VNUHCM Journal of Science and Technology Development 2026, 29(3):4186-4204

Open Access Full Text Article

Review Article

A Comprehensive Review of Deep Learning Techniques in Image Forensics

Dinh Phuc Do[1,2], Phuong Nghi Tram[1,2], Tu Nguyen[1,2], Kha Tu Huynh[1,2,*]

Use your smartphone to scan this QR code and download this article

ABSTRACT

The rapid proliferation of manipulated visual content has raised critical concerns regarding authenticity, privacy, and public trust in digital media. Traditional forgery techniques such as image splicing and copy–move have long posed challenges for forensic analysis; however, the advent of AI-driven generative models has introduced a new era of sophisticated and realistic forgeries, including deepfakes and image morphing. At the same time, the continuous evolution of deep learning (DL) has allowed for the development of powerful tools to counter these threats, making it a cornerstone in modern image forgery detection. This review focuses on the application of DL methods to detect image forgeries over the past eight years. The paper covers a wide range of approaches, addressing both traditional and modern forgery types and discussing representative models, methodologies, datasets, performance results, and their respective contributions and limitations. The goal of this review is to provide a clear understanding of the current progress, key challenges, and opportunities for future research in DL-based image forgery detection.

Key words: Forgery detection, copy-move, splicing, deepfakes, morphing, deep learning, neural networks, CNNs, transformers

INTRODUCTION

The manipulation of digital images has become an urgent issue due to its impact on evidence reliability, scientific credibility, and public trust. Traditional forgery techniques such as image splicing and copy–move have been used repeatedly to alter visual content in domains such as media and online communications, raising concerns about information integrity. For example, a 2024 investigation into image splicing in scientific publications reported that numerous figures contained composite regions and copy–move artifacts, highlighting how splicing remains a major manipulation technique in recent scholarly work [1].

At the same time, modern forgery techniques have advanced rapidly with the rise of AI-based image generation. Recent work has shown that face morphing—in which two or more facial images are blended into a single image that may correspond to multiple identities—poses a concrete threat to biometric systems and ID-issuance workflows; several 2023–2024 studies have analyzed morphing generation methods and their threat to automated recognition systems [2].

Multiple categories and generation techniques of image forgery have been identified. Image forgery is broadly categorized into active approaches, which use pre-embedded information such as watermarks or digital signatures, and passive (blind) approaches, which rely on intrinsic features to detect traces of manipulation. The most commonly studied forgery types include image splicing (pasting content from another image), copy–move (duplicating regions within the same image), removal or inpainting, retouching, and highly realistic synthetic media generated by deep learning (DL) models known as Deepfakes or Generative Adversarial Networks (GANs).

The field of image forensics has seen numerous innovations over the years, particularly in the technological era. Recently, DL and hybrid approaches have increasingly complemented traditional handcrafted feature-based or machine learning (ML) methods (Figure 1). Traditional forensic techniques rely on handcrafted features and signal- or statistical-based analyses to identify physical or structural traces left during image acquisition or manipulation. ML-based forensics builds upon these manually extracted features—such as noise residuals or frequency-domain descriptors—by employing classifiers to distinguish authentic and tampered regions. In contrast, DL-based forensics learns discriminative representations directly from data using convolutional neural networks (CNNs) or transformers, offering superior accuracy and robustness. Hybrid approaches integrate handcrafted forensic cues or residual maps with deep models, effectively combining traditional domain knowledge with the representational

[1]*International University, Ho Chi Minh City, Vietnam*

[2]*Vietnam National University, Ho Chi Minh City, 700000, Vietnam*

Correspondence

Kha Tu Huynh, International University, Ho Chi Minh, Vietnam

Vietnam National University, Ho Chi Minh City, 700000, Vietnam

Email: hktu@vnuhcm.edu.vn

History

- Received: 04-01-2026
- Revised: 19-03-2026
- Accepted: 27-03-2026
- Published Online: 30-07-2026

DOI : 10.32508/vnuhcmj-std.v29i3.4649

Check for updates

Cite this article : Phuc Do D, Tram P N, Nguyen T, Huynh K T. **A Comprehensive Review of Deep Learning Techniques in Image Forensics**. *VNUHCM J. Sci. Technol. Dev.* 2026; 29(3):4186-4204.

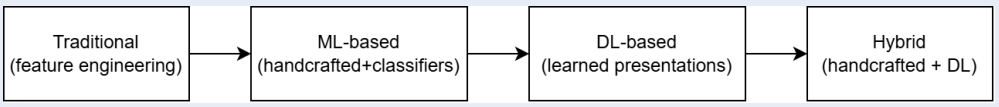

Figure 1: Evolution of image forensics methodologies.

power of DL.

Recent advancements in pure DL have spurred the development of related applied fields, including image forensics. According to statistics from IEEE and Science Direct, the number of publications on image authentication using DL-related methods has increased significantly, particularly over the past two years (see Figure 2). This suggests that applying deep-learning-related methods to detect various types of forged images is a growing and persistent trend.

Despite this rapid growth, no comprehensive statistical summary compares these DL methods across forgery types, datasets, and evaluation metrics. This review aims to provide readers with a clear overview of DL-based image forgery detection over the past eight years, highlighting the main detection approaches for each forgery category and comparing their strengths, weaknesses, data applicability, and performance. The literature included in this survey was collected from major academic databases such as IEEE Xplore, ScienceDirect, and Google Scholar. Relevant papers were identified using keywords, including image forgery detection, image manipulation detection, and deep learning for image forensics. The selection focused on studies published between 2018 and 2025 that proposed DL-based detection approaches and reported experimental results conducted on publicly available datasets. Since each type of forgery has its own scope, datasets, and DL strategies, this paper is organized by forgery category. Within each category, methods are presented in chronological order to reflect their evolution. However, in the modern generative forgery section, models will be introduced in related methodology groups for convenience in analyzing related models.

This review is organized by forgery type and method group. It begins with a detailed analysis of traditional region-based forgeries (splicing, copy–move, and inpainting or removal); these are discussed chronologically to reflect the evolution of DL methods. The paper then transitions to cover modern generative forgeries, examining advanced techniques for detecting generative fingerprints, mitigating generalization gaps, and the emergence of contrastive language-image pre-training (CLIP)-based and zero-shot detectors. Finally, the document concludes with a comparative discussion of contributions, weaknesses, and future challenges across all forgery categories.

REVIEW

Image forgery is broadly categorized into two main groups: passive (blind) approaches, which rely on intrinsic features to detect manipulation traces, and active approaches, which use pre-embedded information such as watermarks or digital signatures. This section primarily focuses on passive DL methods applied to traditional region-based forgeries and modern generative forgeries.

Region-based Forgery

Region-based image forgery involves the manipulation or alteration of specific areas within a digital image, typically through the insertion, deletion, or relocation of pixel content [3]. These methods primarily focus on detecting inconsistencies within localized regions, often modeled as a pixel-wise segmentation problem [4]. Detection relies on analyzing subtle indicators such as noise residuals and resampling features, statistical inconsistencies from combining disparate sources (e.g., varying noise patterns or mismatched lighting), or generative anomalies introduced during reconstruction. A key emphasis in DL-based image forensics for this category is forgery localization, frequently modeled as a pixel-wise segmentation problem. DL methods commonly use CNNs, often customized with high-pass filters, to capture forensic artifacts like noise residuals and resampling features resulting from these classical manipulations. The three main manipulation types in region-based forgery are illustrated in Figure 3.

Image Splicing (Image Composition or Joining from Multiple Sources)

Image splicing, also known as image composition or cut-paste forgery, involves copying content from one source image and pasting it into a different destination image [5]. The primary goal of splicing detection and localization is to identify the statistical inconsistencies that arise from combining content from disparate sources [6]. The forensic challenge lies in detect-

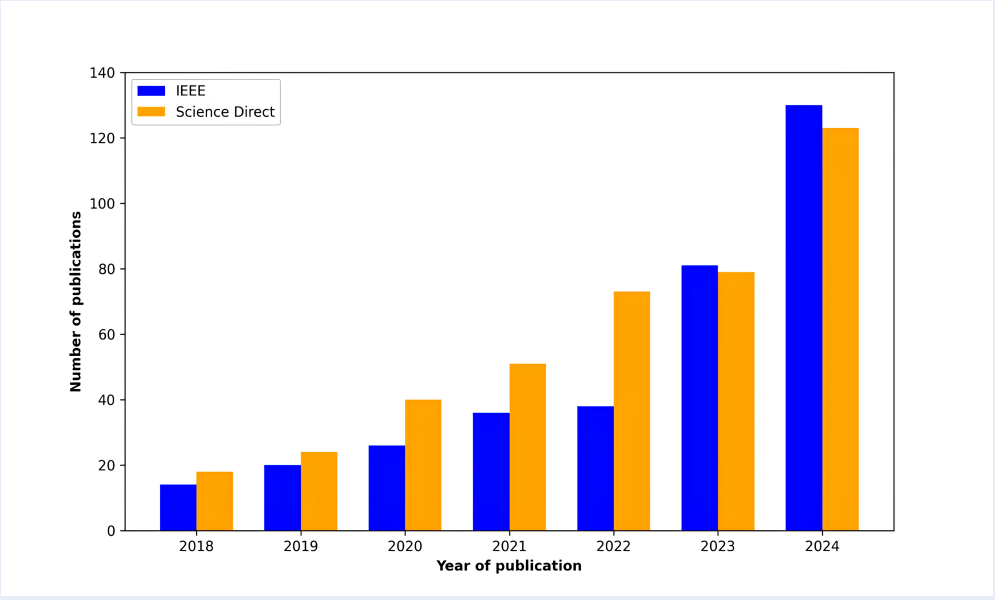

Figure 2: Statistics on publications related to "Deep Learning in Image Forgery" in IEEE Xplore and ScienceDirect.

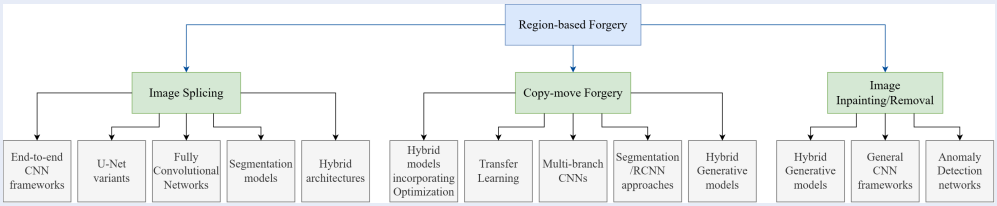

Figure 3: Structure of region-based forgery detection methods.

ing the statistical and physical inconsistencies introduced when merging pixels from two distinct sources, such as varying noise patterns, mismatched lighting conditions, or abrupt changes in color filter array (CFA) artifacts[7]. Specifically, illumination direction, noise patterns, and contrast are generally inconsistent between the spliced region and the host image, making the tampered traces relatively easier for CNNs to detect compared to copy–move forgery (CMF).

DL architectures for splicing often directly incorporate forensic domain knowledge, such as restricting initial layers to focus on residual signals or using sophisticated segmentation models to accurately define tampered boundaries[8]. Common DL architectures include CNNs[9], U-Net and its variants[10], fully convolutional networks (FCNs)[11], and hybrid models incorporating long short-term memory (LSTM) units[12]. Transfer learning, which involves using pre-trained networks such as ResNet, has also been shown to be effective for image classification tasks[6].

Early CNN approaches focused on end-to-end processing. The End-to-End-Trainable CNN Framework (E2E-Fusion) proposed by Marra et al[13]. offers joint image-level detection and localization and generalizes well across forgery types, with performance significantly improved through fine-tuning on challenging NIST datasets. This framework is designed to detect localized manipulations such as splicing and utilizes a cascade of three blocks: patch-level feature extraction, feature aggregation, and global decision (Figure 4). The core methodology relies on end-to-end training, optimizing all model parameters jointly using weak (i.e., image-level) supervision on full-size images. To manage the memory demands of full-resolution analysis, the framework trades memory for computation using a gradient checkpointing strategy. However, its performance is vulnerable to the increased quality of computer-generated fakes.

Qazi et al.[6] introduced a similar DL approach using transfer learning on pre-trained weights from a CNN

(ResNet50v2), achieving high accuracy on the comprehensive CASIA-v2 dataset. A significant performance gap was observed between results on CASIA-v1 and CASIA-v2.

Ali et al[14]. introduced specific modeling of the image quality factor (QF) in a CNN approach, leading to improved performance. This specialized model, however, focuses on image-level detection rather than localization and requires a minimum input size of 128 $\times$ 128.

Another DL framework, DSNI-Net[15], targets salt-and-pepper (S&P) noise, demonstrating high robustness to JPEG compression and noise artifacts. Potential drawbacks include damaging forgery traces in noise-free images, reduced effectiveness against defocus blur, and a lack of superior robustness to non-S&P noise.

Segmentation models are crucial for accurate localization, often addressing the task using regional proposal networks. Previous research resulted in the development of a dual-stream faster R-CNN approach[16]; this architecture effectively incorporates noise residuals by utilizing separate streams for RGB content and noise residuals to distinguish forged regions. Yang et al.[17] introduced a similar Modified R-CNN architecture that incorporated forensic constraints, achieving high localization ability, particularly on NIST16. However, its performance was notably lower on CASIA compared to other datasets. Unlike the R-CNN models focused on region proposal and classification, the Ringed Residual U-Net (RRU-Net)[10] is a powerful segmentation architecture utilizing a residual structure to achieve excellent segmentation accuracy. Generally, U-Net-based architectures demonstrate superior performance for localization, but deeper network designs can suffer from the loss of contextual spatial information and the gradient degradation problem.

Hybrid architectures incorporate contextual learning components to fuse spatial and temporal features. The Hybrid LSTM and Encoder-Decoder architecture (specifically a J-Conv-LSTM-Conv variant) combines spatial features with temporal features for enhanced localization[12]. However, this method is limited to learning correlations between blocks within a patch and cannot capture broader inter-patch correlations. Unlike models requiring extensive training, Wu et al.[18] proposed a general-purpose forgery localization method based on LSTM anomaly detection, which generalizes across multiple manipulation types without prior fine-tuning. A limitation of this approach is its reduced effectiveness when multiple regions are manipulated differently within the same image. Furthermore, its performance significantly improves when external features, such as metadata (EXIF-SC) or camera model forensic similarity, are available. Table 1 details key DL approaches for image splicing detection and localization in chronological order.

DL models demonstrate superior performance in image splicing classification, achieving high reported accuracies. A key trend in splicing detection is the shift toward powerful segmentation architectures and forensic feature fusion. U-Net-based architectures, such as RRU-Net, exhibit superior F-measures for localization, with some attaining a 0.915 F-measure on the COLUMB dataset. Performance significantly improves when external features, such as metadata (EXIF-SC with a 0.98 AUC on Columbia) or camera model forensic similarity (0.95 AUC on Columbia), are incorporated.

Copy-Move Forgery

CMF involves duplicating content within the same image, making it very difficult to detect since the copied patch is inherently consistent with the host image in terms of noise, illumination, and compression history[6]. This manipulation method is particularly challenging because the duplicated (target) region shares identical noise characteristics, illumination, and compression artifacts with the original (source) region, providing a high degree of consistency that evades many detection mechanisms. The goal is thus to detect similarity between the copied (source) and pasted (target) regions. Deep copy-move forgery detection (CMFD) approaches primarily rely on algorithms that perform deep matching and self-correlation of high-level features to locate duplicated regions.

CMFD has transitioned from block-based or keypoint-based methods to comprehensive end-to-end DL systems. These architectures frequently employ multi-branch CNNs or hybrid models designed to detect subtle similarities and subsequent post-processing traces. DL approaches utilize networks such as VGG, ResNet, AlexNet, and variants of Mask R-CNN, often combined with optimizers or clustering techniques.

Early work utilized transfer learning, such as using a VGG-16 backbone fine-tuned on synthetic forged images[19], which achieved high detection accuracy on artificial data. This network classifies image fragments or patches as either ”original” or ”forgery”. To address computational challenges and limited data availability in existing forgery datasets, the strategy involves

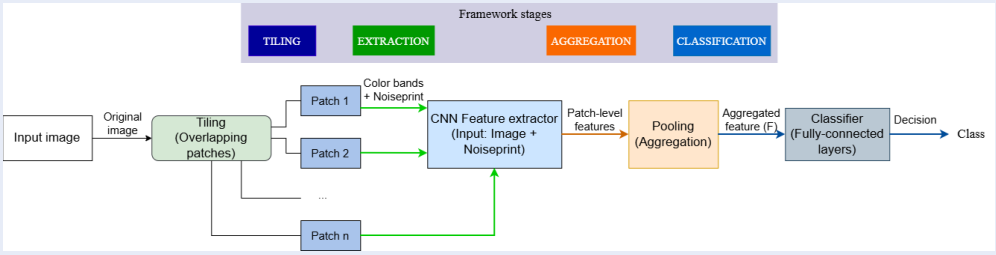

Figure 4: An End-to-End Trainable CNN framework.

Table 1: Key DL approaches for image splicing detection and localization.

Deep Learning Architecture	Datasets Used	Quantitative Result
CNN Framework [13]	Synthetic Vision/UCID dataset, NC2017, MFC2018 (NIST)	AUC (E2E-Fusion, NC2017) grew from 0.846 to 0.932 (with fine-tuning); AUC (E2E-Fusion, MFC2018) grew from 0.838 to 0.902 (with fine-tuning)
Hybrid (LSTM + Encoder-Decoder) [12]	NIST16, IEEE Forensics Challenge, COVERAGE	AUC (NIST16): 84.60%; AUC (IEEE): 77.67%; AUC (COVERAGE): 81.14%
Dual-stream CNN (Faster R-CNN) [16]	NIST16, CASIA	AUC (NIST16): 0.937; F1 score (NIST16): 0.722
LSTM [18]	NIST, COVERAGE, CASIA, Columbia	AUC (NIST): 79.5%; AUC (Columbia): 82.4%; AUC (COVERAGE): 81.9%; AUC (CASIA): 81.7%
U-Net Variant (RRU-Net) [10]	CASIA, COLUMB	F-measure (CASIA): 0.841; F-measure (COLUMB): 0.915; Image-level Accuracy (CASIA): 76%
Constrained R-CNN [17]	NIST16, COVERAGE, Columbia, CASIA	F1 score (NIST16): 0.927; F1 score (COVER): 0.757; F1 score (Columbia): 0.790; F1 score (CASIA): 0.475
Transfer Learning (ResNet50v2) [6]	CASIA-v1, CASIA-v2	Accuracy (CASIA-v2): 99.3%; Accuracy (CASIA-v1): 81%
CNN [14]	CASIA-v2	Accuracy (96% QF): 92.15%; F-measure (96% QF): 90.99%
DL framework [15]	DSO-1, Korus	QF estimation accuracy: 100% (for both DSO and Korus)

classifying input signals as fixed-size $40 \times 40 \times 3$ patches. Training is performed by selecting patches from regions in the original image and from the borders of embedded splicing. The approach leverages transfer learning by implementing a fine-tuned model using COCO pretrained weights rather than starting training from scratch. However, this accuracy depends heavily on synthetic dataset alignment. Similar standalone models, such as ResNet-18 [20], achieved high precision results on CASIA datasets. When used alone, the deep textural features of ResNet-18 yielded lower detection accuracy compared to the final proposed fusion model. More recently, lightweight architectures have emerged that improve efficiency, such as a model based on MobileNetV2 [21], which offers a computationally efficient solution but suffers from a low detection rate.

Since CMF requires identifying a source and target region and locating subtle similarity cues, multi-branch and segmentation frameworks are critical. An end-to-end deep neural network (DNN) solution utilized a dual-branch CNN (Mani-det and Simi-det) that was specifically designed for localization and source/target distinction, marking it as the first end-

to-end CMFD solution [22]. While innovative, its overall F1 performance score on the CoMoFoD dataset was relatively low. Like approaches used in splicing detection, segmentation-based methods were subsequently adopted. In addition, hybrid region-based convolutional neural network (RCNN) approaches were developed using lightweight backbones such as MobileNetV1 or DenseNet-41 to combine instance segmentation capability with computational efficiency [23].

Current CMFD trends emphasize hybrid approaches that combine powerful feature extractors with mechanisms designed for correlation finding or optimization. For example, a complex approach involving a GAN + CNN hybrid architecture employing a three-branch system and leveraging adversarial learning for robust detection was developed [24]; while robust, this approach involved high implementation complexity. Another study proposed an AlexNet-based hybrid model utilizing the Grey Wolf Optimizer (GWO) and superpixel clustering [25], which achieved exceptionally high accuracy through optimization. Notably, the model performed poorly without this optimization. Hybrid models incorporating optimization algorithms demonstrate superior performance over single DL architectures. The key DL methodologies for CMFD, including reported quantitative performance metrics, are summarized in Table 2.

Recent CMFD methods have achieved high detection accuracy. Current trends emphasize hybrid approaches that combine powerful feature extractors with mechanisms designed for correlation finding or optimization. Hybrid DL models tend to perform better than pure DL models in terms of both accuracy and efficiency. The primary challenges in CMFD stem from the non-trivial implementation complexity required for sophisticated hybrid architectures and the ongoing need for models to handle geometric transformations, such as rotation and scaling, that forgers apply to the copied region to further disguise the manipulation.

Image Inpainting or Removal

Image inpainting or removal involves erasing a selected region from an image (in order to hide an object, for example) and filling the void with newly generated pixel values that are typically synthesized from the background using sophisticated generative models. This forensic task focuses on detecting subtle traces or inconsistencies left by the generative process or the filling algorithm. Specifically, detection approaches target telltale generative artifacts or statistical anomalies introduced during reconstruction, often through analysis of high-frequency residuals.

The detection of image inpainting and removal depends heavily on general manipulation localization methods designed for anomaly detection or segmentation. Commonly used methods include FCNs, LSTM-based approaches, and general-purpose CNN frameworks.

Specialized models are designed to target unique artifacts generated during reconstruction processes. For example, a High-Pass FCN was developed for localizing deep inpainting traces (see Figure 5) [26]. This FCN achieved high accuracy, particularly when image quality factors were high, though performance degraded significantly under middle-level JPEG recompression.

General manipulation frameworks often prove effective in removal detection. The E2E-Fusion framework [13], previously noted for splicing, demonstrated high robustness and significant performance improvements when fine-tuned on challenging NIST datasets. The Full-Image Full-Resolution End-to-End-Trainable CNN framework proposed by Marra et al [13]. has been validated for detecting localized manipulations. This approach leverages the preservation of high-frequency details and micro-textures that reflect different digital histories. Even when trained solely on splicing examples, this CNN was demonstrably effective at detecting inpainting manipulation in subsequent tests. However, performance declined when confronted with increasingly sophisticated fakes.

Similarly, a general-purpose forgery localization method based on LSTM anomaly detection treats removal as a local anomaly [18], achieving good generalization. However, its performance remains qualitatively lower than that of specialized fine-tuned DNN methods. Detecting inpainting and removal—particularly with modern GANs—presents unique challenges because forgery techniques are constantly evolving, leading to highly photorealistic results.

Table 3 summarizes key DL methodologies for image inpainting and object removal, along with their reported quantitative performance metrics.

Discussion: Splicing, Copy-Move, and Inpainting

The field of region-based forgery detection is characterized by specialized approaches built upon common DL frameworks. Image splicing detection relies on identifying inter-image inconsistencies, such as

Table 2: Key DL methodologies in CMFD with reported quantitative performance metrics

Deep Learning Architecture	Datasets Used	Quantitative Result
CNN (VGG-16) + Transfer Learning [19]	CASIA	Accuracy (fine-tuned model): 97.8%
Dual-branch CNN [22]	CoMoFoD	F1-score (Overall Average): 0.4926
Hybrid (GAN + CNN + SVM) [24]	CIFAR-10, MNIST (Training); MICC-F600 (Testing)	F1-score (Skeleton metrics): 0.8835; Mean Precision: 0.6963; Mean Recall: 0.8042
CNN (ResNet-18) [20]	CASIA-v1, CASIA-v2	Precision (CASIA I): 99.3%; Precision (CASIA II): 97.94%
Hybrid (RCNN + MobileNet/DenseNet) [23]	COVERAGE, MICC F600, MICC-F2000, CASIA-v2	Mask R-CNN F1-score: 70% (MICC F600); Mask R-CNN Accuracy: 90% (MICC F2000/COVERAGE average); RCNN Acc: 98.12% (CoMoFoD), 99.02% (MICC-F2000), 83.41% (CASIA-v2)
Hybrid (AlexNet + GWO) [25]	MICC-F600, MICC-F2000, GRIP	Maximum Accuracy: 99.66% (MICC-F600); 99.75% (MICC-F2000); 98.48% (GRIP)
CNN (MobileNetV2) [21]	CoMoFoD, MICC-F2000	Maximum improvement rate: 0.4% in accuracy and 0.2% in F1-score

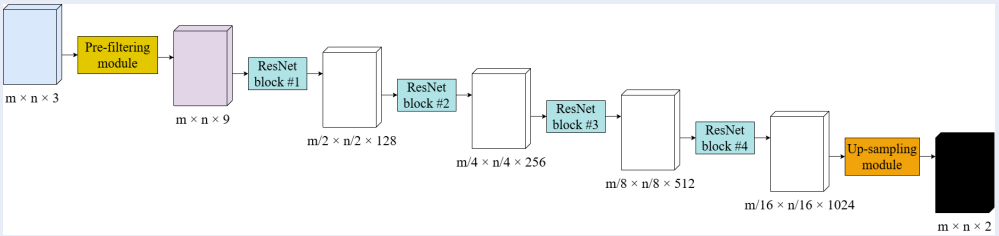

Figure 5: Basic framework of the proposed method.

Table 3: Key DL methodologies forinpainting and removal with reported quantitative performance metrics.

Deep Learning Architecture	Datasets Used	Quantitative Result
LSTM [18]	COVERAGE (contains removal)	AUC: 81.9%
FCN [26]	ImageNet	F1-score: higher than 0.97 (QF=96)
CNN Framework [13]	NC2017, MFC2018	AUC (E2E-Fusion, NC2017) grew to 0.932 (with fine-tuning); AUC (E2E-Fusion, MFC2018) grew to 0.902 (with fine-tuning)

noise, CFA artifacts, and lighting discrepancies. These approaches often employ segmentation architectures like U-Net [10] or hybrid LSTM models [12] that achieve high area under the curve performances (AUC; up to 99.6% on NIST16). CMF detection is inherently more challenging due to intra-image consistency, requiring complex deep matching and self-correlation mechanisms within specialized multi-branch CNNs or hybrid GANs. While high CMF classification accuracies can be achieved (up to 99.75% ACC), achieving precision in pixel-level localization remains a significant challenge (F1-scores often around 0.5 to 0.7 on complex datasets). Image inpainting and removal detection focuses on locating generative anomalies left by reconstruction algorithms, with high-pass FCNs [26] or universal anomaly detection networks proving to be the most successful; for example, some models achieve F1 scores higher than 0.97 when detecting deep inpainting.

All three categories are heavily reliant on CNNs, which are typically integrated into encoder-decoder structures for localization, reflecting the core need

for pixel-wise segmentation. There has been a pronounced shift toward hybrid architectures, combining CNNs with contextual components such as LSTMs (for artifact analysis) or integrating adversarial training via GANs to enhance feature robustness and correlation matching.

Image splicing detection is generally considered easier than CMF detection because the forensic task involves identifying inconsistencies between source and target regions. DL models offer automatic feature learning, allowing them to robustly localize tampered regions even after mild post-processing operations such as blurring and resampling. These DL models are particularly effective when constrained to the extraction of forensic features such as noise residuals or quantization inconsistencies. Splicing detection succeeds by exposing the statistical inconsistencies that arise when combining pixels from disparate sources, such as mismatched lighting conditions, varying noise patterns, or abrupt changes in CFA artifacts. However, performance often degrades significantly with heavy post-processing or highly photorealistic synthetic splices. For example, a Constrained R-CNN model[17] exhibited a markedly weaker performance on certain datasets, achieving a low F1 score of 0.475 on CASIA despite registering a strong F1 score of 0.927 on NIST16.

CMF is widely regarded as the most challenging form of region-based manipulation to detect. This difficulty arises from the identical noise characteristics inherent in both the source and target regions, a property known as intra-image consistency; this creates a high degree of similarity that evades many detection mechanisms. A key advantage of DL models is their ability to mitigate susceptibility to post-processing, a vulnerability that has plagued traditional CMF methods. Furthermore, hybrid models incorporating optimization algorithms, such as the GWO-based AlexNet[25], exhibit superior performance metrics compared with single DL architectures, achieving an exceptional accuracy of up to 99.75% on the MICC-F2000 dataset. Advanced CMF detection methods also focus on distinguishing source and target regions. Despite these advances, the fundamental challenge remains: differentiating identical copied regions from authentic content. Models still struggle to achieve uniformly high pixel-level F1 scores across diverse datasets and complex transformations (e.g., the dual-branch CNN approach[22] achieved an average F1 score of only 0.4926 on the CoMoFoD dataset).

Inpainting and removal detection depend on identifying generative anomalies introduced by advanced reconstruction algorithms. DL models excel at automatically detecting the subtle high-frequency forensic features that are characteristic of these algorithms. Advanced localization networks, such as specialized FCNs[26], achieve high accuracy with F1 scores exceeding 0.97 for deep inpainting when image quality is high. General frameworks like E2E-Fusion are also demonstrably effective at detecting localized manipulations by exploiting high-frequency details[13]. However, the robustness of detection methods is challenged by improvements in generative models (e.g., diffusion models) and adversarial post-processing. Consequently, inpainting detection is effectively a competition against increasingly sophisticated generative models, resulting in noticeable declines in AUC performance when tested against newer computer-generated fakes compared to simpler manipulations.

The need for accurate pixel-wise segmentation is a critical component of traditional forgery detection methods. All three categories of forgery—splicing, copy-move, and inpainting/removal—rely on CNNs integrated into encoder-decoder structures to achieve localization. Region-based image forgery is fundamentally modeled as a pixel-wise segmentation problem designed to generate a mask that precisely identifies tampered regions. DL architectures for splicing detection often employ advanced segmentation models to accurately define tampered boundaries. Foundational to these methods are the U-Net variants, such as RRU-Net, which use a residual structure to achieve excellent segmentation accuracy, and specialized FCNs—like the High-Pass FCN—used for localizing deep inpainting traces.

There is a significant and effective trend toward using hybrid architectures, which outperform models based solely on standard DL architectures. These hybrid models combine CNNs with contextual components, such as LSTM units, to fuse spatial and temporal features for enhanced localization and contextual artifact analysis. For techniques like CMFD, which are highly challenging due to intra-image consistency, complex hybrid approaches are critical. This includes the integration of adversarial training (GANs) or optimization algorithms, such as GWO, to enhance feature robustness and correlation matching, resulting in exceptionally high accuracy in specific CMFD tasks. Hybrid models incorporating optimization algorithms have demonstrated superior performance compared to single DL architectures.

Generative Forgery

Artificial intelligence (AI)-generated image detection involves forensic analysis to determine whether a digital image was created or modified by a generative AI

model. Unlike traditional image forgery detection, which focuses on manipulation techniques such as splicing or copy-move edits, AI-generated image detection identifies systemic, often imperceptible signs that vary across different instances of models[27]. This is in contrast to traditional forgeries, which typically exhibit inconsistencies in lighting or compression levels between different parts of an image.

AI-generated image detection is a rapidly evolving field of research. In the early stages of artificial intelligence-generated content (AIGC), there were fewer generative models, and existing models were relatively simple, such as the initial versions of GANs. Detection solutions initially achieved high accuracy by targeting low-level forensic traces or "fingerprints" characteristic of specific, less sophisticated architectures. However, advancements in highly sophisticated models—including modern GAN variants and diffusion models (DMs) such as Stable Diffusion and DALL-E2 have introduced significant challenges for forgery detection. The current generation of synthetic images is more natural and may originate from an extensive and constantly growing diversity of generators, including newly emerging or unseen architectures. Consequently, a primary goal is to develop robust methods that achieve not only high accuracy and resilience against image impairments such as compression or resizing but also superior generalization performance across diverse generative models.

Figure 6 presents a chronological overview of the evolution of AI-generated image detection methodologies, organized into four distinct phases based on their underlying technological approaches and the generative models they were designed to detect. The timeline begins with Phase one (starting in 2017), characterized by the dominance of GANs such as ProGAN, StyleGAN, CycleGAN, and BigGAN, where detection methods primarily focused on identifying model-specific fingerprints left in synthetic images. Phase two (starting in 2021) marks the emergence of DMs and the subsequent generalization crisis, shifting the technological focus toward mitigating performance gaps when detectors encounter unseen generative architectures. Phase three (starting in 2022) represents a paradigm shift toward leveraging large pretrained models, particularly CLIP-based detectors, to achieve universal detection capabilities across diverse generation methods. Most recently, Phase four (starting in 2023) introduces a revolutionary approach to detector training that eliminates their dependency on synthetic training data, allowing for the development of detectors that can identify AI-generated content without requiring access to fake images during the training process. This progression illustrates the field's continuous adaptation to the rapidly evolving landscape of generative AI technologies.

Detecting Fingerprints

Early detection techniques were based on the hypothesis that generative models, particularly GANs, inadvertently introduce artifacts into the images they create. These artifacts, known as "generative fingerprints," serve as indicators of a synthetic origin. Research has also shown that these fingerprints are unique to the specific generative process used to create them.

Yu et al.[27] demonstrated that even minor differences in the training of a GAN can result in distinct and identifiable fingerprints. These unique characteristics arise from variations in model architecture, training datasets, and random initialization seeds. Different GAN architectures—such as ProGAN versus StyleGAN—produce inherently different statistical patterns. The data used to train a model subtly influences the properties of its output and traces. In addition, different random seeds during initialization can produce unique fingerprints, even when the same architecture and dataset are employed. This study was foundational because it proved that these fingerprints were not merely generic signs of synthesis but specific enough to link an image back to a unique model instance[27]. This established the feasibility of source attribution for intellectual property protection and forensic investigation. The discovery motivated the development of various methodologies aimed at capturing and classifying these generative artifacts.

CNN-based approaches are among the most extensively researched strategies for the detection of AI-generated images[28–33]. CNNs such as ResNet, DenseNet, and XceptionNet can be trained to automatically learn hierarchical features that distinguish between real and GAN-generated images. These models demonstrate high effectiveness when training and testing data originate from the same or similar GANs[28–30].

Frequency domain analysis is a highly effective approach for detecting generative fingerprints. This method leverages the fact that GAN-generated images contain periodic, grid-like artifacts in the frequency domain, often resulting from upsampling operations in transposed convolution layers. Frank et al.[34] demonstrated that transforming an image from the spatial (pixel) domain to the frequency domain using methods such as the Discrete Cosine Transform or Fast Fourier Transform can reveal underlying patterns and high-frequency artifacts that are typically

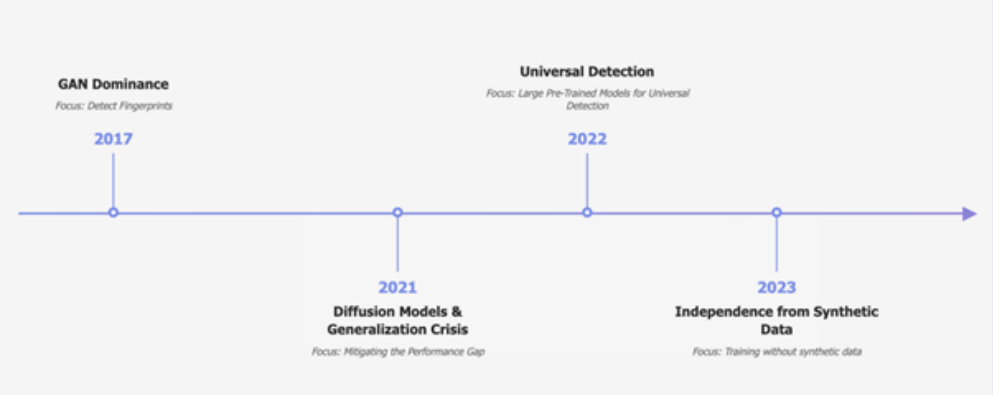

Figure 6: Evolution of AI-Generated content detection.

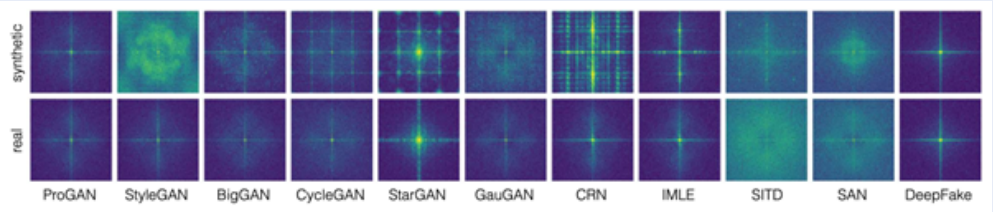

Figure 7: Frequency observations across various datasets. Adapted from Wang et al[30].

invisible to the naked eye. The classifiers achieved higher accuracy (up to 99.61%) and required fewer parameters than other CNN-based methods. They also converged significantly faster and required less training data. However, their performance depended on structural issues (upsampling) in GAN construction and was vulnerable to adversarial examples. Similarly, Wang et al.[30] observed periodic patterns in the average frequency spectra of synthetic images (Figure 7), which they identified as consistent with aliasing artifacts. More recently, hybrid approaches have demonstrated the value of combining spatial and frequency analysis. The proposed models, trained with augmentation, demonstrated high robustness to common post-processing operations and achieved a success rate of 94.54% accuracy. However, their generalizability decreased significantly on DMs[35]. Furthermore, their performance was sometimes negatively affected by augmentation, and the models did not perform well against manipulation techniques using "shallow" methods like Photoshop. In another study, Ahire et al.[36] explicitly incorporated frequency features into their transformer-based SFANet and SFPnet models by applying the Fast Fourier Transform to image patches alongside traditional spatial feature extraction. This dual-domain approach highlights the enduring value of spectral analysis as a powerful signal for forgery detection. The method combines the strengths of global (transformers) and frequency-based (texture/artifacts) feature extraction, achieving 96.13% accuracy, albeit in an approach that involved high computational and memory requirements[36].

Beyond frequency-domain analysis, other traces left by generative models in synthetic images can be leveraged to develop detection tools. One robust statistical method is the pixel co-occurrence matrix, which quantifies spatial relationships between pixel values. Nataraj et al.[31] devised an approach that computes these matrices across the three RGB channels of an image and uses them as input for a deep CNN. This method achieved impressive accuracy, with 99.71% on CycleGAN and 99.37% on StarGAN datasets. A key advantage was its high precision, and its generalizability was also evaluated. In a cross-dataset experiment, a model trained on StarGAN images attained 93.42% accuracy when tested on CycleGAN images, indicating reasonable generalization. However, the researchers acknowledged a notable limitation: model performance significantly declined when images were subjected to JPEG compression[31].

Xi et al.[29] noted that forgeries were often hidden in high-frequency noise and designed a novel dual-

stream network to address the challenge of detecting modern text-to-image models such as DALL-E 2 and DreamStudio. The architecture's "residual stream" uses a set of high-pass filters from the spatial rich model (SRM) to explicitly extract high-frequency noise and texture information. This is complemented by a "content stream" that captures lower-frequency traces, with a cross-attention mechanism that fused information from both streams. Their model achieved 99.2% accuracy on a dataset of images generated by DALL-E 2. Their architecture was noted for its novelty and robustness to changes in image resolution; however, like many other methods, it faced difficulties in maintaining its robustness against post-processing operations such as image rotation and blurring.

Aggarwal et al.[32] found that error level analysis—a technique used to highlight differences in JPEG compression levels within an image—could serve as a useful preprocessing step for forgery detection. They applied error level analysis for image preprocessing before feeding them into a simple CNN architecture. By emphasizing compression inconsistencies common in manipulated images, their model achieved 91.83% accuracy and converged in just nine epochs. The primary advantages of this method are its computational efficiency and fast training time. However, it was tested on a relatively small dataset of fewer than 5,000 images, which may limit the generalizability of the findings. In addition, its robustness against countermeasures that mimic uniform compression remains a concern. Methods to detect AI-generated images via fingerprint detection have proven extremely efficient against the generative models for which they are trained. These approaches are summarized in Table 4.

Mitigating Performance Gaps

The critical challenge in the field of AI-generated image detection is generalization. A detector is only practically useful if it can reliably identify fake images produced by new or unseen generative models. However, research consistently shows that most detectors trained on images from one family of generators (e.g., GANs) perform poorly when tested against images from architecturally different families, such as DMs[35].

This performance drop primarily occurred because early models often overfitted to specific artifacts generated by GANs in their training data. Ojha et al.[35] noted that a deep model trained exclusively on GAN-generated images learned to identify GAN-specific fingerprints so narrowly that it classified nearly all non-GAN images—including those from DMs—as real. This created a failure mode in which the detector classified all unknown types of fakes as "real". To address this, researchers developed several strategic approaches to improve cross-domain (cross-generator and cross-scene) performance.

The Artifact Purification Network proposed by Meng et al.[37] isolates domain-agnostic artifacts common across different generators using a dual purification process. In the frequency domain, a "suspicious frequency-band proposal" method identifies and extracts potential artifact-related features, while in the spatial domain, an orthogonal feature decomposition method separates artifact-related features from image content. This explicit process is complemented by an implicit purification step that uses mutual information to ensure the model focuses on domain-agnostic forgery traces rather than generator-specific artifacts. This approach improved average accuracy on the GenImage dataset by at least 5.6% compared with competitors and maintained high cross-scene performance (only a 0.1% drop on the LSUN-Bedroom scene) by purifying domain-agnostic artifacts. However, the architecture is complex, requiring two purification levels.

Ensemble methods combine the strengths of multiple specialized models to create a more robust and adaptable detector. For example, the Gated Expert CNN proposed by Ahmad et al.[38] trains several "expert" CNNs, each specialized in a different generator (e.g., BigGAN, GLIDE, VQDM). A separate "gating network" then learns to weigh the contributions of each expert based on the input image. This architecture allows the system to dynamically use the most relevant expert for a given image, resulting in strong adaptability across varied datasets and generative sources: this model achieved at least 92% accuracy across tested generators (BigGAN, GLIDE, VQDM, Stable Diffusion) due to its specialized expert contributions.

The Multi-scale Texture Difference model (MTD-Net) aims to detect subtle texture inconsistencies indicative of synthetic images. Yang et al.[39] employed central difference convolution (CDC) to integrate pixel intensity and gradient information, thereby enhancing local texture representation. Specifically, CDC was combined with Atrous Spatial Pyramid Pooling (ASPP) to merge these texture features across multiple scales (Figure 8). The study also found that CNNs tended to focus on different facial regions when distinguishing real versus fake images; for instance, a detector might prioritize the mouth and nose for real images but shift attention to the eyes and eyebrows

Table 4: Representative DL-based fingerprint detection methods and their reported quantitative performance metrics.

Deep Learning Architecture	Datasets Used	Quantitative Result
Ridge Regression, Shallow CNN (4 convolution layers) [34]	Images generated by StyleGAN, ProGAN, SN-DCGAN, Cramer-GAN, and MMDGAN FFHQ, Stanford dog, LSUN bedrooms, CelebA	Accuracy (FFHQ): 100.00% Accuracy (CelebA): 99.91% Accuracy (LSUN): 99.61%
CNN (ResNet50) [30]	ForenSynths, images generated by GANs	Accuracy: 94.54%
SFANet: combining Transformer-based architectures (Swin/ViTs) and CNN-based methods [36]	DFWild-Cup	Accuracy: 96.13% AUC: 0.9822
CNN trained directly on pixel co-occurrence matrices computed on the RGB channels [31]	Images generated by CycleGAN and StarGAN	Accuracy (CycleGAN): 99.71% Accuracy (StarGAN): 99.37% Accuracy (cross-evaluation): 99.49%
Cross-Attention Enhanced Dual-Stream Network: Residual Stream (SRM + CNN) Content Stream (CNN) [29]	DALL-E2, DreamStudio, SPL2018, DsTok	Accuracy (DALL-E2): 99.2% Accuracy (DreamStudio): 99.5% Accuracy (against post-processing operations): 99.5%
Shallow Convolutional Neural Network [32]	943 images collected from Kaggle	Accuracy: 91.83%.

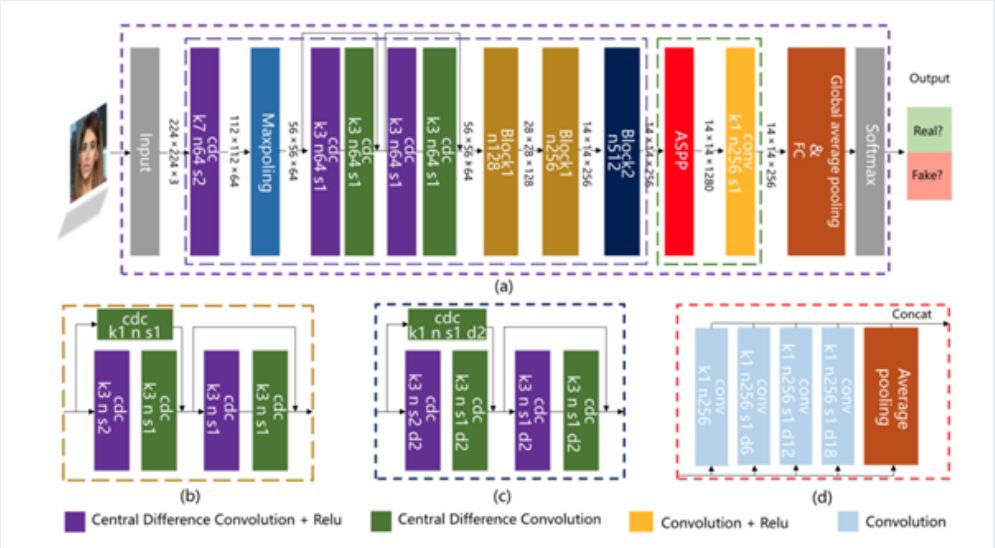

Figure 8: Architecture of MTD-Net [39]. (a) The network architecture includes a texture difference feature module (marked with a purple dotted line) and a multi-scale information module (marked with a green dotted line). (b) Block 1 is part of the texture difference feature module, as is (c) block 2, while the (d) ASPP block belongs to the multi-scale information module. Here, "k" indicates kernel size, "n" represents the number of channels, "s" denotes stride size, and "d" specifies the dilation rate.

for fake images. This observation suggests that textural anomalies are not uniformly distributed. The proposed method offers a robust approach to identifying AI-generated images despite common distortions (e.g., JPEG compression and blur), which is crucial in practical applications. ASPP mitigates the detail loss associated with dilated convolution, allowing for improved performance on low-quality data. However, the model still requires retraining for unfamiliar face manipulation techniques, as performance (AUC) significantly declines when training and testing sets originate from different data distributions (i.e., in cross-database experiments).

Pairwise and contrastive learning involves encouraging models to learn "common fake features" shared among different GANs by establishing relationships between real and fake images. A two-step pairwise learning framework employs triplet loss to train a network, aiming to minimize feature space distances between same-class images (real-real or fake-fake) while maximizing distances between different-class images (real-fake). This approach enhances the model's ability to generalize to fakes from unseen generators, achieving precision rates of 87.6%–98.6% for unseen GANs. These strategies are summarized in Table 5.

The Emergence of CLIP-Based Detectors

CLIP is a vision-language model (VLM) pre-trained with web-scale, general-domain data to understand vast semantic concepts. This paradigm shift was supported by the hypothesis that features learned from web-scale, general-domain data might be more robust for forensic tasks compared to those derived from smaller, specialized real-vs-fake datasets. Researchers proposed that a model like CLIP would develop a feature space capable of implicitly capturing the subtle statistical differences in synthetic media without being explicitly trained for this task.

Ojha et al.[35] proposed UnivFD, a methodology that was elegant, simple, and remarkably effective. The core idea addressed the primary failure mode of previous detectors: deep models trained on GAN images tended to overfit to specific artifacts, leading to the misclassification of all non-GAN-generated images as real. To overcome this, the approach involved using CLIP:ViT and the application of simple classifiers (Figure 9). A frozen, pre-trained CLIP:ViT model extracted feature representations of both real and fake images, ensuring that the feature space was not trained on the fake detection task. Simple, low-complexity classifiers, such as a nearest neighbor or linear Classifier, were then applied to this static feature space.

The impact of this approach on generalization was significant. A model whose feature space was never optimized for fake detection demonstrated strong performance in generalizing from GAN-generated training data to unseen DMs. It significantly outperformed heavily trained deep classifiers, proving that the features needed to distinguish real from synthetic content were already embedded within CLIP's ability to understand visuals.

The forgery awareness module (FAM), based on low-rank adaptation (LoRA), was proposed by Xu et al.[40] for efficient fine-tuning. It employs a semantic feature-guided contrastive learning strategy (SeC) to emphasize training on general differences. FAM effectively mitigates dramatic changes in pre-trained features and prevents overfitting, achieving an average detection accuracy improvement of 14.55% over the state-of-the-art methods (i.e., the UnivFD baseline). This was accomplished despite using smaller training sets: only 0.56% of the training data required by UnivFD models. The primary challenge was how to fine-tune the model with few-shot samples while avoiding overfitting and retaining general features.

Lyu et al.[41] demonstrated that Wasserstein distance compression and dynamic aggregation effectively addressed inefficient training caused by redundant data and poor performance on imbalanced datasets while maintaining high generalization across generative models. This novel method combines dynamic aggregation with information compression. Information compression employs k-means clustering in the Wasserstein space to extract discriminative features, significantly reducing dataset size (e.g., from >720k to 10k samples). A dynamic aggregation mechanism enhances robustness to imbalanced datasets by adaptively aggregating diverse linear kernels, each consisting of static and dynamic components. The method accelerated training speed (9 times faster than a linear classifier) and conserved computational resources (storage reduced from 2.3G to 63.1MB) due to information compression, while also improving robustness to imbalanced datasets[41].

Research by Cozzolino et al.[42] further validated this paradigm by demonstrating that a simple SVM, trained solely on CLIP features extracted from a handful (N) of paired real/fake images in a reference set, could outperform many state-of-the-art DL detectors. This established the use of CLIP features as a powerful and essential baseline for evaluating any new detection method. However, accuracy is strongly reliant on the decision threshold (often fixed at 0.5) in truly OOD scenarios where no calibration data is available.

Table 5: Representative methods to improve the performance gap, along with reported quantitative performance metrics.

Deep Learning Architecture	Datasets Used	Quantitative Result
Swin-Transformer-tiny and ResNet-18 [37]	GenImage, DiffusionForensics, ADM-generated images.	Accuracy (cross-generator - GenImage): 80.4% (5.6% to 16.4% higher than the baseline). Accuracy (cross-generator - DF): 80.4% (1.7% higher than the baseline).
Gated Expert Convolutional Neural Network. Expert models use ResNet-50 pre-trained on ImageNet [38]	GenImage, CIFAKE, images generated by BigGAN, GLIDE, stable diffusion.	Achieves at least 92% accuracy
18-layer ResNet using CDC, Atrous Spatial Pyramid Pooling (ASPP) [39]	FaceForensics++. DeeperForensics-1.0. Celeb-DF. DFDC.	Accuracy: 71.4% AUC: 84.1%
Coupled Deep Neural Network (CDNN)49	CelebA, images generated by DCGAN, WGAN, WGAN-GP, LSGAN, and PGGAN.	Precision/Recall (LSGAN): 0.981/0.956 Precision/Recall (DCGAN): 0.986/0.986. Precision/Recall (PGGAN): 0.951/0.936

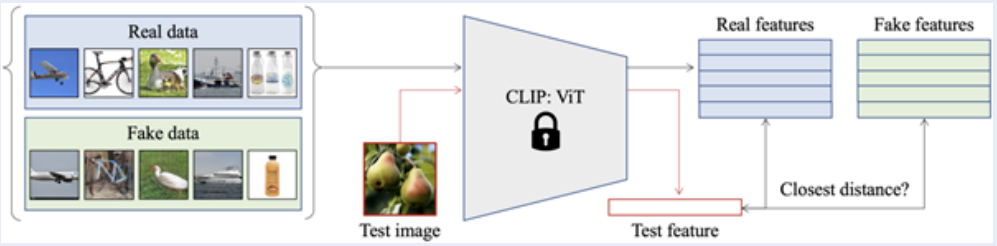

Figure 9: Methodology proposed by Ojha et al.[35] Nearest neighbors are employed for real-vs-fake classification. Real and fake images are mapped to their corresponding feature representations using the CLIP:ViT network. A test image is then projected into the same feature space, and the cosine distance is calculated to identify the nearest neighbor within the feature bank.

The SIDA framework developed by Huang et al.[43] represents the next evolutionary step in forgery detection, leveraging a large multimodal model not only for detection (classifying an image as real or fake) but also for localization (predicting a mask of the tampered regions) and providing textual explanations for its reasoning. This marks a significant move toward interpretable and trustworthy AI forensics. The method demonstrates resilience to low-level social media distortions such as JPEG compression and resizing, though its sole reliance on the FLUX generative model for synthetic images may introduce data skew, and localization performance could be improved in cases of complex or subtle tampering. CLIP-based detectors are summarized in Table 6. Despite these advancements, researchers continue to explore methods that reduce dependency on fake images for training.

Detectors That Do Not Require Fake Images for Training

A major challenge in creating reliable detection systems is the requirement to gather extensive, varied datasets of synthetic images, a task that grows more difficult with the release of each new generative model. This has led researchers to investigate alternative methods that could bypass the need for synthetic training data. Two distinct, promising approaches have emerged in this field.

One-Class Classification (OCC) frames Deepfake detection as an anomaly detection problem, employing a model called OC-FakeDect based on a One-class Variational Autoencoder[44]. This approach trains exclusively on real face images to learn the underlying distribution of authentic data. Fake images, which deviate from this learned distribution, are identified as anomalies due to their high reconstruction error

Table 6: Representative CLIP-based methods with reported quantitative performance metrics.

Deep Learning Architecture	Datasets Used	Quantitative Result
Frozen, pre-trained CLIP:ViT-L/14 [35]	ProGAN real/fake images, fake images from Pro-GAN, StyleGAN, BigGAN, CycleGAN, StarGAN, GauGAN, Guided Diffusion Model, LDM, Glide, DALL-E.	Accuracy (unseen models): 81.52% - 84.25% AUC (unseen models): 90.58 Improves upon SoTA by +15.07 AUC and +25.90% accuracy
CLIP:ViT-L/14 + LoRA [40]	ForenSynths, ProGAN, LSUN, FAMSeC, ForenSynths, UniversalFakeDetect, and GenImage.	Accuracy (ForenSynths): 96.52% Accuracy (UniversalFakeDetect): 97.85% Accuracy (GenImage): 90.90% Improved 14.55% accuracy over the baseline (UniFD)
CLIP:ViT + dynamic aggregation-based classifier [41]	UnivFD, images genegrated by GANs, Deepfakes, Diffusion models	Accuracy: 86.11% Precision: 93.52% Increased of +1.86% accuracy and +0.14% precision over the SoTA. Increased +14.46% accuracy on imbalanced datasets.
CLIP ViT L/14 + linear SVM classifier [42]	COCO, Latent Diffusion , images from GAN-based, DM-based, and commercial tools (DALL-E 3, Midjourney v5, Adobe Firefly).	AUC (original images): 90.0% AUC (post-processed images): 85.2%
LISA + LoRA [43]	SID-Set, fake images generated by FLUX, DMimage	Accuracy (SID-Set): 93.5% Accuracy (DMimage): 90.3%

when passed through the autoencoder. The proposed system includes two variants: OC-FakeDect-1 and the superior OC-FakeDect-2. While OC-FakeDect-1 computes the reconstruction score directly from the input and output images, OC-FakeDect-2 uses an additional encoder block to compute the score between the latent representations of the input and output, resulting in improved anomaly scoring and higher performance. The proposed models achieved 97.5% accuracy in detecting Deepfakes, outperforming the two-class MesoNet and demonstrating comparable performance to XceptionNet; it even surpassed the latter on the NeuralTextures (NT) dataset despite using only real images for training. However, its reliance on the RMSE function to compute the reconstruction score is noted as a limitation, and performance needs improvement for other datasets.

Zero-shot detection uses a model trained on real images, such as a lossless compressor (SReC), to predict the probability distribution of pixels. It detects fakes by measuring the "coding cost gap": the difference between the negative log-likelihood and entropy (H) of an image. For a real image that conforms to the learned statistical model, these values are closely aligned, while large gaps appear in AI-generated images (Figure 10), indicating that the image's pixel distributions do not fit the learned model of real-world images [45]. The core advantages of this method include independence from generator architectures and the elimination of the need for synthetic images during training. It is also computationally efficient and achieves state-of-the-art performance, with an average accuracy improvement of more than 3% over previous state-of-the-art methods across a wide variety of unseen GANs and DMs (e.g., DALL-E, Midjourney V5, Stable Diffusion XL). The AUC exceeded 85% on many models, demonstrating consistent, high performance. However, performance varied significantly in specific cases (e.g., GLIDE-R dataset) depending on the decision statistic used. In addition, detector performance was significantly impacted by the quality and choice of the real dataset used to train the lossless coder. These strategies are summarized in Table 7.

Discussion

The proliferation of AI-generated content necessitates robust and adaptable detection mechanisms capable of distinguishing synthetic images from authentic ones. Techniques developed to address this challenge range from traditional CNNs, which focus on detecting specific artifacts, to advanced vision transformers. These methods aim to overcome challenges such as poor generalization across different generative models, sensitivity to image perturbations like compression, and the high computational costs associated

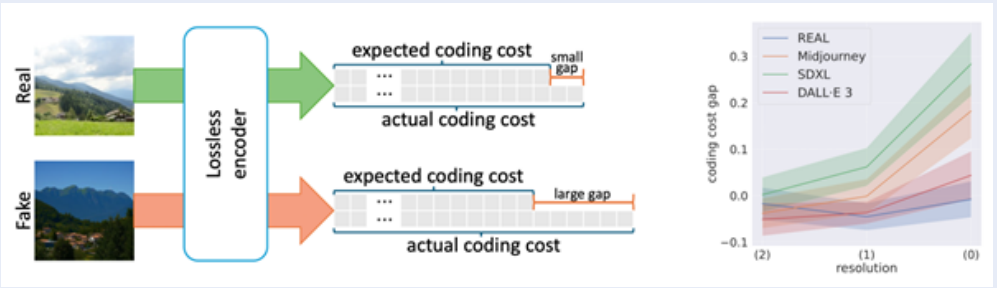

Figure 10: Zero-shot detection mechanism using a coding cost gap[45]. Left: Actual code cost for real images closely matches the expected value of the Zero-Shot Detector, while synthetic images show significant differences. Right: How the coding cost gap increases significantly for synthetic images compared to real images when predicting high-resolution details from low-resolution data.

Table 7: Representative methods trained without needing fake images and their reported quantitative performance metrics

Deep Learning Architecture	Datasets Used	Quantitative Result
One-class Variational Autoencoder[44]	FaceForensics++, Deepfake Dete tion Dataset (DFD).	Average accuracy: 97.5% Accuracy (NT): 97.50% Accuracy (DF): 88.40% Accuracy (FS): 86.10% Accuracy (F2F): 71.20% Accuracy (DFD): 98.20%
Super-resolution based lossless Compressor (SReC)[45]	Open Images, LAION, COC fale images from GauGAN, Styl GAN2Stable Diffusion, DALL- Midjourney	AUC (StarGAN): 100 AUC (BigGAN): 92.6 AUC (GLIDE): 65.1 AUC (SDXL): 99.2 AUC (DALL-E) 98.2 +3.4% of accuracy over the SoTA

with training on large, evolving datasets.

Approaches to artificial fingerprint detection were highly effective in the early stages of AIGC development[30,31,34,46]. This efficacy stemmed from their ability to detect images originating from GAN models with high accuracy and identify their source generators due to unique, specific artificial fingerprints inherent in the synthesis process[47]. However, a critical weakness in the generalization of these models was soon revealed as the set of generative models constantly expanded[41]. Their low resilience to common processing techniques, often termed laundering operations, such as compression and resizing, which tend to destroy the subtle forensic evidence crucial for detection, was another key limitation[47]. Furthermore, they were susceptible to specialized tools designed to erase these fingerprints, such as GANprintR[48].

Methods aimed at mitigating the performance gap were developed in response to the severe limitations of early detection techniques[37–39,49], representing a transition from reliance on source-specific forensic fingerprints to extracting domain-agnostic or universal fake features that describe general properties of synthesized images across architectures. This approach aimed to achieve robust cross-generator and cross-scene performance. Strategies included advanced texture and feature analysis[39], purification and isolation of artifacts[37], use of ensemble and gated expert models[38], and pairwise and contrastive learning[49]. These collective strategies led to a new generation of detectors demonstrating superior performance and generalization, although challenges remained in detecting images from entirely new, unknown generative architectures.

Approaches using CLIP-based detectors leverage the feature space of a large pre-trained VLMs (typically the CLIP:ViT-L/14 image encoder)[35,42]. This fundamentally addresses a major vulnerability of older supervised methods by preventing the network from overfitting to specific low-level fake artifacts and mistakenly classifying unseen generative content into the "real" sink class[35]. Consequently, CLIP-based de-

tectors exhibit state-of-the-art generalization ability across various generative models and architectures [42]. This capability allows the method to achieve superior performance when tested on entirely unseen generative model families or with extremely limited training data. Although the CLIP-based approach is also robust against common post-processing operations, its accuracy degrades on Gaussian-blurred images [50]. Furthermore, the architecture's accuracy relies heavily on the quality and size of the CLIP pre-training data. OCC [44] and ZSD [45] mark a significant departure from conventional binary supervised detection techniques, which rely on extensive and continually expanding datasets of fake images. These methods were designed to achieve broad generalization by functioning independently of synthetic training data. They reframe the task as an anomaly detection problem, where models are trained solely on real, pristine images to learn the statistical properties and underlying distribution of authentic content. However, their effectiveness depends on the accuracy of the real image distribution and the threshold used for score computation to classify images.

The emergence of AI-generated content necessitates continuous advancements in detection methodologies, shifting from dependency on specific forgery traces to the pursuit of universal generalization capabilities. Early approaches targeted artificial fingerprints—subtle, model-dependent artifacts produced by generative processes—which demonstrated high initial effectiveness but exhibited significant declines in performance when confronted with unseen generative architectures or common real-world distortions like compression and resizing. These operations often obscured the faint traces left behind.

Subsequent supervised methods employed complex architectures to enhance robustness but frequently struggled to generalize across different families of generative models due to overfitting on specific training artifacts. A notable breakthrough came with CLIP-based detectors, which utilized the fixed, information-rich feature space of large pre-trained VLMs to achieve robust detection and strong generalization, even when trained on limited examples.

The current frontier is defined by methods such as OCC and ZSD. By focusing solely on the characteristics of real images—such as identifying deepfakes as anomalies or leveraging expected pixel probability distributions derived from lossless encoders—these approaches fundamentally shift the paradigm away from reliance on extensive, ever-expanding datasets of fake images. This offers a critical pathway toward genuinely robust, generator-agnostic, and future-proof detection systems.

CONCLUSIONS

Both region-based and generative forgery detection continue to face a persistent generalization gap, with models frequently failing to recognize novel manipulation or generation techniques. Future research should prioritize robustness against adversarial post-processing, improved pixel-level localization, and the ability to generalize to unseen generative models. The field is becoming an ongoing algorithmic arms race, necessitating adaptive and explainable systems capable of detecting statistical anomalies rather than specific forgery types. Progress will depend on developing lightweight, scalable, and interpretable detectors resilient to evolving anti-forensic strategies and real-world degradations.

LIST OF ABBREVIATIONS

GAN – Generative Adversarial Network
DL – Deep Learning
ML – Machine Learning
CNN – Convolutional Neural Network
CFA – Colour Filter Array
FCN – Fully Convolutional Networks
LSTM – Long Short-Term Memory
E2E-Fusion – End-to-End-Trainable CNN Framework
QF – quality factor
RRU-Net – Ringed Residual U-Net
CMF – Copy-Move Forgery
CMFD – Copy-Move Forgery Detection
GWO – Grey Wolf Optimizer
AIGC – Artificial Intelligence Generated Content
DM – Diffusion Model
SRM – Spatial Rich Model
MTD-Net – Multi-scale Texture Difference model
FAM – Forgery Awareness Module
CDC – Central Difference Convolution
ASPP – Atrous Spatial Pyramid Pooling
VLM – Vision-Language Models
ViT – Vision Transformer
LoRA – Low-Rank Adaptation
SeC – Semantic feature-guided Contrastive learning
SVM – Support Vector Machine
OCC–One-Class Classification

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Dinh Phuc Do conducted the literature review, synthesized the findings, and drafted the manuscript. Phuong Nghi Tram and Tu Nguyen contributed to the literature review process, provided critical feedback on the reviewed content, and helped refine the manuscript. Kha Tu Huynh conceived the research idea, defined the review scope and paper structure, guided the writing process, and critically reviewed and supervised the work. All authors reviewed and approved the final manuscript.

ACKNOWLEDGEMENTS

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number B2026-28-18

REFERENCES

1. Lin X, Tang W, Wang H, Liu Y, Ju Y, Wang S, et al.. Exposing image splicing traces in scientific publications via uncertainty-guided refinement; 2023.
2. Caldeira E, Neto PC, Gonçalves T, Damer N, Sequeira AF, Cardoso JS. Disentangling morphed identities for face morphing detection. Sci Talks. 2024;10(4).
3. Pandey A, Mitra A. Detecting and Localizing Copy-Move and Image-Splicing Forgery. 2022.
4. Hao J, Zhang Z, Yang S, Xie D, Pu S. TransForensics: Image Forgery Localization with Dense Self-Attention. ICCV2021. 2021;p. 15035–15044.
5. Kwon MJ, Yu IJ, Nam SH, Lee HK. CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing. IEEE. 2021;p. 375–384.
6. Qazi E, Zia T, Almorjan A. Deep Learning-Based Digital Image Forgery Detection System. Appl Sci (Basel). 2022;12(6):2851.
7. Gokhale A, Pramod D, Thepade S, Kulkarni R. MissMarple: A Novel Socio-inspired Feature-transfer Learning Deep Network for Image Splicing Detection. 2021.
8. Zhuang P, Li H, Tan S, Li B, Huang J. Image Tampering Localization Using a Dense Fully Convolutional Network. IEEE Trans Inf Forensics Security. 2021;16:2986–2999.
9. Gazzah S, Haddada LR, Shallal I, Amara NE. Digital Image Forgery Detection with Focus on a Copy-Move Forgery Detection: A Survey. In: International Conference on Cyberworlds (CW); 2023. p. 240–247.
10. Bi X, Wei Y, Xiao B, Li W. RRU-Net: The Ringed Residual U-Net for Image Splicing Forgery Detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019. p. 30–9.
11. Kadam K, Ahirrao S, Kotecha K, Sahu S. Detection and Localization of Multiple Image Splicing Using MobileNet V1. IEEE Access. 2021;9:162499–162519.
12. Bappy JH, Simons C, Nataraj L, Manjunath BS, Roy-Chowdhury AK. Hybrid LSTM and Encoder-Decoder Architecture for Detection of Image Forgeries. IEEE Trans Image Process. 2019;28(7):3286–3300.
13. Marra F, Gragnaniello D, Verdoliva L, Poggi G. A Full-Image Full-Resolution End-to-End-Trainable CNN Framework for Image Forgery Detection. 2019.
14. Ali SS, Ganapathi II, Vu NS, Ali SD, Saxena N, Werghi N. Image Forgery Detection Using Deep Learning by Recompressing Images. Electronics (Basel). 2022;11(3):403.
15. Shan W, Xie Y, Wang P, Zou D, Qiu J, Li J. DSNI-Net: Forensic Network for Detecting Splicing in Salt-and-Pepper Noisy Images. IEEE Access. 2025;13:38325–38341.
16. Zhou P, Han X, Morariu V, Davis L. Learning Rich Features for Image Manipulation Detection. 2018;.
17. Yang C, Li H, Lin F, Jiang B, Zhao H. Constrained R-Cnn: A General Image Manipulation Detection Model. In: IEEE International Conference on Multimedia and Expo (ICME); 2020. p. 1–6.
18. Wu Y. AbdAlmageed W, Natarajan P. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 9535–9544.
19. Kuznetsov A. Digital image forgery detection using deep learning approach. J Phys Conf Ser. 2019;1368(3).
20. Walia S, Kumar K, Kumar M, Gao XZ. Fusion of Handcrafted and Deep Features for Forgery Detection in Digital Images. IEEE Access. 2021;9:99742–55.
21. Lalli K, Shrivastava V, Shekhar R. Detecting Copy Move Image Forgery using a Deep Learning Model: A Review. 2023;p. 1–7.
22. Wu Y, Abd-Almageed W, Natarajan P. BusterNet: Detecting Copy-Move Image Forgery with Source/Target Localization. In: Proceedings of the European Conference on Computer Vision (ECCV). vol. 11207. Springer; 2018. p. 170–186.
23. Nazir T, Nawaz M, Masood M, Javed A. Copy move forgery detection and segmentation using improved mask region-based convolution network (RCNN). Appl Soft Comput. 2022;131.
24. Abdalla Y, Iqbal MT, Shehata M. Copy-Move Forgery Detection and Localization Using a Generative Adversarial Network and Convolutional Neural-Network. Information (Basel). 2019;10(9):286.
25. Tinnathi S, Sudhavani G. Copy-Move Forgery Detection Using Superpixel Clustering Algorithm and Enhanced GWO Based AlexNet Model. Cybern Inf Technol. 2022;22(4):91–110.
26. Li H, Huang J. Localization of Deep Inpainting Using High-Pass Fully Convolutional Network. In: IEEE/CVF International Conference on Computer Vision (ICCV); 2019. p. 8300–9.
27. Ning Y, Larry D, Mario F. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In: IEEE/CVF International Conference on Computer Vision (ICCV); 2019.
28. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, NieM. FaceForensics++: Learning to Detect Manipulated Facial Images. 2019.
29. Xi Z, Huang W, Wei K, Luo W, Zheng P. AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network. In: and others, editor. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); 2023. p. 1463–1470.
30. Wang SY, Wang O, Zhang R, Owens A, Efros AA. CNN-Generated Images Are Surprisingly Easy to Spot… for Now. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. p. 8692–701.
31. Nataraj L, Mohammed TM, Manjunath B, Chandrasekaran S, Flenner A, Bappy MJ, et al. Detecting GAN generated Fake Images using Co-occurrence Matrices. Electronic Imaging. 2019;2019(5):532–537.
32. Aggarwal G, Srivastava AK, Jhajharia K, Sharma NV, Singh G. Detection of Deep Fake Images Using Convolutional Neural Networks. In: and others, editor. 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS); 2023. p. 1083–1087.
33. Raza SA, Habib U, Usman M, Cheema AA, Khan MS. MMGAN-Guard: A Robust Approach for Detecting Fake Images Generated by GANs Using Multi-Model Techniques. IEEE Access. 2024;12:104153–64.
34. Frank JC, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, Holz T. Leveraging Frequency Analysis for Deep Fake Image Recognition. In: and others, editor. Computer Vision and Pattern Recognition; 2020.
35. Ojha U, Li Y, Lee YJ. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 24480–9.
36. Ahire V, Muley A, Zample S, Verma S, Menon P, Madan S, et al. SFANet: Spatial-Frequency Attention Network for Deepfake

Detection. 2025.
37. Meng Z, Peng B, Dong J, Tan T, Cheng H. Artifact feature purification for cross-domain detection of AI-generated images. Computer Vision and Image Understanding. 2024;247.
38. Saskoro RAF, Yudistira N, Fatyanosa TN. Detection of AI-Generated Images From Various Generators Using Gated Expert Convolutional Neural Network. IEEE Access. 2024;12:147772–147783.
39. Yang J, Li A, Xiao S, Lu W, Gao X. MTD-Net: Learning to Detect Deepfakes Images by Multi-Scale Texture Difference. IEEE Trans Inf Forensics Security. 2021;16:4234–4245.
40. Xu J, Yang Y, Fang H, Liu H, Zhang W. FAMSeC: A Few-shot-sample-based General AI-generated Image Detection Method. In: and others, editor. IEEE Signal Processing Letters; 2024. p. 1–5.
41. Lyu Z, Xiao J, Zhang C, Lam KM. AI-Generated Image Detection With Wasserstein Distance Compression and Dynamic Aggregation. In: IEEE International Conference on Image Processing (ICIP); 2024. p. 3827–3833.
42. Cozzolino D, Poggi G, Corvi R, NieM, Verdoliva L. Raising the Bar of AI-generated Image Detection with CLIP. 2024;p. 4356–4366.
43. Huang Z, Hu J, Li X, He Y, Zhao X, Peng B, et al. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. 2025;p. 28831–28841.
44. Khalid H, Woo SS. OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020. p. 2794–803.
45. Cozzolino D, Poggi G, NieM, Verdoliva L. Zero-Shot Detection of AI-Generated Images. 2024;p. 54–72.
46. Guarnera L, Giudice O, Battiato S. DeepFake Detection by Analyzing Convolutional Traces. 2020;p. 2841–2850.
47. Gragnaniello D, Cozzolino D, Marra F, Poggi G, Verdoliva L. Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art. In: IEEE International Conference on Multimedia and Expo (ICME); 2021. p. 1–6.
48. Neves JC, Tolosana R, Vera-Rodriguez R, Lopes V, Proença H, Fierrez J. GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection. IEEE J Sel Top Signal Process. 2020;14(5):1038–1048.
49. Zhuang YX, Hsu CC. Detecting Generated Image Based on a Coupled Network with Two-Step Pairwise Learning. In: IEEE International Conference on Image Processing (ICIP); 2019. p. 3212–3216.
50. Park D, Na H, Choi D. Performance Comparison and Visualization of AI-Generated-Image Detection Methods. In: and others, editor. IEEE Access; 2024. p. 62609–62627.