report using 32,000 images generated by GANs, diffusion models, and commercial text-to-image models for training and evaluation, demonstrating the good performance of their proposed approach across diverse datasets.

2.2.8. Representation from encoder-decoder for image detection by Koutlis: 2025

Koutlis and Papadopoulos [44] propose Representations from Intermediate Encoder-Blocks (RINE) to improve synthetic image detection by extracting low-level features from multiple layers of CLIP's Vision Transformer (ViT). The method captures both low- and high-level visual semantics by concatenating CLS tokens from each intermediate transformer block into a comprehensive feature representation. The extracted CLS tokens from each block Z_l are aggregated:

$$K = \bigoplus_{l=1}^{n} Z_l^{[0]} \in \mathbb{R}^{b\times n\times d},$$

where $Z_l^{[0]}$ denotes the CLS token from the l-th transformer block. To improve feature selection, a Trainable Importance Estimator (TIE) dynamically assigns weights to these representations:

$$\widetilde{K}_{ik} = \sum_{l=1}^{n} S(A_{lk}) \cdot K_{ilk},$$

where $S(A_{lk})$ represents softmax-activated importance scores for each block.

The features are processed through a projection network, then passed to a classification head with ReLU-activated dense layers and a final sigmoid output for binary classification. The framework optimizes performance using Binary Cross-Entropy (BCE) and Supervised Contrastive Loss:

$$L = L_{CE} + \xi L_{Cont.},$$

where ξ balances the contributions of both objectives. The authors demonstrate that RINE surpasses state-of-the-art methods on 20 test datasets, achieving a +10.6% accuracy improvement, with training requiring only one epoch (approximately 8 min). Additionally, the model is robust to image perturbations, maintaining strong performance across GAN, diffusion, and other synthetic image types.

2.2.9. Towards explainable fake image detection with multi-modal large language models by Ji et al., 2025

Ji et al. [111] present a study on explainable AI-generated image detection leveraging Multi-Modal Large Language Models (MLLMs) and prompt engineering. The core idea is to enhance interpretability and robustness by designing six specialized prompt paradigms (P1–P6), each interrogating a distinct visual or logical aspect of an image. These include: (P1) *Defect Query* for identifying artifacts such as abnormal lighting or unrealistic textures, (P2) *Regional Analysis* focusing on region-of-interest cues extracted using DINOv2, (P3) *Common Sense Reasoning* to detect physical or spatial inconsistencies, (P4) *Few-Shot Prompting* with exemplar comparisons, (P5) *Structural Analysis* assessing missing or misplaced components, and (P6) *Stereotype Matching* examining exaggerated or overly uniform features. The outputs of these six paradigms are then fused—either sequentially or via majority voting—to yield a final decision supported by coherent reasoning. The authors constructed a diverse dataset of 2000 images (1000 real and 1000 AI-generated) spanning diffusion, GAN-based, and other generative architectures, ensuring cross-model generalization. They benchmarked four major MLLMs—GPT-4o, GPT-4o-mini, Llama-3.2-Vision-Instruct, and LLaVA-CoT—against advanced but non-reasoning detectors such as AEROBLADE [115], CNNSpot [87], and so on. The fusion strategy using GPT-4o achieved the highest accuracy (93.4%), outperforming CNNSpot (91.8%) and even the best human annotator (86.3%). The study also revealed that replacing the term "fake" with "generated" in prompts improved model acceptance and reduced rejections, enhancing stability and interpretability.

2.2.10. Interpretable and reliable detection via grounded reasoning in MLLMs by Ji et al: 2025

Ji et al. [112] propose an interpretable and reliable detection framework that advances AI-generated image forensics through grounded multimodal reasoning. The study introduces the FakeXplained dataset, which contains 8772 high-quality AI-generated images annotated with bounding boxes and descriptive captions highlighting synthesis artifacts such as logical inconsistencies, unnatural textures, and structural errors. Using this dataset, the authors fine-tune the Qwen2.5-VL-32B [104] model through a two-stage process involving supervised fine-tuning (SFT) followed by Reinforcement Learning from Human Feedback (RLHF) with progressive Group Relative Policy Optimization (GRPO). This progressive training employs structured rewards for classification accuracy, grounding precision, and output format validity, guiding the model to produce coherent and human-aligned reasoning. The resulting model classifies images as real or synthetic localizes flawed regions and provides textual justifications aligned with visual

evidence. Furthermore, it demonstrates strong robustness under compression, cropping, and downsampling, as well as generalization to unseen datasets such as FaceForensics++ [74] and GPT-4o-generated images. The approach establishes a significant step toward explainable and human-trustworthy AI-generated image detection through grounded reasoning in large multimodal models.

2.2.11. ThinkFake: reasoning in multimodal large language models by Huang et al: 2025

The most important aspect of ThinkFake introduced by Huang et al. [110] is the integration of reasoning into AI-generated image detection. The ThinkFake framework introduces a reasoning-based approach to identifying synthetic imagery using Multimodal Large Language Models (MLLMs/VLMs). Built upon the Qwen2.5-VL-7B model [104], ThinkFake performs deliberate and interpretable detection through a Forgery Reasoning Prompt that follows a structured chain-of-thought pipeline—quick scan, detailed observation, technical tracing, auxiliary inspection, and final decision—to produce human-readable explanations. The framework incorporates Group Relative Policy Optimization (GRPO) reinforcement learning, where four reward functions—reasoning format, JSON structure, accuracy, and agentic reward—guide structured reasoning and enhance generalization. The agentic reward leverages auxiliary expert agents (UnivFD [63], AIDE [92], and so on) to provide semantic, frequency, and dual-stream feedback for more robust detection. Training follows the GenImage protocol, using 6000 balanced samples of real and AI-generated images (from ImageNet and Stable Diffusion v1.4, respectively). A small supervised fine-tuning (SFT) set of 638 samples is created using Gemini-1.4-pro for reasoning annotations, while a larger reinforcement learning (RL) set of 5000 samples is used for GRPO-based training. ThinkFake achieves 84.0% mean accuracy on GenImage and 75.4% on LOKI, surpassing all baseline detectors while maintaining strong interpretability and generalization capability.

2.2.12. ForenX: towards explainable AI-generated image detection with MLLMs by Tan et al., 2025

ForenX [116] proposes an explainable AIGI detector that augments an MLLM (LLaVA-8B [113] with Llama-3) using a forensic prompt constructed from CLIP-ViT [27,67] features and a lightweight forensics projector; the projector creates a forensics embedding supervised by an auxiliary detection loss and mapped into the LLM's token space, so the LLM consumes three inputs—text tokens, visual content tokens, and the forensic prompt—to produce a yes/no decision with human-readable reasons. The dataset pipeline has two parts: (i) large-scale machine-generated conversations from LLaVA over GenImage and ForenSynths (content Q&A + detection labels), and (ii) ForgReason, a human-aligned set composed of 2215 realistic Midjourney fakes with box-level artifact descriptions summarized via GPT-4V [103], plus 5000 real and 1000 fake samples from GenImage to balance fine-tuning. Training proceeds in two stages with LoRA: Stage-1 jointly tunes CLIP-ViT and the MLLM on GenImage/ForenSynths; Stage-2 freezes CLIP-ViT and further instruction-tunes the MLLM on ForgReason to strengthen explanation quality without degrading recognition. ForenX achieves strong cross-source detection (e.g., 97.8% mAcc on GenImage; 94.4% on ForenSynths) and produces grounded, user-study–preferred explanations; it also generalizes to SD-v3 and FLUX images (97.7–97.8% mAcc).

2.2.13. AIGI-holmes: towards explainable and generalizable AI-generated image detection via MLLMs by Zhou et al., 2025

AIGI-Holmes [118] introduces an explainable and generalizable AI-generated image (AIGI) detector that builds on LLaVA-1.6-Mistral-7B with a CLIP-ViT/L-14 visual encoder augmented by an NPR [119] low-level artifact expert. It constructs the Holmes-Set, a large-scale dataset comprising Holmes-SFTSet (65K images with textual explanations across semantic and artifact-level features) and Holmes-DPOSet (4K contrastive preference pairs refined through expert feedback). The dataset is annotated via a Multi-Expert Jury system combining four MLLMs (Qwen2VL-72B, InternVL2-76B, InternVL2.5-78B, and Pixtral-124B) with structured prompting and human preference refinement. The proposed Holmes Pipeline consists of three training stages: (1) Visual Expert Pre-training using CLIP-ViT and a low-level artifact detector NPR [119] to establish domain-specific perception; (2) Supervised Fine-Tuning (SFT) to align multimodal representations with textual explanations of real/fake reasoning; and (3) Direct Preference Optimization (DPO) to refine reasoning and align explanations with human judgment. During inference, a Collaborative Decoding strategy fuses predictions from the visual expert and MLLM for robust decision-making. Trained on LLaVA-1.6-Mistral-7B AIGI-Holmes achieves 99.2% mean accuracy and 99.9 A.P. on unseen diffusion and autoregressive models (e.g., VAR, FLUX), outperforming prior methods such as RINE, AIDE, and NPR. Its explanations also surpass GPT-4o and Pixtral-124B in BLEU (0.622) and ELO (11.42), demonstrating state-of-the-art interpretability and robustness against perturbations (JPEG, blur, downsampling).

2.2.14. Fake-GPT: detecting fake image via large language model by Fan et al., 2024

Fake-GPT [108] introduces a paradigm shift in fake image detection by employing a pure Large Language Model (LLM) for AI-generated image detection, without any vision encoders, position embeddings, or multimodal alignment. Instead of visual feature extraction, the authors reformulate the problem as a text-sequence prediction task, converting each RGB image into a serialized textual representation of pixel values (e.g., 32×32 RGB flattened as a string) and fine-tuning an LLM to classify it as real or fake. The proposed system uses Qwen-7B-Chat [104] as the base model and applies Low-Rank Adaptation (LoRA) fine-tuning to adapt the pre-trained LLM for this unconventional visual task. The model receives a prompt such as "You are a trained fake image detector. Given a string of RGB pixel values, determine if this image is fake."—allowing the LLM's sequence reasoning ability to distinguish subtle pixel-level patterns indicative of generative artifacts. For training, real samples were drawn from LSUN, ImageNet, CelebA, CelebA-HQ, and COCO datasets, while fake images were generated using ProGAN and evaluated across StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, and DeepFake generators. All inputs were resized to 32×32 to fit within the token limit, showcasing the model's robustness under low-resolution settings. Experiments demonstrated competitive accuracy—88.2% mean across cross-model evaluations and up to 97.4% on CycleGAN images—outperforming several CNN-based baselines like BiHPF, LGrad, and FrePGAN. When extended to diffusion-generated datasets, Fake-GPT maintained a mean accuracy of 83.8%, confirming strong cross-model generalization. In summary, Fake-GPT pioneers the application of pure LLMs for image authenticity detection, reframing fake image detection as a language modeling task. It eliminates vision-specific modules, simplifies training, and demonstrates that large-scale sequence models can serve as universal detectors for AI-generated imagery across both GAN and diffusion domains.

Comparative analysis of vision-language models. The earlier generation of vision-language approaches focuses primarily on leveraging pre-trained VLMs such as CLIP to improve the generalization of AI-generated image detection through robust visual–textual feature alignment. Ojha et al. [63] initiated this direction by mapping images into CLIP's semantic space to enable cross-model detection, while Wu et al. [90] extended it with language-guided textual prompts to refine semantic supervision. Subsequent works, including Zhu et al. [101], Liu et al. [53], and Cao et al. [9], progressively enhanced adaptability through feature-level regularization, low-rank adapters, and hypernetwork-driven parameterization, thereby strengthening cross-generator robustness. FatFormer [51] advanced this paradigm by incorporating frequency-domain features and language-guided alignment, while Cozzolino et al. [21] simplified CLIP-based detection through a linear SVM classifier to isolate higher-level semantics less sensitive to low-level perturbations. Finally,

Table 2
Evaluation of detection models from pretrained vision-language methods (and/or multi-modal large language model) category on cross-family generators, cross-category, and cross-scene generalization.

Models	Cross-family generators	Cross-category	Cross-scene
Ojha et al. [63]	✓	✓	✗
Wu et al. [90]	✓	✓	✗
Zhu et al. [101]	✓	✓	✗
Liu et al. [53]	✓	✓	✗
Cao et al. [9]	✓	✓	✗
Liu et al. [51]	✓	✓	✗
Cozzolino et al. [21]	✓	✓	✗
Koutlis et al. [44]	✓	✓	✗
Ji et al. [111]	✓	✓	✗
Ji et al. [112]	✓	✓	✗
Huang et al. [110]	✓	✓	✗
Tan et al. [116]	✓	✓	✗
Zhou et al. [118]	✓	✓	✓
Fan et al. [108]	✓	✓	✗

Koutlis and Papadopoulos [44] further optimized representation learning by aggregating intermediate CLIP features to capture both global and local semantics efficiently. Collectively, these studies form a coherent progression—from feature-space distance learning to parameter-efficient fine-tuning and adaptive multimodal integration—marking a shift from handcrafted forensic cues toward transferable, foundation-model-based visual-text alignment for synthetic image detection.

Comparative analysis of MLLM models. All studies converge toward a shared goal of enhancing the interpretability, reliability, and generalizability of AI-generated image detection through language-based reasoning frameworks. The earliest efforts (from 2024 to 2025), such as Fake-GPT [108] and ForenX [116], establish two divergent yet complementary directions—Fake-GPT reformulates image forensics as a pure sequence modeling task using only a textual representation of pixels, while ForenX introduces the first multimodal prompting mechanism that explicitly injects forensic embeddings into an MLLM for explainable decisions. Ji et al. [111,112] advance the field with Towards Explainable Detection and Grounded Reasoning, which evolves from handcrafted forensic prompts to structured, grounded reasoning guided by reinforcement learning and visual grounding, deepening interpretability and human alignment. ThinkFake [110] continues this trajectory by explicitly modeling a reasoning chain-of-thought through GRPO optimization and multi-agent feedback, bridging MLLM interpretability with robustness. Finally, AIGI-Holmes [104] represents the culmination of these trends, integrating multi-expert jury annotation, large-scale instruction tuning, and collaborative decoding to achieve both explainability and cross-domain generalization. Collectively, these works demonstrate a clear methodological progression—from textual reformulation (Fake-GPT), to multimodal prompting (ForenX), to reasoning-driven and human-aligned explanation systems (Ji et al., ThinkFake), culminating in a comprehensive, expert-supervised reasoning pipeline (AIGI-Holmes). The principal differences thus lie in their modality design (pure LLM vs. MLLM), training paradigm (prompt engineering vs. SFT/RLHF/DPO), and explanation grounding (textual, visual, or collaborative), marking a coherent evolution toward explainable and reasoning-based forensic detection (Table 2).

2.3. Training-free methods

Training-Free Methods refer to detection approaches that do not require any supervised training on synthetic or real datasets. Instead of learning discriminative boundaries through model optimization, these methods rely on intrinsic analytical cues, statistical inconsistencies, or reconstruction behaviors inherent to the image or its representation. They typically exploit properties observable through pretrained encoders, perturbation responses, or entropy measures to distinguish

Table 3
Evaluation of multimodal vision-language methods on GAN and diffusion-generated images using the UnivFD dataset [63]. Results are reported in classification accuracy (%). Methods from [21,44] were trained on only four classes of ProGAN-generated images, while others were trained on the full ProGAN dataset from [87].

Method	Generative adversarial networks						Low-level vision		Perceptual loss		Diffusion models								Total
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN	Deep-fakes	SITD	SAN	CRN	IMLE	Guided	LDM			GLIDE			DALL-E
													200 steps	200s w/CFG	100 steps	100 27	50 27	100 10	
													steps	w/CFG	steps	27	27	10	
Ojha et al. [63]	100.0	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.5	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
Zhu et al. [101]	99.00	99.50	99.30	99.05	99.00	96.75	88.20	63.50	67.50	93.90	98.75	98.70	98.80	98.60	98.75	98.75	98.75	98.75	98.45	94.42
Liu et al. [53]	100.0	99.33	99.67	99.46	99.83	97.07	77.53	81.11	65.50	82.32	96.79	90.70	98.30	95.90	98.75	92.40	93.95	93.00	94.90	92.45
Cao et al. [9]	100.0	97.40	97.50	97.50	96.20	98.65	73.85	93.00	75.00	92.75	93.20	77.35	98.70	96.60	98.80	87.75	89.95	88.70	97.00	92.10
Liu et al. [51]	99.90	99.30	99.50	97.20	99.40	99.80	93.20	-	-	-	-	76.10	98.60	94.90	98.70	94.40	94.70	94.20	98.80	-
Koutlis et al. [44]	100.0	99.30	99.60	88.90	99.80	99.50	80.60	90.60	68.30	89.20	90.60	76.10	98.30	88.20	98.60	88.90	92.60	90.70	95.00	-

genuine from generated content. By eliminating dependence on dataset-specific training, training-free methods emphasize generalization, interpretability, and robustness to unseen generative models.

2.3.1. AEROBLADE: training-free detection of latent diffusion images using autoencoder reconstruction error by Ricker et al., 2024

AEROBLADE [115] presents a training-free detection framework for identifying images generated by latent diffusion models (LDMs) through their inherent autoencoder (AE) structure. The core idea is that an LDM's AE reconstructs generated images with lower error than real ones because synthetic samples lie within the learned latent manifold, while real images fall slightly outside it. Consequently, the reconstruction error between an input image and its AE-reconstructed version serves as a reliable indicator of synthesis. This reconstruction-based approach is comparable to DIRE [89] but differs fundamentally in that AEROBLADE requires no training or classifier tuning. Formally, AEROBLADE computes the LPIPS [117] distance between an image x and its reconstruction $D(E(x))$. To generalize across generators, reconstruction errors are evaluated over multiple AEs, and the minimum reconstruction error is used as the detection criterion. The framework was evaluated on seven LDMs and achieved near-perfect detection performance (mean AP up to 0.999), generalizing effectively even to closed-source generators such as Midjourney [60]. Furthermore, it enables qualitative forensics by visualizing reconstruction error maps that reveal inpainted or manipulated regions within real images. The approach remains robust to perturbations such as JPEG compression, cropping, blur, and noise.

2.3.2. RIGID: a training-free and model-agnostic framework for robust AI-generated image detection by He et al., 2024

RIGID [109] introduces a training-free and model-agnostic method for detecting AI-generated images by exploiting the sensitivity of image representations to small noise perturbations. The central observation is that real images exhibit greater stability under minor perturbations than AI-generated images in the feature space of large vision models. RIGID compares the cosine similarity between an image's feature embedding and that of its noise-perturbed version using pretrained extractors such as DINOv2 [114]. Images showing lower similarity (determined by the threshold) under perturbation are classified as AI-generated. Unlike training-based detectors such as DIRE [89] or training-free detectors such as AEROBLADE [115], RIGID operates during inference, requiring no fine-tuning or prior knowledge of the generative model. Its detection score is derived as:

$$S(x) = 1\{sim(f(x), f(x + \lambda\delta)) \leq \epsilon\}, \quad \delta \sim N(0, I),$$

where f is the feature extractor and λ controls perturbation intensity. The framework generalizes across backbones, with DINOv2 achieving the best trade-off between global and local representation. Evaluated on diverse datasets including ImageNet, LSUN-Bedroom, and GenImage—covering diffusion, GAN, and VAE models—RIGID achieved mean AUC/AP improvements exceeding 25% over AEROBLADE and frequently surpassed trained detectors such as DIRE [89] and Corvi et al. [20].

2.3.3. HFI: a unified framework for training-free detection and implicit watermarking of latent diffusion model generated images by Choi et al., 2024

HFI [106] proposes a training-free approach that improves the robustness of autoencoder-based detection of latent diffusion model (LDM) images. It addresses the key limitation of AEROBLADE [115], which relies on reconstruction error but tends to overfit to background information, failing on images with simple or uniform backgrounds. HFI instead measures the aliasing effect—distortions of high-frequency components introduced by the LDM autoencoder by quantifying how much high-frequency information is lost during reconstruction. HFI computes the directional derivative of the reconstruction distance in the direction of high-frequency components filtered by a low-pass kernel, capturing subtle distortions beyond global reconstruction loss. A numerical approximation via first-order Taylor expansion allows efficient evaluation without training or model fine-tuning. When multiple autoencoders are available, HFI adopts an ensemble strategy that selects the minimum response across them, improving generalization to unknown generators. Compared to prior training-free methods, HFI achieves clear architectural advancement: unlike RIGID [109], which relies on feature-space perturbation stability, and AEROBLADE, which measures pixel-space reconstruction consistency, HFI directly exploits the frequency-space degradation intrinsic to LDM autoencoders. This design allows both detection and model attribution (implicit watermarking) by associating distinct aliasing signatures with specific diffusion models. Empirical evaluation across GenImage, SynthBuster, and DiffusionFace benchmarks shows that HFI consistently surpasses AEROBLADE and RIGID, achieving competitive performance to training-based methods.

2.3.4. Zero-shot detection of AI-generated images by Cozzolino et al., 2024

Zero-Shot Entropy-based Detector (ZED) [107] introduces a training-free and generator-independent method for detecting AI-generated images, similar in spirit to AEROBLADE [115] and RIGID [109], but differing in that it requires training only on real images. Instead of relying on reconstruction or perturbation cues, ZED measures the surprise of an image under a probabilistic model of real imagery learned through a lossless image encoder. The key idea is that real images conform to the learned distribution, while synthetic ones deviate from it, resulting in a higher coding cost. Using the Super-Resolution based lossless Compressor (SReC) [105], ZED predicts conditional probability distributions of pixels across multiple resolutions and computes the negative log-likelihood (NLL) and entropy maps. The difference between the actual and expected coding costs—termed the coding cost gap—serves as the detection statistic. Synthetic images exhibit larger gaps, especially at higher resolutions, reflecting inconsistencies with the statistical model of natural images. This approach generalizes across both GAN- and diffusion-generated images without dependence on any generator-specific artifacts.

Comparative analysis of training-free methods. Training-free detection methods share the common objective of identifying AI-generated content without requiring supervised learning, instead leveraging analytical properties inherent to pretrained models or statistical characteristics of natural images. Despite this shared foundation, their core mechanisms diverge in various perspectives. AEROBLADE [115] pioneers reconstruction-based detection by exploiting the autoencoder behavior in latent diffusion models, demonstrating that generated images exhibit lower reconstruction error than real ones. RIGID [109] extends this paradigm by evaluating the stability of image embeddings under random perturbations, introducing feature-space sensitivity as a discriminative signal independent of any generator. HFI [106] advances the idea further through frequency-domain analysis, identifying aliasing artifacts in high-frequency components of reconstructed images, thus unifying detection and implicit watermarking. ZED [107] reinterprets detection as a coding-cost analysis problem, employing a probabilistic model of real images to measure deviations in entropy and compression cost, offering a fully zero-shot and generator-agnostic solution. Collectively, these approaches trace a clear methodological evolution—from pixel and latent reconstruction analysis (AEROBLADE), to perturbation and representation stability (RIGID), to spectral and aliasing cues (HFI), and finally to probabilistic learning (ZED). Importantly, these methods converge toward efficient forensic detection without relying on extensive training(Table 4).

2.4. Frequency domain analysis methods

Frequency domain analysis transforms image data into the spectral domain, facilitating the detection of periodic artifacts, noise distributions, and variations in frequency components often associated with synthetic image generation. Techniques such as the Discrete Fourier

Table 4
Evaluation of detection models from training-free methods on cross-family generators, cross-category, and cross-scene generalization.

Models	Cross-family generators	Cross-category	Cross-scene
AEROBLADE [115]	✓	✗	✗
RIGID [109]	✓	✓	✓
HFI [106]	✓	✗	✗
ZED [107]	✓	✗	✗

Transform (DFT), Discrete Wavelet Transform (DWT), and Discrete Cosine Transform (DCT) are commonly used to extract these spectral features. The Fourier Transform, proposed by Fourier [5], decomposes signals into their constituent frequencies and is effective in identifying global periodic patterns, including checkerboard artifacts. The two-dimensional Discrete Fourier Transform (DFT) of an image $f(x, y)$ of size $M \times N$ is given by:

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{1}$$

where $F(u, v)$ represents the frequency domain coefficients, and the inverse DFT (IDFT) reconstructs the image as:

$$f(x, y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u, v) e^{j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{2}$$

To improve computational efficiency, the Fast Fourier Transform (FFT), introduced by Cooley and Tukey [19], accelerates the DFT process, enabling the rapid detection of periodic patterns and aliasing artifacts commonly found in synthetic images. The Wavelet Transform, introduced by Haar [34], provides both spatial and frequency localization, making it suitable for detecting transient and localized artifacts. Unlike the Fourier Transform, which represents signals in terms of sinusoidal waves, the Discrete Wavelet Transform (DWT) uses wavelet basis functions to analyze variations in different frequency bands. The one-level decomposition of an image using DWT is given by:

$$f(x, y) = \sum_{m} \sum_{n} c_{m,n} \phi_{m,n}(x, y) + \sum_{m} \sum_{n} d_{m,n} \psi_{m,n}(x, y) \tag{3}$$

where $\phi_{m,n}(x, y)$ represents the approximation coefficients (low-frequency components), and $\psi_{m,n}(x, y)$ captures the detailed coefficients (high-frequency components) at different scales. The Discrete Cosine Transform (DCT), widely applied in image compression (e.g., JPEG), is particularly effective for analyzing energy distribution in smooth regions and identifying compression-induced anomalies [2]. The 2D DCT of an image block $f(x, y)$ of size $N \times N$ is computed as:

$$F(u, v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2N}\right] \tag{4}$$

where $\alpha(u), \alpha(v)$ are normalization factors:

$$\alpha(u) = \begin{cases} \frac{1}{\sqrt{N}}, & \text{if } u = 0 \\ \frac{\sqrt{2}}{\sqrt{N}}, & \text{if } u > 0 \end{cases} \tag{5}$$

The Inverse DCT (IDCT) reconstructs the image block from its frequency coefficients. Since DCT concentrates most of the signal energy in a few low-frequency components, it is useful for detecting compression artifacts and synthetic image inconsistencies (Fig. 4).

2.4.1. Fourier spectrum discrepancies: Dzanic 2020

Dzanic et al. [28] presented a method for analyzing high-frequency Fourier models to highlight the limitations of generative models, such as GANs and VAEs, in reconstructing high-frequency components. Their approach involves applying the Fourier Transform to images to obtain a reduced spectrum, which is then modeled using two decay parameters: b_1, representing the magnitude of high-frequency content, and b_2, representing the decay rate. These parameters were used to train a KNN classifier capable of distinguishing synthetic images from real ones, achieving 99.2% accuracy on uncompressed high-resolution images. The process involves normalizing the Fourier Transform by the DC gain, converting the data to normalized polar coordinates, binning and averaging Fourier coefficients to create a reduced spectrum, and fitting the decay parameters above a threshold wavenumber. This comprehensive method underscores the effectiveness of frequency domain features in identifying the discrepancies characteristic of synthetic image generation.

2.4.2. Liu's detection method derived from analysis on real images: 2022

Liu et al. [50] proposed a novel approach to detecting synthetic images by focusing on the inherent noise patterns of real images, deviating from existing methods that analyze artifacts in generated images. They introduced the concept of Learned Noise Patterns (LNP), a high-dimensional spatial mapping derived from neural networks, to characterize the noise properties of real images. By comparing these learned patterns with the noise present in synthetic images, the method identifies discrepancies that indicate image generation. Leveraging both spatial and frequency domain representations, this approach demonstrated improved accuracy in detecting synthetic images across multiple domains.

2.4.3. Two-stream convolutional network for fake content detection by Yousaf: 2022

Yousaf et al. [94] proposed TwoStreamNet, a two-stream convolutional neural network designed to enhance the generalizability of fake visual content detection by jointly analyzing spatial and frequency features. The network comprises two main modules: the Spatial Stream and the Frequency Stream, which independently process input images and fuse their outputs at the classification stage to improve detection accuracy.

The Frequency Stream captures frequency domain artifacts by first converting images to the YCbCr color space to decorrelate color channels. Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) (see Eq. (3)) are applied to each channel. DFT decomposes image signals into real and imaginary components, while DWT captures both low- and high-frequency sub-bands. The resulting frequency features, represented as $H \times W \times 18$ feature maps, are processed using ResNet-50 [35] to extract discriminative frequency patterns. The Spatial Stream focuses on spatial domain features by processing original RGB images augmented with Gaussian blur and JPEG compression, similar to the approach in [87]. These augmented images are passed through a ResNet-50 network to extract spatial features. Finally, the outputs of the two streams are fused via probability averaging, ensuring equal contributions from spatial and frequency domains to the final decision. This combined framework highlights the importance of integrating frequency features for robust detection of synthesized visual content.

2.4.4. Synthbuster by Bammey: 2023

Bammey [3] introduced Synthbuster, a forensic method for detecting synthetic images generated by state-of-the-art diffusion models. The method begins by applying a cross-difference filter, originally defined by Chen et al. [16], to highlight periodic frequency artifacts in synthetic images. The filter acts as a high-pass filter, generating residual images using the operation:

$$C(x, y) = |I(x, y) + I(x+1, y+1) - I(x+1, y) - I(x, y+1)|,$$

where $I(x, y)$ represents pixel intensity.

The Fast Fourier Transform (FFT) is then applied to the residual image, extracting spectral components corresponding to periodicities

Table 5
Diffusion-based evaluation of training-free methods and an MLLM method (ForenX [116]) on the GenImage dataset [102]. Accuracy (Acc) is reported.

Method	ADM	BigGAN	GLIDE	Midjourney	SD1.4	SD1.5	VQDM	wukong	Mean
RIGID [109]	0.790	0.976	0.964	0.797	0.698	0.699	0.860	0.708	0.812
AEROBLADE [115]	0.838	0.986	0.990	0.988	0.983	0.984	0.723	0.984	0.935
HFI [106]	0.923	0.996	0.995	0.998	0.998	0.998	0.905	0.999	0.977
ForenX [116]	0.974	0.978	0.980	0.979	0.978	0.977	0.977	0.980	0.978

Notes: All methods are training-free diffusion image detectors evaluated on the GenImage dataset [102], with an additional method from the MLLM category, ForenX [116]. Unlike training-free approaches, ForenX was trained on SDv1.4-generated images and provides reasoning to explain why an image is classified as generated. The results for training-free methods are reported from the study presented by Choi et al. [106].

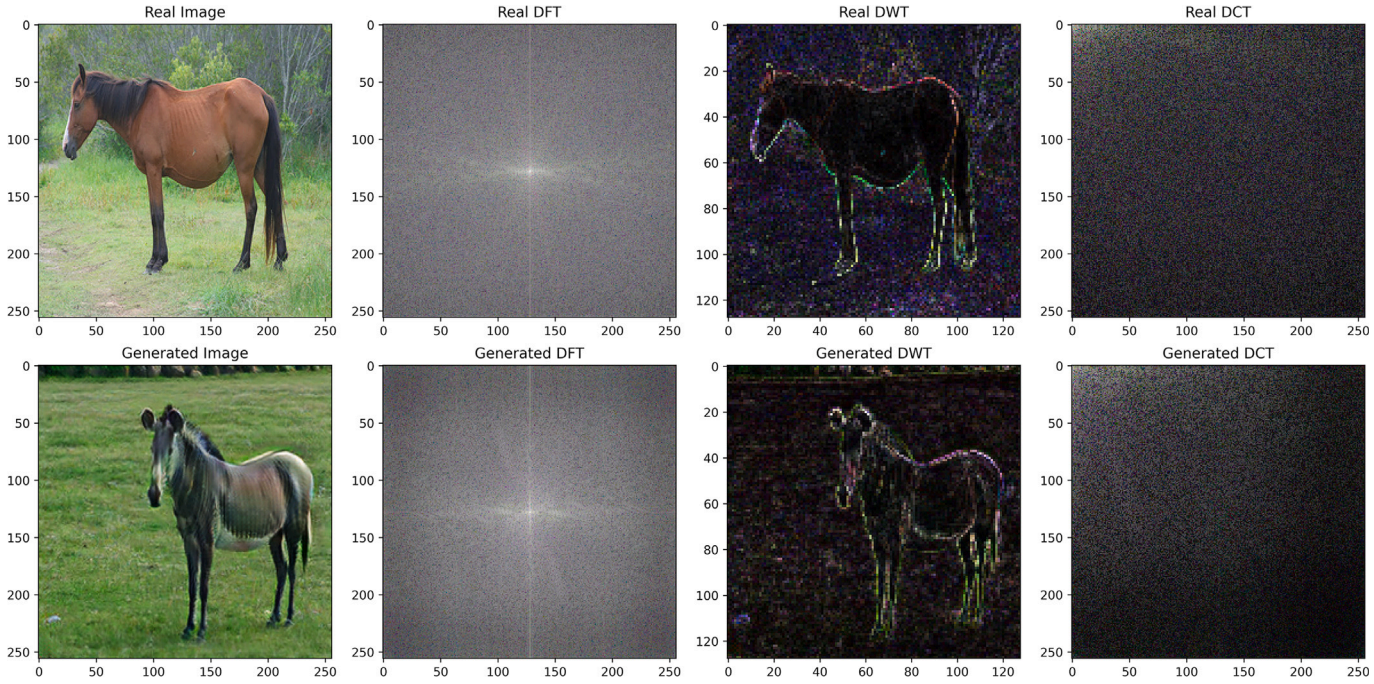

Fig. 4. Comparison of frequency domain transformations for real and generated images. The first column presents the input images, followed by their respective DFT, DWT, and DCT representations. The images were obtained from ForenSynth [87], and the generated image was produced using CycleGAN [100]. These images are reproduced under the Attribution-NonCommercial-ShareAlike 4.0 International license.

$0, 2, 4$, and 8 in both vertical and horizontal directions. The resulting features form a 135-dimensional magnitude peak vector for the RGB channels. These features are classified using a histogram-based gradient boosting tree classifier (HBGB) [42], trained specifically to distinguish synthetic images from real ones.

Synthbuster adapts techniques traditionally used for detecting JPEG compression artifacts to address artifacts from diffusion models, highlighting potential overlaps between the two. The method was validated on JPEG-compressed images to mitigate misclassification risks. Despite its simplicity, Synthbuster outperformed state-of-the-art methods for diffusion image detection. Additionally, the authors have publicly released their dataset, as detailed in the dataset section.

2.4.5. Meng's artifact feature purification: 2024

Meng et al. [59] addressed two key limitations in existing detection methods: poor generalization across generative models and limited effectiveness on images from diverse large-scale datasets. To overcome these challenges, they proposed the Artifact Purification Network (APN), which extracts generalizable artifact features through explicit and implicit purification processes. Explicit purification isolates artifact features in spatial and frequency domains by employing feature decomposition and frequency-band proposals to detect suspicious patterns. Implicit purification, guided by a classifier, further refines these features using mutual information estimation, enhancing the robustness of detection across various generators and datasets. This dual purification approach significantly improves generalization and detection accuracy.

2.4.6. An image transformation perspective by Li: 2024

Li et al. [47] identified biases in existing detection methods, including weakened and overfitted artifact features, which limit their generalization capability. To address these challenges, the authors proposed the SAFE (Simple Preserved and Augmented Features) framework, designed to improve generalizable artifact detection by integrating effective transformations with preprocessing and augmentations. The framework employs several key strategies: (1) replacing conventional down-sampling with a crop operator (RandomCrop for training and CenterCrop for inference) to preserve local correlations and prevent artifact distortion; (2) applying invariant augmentations, such as ColorJitter and RandomRotation, to mitigate color discrepancies and irrelevant rotation-related features; (3) using a patch-based random masking strategy to enhance sensitivity to local image regions and subtle artifacts; and (4) incorporating Discrete Wavelet Transform (DWT) to extract high-frequency features critical for distinguishing synthetic from real images. The processed data were used to train a ResNet model [35]

with 1.44M parameters, achieving strong detection accuracy during testing. This approach demonstrated the effectiveness of integrating diverse transformations to improve detection robustness and generalization.

2.4.7. Yan's AIDE method for synthetic image detection (2024)

Yan et al. [92] introduced the Chameleon dataset, a collection of AI-generated images designed to closely resemble real-world scenes, highlighting the limitations of existing detectors, which frequently misclassify these images as real. To address these challenges, the authors proposed AIDE (AI-generated Image Detector with Hybrid Features), a method that combines low-level frequency-based features with high-level semantic embeddings from CLIP. AIDE processes input images by dividing them into patches and applying discrete cosine transform (DCT) (see Eq. (4)) to extract frequency-domain features. Patches are sorted based on computed scores, with the highest- and lowest-frequency patches selected, resized, and processed through SRM [29] to capture noise patterns. These frequency-based features are embedded using a ResNet-50 network, while high-level semantic features are extracted via ConvNeXt-based OpenCLIP. The outputs from both networks are concatenated and passed through a multi-layer perceptron (MLP) for final classification. This hybrid approach demonstrated better performance across benchmarks, including improvements on *AIGCDetectBenchmark* and GenImage [102], while achieving competitive results on the challenging Chameleon dataset.

Comparative analysis of frequency domain analysis methods. Frequency-domain detection methods analyze the spectral signatures and periodic inconsistencies introduced by generative models, offering an orthogonal perspective to spatial feature analysis. Early work by Dzanic et al. [28] demonstrated that GANs fail to reproduce high-frequency spectral content accurately, revealing that the Fourier spectrum decay parameters could effectively distinguish real from generated images. However, their approach was limited to GAN-generated data and relied on handcrafted frequency modeling. Liu et al. [50] shifted focus toward the intrinsic noise characteristics of real images, improving robustness by learning noise priors rather than relying solely on synthetic artifacts, thereby addressing over-dependence on generator-specific patterns. Yousaf et al. [94] advanced this direction with TwoStreamNet, which fused spatial and frequency representations through parallel CNN streams, overcoming the modality isolation problem seen in prior methods that treated spatial and frequency cues independently. Bammey's SynthBuster [3] further extended frequency-based detection to diffusion models, addressing the prior gap in diffusion image forensics. By employing a cross-difference filter and spectral peak analysis, SynthBuster generalized frequency-based techniques beyond GANs and achieved robustness under compression. Building on these insights, Meng et al. [59] introduced the Artifact Purification Network (APN), which directly tackled two recurring challenges—poor cross-model generalization and dataset dependency—by purifying spatial and frequency artifacts through explicit and implicit feature separation. Li et al. [47] improved upon this with the SAFE framework, emphasizing preserved and augmented transformations, including DWT-based feature extraction, to strengthen generalization and mitigate overfitting observed in prior frequency-based CNNs. Yan et al. [92] then unified these developments by combining frequency and semantic embeddings in AIDE, bridging low-level artifact cues with high-level CLIP features, and achieving superior robustness on challenging datasets like Chameleon, where earlier models failed due to limited semantic understanding. Collectively, these methods demonstrate a clear evolution—from handcrafted spectral analysis (Dzanic) and real-noise modeling (Liu) to dual-domain fusion (Yousaf), diffusion-oriented adaptation (Bammey), feature purification and augmentation (Meng, Li), and finally hybrid semantic–frequency reasoning (Yan). This trajectory reflects a shift from model-specific artifact detection toward unified, semantically informed, and cross-domain generalization in frequency-domain image forensics (Table 6).

Table 6
Evaluation of detection models from frequency domain analysis category on cross-family generators, cross-category, and cross-scene generalization.

Models	Cross-family generators	Cross-category	Cross-scene
Dzanic et al. [28]	✗	✗	✗
Liu et al. [50]	✓	✓	✓
Yousaf et al. [94]	✓	✓	✗
Bammey [3]	✓	✓	✗
Meng et al. [59]	✓	✓	✓
Li et al. [47]	✓	✓	✓
Yan et al. [92]	✓	✓	✓

2.5. Fingerprint analysis methods

Fingerprint analysis in digital forensics refers to techniques that detect unique, traceable features left behind by devices or processes during image generation. Traditional approaches focused on detecting handcrafted features, including device fingerprints and postprocessing fingerprints. Device fingerprints, such as the photo-response nonuniformity (PRNU) pattern [56], arise from manufacturing imperfections in imaging sensors, leaving a unique and stable mark on each captured image. Postprocessing fingerprints, on the other hand, originate from in-camera processing pipelines, including operations like demosaicking and compression, which embed specific patterns into images [22]. While fingerprint analysis often overlaps with frequency domain methods, its primary focus is on extracting distinct fingerprints specific to generative models, such as GANs and diffusion models, rather than general spectral artifacts. The following methods are categorized under fingerprint analysis, as they aim to identify the unique traces of generative techniques.

2.5.1. Disentangling GAN fingerprints by Yang et al. (2021)

Yang et al. [93] proposed the GAN Fingerprint Disentangling Network (GFD-Net), advancing previous methods by focusing on disentangling content-irrelevant and GAN-specific fingerprints in synthetic images. Unlike earlier approaches that employ classification setups, GFD-Net integrates a generator, a discriminator, and an auxiliary classifier within an extended GAN framework to explicitly learn and separate these fingerprints. The generator adopts a U-Net architecture with an encoder–decoder structure. A classifier is added to the encoder's output to predict the GAN source, enhancing the generator's feature learning capability. The generator outputs a fingerprint f, which is added to a real image x_{real} to produce a fingerprinted synthetic image x_{fp}:

$$x_{\text{fp}} = x_{\text{real}} + f.$$

The discriminator utilizes a PatchGAN architecture to classify images by evaluating smaller patches and averaging their authenticity scores, encouraging the generator to focus on GAN-specific features. Consequently, a ResNet-50 classifier is employed to differentiate fingerprints from different GANs. The training process alternates between: 1. Training the generator (G) with fixed discriminator (D) and classifier (C). 2. Training the discriminator (D) and classifier (C) with a fixed generator (G). The generator is trained using a combination of loss functions:

$$L_G = \omega_1 L_{z_G} + \omega_2 L_{\text{adv}_G} + \omega_3 L_{\text{cls}_G} + \omega_4 L_{\text{percept}_G},$$

where L_{z_G} ensures the generator correctly predicts the image source, L_{adv_G} enforces realistic fingerprinted images, L_{cls_G} ensures the fingerprint represents the GAN source, and L_{percept_G} minimizes perceptual differences between x_{fp} and x_{real}. The discriminator and classifier are trained with two loss functions:

$$L_{\text{adv}_D} = \mathbb{E}[\log(1 - D(x_{\text{fp}}))] + \mathbb{E}[\log(D(x_{\text{real}}))],$$

$$L_{\text{cls}_C} = \mathrm{L}_{\text{CE}}(C(x_{\text{real}}), y) + \mathrm{L}_{\text{CE}}(C(x_{\text{fp}}), y),$$

where L_{adv_D} ensures the discriminator distinguishes real and fingerprinted images, and L_{cls_C} trains the classifier for accurate source attribution. By explicitly disentangling GAN-specific fingerprints from image content, GFD-Net demonstrates improved robustness and generalizability compared to earlier methods. Its design builds on prior GAN fingerprinting approaches by incorporating both architectural and training innovations, addressing limitations in source attribution across diverse GANs.

2.5.2. FingerprintNet: synthesized fingerprints for generalized GAN detection by Jeong: 2022

Jeong et al. [37] proposed FingerprintNet, a framework aimed at generalizing GAN-generated image detection by synthesizing diverse GAN fingerprints. This approach addresses the challenge of detecting images from unseen GAN architectures without relying on GAN-specific training datasets. FingerprintNet employs an autoencoder-based fingerprint generator, incorporating random layer selection, multi-kernel deconvolution, and feature blending modules to create diverse and robust fingerprints. The generator is trained using a combination of reconstruction loss and similarity loss, ensuring that synthesized fingerprints accurately represent the characteristics of GAN-generated images.

For detection, FingerprintNet applies a Fast Fourier Transform (FFT) to the generated images to extract 2D spectra, where GAN fingerprints are more evident. A ResNet-50 [35] classifier is then used to distinguish real from generated images based on these fingerprint-highlighted spectra. To address dataset imbalance, the generator creates three fake images for each real image, and a mixed-batch strategy is applied during training to maintain balanced mini-batches. By synthesizing fingerprints and avoiding reliance on specific GAN datasets, FingerprintNet demonstrates improved generalization for detecting images from unseen GAN architectures, advancing robustness in GAN detection tasks.

2.5.3. Learning on gradients by Tan: 2023

Tan et al. [81] proposed a novel detection framework, Learning on Gradients (LGrad), which leverages gradient-based representations as generalized artifacts. The framework begins by employing a transformative model, T, a pretrained CNN, to convert real and generated images into feature vectors. The gradient of $\text{sum}(l)$ with respect to the input image is then computed to capture generalized artifacts:

$$G = \frac{\partial \text{sum}(l)}{\partial I_i},$$

where l represents the feature vector output by T for an input image I_i.

Following this, a classification model, ResNet-50 [35], pretrained on the ImageNet dataset [75], is trained on the computed gradients to learn underlying artifacts. Notably, T remains fixed during gradient computation and is reused during inference to first extract gradients, which are then classified by the trained model. The authors highlight that training the classifier on computed gradients enables generalized learning of artifacts common across GAN models. The paper also compares the performance of various transformative models, including pretrained classifiers and GAN discriminators, showcasing their impact on detection performance. Readers are encouraged to refer to the main paper for a detailed analysis.

2.5.4. Data augmentation in fingerprint domain by Wang: 2023 (Scaling & mixup)

Wang et al. [85] proposed a framework to enhance the generalizability of GAN-generated image detectors by augmenting synthetic data in the fingerprint domain. The method involves two key contributions: (1) extraction of fingerprints from synthetic images across different scenes using an encoder trained with Mean Squared Error (MSE) and adversarial losses, and (2) improved cross-GAN generalization through fingerprint perturbation. The framework utilizes an autoencoder trained on real images to extract residual fingerprints from GAN-generated images. These fingerprints, which represent GAN-specific artifacts, are made generalizable across scenes by incorporating a Gradient Reversal Layer (GRL) alongside MSE loss.

To address architecture dependency and simulate fingerprints of unseen GANs, two augmentation strategies were introduced: scaling, which applies a random scaling factor α to modify fingerprint intensity, and Mixup, which combines fingerprints from multiple samples using weighted sums to generate diverse synthetic fingerprints. After augmentation, perturbed fingerprints are added back to the reconstructed images to create augmented fake images. These augmented samples are used to train a binary classifier with cross-entropy loss, improving detection accuracy across unseen GAN architectures. This approach demonstrates the effectiveness of fingerprint augmentation in enhancing the generalization of GAN detectors, addressing the variability of GAN fingerprints across architectures and scenes.

2.5.5. Corvi's analysis on trending detection methods to detect generated images by diffusion models: 2023

Corvi et al. [20] examined the effectiveness of current forensic detection methods in identifying synthetic images generated by diffusion models. Building on the methodology of Marra et al. [58], they employed a denoising filter [97] to isolate noise residuals by subtracting the scene content from images. Averaging the residuals across multiple images produced a synthetic fingerprint, capturing the artifacts specific to the generation process. Additionally, Corvi et al. conducted spectral analysis by applying the Fourier transform to averaged residuals from 1000 images. Their findings revealed distinct spectral patterns for models like Stable Diffusion and Latent Diffusion, whereas weaker artifacts were observed in ADM and DALL.E 2, posing challenges for detection. The study highlighted the limitations of current detectors, particularly in handling resized and compressed images, and emphasized the need for robust methods tailored to diffusion models. Their findings contribute to understanding the unique challenges of detecting diffusion-generated images, with datasets and code available for further research.

2.5.6. MaskSim (Li, 2024)

Li et al. [49] proposed MaskSim, a forensic framework for detecting synthetic images generated by diffusion models by focusing on extraction of artifacts. The method leverages traceable artifacts in the Fourier Transformed Spectrum, selectively amplifies these artifacts, and uses a simple linear classifier to achieve competitive detection performance. The framework begins by preprocessing input images with a DnCNN denoiser [97] to suppress textures and enhance residual synthesis artifacts building on approaches introduced in [58] and later adapted in [20] for diffusion-generated images. The residual image undergoes DFT to compute the logarithmic magnitude spectrum, which is then refined using a trainable 1×1 convolutional layer and an element-wise masking procedure. A trainable mask is applied to the spectrum, followed by Batch Normalization, and a normalized reference spectrum is computed for comparison. Detection is performed by computing the cosine similarity between the masked spectrum and the reference spectrum. For synthetic images, cosine similarity values are maximized, while for real images, the absolute cosine values are minimized. A logistic regression classifier, trained using cross-entropy loss, predicts the probability of an image being synthetic. During testing, only regular cosine similarity is used, ensuring robustness and avoiding overfitting. The framework was validated on a dataset of diffusion-generated images [3], achieving strong detection performance and demonstrating its effectiveness in leveraging frequency artifacts for synthetic image detection.

Comparative analysis of fingerprint analysis methods. Fingerprint-based methods aim to capture intrinsic traces left by generative models, evolving from GAN-specific fingerprints to diffusion-aware and hybrid artifact extraction techniques. Yang et al. [93] initiated this line of research with GFD-Net, which disentangled content-irrelevant and GAN-specific fingerprints, improving robustness over earlier classifier-based approaches. However, its dependence on model-specific training limited

Table 7
Fingerprint analysis methods on cross-family generators, cross-category, and cross-scene generalization.

Models	Cross-family generators	Cross-category	Cross-scene
Yang et al. [93]	✗	✓	✓
Jeong et al. [37]	✓	✓	✗
Corvi et al. [20]	✓	✗	✗
Tan et al. [81]	✗	✓	✓
Wang et al. [85]	✗	✓	✗
Li et al. [49]	✗	✓	✓

scalability to unseen architectures. Jeong et al. [37] addressed this constraint through FingerprintNet, which synthesized diverse fingerprints via an autoencoder-based generator, enhancing generalization to unseen GANs without retraining. Tan et al. [81] advanced the paradigm with LGrad, which replaced explicit fingerprint synthesis with gradient-based representations, capturing universal generation artifacts across models and reducing reliance on handcrafted fingerprint features. Wang et al. [85] further improved generalization by augmenting fingerprints through scaling and Mixup in the latent fingerprint domain, mitigating variability across architectures and scenes. Corvi et al. [20] shifted focus toward diffusion models, demonstrating that residual-based fingerprints could capture distinctive frequency-domain signatures while highlighting limitations under compression and resizing. Li et al. [49] proposed MaskSim, which unified prior residual-based and spectral fingerprinting techniques using a masked similarity approach in the Fourier domain, achieving robust detection of diffusion-generated imagery without extensive retraining. Collectively, these studies outline a coherent progression—from explicit fingerprint disentanglement and synthesis to generalized and gradient-based learning, and finally to diffusion-oriented spectral fingerprinting. This trajectory reflects the transition from architecture-specific fingerprint extraction to universal, spectrum-informed, and model-agnostic forensic analysis(Table 7).

2.6. Patch-based analysis methods

Patch-Based Analysis methods focus on identifying synthetic images by analyzing localized patches rather than the entire image. These methods exploit inconsistencies or artifacts that may appear within smaller regions, enabling the detection of subtle generative patterns. By dividing an image into patches and examining features such as texture, edge coherence, or pixel-level anomalies, these approaches enhance detection granularity. Patch-based strategies are particularly effective in scenarios where global analysis may miss fine-grained artifacts introduced by generative models.

2.6.1. Patch-based classification by Chai et al.: 2020

Chai et al. [10] proposed a patch-based classification framework using truncated ResNet and Xception backbones to classify localized patches of an image as real or fake. By limiting the model's receptive field, the framework focuses on local artifacts rather than global image structure, enhancing generalization across datasets and generative model types. Each patch is processed independently using a 1×1 convolution layer appended to the truncated backbone, and cross-entropy loss is applied. The final output is obtained by averaging the softmax predictions across patches. This increases the data-to-parameter ratio, improving generalization to unseen data. The framework includes a visualization mechanism that generates heatmaps, highlighting regions contributing to the classifier's predictions. These heatmaps reveal that complex textures, such as hair and facial boundaries, are key to distinguishing real from generated images.

2.6.2. Orthogonal training in detecting GAN-generated images: 2022

Mandelli et al. [57] proposed a compact detection framework designed to generalize to unseen GAN architectures. This method employs orthogonal training, where multiple CNNs with EfficientNet-B4 [82] backbones are trained on datasets differing in semantic content, GAN models, and post-processing operations. The framework divides input images into 128×128 RGB patches, which each CNN analyzes independently. These patch-level scores are aggregated to classify the entire image. Synthetic images receive the highest patch scores, while real images receive the lowest. The final decision is derived from the average of all CNNs' image-level scores. Experimental results confirmed the method's strong performance in detecting StyleGAN3-generated images without prior exposure during training.

2.6.3. Fusing global and local information by Ju: 2022

Ju et al. [39] proposed a detection method that combines global and local features to improve generalization for synthetic image detection. The framework uses a ResNet-50 backbone to extract global feature maps from input images. These global features are complemented by a Patch Selection Module (PSM), which identifies and processes the most informative patches to capture subtle, localized artifacts. The PSM selects patches by sliding windows of sizes 3×3 and 2×2 over the activation maps, scoring each patch based on its aggregated activation. The top patches ($k = 6$) are mapped back to the original image and reprocessed through ResNet-50 for local feature extraction. Subsequently, an Attention-based Feature Fusion Module (AFFM) combines the global and local features through multi-head attention, generating a unified representation for classification. The authors evaluated their method using a dataset synthesized by 19 different models, including GANs and autoencoder-based DeepFakes. Experimental results showed improved performance over prior methods, particularly under diverse post-processing conditions such as Gaussian blur and JPEG compression.

2.6.4. Zhong's PatchCraft: 2024

Zhong et al. [99] introduced PatchCraft, a novel framework emphasizing texture patches over global semantic features for detecting synthetic images. The key innovation is a preprocessing technique, Smash and Reconstruction, which segments an image into patches, ranks them by texture diversity—measured as the sum of pixel differences in horizontal, vertical, and diagonal directions—and reconstructs two images: one enriched with rich-texture patches and another with poor-texture patches. Both reconstructed images are processed through Spatial Rich Model (SRM) filters [29] to extract high-frequency noise patterns, which are further analyzed using a learnable convolutional block. The residual noise patterns between rich and poor texture regions capture inter-pixel correlations, forming a fingerprint for generative models. This fingerprint is leveraged by a convolutional neural network (CNN) classifier, trained with cross-entropy loss, to distinguish between real and synthetic images. During inference, the same pipeline extracts the fingerprint, enabling an accurate classification of input images across diverse generative models.

2.6.5. Chen's single patch method for AI-generated image detection (2024)

Chen et al. [13] proposed the Single Simple Patch (SSP) network, designed to identify whether an image is real or AI-generated by analyzing a single low-texture patch. The selected patch, with dimensions $M \times M$, is determined by having the lowest texture diversity among all image patches, calculated using the method in [99]. The selected patch is resized to match the original image dimensions and processed using the SRM [29] to extract high-frequency noise patterns. These noise patterns are then fed into a ResNet-50 [35] classifier, which is trained with binary cross-entropy loss to distinguish real images from generated ones. To address performance challenges with low-quality images (e.g., those affected by blur or compression artifacts), Chen et al. introduced two additional modules: an enhancement module and a perception module. The perception module, a lightweight three-class classifier, identifies whether a patch is blurry, compressed, or intact. Its predictions guide the enhancement module, which uses a U-Net architecture to improve the patch quality by performing deblurring, decompression,

Table 8
Evaluation of fingerprint analysis methods on GAN and diffusion-generated images.

GAN-Based Evaluations using Forensynths dataset [87]). Accuracy (Acc) and Average Precision (AP) metrics are reported.
Tan et al. [81] was trained on class Bedroom from ProGAN-generated image, while remaining were trained on class Horse.

Method	StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
Jeong [37]	74.1	85.3	89.5	96.1	85.0	94.8	71.2	96.9	99.9	100.0	75.9	90.9	82.6	94.0
Tan et al. [81]	82.60	95.60	83.30	98.40	76.20	81.80	82.30	90.60	99.70	100.0	71.70	75.00	80.90	87.40
Wang et al. (Scaling) [85]	85.7	98.6	83.8	98.2	81.2	85.3	83.3	93.9	99.1	100.0	75.1	81.3	84.7	92.9
Wang et al. (Mixup) [85]	82.2	98.7	78.0	98.1	79.1	84.8	86.4	95.6	98.8	100.0	83.4	90.3	84.7	94.6

Diffusion-Based Evaluations on the dataset as mentioned in (Li et al. [49]). Area Under the Curve (AUC) and Accuracy (Acc) metrics are reported.

Method	SD-1		SD-2		SD-XL		DALL.E 2		DALL.E 3		Midjourney		Firefly		Mean	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
Corvi et al. [20]	100.0	99.6	99.5	97.2	98.9	80.4	48.8	49.9	54.9	49.7	99.8	95.0	86.2	52.4	84.0	74.9
Li et al. [49]	89.4	75.5	99.1	95.9	96.6	90.0	68.2	55.4	90.2	75.3	96.4	90.9	76.0	64.0	88.3	79.4

Table 9
Evaluation of patch-based analysis methods on cross-family generators, cross-category, and cross-scene generalization.

Models	Cross-family generators	Cross-category	Cross-scene
Chai et al. [10]	✗	✗	✗
Mandelli et al. [57]	✗	✓	✗
Ju et al. [39]	✓	✓	✗
Zhong et al. [99]	✓	✓	✗
Chen et al. [13]	✓	✓	✗
Chen et al. [12]	✓	✓	✗

or reconstruction. The improved patch is then processed by the SSP network.

2.6.6. Inter-patch dependencies for AI-generated image detection by Chen: 2024 (IPD-Net)

Chen et al. [12] introduced IPD-Net, a detection framework that leverages inter-patch dependencies to improve generalization for AI-generated image detection. This work builds on previous research by Zhong et al. [99], which demonstrated that inconsistencies in interpixel relations between rich and poor texture regions can serve as key features for detection. IPD-Net extends this by modeling dependencies between patches using a self-attention mechanism. The framework consists of two main modules: Inter-Patch Dependencies Extraction and Inter-Patch Dependencies Classification. In the extraction module, input images are preprocessed with operations such as Gaussian blur and JPEG compression before being passed through SRM filters [29] to extract noise patterns. A non-trained ResNet-50 backbone generates feature maps, and patch dependencies are computed using dot-product similarity across all patches. In the classification module, the dependency matrix undergoes two-dimensional average pooling to reduce dimensionality. A linear classification layer with sigmoid activation then predicts whether an image is real or synthetic. The model is trained using binary cross-entropy loss, and the architecture supports end-to-end training. Experimental evaluations on the Forensynths [87] and GenImage [102] datasets demonstrated that IPD-Net outperforms state-of-the-art baseline models in both in-dataset and cross-dataset evaluations, showcasing strong generalization capabilities.

Comparative analysis of patch-based analysis methods. Patch-based detection approaches aim to enhance sensitivity to localized generative artifacts by focusing on smaller image regions rather than global context. Chai et al. [10] demonstrated that limiting the receptive field to local patches enables the network to detect fine-grained generative inconsistencies, improving generalization across GAN architectures. Mandelli et al. [57] extended the concept by employing orthogonal training across semantically diverse datasets and generative sources, enhancing robustness to unseen models and post-processing operations. Ju et al. [39] further improved representation quality by integrating both global and local cues through a patch selection and feature fusion mechanism, achieving balanced detection performance under varied degradation scenarios. Zhong et al. [99] emphasized texture-based evidence using PatchCraft, which distinguished between rich- and poor-texture regions to extract residual fingerprints indicative of generative artifacts. Chen et al. [13] refined this perspective with the Single Simple Patch (SSP) network, showing that even a single low-texture region can encode sufficient forensic information when enhanced for degradations such as blur or compression. Finally, Chen et al. [12] proposed IPD-Net, which modeled inter-patch dependencies through self-attention, capturing relational inconsistencies across image regions and achieving strong generalization across datasets. Together, these works trace a clear methodological evolution—from independent local patch analysis to integrated global–local reasoning and inter-patch dependency modeling—demonstrating how localized feature extraction has matured into a robust and generalizable paradigm for synthetic image forensics (Table 9).

2.7. Commercial detection methods

Watermarking by DeepMind [25] is a commercial technology that introduces watermarking to the generated image invisible to human eyes, which works as basis for the proper detection of the synthesized image.

3. Discussion and limitation

Each category of forensic detection contributes distinct insights into identifying AI-generated imagery, forming a complementary ecosystem of methodologies. Spatial-domain and frequency-domain analyses remain the foundation of pixel- and signal-level forensic understanding. Spatial detectors such as DIRE [89] and SelfCon [7] have moved beyond early CNN-based designs by integrating reconstruction and contrastive learning for improved generalization, whereas frequency-based frameworks like SAFE [47] and AIDE [92] extend robustness by coupling spectral cues with semantic embeddings. Together, these advances demonstrate that spatial–frequency fusion provides a more holistic view of generation artifacts, though both remain limited under heavy compression and geometric transformations. Fingerprint and patch-based methods address these weaknesses from orthogonal perspectives. Fingerprint-based detectors such as MaskSim [49] and LGrad [81] mitigate model dependence through spectral and gradient-informed fingerprints, achieving cross-generator consistency absent in earlier handcrafted designs. Patch-based frameworks, including IPD-Net [12] and PatchCraft [99], enhance sensitivity to localized texture inconsistencies while maintaining contextual awareness through inter-patch relationships. These directions collectively narrow the gap between local artifact discovery and global scene understanding. Successful yet, these methods require extensive training or finetuning, which is computationally expensive, as well as limiting in scope since they require retraining

Table 10
Evaluation of patch-based analysis methods on GAN and diffusion-generated images.

GAN-Based Evaluations using Forensynths dataset [87]. Accuracy (Acc) is reported.							
Method	ProGAN	StyleGAN	StyleGAN2	BigGAN	CycleGAN	StarGAN	GauGAN
Chai et al. [10]	75.03	79.16	_	_	_	_	_
Ju et al. [39]	100.0	85.20	83.30	77.40	87.00	97.00	77.00
Zhong et al. [99]	100.0	92.77	89.55	95.8	70.17	99.97	71.58
Chen et al. [13]	97.05	96.05	_	68.65	83.25	95.00	57.85
Chen et al. [12]	99.98	95.19	_	81.02	86.57	99.08	68.67

Diffusion-Based Evaluations on the GenImage dataset [102]. Accuracy (Acc) metrics are reported.
Chen et al. [13] and Chai et al. [10] were trained on SD V1.4, while all other methods were trained on ProGAN from the ForenSynths dataset.
The performance metrics for Chai et al. were recorded from the work of Meng et al. [59]

Method	SDv1.4	SD-1.5	ADM	Glide	Midjourney	VQDM	wukong	DALLE2	SDXL
Chai et al. [10]	99.70	99.40	51.00	54.10	66.20	54.10	96.70	_	_
Ju et al. [39]	51.00	51.40	49.00	57.20	52.20	55.10	51.70	52.80	55.60
Zhong et al. [99]	95.38	95.30	82.17	83.79	90.12	88.91	91.07	96.60	98.43
Chen et al. [13]	99.20	99.30	78.90	88.90	82.60	96.00	98.60	_	_
Chen et al. [12]	80.03	79.70	84.38	95.05	72.19	79.37	77.19	_	_

when new and advanced generative models enter the market. In recent years, these limitations have been addressed by training-free approaches such as RIGID [109], HFI [106], and ZED [107], which depart from traditional learning paradigms by requiring no training on AI-generated images, instead leveraging statistical stability, frequency aliasing, or entropy deviation. While these methods excel in scalability and generator-agnostic detection with low computational cost, they still lack interpretability and resilience to adversarial perturbations.

Even more recently, vision-language and reasoning-based multimodal models—spanning CLIP-based frameworks like HyperDet [9] to reasoning-driven systems such as ForenX [116] and ThinkFake [110]—push the field toward semantically aligned and explainable forensics with reasoning. These approaches overcome the semantic blindness of classical detectors, offering text-grounded rationales for model decisions, though they rely heavily on computationally expensive foundational backbones and curated instruction data. Importantly, Fan et al. [108] present detection that is fully dependent on LLMs without integrating understanding between text–vision data and still provide comparable results. While this is a great perspective on the usage of LLMs, it might suffer in generalizability or from hallucination, as LLMs may provide contextual reasoning with incorrect information during detection.

Finally, commercial detection frameworks unify several of these perspectives by prioritizing scalability and usability over transparency. They often integrate pretrained multimodal backbones and frequency cues for enterprise-level deployment, but their proprietary nature limits academic reproducibility and benchmarking.

For clarity, we present a comparative analysis of methods across six distinct categories, following a general review of their reported performances in the literature. The results are detailed in Tables 3, 8, 10, and 5. Additionally, a comprehensive performance comparison across all six categories using the UnivFD dataset [63] is provided in Table 11 for classification accuracy and Table 12 for average precision. The evaluation results on the UnivFD dataset [63] are selected due to its inclusion of state-of-the-art detection methods, particularly those utilizing vision-language approaches, which have demonstrated strong forensic capabilities in synthetic image detection. The recent perspective of reasoning provided by Multimodal Large Language Models (MLLMs) helps in moving from a black box to a grey box understanding—explaining why content is generated for the general public. A limitation we faced for these methods' quantitative comparative analysis is that they are recent and present new datasets, often compared with general MLLM models that were not designed for forensic detection; such comparisons differ from the scope of this study. Therefore, we only include the quantitative evaluation of ForenX [116] by merging it with training-free methods in Table 5.

While previous methods from spatial-domain, frequency-domain, fingerprint-based, and patch-based categories provide good detection results on different datasets, they may suffer from generalizability issues and computational expense due to dependence on the training of generative methods. Training-free methods could provide generalizability and computational efficiency, yet they may fail when the generated images exhibit very high fidelity. MLLM (VLM) methods could offer strong detection performance with reasoning—aligning with recent trends in the AI field—but they may be limited by extensive computational requirements and may suffer from the inherent issue of LLM hallucination. A hybrid structure leveraging the cross-category functions of MLLM and training-free methods may provide a new perspective on robustness, reasoning, generalizability, and efficiency.

4. Datasets

4.1. ForenSynths by Wang: 2020

The dataset used by Wang et al. [87] is publicly available on their GitHub page, as referenced in the main paper. The training set consists of 724,000 images, including 362,000 real images and 362,000 fake images generated by ProGAN [40], trained on the LSUN dataset [95] across 20 different object categories. The testing dataset, which includes images generated by various other GAN models, is also accessible on the GitHub page.

4.2. Artifact dataset by Rahman: 2023

Rahman et al. [68] introduced the Artifact dataset, containing 2,496,738 images, including 964,989 real images and 1,531,749 synthetic images generated using 25 different models (both GANs and diffusion models). The dataset includes diverse object categories representative of general real-world content.

4.3. SynthBuster by Bammey: 2023

Bammey et al. [3] introduced the SynthBuster dataset, consisting of 9000 synthetic images, with 1000 images generated by each of nine different diffusion models, namely: DALL.E 2 [70], DALL.E 3 [70], Adobe Firefly [1], Midjourney v5 [60], Stable Diffusion [73] 1.3, 1.4, 2, and XL, and GLIDE [62]. The dataset is publicly available, with access details provided in the main paper. For real images, the authors recommend using the RAISE-1k dataset [24], accessible via the same portal.

4.4. DiffusionForensics dataset by Wang et al.: 2023

Wang et al. [89] introduced the DiffusionForensics dataset, specifically designed to evaluate forensic detectors for diffusion-generated

Table 11
Comprehensive evaluation of methods on GAN and diffusion-generated images using the UnivFD dataset [63]. Results are reported in classification accuracy (%).

Method	Generative adversarial networks							Low-level vision		Perceptual loss		Diffusion models								avg.
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN	Deep-fakes	SITD	SAN	CRN	IMLE	Guided	LDM			GLIDE			DALL-E	
													200 steps	200s w/CFG	100 steps	100 27 27	50 27 27	100 10 10		
Wang et al. [87]	100.0	80.49	55.77	64.14	82.23	80.97	50.66	56.11	50.00	87.73	92.85	52.30	51.20	52.20	51.40	53.45	55.35	54.30	52.60	64.41
Chai et al. [10]	68.81	53.02	55.76	59.24	52.64	77.49	55.78	59.65	48.80	65.57	61.69	52.26	58.53	60.72	58.21	55.78	56.58	55.05	61.24	58.78
Ojha et al. [63]	100.0	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.5	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
Wang et al. [89]	100.0	67.73	64.78	83.08	65.30	100.0	94.75	57.62	60.96	62.36	62.31	83.20	82.70	84.05	84.25	87.10	90.80	90.25	58.75	77.89
Tan et al. [81]	99.90	85.10	83.00	94.80	72.50	99.60	56.40	47.80	41.10	50.60	50.70	74.20	94.20	95.90	95.00	87.20	90.80	89.80	88.40	_
Corvi et al. [20]	100.00	92.00	96.90	99.40	94.80	99.50	54.10	90.60	55.50	100.00	100.00	53.90	58.00	61.10	57.50	56.90	59.60	58.80	71.70	_
Zhu et al. [101]	99.00	99.50	99.30	99.05	99.00	96.75	88.20	63.50	67.50	93.90	98.75	98.70	98.80	98.60	98.75	98.75	98.75	98.75	98.45	94.42
Liu et al. [53]	100.0	99.33	99.67	99.46	99.83	97.07	77.53	81.11	65.50	82.32	96.79	90.70	98.30	95.90	98.75	92.40	93.95	93.00	94.90	92.45
Cao et al. [9]	100.0	97.40	97.50	97.50	96.20	98.65	73.85	93.00	75.00	92.75	93.20	77.35	98.70	96.60	98.80	87.75	89.95	88.70	97.00	92.10
Liu et al. [51]	99.90	99.30	99.50	97.20	99.40	99.80	93.20	_	_	_	_	76.10	98.60	94.90	98.70	94.40	94.70	94.20	98.80	_
Koutlis et al. [44]	100.0	99.30	99.60	88.90	99.80	99.50	80.60	90.60	68.30	89.20	90.60	76.10	98.30	88.20	98.60	88.90	92.60	90.70	95.00	_

Table 12
Comprehensive evaluation of methods on GAN and diffusion-generated images using the UnivFD dataset [63]. Results are reported in average precision (%).

Method	Generative adversarial networks							Low-level vision		Perceptual loss		Diffusion models								mAP
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN	Deep-fakes	SITD	SAN	CRN	IMLE	Guided	LDM			GLIDE			DALL-E	
													200 steps	200s w/CFG	100 steps	100 27 27	50 27 27	100 10 10		
Wang et al. [87]	100.0	96.36	85.34	98.10	98.48	96.97	60.33	82.95	54.22	99.61	99.81	69.93	66.17	67.68	66.13	71.18	76.37	72.13	67.66	80.50
Chai et al. [10]	68.44	55.59	64.37	64.10	58.74	84.48	59.92	72.08	47.63	73.05	68.38	58.98	77.05	76.87	76.35	75.97	77.41	74.68	71.91	68.74
Ojha et al. [63]	100.0	99.46	99.59	97.24	99.98	99.60	82.45	61.32	79.02	96.72	99.00	87.77	99.14	92.15	99.17	94.74	95.34	94.57	97.15	93.38
Wang et al. [89]	100.0	76.73	72.80	97.06	68.44	100.0	98.55	54.51	65.62	97.10	93.74	94.29	95.17	95.43	95.77	96.18	97.30	97.53	68.73	87.63
Corvi et al. [20]	100.00	98.60	99.80	100.0	99.80	100.0	94.70	99.80	87.70	100.00	100.00	73.00	86.80	89.40	87.30	86.50	89.90	89.00	96.10	_
Zhu et al. [101]	99.95	99.95	99.92	99.92	99.92	99.25	91.38	61.23	72.66	97.90	98.88	99.30	99.85	99.51	99.85	99.50	99.46	99.19	99.47	95.64
Liu et al. [53]	100.0	99.85	99.88	99.69	100.0	99.68	87.38	88.26	84.48	98.82	99.84	93.39	99.81	96.80	99.88	98.71	98.84	98.60	98.81	96.99
Cao et al. [9]	100.0	99.96	99.89	99.73	99.93	100.0	88.38	97.12	89.22	98.82	99.98	95.31	99.86	99.14	99.90	97.20	97.99	98.02	99.65	97.90
Tan et al. [81]	100.0	94.00	90.70	99.90	79.30	100.0	67.90	_	_	_	_	100.0	99.10	99.20	99.20	93.20	95.10	94.90	97.30	_
Liu et al. [51]	100.0	100.0	100.0	99.80	100.0	100.0	98.00	_	_	_	_	92.00	99.80	99.10	99.90	99.10	99.40	99.20	99.80	_
Koutlis et al. [44]	100.0	100.00	99.90	99.40	100.0	100.0	97.90	97.20	94.90	97.30	99.70	96.40	99.80	98.30	99.90	98.80	99.30	98.90	99.30	_

images. The dataset includes images sourced from LSUN-Bedroom [95], ImageNet [75], and CelebA-HQ [40], with the following distributions: 42,000 synthetic images generated from LSUN-Bedroom, - 50,000 synthetic images generated from ImageNet, and - 42,000 synthetic face images generated using Stable Diffusion V2 (SD-V2), paired with 42,000 real images from CelebA-HQ.

4.5. UnivFD dataset by Ojha et al.: 2023

Ojha et al. [63] expanded the ForenSynths dataset by integrating images generated from various models, including ProGAN [40], CycleGAN [100], BigGAN [8], StyleGAN [41], GauGAN [66], StarGAN [17], Deepfakes [74], SITD [11], SAN [23], CRN [14], and IMLE [46]. In addition, diffusion-based models such as the Guided diffusion model (ADM) [26], LDM [73], GLIDE [62], and DALL-E [71] were included, increasing its diversity for forensic detection studies.

4.6. Community forensics by Park: 2024

Many existing datasets contain images generated by diverse generative models; however, they often lack diversity and generalization capability for many-to-many scenarios. These scenarios include testing images generated from multiple diffusion models against those from various GANs, as well as incorporating generative models such as VAEs. Recognizing the need for a large, diverse dataset capable of detecting synthetic images from unseen generative models, Park and Owens [65] introduced the Community Forensics Dataset. This dataset comprises 2.4 million images, generated by 4803 distinct generative models, alongside an equal number of real images, making it more diverse than previous datasets. To evaluate its effectiveness, the authors trained and tested state-of-the-art models on this dataset, utilizing a binary classification setup with pre-trained models ViT-S [27] and ConvNeXt-S [52] as backbone architectures. Unlike previous methods that freeze backbone layers, the authors fine-tuned the entire backbone end-to-end, achieving better performance compared to existing state-of-the-art methods on both the Community Forensics dataset and other publicly available benchmarks.

4.7. GenImage by Zhu: 2024

Zhu et al. [102] introduced a large-scale dataset containing synthetic images generated by advanced state-of-the-art GANs and diffusion models (Wukong [91] and VQDM [33], including commercial models, alongside real images. The dataset comprises approximately 1.3 million synthetic images and 1.3 million real images, covering diverse general image content. The authors also propose two real-world analysis factors to assess detection performance: (a) Cross-Generator Image Classification: training detectors on images generated by one model and evaluating them on images from different generators, and (b) Degraded Image Classification: testing detectors on degraded images affected by factors such as low resolution, JPEG compression, and Gaussian blur.

4.8. CIFAKE dataset by bird and Lotfi: 2024

Bird and Lotfi [6] released a dataset comprising 120,000 images, equally divided between 60,000 real images and 60,000 synthetic images. The real images originate from CIFAR-10 [45], which includes ten object categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The synthetic images were generated using diffusion models, including commercially available ones.

4.9. AIGCDetection benchmark dataset by Zhong et al.: 2024

Zhong et al. [99] introduced a curated benchmark dataset incorporating images generated by 17 different generative models, including GANs and diffusion models. The dataset aggregates samples from ForenSynths [87] and GenImage [102], and includes additional images from recent diffusion models, offering a comprehensive benchmark for generative image forensics.

4.10. ImagiNet dataset by Boychev: 2024

Boychev et al. [7] presented the ImagiNet dataset, comprising 200,000 images equally split between real and synthetic categories. The dataset further categorizes images into photos, paintings, faces, and uncategorized types. The synthetic images are generated using GANs, diffusion models, and proprietary generative models. For evaluation, the dataset is divided into 160,000 training images and 40,000 test images.

4.11. Chameleon dataset by Yan et al.: 2024

Yan et al. [92] introduced the Chameleon dataset, which focuses on realistic AI-generated image detection. The dataset consists of 150,000 synthetic images covering content categories such as humans, animals, scenes, and objects, generated using GANs and diffusion models. In addition, it includes 20,000 real images to facilitate forensic evaluation.

4.12. Metrics

The performance of synthetic image detection is primarily evaluated using classification accuracy (Acc) and average precision (AP), which are mathematically expressed as follows:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{AP} = \sum_{n} (R_n - R_{n-1}) P_n \tag{7}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. The AP is computed as the weighted sum of precision P_n at different recall levels R_n.

5. Conclusion

The advancement of generative AI in image synthesis and forensic detection has been an ongoing and increasingly active research area, driven by both the successes of AI and the ethical concerns it introduces. From a practical standpoint, forensic detection must remain ahead to mitigate potential human-related harm. While several reviews exist in the current literature, they do not comprehensively cover detection methods capable of identifying images generated by state-of-the-art generative models, nor do they adequately address the role of multimodal approaches in detection. To bridge this gap, we have conducted a comprehensive survey of detection methodologies. First, we categorize these methods into seven primary groups based on their underlying approaches. We then describe each method in detail, followed by an analysis of their comparative performance on publicly available datasets, and assess whether they satisfy three key criteria for evaluating their generalizability. Our findings indicate that detection methods leveraging multimodal frameworks tend to exhibit greater robustness and adaptability across different generative models. Building upon this comparative analysis, we observe that spatial- and frequency-domain methods form the foundational layer of forensic understanding by identifying low-level pixel and spectral inconsistencies, yet they often struggle under compression or strong generative realism. Fingerprint- and patch-based methods address these weaknesses through architecture-agnostic learning and localized texture analysis, improving sensitivity and cross-generator robustness. Training-free approaches further enhance scalability by eliminating retraining needs, although interpretability and adversarial resilience remain open challenges. Finally, multimodal and reasoning-based models, particularly those integrating vision-language and large language models, bring explainability and semantic grounding but at the expense of heavy computational requirements. Looking ahead, the convergence of these paradigms offers a promising pathway. Hybrid frameworks that combine the reasoning and semantic grounding of multimodal models with the lightweight efficiency of training-free detection could lead to robust, interpretable, and real-time forensic systems. Future research should also emphasize cross-modal benchmarking and unified evaluation standards.

Through these directions, the field can advance toward trustworthy and transparent generative AI systems capable of safeguarding authenticity in the digital era.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Adobe Inc, Adobe firefly, 2024, https://www.adobe.com/products/firefly (Accessed 8 December 2024).
[2] N. Ahmed, T_ Natarajan, K.R. Rao, Discrete cosine transform, IEEE Trans. Comput. 100 (1) (1974) 90–93.
[3] Q. Bammey, Synthbuster: towards detection of diffusion model generated images, IEEE Open J. Signal Process. 5 (2024) 1–9, https://doi.org/10.1109/OJSP.2023.3337714
[4] S.S. Baraheem, T.V. Nguyen, AI VS. AI: can AI detect AI-generated images? J. Imaging 9 (10) (2023) 199.
[5] J.B.J.B. Fourier, et al., The Analytical Theory of Heat, Courier Corporation, 2003.
[6] J.J. Bird, A. Lotfi, CIFAKE: image classification and explainable identification of AI-generated synthetic images, IEEE Access (2024).
[7] D. Boychev, R. Cholakov, ImagiNet: a multi-content dataset for generalizable synthetic image detection via contrastive learning, arXiv preprint arXiv:2407.20020, 2024.
[8] A. Brock, Large scale GAN training for high fidelity natural image synthesis, arXiv preprint arXiv:1809.11096, 2018.
[9] H. Cao, Y. Wang, Y. Liu, S. Zheng, L. Kangtao, Z. Zhang, B. Zhang, X. Ding, W. Fei, HyperDet: generalizable detection of synthesized images by generating and merging a mixture of hyper LoRAs, arXiv preprint arXiv:2410.06044, 2024.
[10] L. Chai, D. Bau, S.-N. Lim, P. Isola, What makes fake images detectable? Understanding properties that generalize, in: Computer Vision–ECCV 2020: 16th European Conference Proceedings, Part XXVI, August 23–28, Glasgow, UK. Springer, 2020, pp. 103–120, *16*).
[11] C. Chen, Q. Chen, X. Jia, V. Koltun, Learning to see in the dark, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3291–3300.
[12] J. Chen, L. Mengtin, H. Liao, T. Huang, IPD-Net: detecting AI-generated images via Inter-Patch dependencies, Int. J. Adv. Comput. Sci. Appl. 15 (7) (2024).
[13] J. Chen, J. Yao, L. Niu, A single simple patch is all you need for AI-generated image detection, arXiv preprint arXiv:2402.01123, 2024.
[14] Q. Chen, V. Koltun, Photographic image synthesis with cascaded refinement networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1511–1520.
[15] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning. PMLR, 2020, pp. 1597–1607.
[16] Y.-L. Chen, C.-T. Hsu, Image tampering detection by blocking periodicity analysis in JPEG compressed images, in: 2008 IEEE 10th Workshop on Multimedia Signal Processing, IEEE, 2008, pp. 803–808.
[17] Y. Choi, M. Choi, M. Kim, H. Jung-Woo, S. Kim, J. Choo, StarGAN: unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[18] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[19] J.W. Cooley, J.W. Tukey, An algorithm for the machine calculation of complex Fourier series, Math. Comp. 19 (90) (1965) 297–301.
[20] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic images generated by diffusion models, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.
[21] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, L. Verdoliva, Raising the bar of AI-generated image detection with CLIP, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4356–4366.
[22] D. Cozzolino, L. Verdoliva, Noiseprint: a CNN-based camera model fingerprint, IEEE Trans. Inf. Forensics Secur. 15 (2019) 144–159.
[23] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, L. Zhang, Second-order attention network for single image super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11065–11074.
[24] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, G. Boato, Raise: a raw images dataset for digital image forensics, in: Proceedings of the 6th ACM Multimedia Systems Conference, 2015, pp. 219–224.
[25] DeepMind, Identifying AI-generated images with SynthID, 2024, https://deepmind.google/technologies/synthid/ (Accessed 12 December 2024).
[26] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.
[27] A. Dosovitskiy, An image is worth 16×16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.
[28] T. Dzanic, K. Shah, F. Witherden, Fourier spectrum discrepancies in deep network generated images, Adv. Neural Inf. Process. Syst. 33 (2020) 3022–3032.
[29] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Trans. Inf. Forensics Secur. 7 (3) (2012) 868–882.
[30] M. Goebel, L. Nataraj, T. Nanjundaswamy, T.M. Mohammed, S. Chandrasekaran, B.S. Manjunath, Detection, attribution and localization of GAN generated images, arXiv preprint arXiv:2007.10466, 2020.
[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. (2014) 27.
[32] Google DeepMind, Imagen 3, 2024, https://deepmind.google/technologies/imagen-3 (Accessed 8 December 2024).
[33] G. Shuyang, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, B. Guo, Vector quantized diffusion model for text-to-image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10696–10706.
[34] A. Haar, Zur theorie DER orthogonalen funktionensysteme, Math. Ann. 71 (1) (1911) 38–53.
[35] H. Kaiming, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[36] H. Jonathan, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[37] Y. Jeong, D. Kim, R. Youngmin, P. Kim, J. Choi, Fingerprintnet: synthesized fingerprints for generated image detection, in: European Conference on Computer Vision, Springer, 2022, pp. 76–94.
[38] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, LayerCAM: exploring hierarchical class activation maps for localization, IEEE Trans. Image Process. 30 (2021) 5875–5888.
[39] J. Yan, S. Jia, K. Lipeng, H. Xue, K. Nagano, S. Lyu, Fusing global and local features for generalized AI-synthesized image detection, IEEE Int. Conf. Image Process. (2022) 3465–3469 IEEE.
[40] T. Karras, Progressive growing of GANs for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196, 2017.
[41] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.
[42] K. Guolin, Q. Meng, T. Finley, T. Wang, W. Chen, M. Weidong, Y. Qiwei, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. (2017) 30.
[43] D.P. Kingma, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114, 2013.
[44] C. Koutlis, S. Papadopoulos, Leveraging representations from intermediate encoder-blocks for synthetic image detection, in: European Conference on Computer Vision, Springer, 2025, pp. 394–411.
[45] A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report, University of Toronto, 2009.
[46] K. Li, T. Zhang, J. Malik, Diverse image synthesis from semantic layouts via conditional IMLE, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4220–4229.
[47] L. Ouxiang, J. Cai, Y. Hao, X. Jiang, H. Yao, F. Feng, Improving synthetic image detection towards generalization: an image transformation perspective, arXiv preprint arXiv:2408.06741, 2024.
[48] L. Weichuang, H. Peisong, L. Haoliang, H. Wang, R. Zhang, Detection of GAN-generated images by estimating artifact similarity, IEEE Signal Process. Lett. 29 (2022) 862–866, https://doi.org/10.1109/LSP.2021.3130525
[49] L. Yanhao, Q. Bammey, M. Gardella, T. Nikoukhah, J.-M. Morel, M. Colom, R.G.V. Gioi, MaskSim: detection of synthetic images by masked spectrum similarity analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3855–3865.
[50] B. Liu, F. Yang, B. Xiuli, B. Xiao, L. Weisheng, X. Gao, Detecting generated images by real images, in: European Conference on Computer Vision, Springer, 2022, pp. 95–110.
[51] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, Y. Zhao, Forgery-aware adaptive transformer for generalizable synthetic image detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10770–10780.
[52] Z. Liu, H. Mao, W. Chao-Yuan, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[53] Z. Liu, H. Wang, Y. Kang, S. Wang, Mixture of low-rank experts for transferable AI-generated image detection, arXiv preprint arXiv:2404.04883, 2024.
[54] P. Lorenz, R.L. Durall, J. Keuper, Detecting images generated by deep diffusion models using their local intrinsic dimensionality, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 448–459.
[55] P. Lorenz, M. Keuper, J. Keuper, Unfolding local growth rate estimates for (almost) perfect adversarial detection, arXiv preprint arXiv:2212.06776, 2022.
[56] J. Lukas, J. Fridrich, M. Goljan, Digital camera identification from sensor pattern noise, IEEE Trans. Inf. Forensics Secur. 1 (2) (2006) 205–214.
[57] S. Mandelli, N. Bonettini, P. Bestagini, S. Tubaro, Detecting GAN-generated images by orthogonal training of multiple CNNS, IEEE Int. Conf. Image Process. (2022) 3091–3095 IEEE.

[58] F. Marra, D. Gragnaniello, L. Verdoliva, G. Poggi, Do GANs leave artificial fingerprints? in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2019, pp. 506–511.
[59] Z. Meng, B. Peng, J. Dong, T. Tan, H. Cheng, Artifact feature purification for cross-domain detection of AI-generated images, Comput. Vis. Image Underst. 247 (2024) 104078.
[60] MidJourney, MidJourney, 2024, https://www.midjourney.com/ (Accessed 8 December 2024).
[61] M. Mylrea, The generative AI weapon of mass destruction: evolving disinformation threats, vulnerabilities, and mitigation frameworks, in: Interdependent Human-Machine Teams, Elsevier, 2025, pp. 315–347.
[62] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, Glide: towards photorealistic image generation and editing with text-guided diffusion models, arXiv preprint arXiv:2112.10741, 2021.
[63] U. Ojha, L. Yuheng, Y.J. Lee, Towards universal fake image detectors that generalize across generative models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24480–24489.
[64] OpenAI, DALLE 3, 2024, https://openai.com/index/dall-e-3/ (Accessed 8 December 2024).
[65] J. Park, A. Owens, Community forensics: Using thousands of generators to train fake image detectors, arXiv preprint arXiv:2411.04125, 2024.
[66] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.
[67] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning. PMLR, 2021, pp. 8748–8763.
[68] M.A. Rahman, B. Paul, N.H. Sarker, Z.I.A. Hakim, S.A. Fattah, ArtiFact: a large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection, IEEE Int. Conf. Image Process. (2023) 2200–2204 IEEE.
[69] H.G. Ramaswamy, et al., Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 983–991.
[70] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125, 2022 1, 2, 3.
[71] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning. PMLR, 2021, pp. 8821–8831.
[72] D. Rezende, S. Mohamed, Variational inference with normalizing flows, in: International Conference on Machine Learning. PMLR, 2015, pp. 1530–1538.
[73] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.
[74] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, FaceForensics++: learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.
[75] O. Russakovsky, J. Deng, S. Hao, J. Krause, S. Satheesh, M. Sean, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[76] L. Sandrini, R. Somogyi, Generative AI and deceptive news consumption, Econ. Lett. 232 (2023) 111317.
[77] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[78] H.T. Sencar, N. Memon, Digital Image Forensics, Springer, 2013.
[79] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International Conference on Machine Learning. PMLR, 2015, pp. 2256–2265.
[80] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv preprint arXiv:2010.02502, 2020.
[81] C. Tan, Y. Zhao, S. Wei, G. Guanghua, Y. Wei, Learning on gradients: generalized artifacts representation for GAN-generated images detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12105–12114.
[82] M. Tan, L. Quoc, EfficientNet: rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning. PMLR, 2019, pp. 6105–6114.
[83] A. Van Den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in: International Conference on Machine Learning. PMLR, 2016, pp. 1747–1756.
[84] M. Laurens Van der, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 11.
[85] H. Wang, J. Fei, Y. Dai, L. Leng, Z. Xia, General GAN-generated image detection by data augmentation in fingerprint domain, in: 2023 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2023, pp. 1187–1192.
[86] H. Wang, Z. Wang, D. Mengnan, F. Yang, Z. Zhang, S. Ding, P. Mardziel, H. Xia, Score-CAM: score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 24–25.
[87] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A.A. Efros, CNN-generated images are surprisingly easy to spot for now, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8695–8704.
[88] T. Wang, X. Liao, K.P. Chow, X. Lin, Y. Wang, Deepfake detection: a comprehensive survey from the reliability perspective, Comput. Surv. 57 (3) (2024) 1–35.
[89] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hezhen, H. Chen, L. Houqiang, Dire for diffusion-generated image detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22445–22455.
[90] W. Haiwei, J. Zhou, S. Zhang, Generalizable synthetic image detection via language-guided contrastive learning, arXiv preprint arXiv:2305.13800, 2023.
[91] Wukong, Wukong, 2022. https://xihe.mindspore.cn/modelzoo/wukong.
[92] S. Yan, L. Ouxiang, J. Cai, Y. Hao, X. Jiang, H. Yao, W. Xie, A sanity check for AI-generated image detection, arXiv preprint arXiv:2406.19435, 2024.
[93] T. Yang, J. Cao, Q. Sheng, L. Lei, J. Jiaqi, L. Xirong, S. Tang, Learning to disentangle gan fingerprint for fake image attribution, arXiv preprint arXiv:2106.08749, 2021.
[94] B. Yousaf, M. Usama, W. Sultani, A. Mahmood, J. Qadir, Fake visual content detection using two-stream convolutional neural networks, Neural Comput. Appl. 34 (10) (2022) 7991–8004.
[95] Y. Fisher, A. Seff, Y. Zhang, S. Song, T. Funkhouser, J. Xiao, LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint arXiv:1506.03365, 2015.
[96] C.J. Zhang, A.Q. Gill, B. Liu, M.J. Anwar, AI-based identity fraud detection: a systematic review, arXiv preprint arXiv:2501.09239, 2025.
[97] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising, IEEE Trans. Image Process. 26 (7) (2017) 3142–3155.
[98] L. Zheng, Y. Zhang, V.LL. Thing, A survey on image tampering and its detection in real-world photos, J. Vis. Commun. Image Represent. 58 (2019) 380–399.
[99] N. Zhong, X. Yiran, L. Sheng, Z. Qian, X. Zhang, Patchcraft: Exploring texture patch for efficient ai-generated image detection, arXiv preprint arXiv:2311.12397, 2024, 1–18.
[100] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[101] M. Zhu, H. Chen, M. Huang, L. Wei, H. Hailin, H. Jie, Y. Wang, GenDet: Towards good generalizations for AI-generated image detection, arXiv preprint arXiv:2312.08880, 2023.
[102] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, L. Wei, T. Zhijun, H. Hailin, H. Jie, Y. Wang, GenImage: a million-scale benchmark for detecting ai-generated image, Adv. Neural Inf. Process. Syst. 36 (2024).
[103] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. arXiv preprint arXiv:2303.08774, 2023.
[104] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. arXiv preprint arXiv:2502.13923, 2025.
[105] S. Cao, C.Y. Wu, P. Krähenbühl, Lossless image compression through super-resolution, arXiv preprint arXiv:2004.02872, 2020.
[106] S. Choi, S. Park, J. Lee, S. Kim, S.J. Choi, M. Lee, HFI: A unified framework for training-free detection and implicit watermarking of latent diffusion model generated images, arXiv preprint arXiv:2412.20704, 2024.
[107] D. Cozzolino, G. Poggi, M. Nießner, L. Verdoliva, Zero-shot detection of ai-generated images, in: European Conference on Computer Vision, Springer, 2024, pp. 54–72.
[108] Y. Fan, D. Yang, J. Zhang, B. Yang, Y. Zou, Fake-GPT: detecting fake image via large language model, in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, 2024, pp. 122–136.
[109] Z. He, P.Y. Chen, T.Y. Ho, RIGID: A training-free and model-agnostic framework for robust AI-generated image detection, arXiv preprint arXiv:2405.20112, 2024.
[110] T.M. Huang, W.T. Lin, K.L. Hua, W.H. Cheng, J. Yamagishi, J.C. Chen, ThinkFake: Reasoning in multimodal large language models for AI-generated image detection, arXiv preprint arXiv:2509.19841, 2025.
[111] Y. Ji, Y. Hong, J. Zhan, H. Chen, H. Zhu, W. Wang, L. Zhang, J. Zhang, et al., Towards explainable fake image detection with multi-modal large language models, arXiv preprint arXiv:2504.14245, 2025.
[112] Y. Ji, H. Yan, J. Lan, H. Zhu, W. Wang, Q. Fan, L. Zhang, J. Zhang, Interpretable and reliable detection of AI-generated images via grounded reasoning in mllms, arXiv preprint arXiv:2506.07045, 2025.
[113] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, Adv. Neural Inf. Process. Syst. 36 (2023) 34892–34916.
[114] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. arXiv preprint arXiv:2304.07193, 2023.
[115] J. Ricker, D. Lukovnikov, A. Fischer, 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, 2024, pp. 9130–9140.
[116] C. Tan, J. Wang, X. Ming, R. Tao, Y. Wei, Y. Zhao, Y. Lu, ForenX: Towards explainable AI-generated image detection with multimodal large language models, arXiv preprint arXiv:2508.01402, 2025.
[117] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[118] Z. Zhou, Y. Luo, Y. Wu, K. Sun, J. Ji, K. Yan, S. Ding, X. Sun, Y. Wu, R. Ji, AIGI-Holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models, arXiv preprint arXiv:2507.02664, 2025.
[119] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, Y. Wei, 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, 2024, pp. 28130–28139.