INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

A Comprehensive Review on Deep Learning Approaches for Classifying Real and AI generated Images

Yashraj Namdeo[1], Mr Nitesh Gupta[2]

[1]Department of Computer Science and Engineering (CSE), NRI Institute of Information and Technology (NIIST), Bhopal, Madhya Pradesh, 480661 India

[2]Assistent Professor, NIIST, Bhopal, Madhya Pradesh, 480661 India

DOI : https://doi.org/10.51583/IJLTEMAS.2026.150100013

Received: 08 January 2026; Accepted: 13 January 2026; Published: 23 January 2026

ABSTRACT:

Due to the realm of the blooming generative models, like GANs, VAE and diffusion-based environment, the trustworthiness of authentication in visual media has been vastly challenged. AI-generated images appearing quite photorealistic go beyond human perceptual boundaries often overlying concerns for disinformation, digital tampering, forensic evidence authentication, and or the security of biometric purposes. This review organizes throughout the chronology of deep learning advancement to segregate real from synthetic images by focusing not much on its generalization snags, extraction of fake image or artifacts from sustainability in the aspect of datasets. Also this review provides systematic analysis on publicly available datasets and those essential research directions in the future being necessary to become a robust detection mechanism for fake images.

Keywords: - Deep Learning, AI-Generated Images, Diffusion Models, Image Forensics, Artifact Detection, Generalization, Convolutional Neural Networks (CNNs), Vision Transformers (ViT)

INTRODUCTION

The rapid progress in artificial intelligence has completely transformed the abilities of machines to create synthetic visual content that almost achieves, and in many cases surpasses, real images. Modern generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), autoregressive models, and cutting-edge diffusion models have been developed to paint high-resolution images with very coherent textures, lighting, geometry, and semantic details. These results in high resolution and high coherence were otherwise out of reach in computational synthesis [1]. Another key advancement is that latent diffusion models—Stable Diffusion, Midjourney, and DALL·E—have farther increased the reality and accessibility of image generation through a combination of high-dimensional denoising, cross-attention mechanisms, and prompt-conditioned semantic control [2]. Consequently, artificial and real visual elements have to some degree started to merge, thereby causing a huge opening for creativity. With this opening, societal and technological issues arise.

The field of malicious AI-generated image detection became all the more pronounced with generative modeling; synthetic images can have both benign and nefarious purposes, the harmless - digital art, entertainment, data augmentation, advertisement design, and conceptual visualization-inbalance the risk-prone purposes - misinformation, social engineering attacks, identity manipulation, fraudulent media production, and tampering with legal evidence. Amongst the various areas where serious concerns have been raised have been deep fake applications, which are able to alter normal facial identities and replicate human expressions with uncanny precision. The attractiveness of the situation is amplified by the fact that generative tools are available as both open-source and cost-effective, so as threats can be brought to bases in the confusion of reality [3]. Substantial capability for these visual hyper-realistic creations is made available to those actors outside professional skill and specialized tools, as a result of which generically robust detection systems become necessary for preserving trustworthiness in visual media, safety of online platforms, and digital trust.

The issues herein become of heightened interest given the speed at which generative models evolve. Traditional forensic features like extreme textures, spatial discontinuities, or boundary artifacts were easy enough to identify

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

early on in GAN-generated images. Their current diffusion model prototypes do away with these recognizable flaws of early generations by using better sampling methods along with noise schedules and improved-complexity latent datasets. The natural depiction of realistic reflections, shadows, skin details, and combined background patterns is impressive in relation to the ability for humans to make accurate assessments visually. Forensic tools available so far are very poor at detection, since the said tools, more often than not, cater to the specific type of generative signatures that no longer exist in the domains of those implementations. So, characteristics with which forensic audits are carried out mutate rapidly or are heavily inclined to extinction within diffusion-constructed images.

While substantial research work has been conducted, state-of-the-art deep learning-based detection methods are facing many challenges. One well-known limitation in particular is that the models fail to generalize across various image generators, architectures, and training regimes. Almost all the detection models have been trained on images produced by a single generator such as a specific version of StyleGAN or a specific implementation of Stable Diffusion which hampers the detection of synthetic images produced by unseen models [5]. These models, however, were badly designed according to different generative models, which have greatly different principles regarding their sampling strategy, architectural design, and noise-characteristic use. Accordingly, a GAN model often performs poorly on diffusion-generated images due to the distinct artifact signature of each type of generator.

Another hurdle is the extraction of fine forensic evidence that modern generative models might be actively hiding. Initially, the detectors detected gross issues due to color or texture distributions but more advanced contrivances had different tricks suitably encoded so no one noticed them. The loyalty of small evidences—anomalies in texture geometry due to micro-irregularities, or unexpected transitions on the edges, in the distorted spatial frequency or interpolation traces in the latent space often calls for the art of multi-scale feature extraction. Definitely, excerpts from the former allow subtle structures to break through: with active convolution over the ViT model, and then added through a wavelet-based decomposition as a way to produce even more finer-grained results during global teaching. Lesser issues may surface when employing such frequency-domain recruits in a real-world model that calls for the accurate capture of latent discrepancies in advanced image generation. However, as models continue to gain sophistication concerning synthetic data generation, emergent challenges stretch into a future of uncertainties [7].

The third challenge has got to do with data, characterized by the absence of a particularly diverse and global set of conditions. The deterrence to detection comes in the form of many datasets comprising highly polished real color images under controlled conditions that do not capture drives for amplification by mobile cameras, social media compression, noise, blur, resizing, or filtering. So, detecting models taught on "clean" datasets realistically stand withdrawn in field scenarios while posting rise to 'real' images in online settings. The yawning chasm that separates laboratory performance from practical implementation only makes the call for larger training datasets that can encompass the needed variation from real-world data, the potential to produce multiple generators, and various resolutions in order to enhance the ability to capture the changing habits of the world.

This review adds to existing literature taunting several good career prospects. First, it consists of a structured synthesis of deep learning applications utilized for detecting synthetic images through CNNs models, transformer approaches, frequency-based methods, and hybrid architectures. Secondly, the current standing on accepted real and synthetic image datasets is considered across their characteristics, limitations, and appropriateness in evaluation of detection performance. Thirdly, the discussion is an appraisal of the existing research gaps in generalization, fine-grained artifact extraction, dataset diversity, explainability, and edge deployment and actional thoughts for future research. Fourth, the review discusses newer ideas such as zero-shot detection, contrastive learning, anomaly-based modeling, and multi-modal approaches, along the broader context of digital image forensics. To sum up, this review covers the pooled value and direction of the impact of deep learning in its role in synthetic image detection in the evolution of the field, the missing points for taking a leap towards building secure-detection setups in the real world.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

RELATED WORK

Generative Models

Generative Adversarial Networks (GANs) have been extensively explored for image generation due to their adversarial training mechanism, which enables the generator to synthesize realistic images by fooling a discriminator. Mohammadjafari et al. [1] (2022) proposed an improved 3D α GAN for generating realistic 3D connected volumes, particularly for medical applications, demonstrating that their approach preserves the moment invariance to a certain extent, which is a very desired characteristic for real volumetric data. Yet, there evaluation was limited to synthetic medical data, so transfer to natural images could not be assessed. Sabnam [2] (2024) revisited GAN applications in medical imaging, with reviews and discussions in the literature, such as articles associating GAN with data augmentation in terms of upwards shift in sensitivity of downstream network in segmentation tasks and GAN-related computational demand issues related to training and hyperparameter tuning." Ruan et al. [3] (2023) unveiled architectural and loss-function improvements for GAN models for negative-value data generation for improved classifier performance in bearing data, but the results were relevant for the domain and may not generalize to broader image-generation tasks. Therefore, the bright side of GANs is that they are good at producing lifelike outputs, while the flip side is that the domain, data, and training stability cause drawbacks.

Diffusion has been evolved as a powerful GAN replacement for creating high-fidelity images with fewer training problems emerging along the way. That is what Wang et al. [4] (2023) set out to demonstrate, with an improved model based on diffusion, so called adversarial training, garnering the robust accuracy data of 70.69% on CIFAR-10 and 42.67% on CIFAR-100 under ℓ_∞ threat models. (Huang et al. [5], 2023) introduced another dimension by having crossed-modal diffusion frameworks, DiffDis [diffusion-disentangled learning], where text was combined with image representations, showing a clear improvement by 2.42 FID when the model was compared against the baseline and by 1.65 percentage, increasing PET accuracy on zero-shot image–text classification across 12 data sets. (Hatamizadeh et al. [6] 2023) married the Vision Transformers with diffusion models in DiffiT; SOT ImageNet-256 FID was calculated to be 1.73 and relinquished 17–19% of the volume unread. The case was shown by Azizi et al. [7] in 2023 to illustrate the situation wherein synthetic data was generated utilizing diffusive model information to improve ImageNet recognition up to 69.24% from 64.96% using models ResNet and ViT, although these gains were underpinned by heavy-handed augmentation complete with question marks.This attests to the perceptual qualities and adaptability of diffusion models, and in the same breath, to their underperformance in terms of sampling speed and domain-specific dependence.

Exemplified mainly by examples from their image-generating prowess, autoregressive structures have taken a strong stand of mastering sequence data. On all fronts of scaling an image, Tian et al. [8] (2024) introduced the approach of VARS, where they predict one level farther in the future. With an FID improvement ranging from 18.65 to 1.73 and an IS from 80.4 to 350.2 running at a speed about 20× faster than diffusion transformers. Tang et al. [9] (2025) put forth a modified hybrid diffusion-autoregressive model seen as beneficial to improvement in stressed reconstruction loss and hard FID choice for finite small-scale datasets as the huge scale and high-complexity images yet remain a massive challenge. In this scenario, Yu et al. [10] (2025) developed Randomized Autoregressive (RAR) models with bidirectional context, one with FID of 1.48 on ImageNet256 by Li et al. [11] (2024) or AR image generation without vector quantization; diffusion loss was under 2.0 FID and inference times were under 0.3 s for every respective image. Nevertheless, the autoregressive methodology is lesser computationally heavy for a high-resolution output when operated with extremely serialized, complex dependencies.

Although Variational Autoencoders (VAEs) are less prominent in recent high-resolution image generation studies, they remain influential in learning latent representations and probabilistic modeling. VAEs allow controlled manipulation of latent codes and have been applied in conditional image generation and hybrid GAN-VAE frameworks [12]. However, VAEs often produce blurrier images compared to GANs and diffusion models due to limitations in reconstruction loss and latent distribution approximation.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

CNN-Based Detectors

Recently, Convolutional Neural Network (CNN) architectures are researched extensively for the detection of AI-generated or tampered images and videos. Lipianina Honcharenko et al. [13] (2024) conducted an evaluation of the ResNet, EfficientNet, and Xception architectures in the case of deepfake video detection and mentioned that Xception and EfficientNet yielded better balance performance than ResNet. Further reporting had that Xception produced an accuracy of 87.7% with an accuracy of 80.3%, making EfficientNet and ResNet much more computationally effective. While admittedly Xception's deeper structure picks up on minute handicraft artifacts, the network is not very well-suited for real-time deployment. Similarly, Yasser et al. [14] (2024) tried EfficientNet-B4 plus XceptionNet with the FaceForensics++ and Celeb-DF (v2) datasets for deepfake detection. The bright neural network—with a mixture of face cropping and normalization—outperformed largely for these pioneering work, thereby showing high AUC values, which indicate strong discriminatory power. While EfficientNet-B4 performed well, XceptionNet outperformed in certain cases. Hao Lin et al. [15] (2021) proposed an improved version of the Xception model employing a dual-attention mechanism and feature fusion. The model enables this Xception model to focus on subtle local artifacts of forgery and global artifacts generated by other deep learning-based methods. The experiments led to improvement in detection with enhanced generalization, though the complexity might have increased training time and cost. Qadir et al. [16] (2024) made the incorporation of a ResNet backlink for a successful deepfake detection method. Utilizing its pre-trained weights for initialization, the model showed marked improvement in accuracy, though convergence was quicker as it stood robust in massive video analysis. Nonetheless, with the ResNet architecture, there is a slight loophole in missing out on micro-level forgery artifacts. Finally, Vaishnavi et al [17] (2025) fused two basic CNNs; a hybrid approach that merged ResNet50 and EfficientNet-B0 for enhanced performance in deepfake detection. The fusion model's final prediction accuracy was 89.08%, with precision = 0.91, recall = 0.87, and F1-score = 0.89, improving on the individual performance of ResNet50 and EfficientNet-B0. This fusion strategy worked for the improvement of the basic detection by pushing the model size towards increasing inference cost with the attendant threshold effect on accuracy and efficiency. These collective works establish one fact in commonality: Xception models are particularly good at seizing subtle forgery fragments due to their very deep convolutional architecture, ResNet models combine unbridled robustness and scalability with effective pre-trained features, and EfficientNet gives the best-fitting trade-off between the possible computational efficiency and detection accuracy. The hybrid approach employing mixed CNN architectures augmented by combining strengths from one to the other implanted enhancement into the detection performance; on the flip side, increased complex systems and resource requirements cause challenges for deployment scenarios in real-time or scaling for video analytics.

Transformer-Based Detectors

Models based on Vision Transformers (ViTs) and hybrid ViT-CNN models are preferred for detecting deepfakes and fake images due to their key ability in aggregating long-range spatial dependencies and global contextual information. One of the initial works, so-called Convolution Vision Transformer (CViT), was proposed for deepfakes by Wodajo and Atnafu [18]. To sum up, the authors developed an initial attempt at combining the CNN feature extractor with the ViT classifier towards developing a good detection method for these fake videos. On this project, an experiment was carried out on the DFDC dataset with ratings as follows: accuracy was 91.5%, area under the curve (AUC) of 0.91, and loss of 0.32. Such results indicated that combining CNN with ViT could really detect the artificial work. Heo et al. [19] in 2012 joined up with the concept by introducing a ViT-detection approach with a sort of knowledge distillation: They fed in CNNers as input to transformers, distilling their treatment to predict. This model hit an AUC of 0.978 and an F1-score of 0.919 in the DFDC dataset, which went on to perform better than earlier CNN-only models at detection and generalization tasks. Al Jallad et al. [20] detected DFDT in another round in 2022 with a multi-stream Vision Transformer detection framework employing re-attention mechanisms instead of core self-attention. Models were tested across FaceForensics++, Celeb-DF (v2), and WildDeepfake; despite incredibly uneven distribution, the DFDT posed on fixed detection probabilities of 99.41%, 99.31%, and 81.35%, impressing results on a cross-dataset generalization basis. Pham Minh Thuan, Lam Bui, and Trung Pham [21] have developed another model named DSViT, or "Spatial Convolutional Vision Transformer," which is an extension of CViT. SD (spatial may be convolution) modules together with CViT for superior feature fusion. This model's improved architectural design better captured

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

falsified artifacts, but the researchers noted the increase in memory and computation. Just lately, Nguyen et al. [22] showed FakeFormer, a transformer-based deepfake detector, for emphasizing the crowds' attention on "artifact-vulnerable" patches. Instead of localizing a mapped noise image, FakeFormer's attention learning focused on "building the focus" inside these noise-carrying patches and detected every local inconsistency, so outperformance of past transformer-based and CNN-based detectors through the train/test discrepancy prevails, thus yielding better generalization for lesser computation costs.

Multi-Scale Feature Fusion

Recent deepfake detection research has ceaselessly grown reliant on multi-scale fusion, the principle applied to join information of various spatial resolution and feature domains (texture and frequency) to better sustain the detection robustness. Based on the good performance across datasets, Zhao et al.'s MFF Net [23] (2021) features directional diversity in learning features for multiple scales through this repository of features from texture and frequency domain fused by ways of Gabor convolution and attention residual blocks. Since the direct comparison only considers the frequency content of the video, trained on FaceForensics++ and Celeb DF, a so-called diversity loss is designed to enforce the learning on predictive diversity to better confront-gether-capably with unseen forgery distortions. Nonetheless, due to their high complexity-a differentiating loss-both cases consumed more resources for training. Henceforth, in 2024, a model for spatial frequency-aware scattering-network multi-scale fusion, SFMFNet, was developed consummating the gathering of intelligence and understanding images and further in-growth designed in the form of a gated regime through cross-attention to combine yet retain textures and high-frequency, by which we was allowed to truly interact with as best an attempt at achieving state-of-the-art AUCs on denoising and benchmark datasets at the least degree of computational cost [24]. From their great efficiency, there remained cases were SFMFNet would not well perceive very slight or adversarially blurred deepfakes. A different line hence used a boosted version of the Xception-in a two-stream fashion-fusion method (Tception) [25] (2023), where two parallel convolutional layers having different kernel sizes are meant to learn features at different scales—on the one hand, while their results will be combined with the prior stage noise space features coming from the SRM filters. In this way at least, a touch of improvement against conventional Xception may be perceived. However, all of these reports have the bottleneck of imposed handmade SRM filters that cannot efficiently belt up with evolving generative artifacts. Likewise persisting transitions were used on the database with another fusion model built as Multi-scale ViT plus CNN; the Multi-scale normally was supposed to garner global features while CNN would rather collect local texture details, and they merged through sparse cross-attention & better spell-out-so with frame-level AUCs of 0.986, 0.984, and 0.988 on Deepfakes, FaceSwap, and Celeb DF (v2), respectively, showing strength with respect to cross compression, but this approach requires large memory and compute of attention maps on high-resolution feature maps [26]. Much later on, D2Fusion for deepfake detection approved such a perspective to on-board the feature fusion strategy in dual-domain using bidirectional attention to build spatially connected features with frequency features, ultimately they opted for the overlaid approach to emphasize the most eminent phase details of counterfeit tokens in deepfakes, out-classing the clusters of prior SOTA works over a variety of deep-fake datasets in the territory; this also implied higher model complexity and slower inference through greater feature contact [27]. These lines all demonstrate the power of multi-scale fusion in detecting subtle generative artifacts via scale-wise spatial, texture, and frequency features, yet also highlight common challenges of these DOA approaches such as computational load, memory requirements, and the choice of durable fusion mechanisms against continuing forgery techniques.

Review of Existing Datasets

Publicly available real image datasets have been widely used in research for image generation, manipulation detection, and computer vision using deep learning. COCO (Common Objects in Context) is a large-scale dataset having over 330,000 images with more than 80 object categories, with annotation in the forms of object segmentation masks, keypoints, and captions [18]. Being distinguishingly diverse with full senses of context, COCO has been popular as a benchmark for object detection and segmentation as well as measuring the reality of model-based synthesized images in scratch images. ImageNet is the most influential dataset in computer vision in which there are above 14 million images across more than 20,000 categories, with the most utilized ILSVRC subset Built with 1.2 million images in 1,000 categories [19]. It has served as one of the main

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

benchmarks for the training of convolutional neural networks, transformer-based models, and hybrid architectures, besides running pre-training on smaller datasets that are task-specific before fine-tuning. Its large number, high variety, and structured hierarchy of classes allow researchers to probe the generalization capacity of both generative and discriminative models. An illustration of FFHQ (Flickr-Faces-HQ), included a collection of stark photos of diverse faces close to 70,000 at 1024×1024, with age, ethnicity, and shade background variations [20]. FFHQ is widely used for face synthesis and deepfake detection because it provides a lot of diversity and high resolution. Thus, the models get to learn critical facial features and expressions. An extensive collection of such real image datasets serves as an excellent backbone for training and evaluating models on generative tasks and the detection of AI-generated images with large, varied, and high-resolution real-world-like images.

AI-Generated Image Datasets

Alongside real image datasets, several AI-generated datasets have been introduced for supporting research on detection of deepfakes or the evaluation of generative models. DeepFake Detection Dataset (DFD) was one of the early large-scale video-based deepfake datasets, constituting several thousand manipulated videos and their real pairs [21]. These videos cover subjects, faces expression, and lighting conditions - indeed, a benchmark value for training and evaluation of CNN-based versus transformer-based deepfake detectors. Whereas StyleGAN and ProGAN datasets are composed of high-quality synthetic images, each mounted using a corresponding GAN architecture that explored human face, object, and scene. These datasets are prominently used to train detectors to recognize unnatural imprints produced in GAN-based antecedence like texture inconsistencies, boundary artifacts, or frequency-domain null. Stable Diffusion image datasets emerged recently, refer to a huge number of images created using latent diffusion models mostly conditioned on textual cues. These datasets are essential for assessing detector-model performance in the context of modern AI-generated pictures exhibiting fewer visible artifacts and almost total photo-realism [23]. Combined, these AI-generated datasets serve as crucial resources for the development of robust detection models separating real from synthetic content and help in studying generative model behavior, artifact patterns, and evaluation metrics across varied environments.

Comparison Table

Table 1 Shows Comparison of AI-Generated Image Datasets

Table 1: Comparison of AI-Generated Image Datasets

Feature	**Style GAN Synthetic Faces Dataset**	**Synthetic ImageNet-100 (Diffusion Models)**
Size	~100,000 to 1,000,000 synthetic face images (varies by release; commonly 100k+)	~130,000 synthetic images (100 classes × ~1,300 images/class)
Number of Classes	**1 (faces only)**	**100 object classes**
Resolution	1024×1024 (common), also 512×512 or 256×256 depending on version	Typically, 256×256 or 512×512 depending on the diffusion model used
Generator Type	**StyleGAN2 / StyleGAN3** (GAN-based generator)	**DDPM, DDIM, or Stable Diffusion variants** (Diffusion-based generator)
Real-World Distortions Modelled	Lighting variations, pose variations, age differences, background artifacts, minor GAN artifacts (blobs/texture inconsistencies)	Noise, blur, occlusion, perspective variation, colour shifts, illumination changes (depending on training set); diffusion artifacts include slight over-smoothing or texture drift

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

Research Gaps & Challenges

Limited generalization: - AI models trained on synthetic or curated datasets often fail to perform well on real-world data due to domain shift. Changes in lighting, background, or noise patterns cause noticeable drops in accuracy**.**

Fine-grained artifact detection: - Even advanced models struggle to capture subtle, micro-level artifacts. Small textural differences, edge inconsistencies, and minute distortions are often overlooked, especially in high-resolution images**.**

Noisy real-world images: - Models trained on clean or synthetic datasets may not handle noise, blur, occlusions, or compression artifacts commonly found in practical applications, reducing reliability in uncontrolled environments**.**

Explain ability gaps: -Deep learning models particularly GANs, transformers, and diffusion architectures operate as "black boxes." Understanding how a model makes a decision is difficult, limiting trust and acceptance in critical fields**.**

Dependency on curated datasets: - AI systems rely heavily on large, high-quality, well-annotated datasets. Any bias, imbalance, or limited variation in these datasets directly affects model performance and fairness**.**

Heavy models not suitable for mobile devices: - Modern models like transformers, diffusion networks, and large CNNs require significant computational power and memory, making them unsuitable for low-resource environments such as smartphones, IoT, and edge devices**.**

Abbreviations and full forms

Table 2 Shows Abbreviations and full forms

Table 2: Abbreviations and full forms

Abbreviation	**Full form**
AI	Artificial Intelligence
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
ViT	Vision Transformer
AR	Autoregressive (model)
DDPM	Denoising Diffusion Probabilistic Model
DDIM	Denoising Diffusion Implicit Model
DFD	Deep Fake Detection Dataset
FFHQ	Flickr-Faces-HQ
COCO	Common Objects in Context

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

AUC	Area Under the Curve
FID	Fréchet Inception Distance
IS	Inception Score
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
SFMFNet	Spatial-Frequency Aware Multi-Scale Fusion Network
MFFNet	Multi-Feature Fusion Network
D2Fusion	Dual-domain Fusion with Feature Superposition
DFDC	Deep Fake Detection Challenge
SOTA	State Of The Art
RAR	Randomized Autoregressive (model)
ResNet	Residual Network
F1	F1-score (harmonic mean of precision and recall)

Declarations

Availability of data and material

This is a review article that synthesizes and analyzes previously published studies and publicly available datasets related to deep learning–based detection of real and AI-generated images. No new experimental data were generated, and all data discussed are available in the cited original publications and datasets.

Competing interests

The authors declare that they have no known financial or non-financial competing interests that could have influenced the work reported in this manuscript.

Funding

The authors did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors for the preparation of this review.

Authors' contributions

All authors contributed to the conception and design of the review, literature search, analysis, and interpretation of the included studies. All authors participated in drafting and revising the manuscript critically for important intellectual content and approved the final version for submission.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Department of Computer Science, NIIST College, Bhopal, for providing academic support and an encouraging research environment during the preparation of this review. The authors also thank the researchers whose work on deep learning, image forensics, and AI-generated image detection forms the basis of this survey.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

CONCLUSION AND FUTURE DIRECTIONS

The study emphasizes the strengths and weaknesses of current AI-generated synthetic content detection/analysis approaches. Even though modern GANs, diffusion models, and transformer-based networked generators can replicate very realistic images, existing detectors fail to generalize enough knowledge, exhibit fine-grained artifact detection in extreme outdoor environments, and even are robust to very complex real-world noise. Also, the dependency on carefully procured datasets and the limited interpretability of deep neural models contribute to the non-adoption of learning algorithms in scenarios of safety-critical applications. Therefore, addressing these deficits is critical in emboldening research for more reliable and trustworthy synthetic-content detection methods. Moreover, more complicated and powerful-than-ever next-level accurate, robust, and bindingly multi-modal detectors which can discover complementarity between images, metadata, sensor data within the video game, and temporal cues are feasible. In that regard, future training pipelines must posses an augmentation that is aware of real-world distortions to handle better the noise, compression, motion blur, and low-light conditions visible in a strict operational setting. Lastly, cross-generator generalization research will provide a critical direction in which applied AI remains effective against newer, more advanced generative models.

REFERENCES

1. S. Mohammadjafari, “Improved 3D α-GAN for Generating Connected Volumes,” arXiv preprint, 2022.
2. S. Sabnam, “Application of Generative Adversarial Networks in Image, Text-to-Image and Medical Imaging,” International Journal of Pattern Recognition and Artificial Intelligence, 2024.
3. D. Ruan, “Improvement of Generative Adversarial Network and Its Application to Bearing Fault Data Augmentation,” MDPI, 2023.
4. Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan, “Better Diffusion Models Further Improve Adversarial Training,” arXiv preprint, 2023.
5. R. Huang, J. Han, G. Lu, X. Liang, Y. Zeng, W. Zhang, and H. Xu, “DiffDis: Empowering Generative Diffusion Model with Cross-Modal Discrimination Capability,” arXiv preprint, 2023.
6. A. Hatamizadeh, J. Song, G. Liu, J. Kautz, and A. Vahdat, “DiffiT: Diffusion Vision Transformers for Image Generation,” arXiv preprint, 2023.
7. S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. Fleet, “Synthetic Data from Diffusion Models Improves ImageNet Classification,” arXiv preprint, 2023.
8. K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, “Visual AutoRegressive Modeling: Scalable Image Generation via Next-Scale Prediction,” NeurIPS, 2024.
9. X. Tang, et al., “Image Generation Method Based on Improved Diffusion Models,” SPIE Conference on Computational Imaging, 2025.
10. Q. Yu, et al., “Randomized Autoregressive Visual Generation,” ICCV, 2025.
11. T. Li, et al., “Autoregressive Image Generation Without Vector Quantization via Diffusion Loss,” NeurIPS, 2024.
12. A. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” ICLR, 2014.
13. K. Lipianina-Honcharenko, M. Telka, and N. Melnyk, “Comparison of ResNet, EfficientNet, and Xception architectures for deepfake video detection,” CEUR Workshop Proc., vol. 3899, 2024.
14. B. Yasser, J. Hani, S. M. Elgayar, and O. Abdelhameed, “Deepfake Detection Using EfficientNet-B4 and XceptionNet,” ICICIS / ResearchGate, 2024.
15. H. Lin, W. Luo, K. Wei, and M. Liu, “Improved Xception with Dual Attention Mechanism and Feature Fusion for Face Forgery Detection,” arXiv preprint, 2021.
16. A. Qadir et al., “An Efficient Deepfake Video Detection Using Pre-trained ResNet CNN,” Journal / Elsevier, 2024.
17. V. D., J. S., G. J., and S. S., “Hybrid Deep Learning Approach for Deepfake Detection Using ResNet50 and EfficientNet-B0,” IROIIP Journal, 2025.
18. D. Wodajo and S. Atnafu, “Deepfake Video Detection Using Convolutional Vision Transformer,” arXiv preprint, 2021.
19. Y.-J. Heo, et al., “Deepfake Detection Scheme Based on Vision Transformer and Distillation,” DeepAI, 2021.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

20. A. Al-Jallad, et al., "DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer," Applied Sciences, vol. 12, no. 6, pp.2953, 2022.
21. P. M. Thuan, B. T. Lam, and P. D. Trung, "DSViT: An Enhanced Transformer Model for Deepfake Detection," Journal of Science and Technology on Information Security, vol. 2, no. 22, 2024.
22. D. Nguyen, M. Astrid, E. Ghorbel, and D. Aouada, "FakeFormer: Efficient Vulnerability-Driven Transformers for Generalisable Deepfake Detection," arXiv preprint, 2024.
23. L. Zhao, M. Zhang, H. Ding, and X. Cui, "MFF-Net: Deepfake Detection Network Based on Multi-Feature Fusion," *Entropy*, vol. 23, no. 12, p. 1692, 2021. MDPI+1
24. "A Spatial-Frequency Aware Multi-Scale Fusion Network for Real-Time Deepfake Detection," Fraunhofer, 2024. deepfake-demo.aisec.fraunhofer.de
25. "Two-Stream Xception Structure Based on Feature Fusion for DeepFake Detection," *Int. J. Computational Intelligence Systems*, vol.16, article 134, 2023. SpringerLink
26. "Multi-scale Deepfake Detection Method with Fusion of Spatial Features," ECICE06 Journal, 2023. ECICE06
27. X. Qiu, X. Miao, F. Wan, H. Duan, T. Shah, V. Ojhab, Y. Long, and R. Ranjan, "D2Fusion: Dual-domain Fusion with Feature Superposition for Deepfake Detection," *arXiv preprint*, Mar. 2025.