Synthetic vs Real Image Detection using Vision Transformers and CNN-based Architectures

Noor ul Huda[1], Ayesha Fayyaz[1], Uswa Asad[1] and Yasir Saleem Afridi[1]

[1]*Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan*

Abstract

The rapid progress of generative AI models such as GANs and diffusion networks has made differentiating between real and computer-generated images increasingly complex. While these systems offer new opportunities for creativity and automation, they also threaten visual authenticity and challenge digital forensics. This paper investigates the use of deep learning methods for classifying authentic and AI-synthesized images using four popular architectures: Vision Transformer (ViT), EfficientNet, ResNet, and VGG19. Each model was fine-tuned and trained with data augmentation to improve generalization across varied image distributions. Experimental results reveal that the Vision Transformer achieved the highest F1-score of 0.86, surpassing CNN-based models. These findings highlight the superior capability of transformer-based methods in identifying subtle structural and textural inconsistencies typical of synthetic imagery.

Keywords

Vision Transformer, EfficientNet, ResNet, VGG19, Synthetic Image Detection, Deep Learning

1. Introduction

The recent surge in generative AI has enabled the production of remarkably realistic synthetic images that are often indistinguishable from genuine photographs. Models such as Generative Adversarial Networks (GANs) and diffusion frameworks have transformed digital media generation, driving innovation in entertainment, design, and visual content creation. However, this technology also introduces significant concerns, including misinformation, identity manipulation, and reduced confidence in visual evidence. Therefore, developing automated methods capable of distinguishing authentic images from AI-generated ones is now a key research focus in digital forensics and media authenticity.

This work was developed as part of the *MediaEval 2025 Synthetic Image Detection task* [1], which aims to advance the automatic detection of generative AI content in real-world imagery. Within this context, we explored deep learning-based approaches for real vs. synthetic image classification.

Early detection methods mainly relied on handcrafted features or frequency-domain patterns to spot artifacts produced during image synthesis. While these approaches offered acceptable performance on specific datasets, they lacked adaptability to new or unseen generative models. With deep learning's evolution, convolutional neural networks (CNNs) such as VGG and ResNet became dominant tools due to their ability to learn hierarchical representations automatically. Recently, transformer-based architectures have gained attention for their ability to model long-range dependencies and global spatial relationships, leading to improved detection accuracy. Several recent studies have also emphasized the growing importance of reliable DeepFake and

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

*Corresponding author.

†These authors contributed equally.

AI-generated image detection frameworks [9, 11, 12, 13].

2. Related Work

The detection of AI-generated imagery has gained considerable attention in recent years. Earlier studies employed pixel-level inconsistencies or statistical irregularities to identify GAN-based content. However, these manually designed approaches often lacked generalization to novel generative models. The emergence of deep learning significantly enhanced this field by enabling data-driven feature extraction. CNN architectures such as VGG [5] and ResNet [4] have been widely adopted in image forensics, leveraging multi-layer feature hierarchies to identify discriminative image characteristics. EfficientNet [3] further improved CNN performance through compound scaling, balancing network width, depth, and resolution. Yet, the local receptive field of CNNs limits their ability to capture holistic spatial information. The Vision Transformer (ViT) [2], in contrast, treats an image as a series of patches and applies self-attention to learn long-distance dependencies between them. ViT models have shown notable results in synthetic image and manipulation detection tasks by analyzing global texture and structural variations that CNNs often fail to recognize. Recent works have also explored diffusion-based synthesis and attention-driven detection frameworks [10, 14, 15].

Several benchmark datasets have been created to facilitate research on synthetic image detection. The dataset proposed by Wang et al. [6] contains samples from major GAN architectures like StyleGAN2, BigGAN, and ProGAN, serving as a foundational benchmark. Corvi et al. [7] extended this dataset by including diffusion-based images and a wide range of real images from datasets such as COCO and SUN, offering greater diversity and realism. Similarly, the CIFAKE dataset [8] provides a well-balanced mix of authentic and AI-generated samples from contemporary generative models, enabling robust evaluation of model generalization. Collectively, these datasets support comprehensive studies of detection performance across various generative paradigms.

3. Approach

3.1. Dataset Preparation

We trained models under two distinct configurations, corresponding to the **constrained** and **open** runs defined by the MediaEval 2025 task.

For the **constrained runs** (Vision Transformer and EfficientNet), models were trained exclusively on the official training set provided by the task organizers.

For the **open runs** (ResNet and VGG19), models were trained exclusively on the CIFAKE dataset [8], a publicly available collection of real and AI-generated images from modern generative models.

All images were resized to 224×224 pixels to match model input requirements. Data augmentation—including rotation, flipping, shifting, and zooming—was applied to enhance model robustness. Normalization was performed using ImageNet mean and standard deviation values.

3.2. Model Architectures

Four deep learning models were evaluated:

Vision Transformer (ViT-B/16): Implemented using the `timm` library, this model divides each image into 16×16 patches, encodes them into embeddings, and processes them through 12 transformer encoder layers. The pretrained ImageNet-21k weights were fine-tuned for binary classification.

EfficientNet-B0: A lightweight CNN architecture designed through compound scaling of model dimensions. The network was initialized with ImageNet weights, and its classification head was replaced with a sigmoid-activated binary output layer.

ResNet50: A deep residual network that incorporates skip connections to maintain gradient stability during training. The convolutional base was initially frozen and then fine-tuned on upper layers to better adapt to the target dataset.

VGG19: A traditional CNN architecture composed of sequential convolutional blocks. The top layers were unfrozen for fine-tuning, with dropout and augmentation techniques applied to improve robustness.

3.3. Training Configuration

All networks were trained using the Adam optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss. Training was conducted for up to 30 epochs with early stopping based on validation performance. A batch size of 32 was used, and the best-performing models were selected via checkpoint callbacks.

4. Results and Discussion

Table 1 summarizes the experimental outcomes for both constrained and open runs. The Vision Transformer achieved the highest F1-score (0.86) under the constrained setting, while EfficientNet-B0 also showed strong results. In the open configuration, CNN-based models such as ResNet and VGG19 demonstrated higher recall but lower precision, reflecting a trade-off between sensitivity and specificity.

Table 1
Performance comparison of evaluated architectures for real vs. synthetic image detection. Constrained runs used only the official training set; open runs included additional data from CIFAKE.

Model	Run Type	Accuracy	Precision	Recall	F1
ResNet50	Open	0.497	0.498	0.975	0.660
VGG19 (fine-tuned)	Open	0.489	0.494	0.931	0.646
EfficientNet-B0	Constrained	0.863	0.910	0.807	0.856
Vision Transformer (ViT-B/16)	Constrained	**0.862**	**0.879**	**0.839**	**0.859**

The constrained models—particularly ViT—showed robust generalization despite limited training data, demonstrating the power of self-attention in learning global dependencies. Open runs benefited from additional CIFAKE samples, achieving higher recall but at the cost of precision. These results align with existing literature emphasizing that transformers capture structural integrity, while CNNs are more responsive to localized artifacts [16, 17, 18].

5. Conclusion

This work presented a comparative evaluation of deep learning methods for detecting AI-generated images within the MediaEval 2025 Synthetic Image Detection task [1]. The Vision Transformer (ViT-B/16) achieved the best overall results under the constrained setting, confirming the benefit of self-attention mechanisms in identifying global visual inconsistencies. EfficientNet-B0 also performed effectively with fewer parameters. In contrast, open-run CNN models benefited from external data (CIFAKE), improving recall but lowering precision. Future work may explore hybrid transformer-CNN architectures and domain adaptation strategies to further enhance robustness across generative models.

Acknowledgments

The authors thank the Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, for providing computational support and resources essential for this research.

References

[1] O. Papadopoulou, M. Schinas, R. Corvi, D. Karageorgiou, C. Koutlis, F. Guillaro, E. Gavves, H. Mareen, L. Verdoliva, and S. Papadopoulos, "Synthetic Images at MediaEval 2025: Advancing Detection of Generative AI in Real-World Online Images," *Proc. of the MediaEval 2025 Workshop*, Dublin, Ireland and Online, 25–26 October 2025.

[2] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

[3] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *ICML*, 2019.

[4] K. He et al., "Deep Residual Learning for Image Recognition," *CVPR*, 2016.

[5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556*, 2014.

[6] S.-Y. Wang et al., "Detecting GAN-generated Imagery Using Color Cues," *arXiv:2002.03745*, 2020.

[7] R. Corvi et al., "Detection of Synthetic Images Generated by Diffusion Models," *CVPRW*, 2022.

[8] Birdy654, "CIFAKE: Real and AI-Generated Synthetic Images Dataset," *Kaggle*, 2023. Available: https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images.

[9] H. Farid, "Digital Image Forensics: Detecting Visual Manipulations," *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2022.

[10] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *NeurIPS*, 2021.

[11] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[12] L. Guarnera, O. Giudice, and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," *CVPRW*, 2020.

[13] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-Generated Fake Images over Social Networks," *MIPR*, 2018.

[14] Y. Liu, X. Zhang, S. Zhang, and C. Xu, "Global and Local Attention for Deepfake Image Detection," *Pattern Recognition Letters*, 2022.

[15] J. Xu, Y. Li, and S. Lyu, "DeepFake Detection with Dual Attention and Frequency Consistency," *IEEE Transactions on Image Processing*, 2023.

[16] F. Guillaro et al., "A Bias-Free Training Paradigm for More General AI-generated Image Detection," *arXiv preprint arXiv:2412.17671*, 2024.

[17] Q. Bammey, "Synthbuster: Towards Detection of Diffusion Model Generated Images," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 1–9, 2023.

[18] H. Qi, P. Zhang, and L. Ke, “DeepFake Detection Using Spatio-Temporal Convolutional Networks,” *ICASSP*, 2020.