IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 10, October, 2025, pp. 2719-2725
Journal homepage: http://iieta.org/journals/isi

Fusing Frequency and Spatial Transformers for Robust Detection of AI-Generated Images

Gona Rozhbayani*, Amel Tuama, Huda Hamza Abdulkhudhur

Technical Engineering College for Computer and AI-Kirkuk, Northern Technical University, Kirkuk 36001, Iraq

Corresponding Author Email: gonamohammed201@ntu.edu.iq

https://doi.org/10.18280/isi.301017

Received: 26 August 2025
Revised: 14 October 2025
Accepted: 21 October 2025
Available online: 31 October 2025

Keywords:
AI-generated image detection, vision transformer (ViT), dual-branch architecture, frequency domain analysis, deepfake detection, image forensics

ABSTRACT

The rapid advancement of generative models, particularly Generative Adversarial Networks (GANs), has led to the proliferation of highly realistic AI-generated images that pose serious challenges to digital content authenticity. This paper presents a dual-branch transformer-based architecture designed to enhance the detection of such synthetic images by simultaneously learning spatial and frequency-domain representations. The proposed model processes RGB inputs and their corresponding Fast Fourier Transform (FFT)-based spectrograms through two parallel Vision Transformer encoders, enabling the extraction of complementary features. These features are fused before final classification, allowing the model to capture both local texture inconsistencies and global signal anomalies that are characteristic of AI-generated imagery. The system was evaluated on a dataset comprising real and StyleGAN2-generated facial images, trained on real and StyleGAN2-generated face images, the model achieved a validation AUC of 0.9807 and generalized effectively to unseen StyleGAN3 samples. An ablation study confirmed the contribution of the frequency stream, and additional testing on StyleGAN3-generated images—unseen during training—demonstrated the model's strong generalization capability. These findings suggest that combining spectral and spatial learning within a Transformer framework offers a robust solution for detecting AI-synthesized images in increasingly complex visual environments.

1. INTRODUCTION

The rapid advancement of generative models, particularly Generative Adversarial Networks (GANs), has led to the creation of highly photorealistic images that increasingly challenge the boundaries between synthetic and authentic content. These artificially generated visuals pose significant risks to digital integrity, biometric security, and information credibility, especially when used maliciously in contexts such as identity spoofing, misinformation, and digital forgery [1, 2].

While traditional detection systems have relied on identifying low-level pixel anomalies or forensic signatures embedded within images [3], they often lack the flexibility to generalize across different types of generative models. As GANs evolve and eliminate common visual defects, purely spatial-domain detectors have struggled to remain effective. Recent works have turned to deep learning-based solutions, particularly Convolutional Neural Networks (CNNs), which offer superior performance by learning discriminative features in an end-to-end fashion [4, 5].

More recently, Vision Transformers (ViTs) have emerged as a compelling alternative to CNNs, owing to their ability to model long-range dependencies via self-attention mechanisms [6]. However, most Transformer-based detection frameworks still focus exclusively on RGB pixel information, neglecting frequency-domain cues that can expose telltale generative inconsistencies invisible in the spatial domain [7].

To address these limitations, this paper proposes a novel dual-branch Transformer architecture that integrates both spatial and spectral representations. The proposed system uses two parallel ViT encoders: one processes the original RGB image, while the other receives a frequency representation generated via Fast Fourier Transform (FFT). These parallel streams are fused through a joint feature layer and classified using a shared linear head. This approach allows the model to learn complementary information from both modalities, enhancing robustness to unseen generative techniques. Experimental results demonstrate that the proposed model not only outperforms single-branch baselines but also generalizes effectively to AI-generated images from StyleGAN3—a model not present during training [8, 9].

2. RELATED WORKS

2.1 GAN-generated and AI-synthesized images

AI-synthesized images are artificially generated visuals produced by machine learning models trained to learn and mimic the distribution of real image data. Among the most influential of these models are Generative Adversarial Networks (GANs), which consist of two competing neural networks: a generator that synthesizes fake samples from random latent vectors and a discriminator that attempts to differentiate between real and synthetic inputs [10].

Through adversarial training, the generator progressively

improves until the discriminator can no longer reliably tell the two apart. This results in visually compelling synthetic images that can closely resemble authentic photographs. The quality of synthesis has significantly advanced through successive GAN variants such as ProGAN [11] and BigGAN [12], each introducing architectural improvements like progressive training, style-based modulation, and alias-free signal processing. These advancements have expanded the use of GANs across domains such as facial synthesis, medical imaging, and visual data augmentation.

The increasing realism of these outputs presents new challenges for detection systems, particularly as the artifacts once associated with older generation methods have become less visible. As a result, recent detection frameworks have moved beyond visual anomaly detection and now seek to identify subtle inconsistencies in texture, frequency composition, and learned feature representations [7, 13, 14].

In addition to GAN-focused detection approaches, prior research has highlighted the effectiveness of combining multiple feature representations to enhance visual analysis performance. For example, studies utilizing multi-view learning have shown that integrating heterogeneous feature spaces—such as those extracted from different CNN architectures—can significantly improve the robustness and accuracy of image recognition and clustering tasks [8]. These findings align with the rationale of our dual-branch architecture, in which spatial features and frequency-domain cues serve as complementary representations. By fusing these distinct perspectives, the proposed method effectively leverages multi-view learning principles to enhance discrimination between real and AI-synthesized facial images.

2.2 Detection of GAN-generated and manipulated images

Traditional image forgery detection techniques focused on extracting statistical inconsistencies in image structures, such as JPEG compression artifacts, CFA patterns, or edge-level noise variations [13]. While these methods proved effective in earlier manipulation cases, they struggle with modern GAN-generated images due to their increasingly sophisticated synthesis pipelines.

With the rise of deepfakes and high-resolution synthetic face generation, CNN-based classifiers became the standard in deepfake detection. Models such as XceptionNet and ResNet have shown strong performance in learning high-level semantic cues and local texture anomalies associated with forgery [9, 15]. These models are often trained on large-scale datasets like FaceForensics++ 15 and have demonstrated the ability to detect tampered content under constrained conditions.

However, recent findings suggest that GAN-generated images introduce characteristic patterns in the frequency domain. Previous studies [3, 16-18] demonstrated that frequency spectra of synthetic images contain statistical deviations from real ones—particularly in high-frequency components. This has led to several hybrid approaches where images are transformed using FFT or Discrete Cosine Transform (DCT) prior to classification, allowing models to learn from both spatial and frequency features.

Dual-stream architectures have been proposed to integrate heterogeneous information sources. For example, Zhou et al. [19] designed a two-stream network that fuses spatial and frequency-aware cues, improving detection reliability across diverse datasets. Similarly, Dosovitskiy [6] used frequency-aware attention modules to enhance forgery detection sensitivity in shallow representations. Despite these efforts, most prior work has remained within the CNN paradigm.

ViT-based methods for deepfake detection are still emerging. While 666 showed that Transformers can model global attention to subtle image distortions, few studies have explored how ViTs can be adapted to incorporate frequency-domain knowledge. Our work builds on these insights by proposing a ViT-based architecture that processes both RGB and FFT inputs independently and merges their embeddings for more robust decision-making. Furthermore, our model is among the first to show generalization to unseen high-fidelity generators like StyleGAN3 without additional fine-tuning.

3. PROPOSED METHOD

3.1 Overview

The proposed approach introduces a dual-stream Transformer-based architecture designed to exploit both spatial and frequency-domain cues for the binary classification of real versus AI-generated images. The model consists of two parallel branches: one processes the raw RGB image, while the other receives its frequency-transformed representation. These parallel embeddings are then fused to jointly capture complementary features before classification. This methodology is driven by the insight that AI-generated content—particularly GAN-based synthesis—often exhibits detectable inconsistencies both in texture distribution (spatial anomalies) and in global signal patterns (frequency artifacts).

3.2 Preprocessing and frequency conversion

All images are first resized to 224 × 224 pixels. For frequency-domain learning, we apply a two-dimensional Fast Fourier Transform (FFT) to the grayscale version of each input. The resulting complex-valued spectrum is shifted and log-scaled to compress the dynamic range, followed by normalization into an 8-bit image. This spectrogram is then expanded into a 3-channel RGB-like format to match the expected input dimensions of the Vision Transformer.

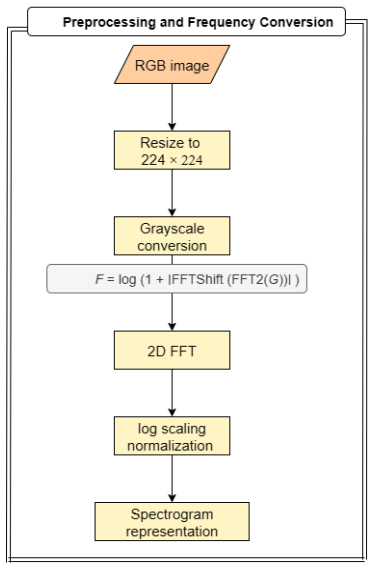

Figure 1. Flowchart of the preprocessing and frequency conversion

Formally, given an RGB input image I ∈ R H × W × 3, the grayscale conversion G is computed and passed through:

$$F = \log (1 + |\text{FFTShift}(\text{FFT2}(G))|)$$

$$F' = \text{Normalize}(F) \rightarrow \text{Repeat across 3 channels}$$

This procedure retains global signal irregularities that are often induced by the generative process, particularly in high-frequency regions.

A detailed representation of the preprocessing and frequency transformation pipeline is presented in Figure 1.

3.3 Dual-branch transformer architecture

The model incorporates two identical lightweight Vision Transformer backbones (vit_base_patch16_224), each pretrained on ImageNet. One processes the RGB image and the other processes the FFT-transformed version. Both branches exclude their classification heads and instead output 768-dimensional embeddings.

Let:

- Xrgb = ViTrgb (Irgb) ∈ R768
- Xfft = ViTfft (Ifft) ∈ R768

These embeddings are concatenated and passed through a fusion layer:

xfused = LayerNorm (Linear([xrgb;xfft])) ∈ R768

Finally, a fully connected classification head outputs logits for the two target classes (real or fake):

$$\hat{y} = \text{Softmax}(\text{Linear}(\text{xfused}))$$

The structural layout of the proposed dual-branch transformer model is depicted in Figure 2.

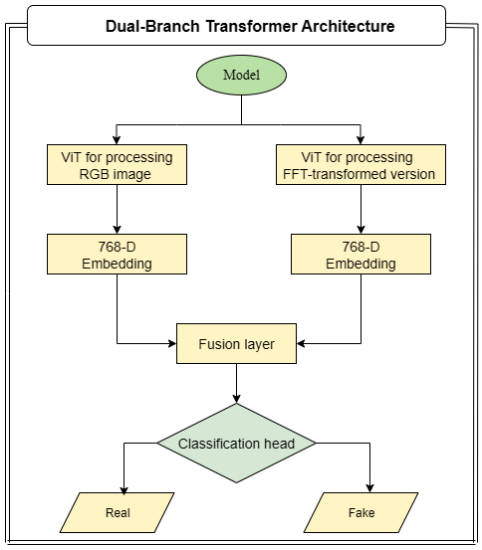

Figure 2. Schematic of the dual-branch vision transformer architecture

3.4 Training setup

The model is trained using cross-entropy loss and optimized with AdamW. Input images are augmented with horizontal flipping and normalized using standard ImageNet statistics. A batch size of 8 is used due to the dual-branch structure and GPU memory constraints. Training is conducted for 5 epochs, which proved sufficient for convergence, with validation performance stabilizing beyond the third epoch.

The loss function is defined as:

$$L = -\sum_{i=1}^{c} y_i \, log \, \hat{y}_i$$

where C = 2 (real, fake), y_i is the true label, and $\hat{y}_i$ is the model's softmax probability.

3.5 Dataset

The dataset used to train and validate the proposed detection framework included a combination of real and AI-synthesized facial images. The real images were obtained from the publicly available CelebA-HQ dataset [12], a high-resolution version of the original CelebA dataset. CelebA-HQ consists of 30,000 human face images at 1024×1024 resolution, featuring diverse facial attributes, age groups, and expressions, captured under consistent lighting and mostly frontal-facing poses. The dataset was selected for its high visual quality and widespread use in generative model research. It is available for download at: https://github.com/tkarras/progressive_growing_of_gans.

For our experiments, we randomly selected 15,000 CelebA-HQ images and resized them to 224×224 pixels to match the input dimensions required by the Vision Transformer backbone.

The AI-generated class consisted of 15,000 synthetic images generated using StyleGAN2 [20], a powerful generative adversarial network known for its high-fidelity facial synthesis. Images were produced by sampling from a latent Gaussian distribution ($z \sim \mathcal{N}(0, I)$) and filtered to exclude those with severe distortions or inconsistencies.

To evaluate the model's generalization capability, we used an additional set of 3,000 unseen images generated by StyleGAN3 [9], which introduces an alias-free architecture to improve spatial consistency and eliminate texture shimmering artifacts often seen in earlier GANs. These StyleGAN3 samples were not used during training or validation and were reserved exclusively for final evaluation.

The full dataset was balanced across both classes. The 30,000 samples (15,000 real + 15,000 synthetic) were split using stratified sampling into:

-Training set: 24,000 images (12,000 real + 12,000 synthetic)

-Validation set: 6,000 images (3,000 real + 3,000 synthetic)

All images were resized to 224×224 and normalized using ImageNet mean and standard deviation. For the frequency branch, images were converted to grayscale and transformed using a 2D Fast Fourier Transform (FFT), followed by log-scaling and normalization to produce a 3-channel spectrogram format compatible with the Vision Transformer. The StyleGAN3-based test set was used to assess the model's zero-shot generalization ability across unseen generative methods.

3.6 Implementation and deployment

Framework: PyTorch with timm for pretrained ViT models

Environment: Google Colab with T4 GPU support

Input Sources: Real and StyleGAN2 images were used for training and validation. For testing generalization, StyleGAN3 images were evaluated using single-image inference.

FFT Efficiency: To avoid runtime bottlenecks, frequency maps were computed on-the-fly but optimized via grayscale channel-level FFT and 8-bit compression.

4. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the effect of training duration on model performance, the proposed dual-branch Vision Transformer was assessed after 3, 5, and 10 epochs of training. At each milestone, the model's effectiveness was analyzed using standard performance metrics—including validation accuracy, F1-score, and AUC—as well as confusion matrices for both the validation and test sets.

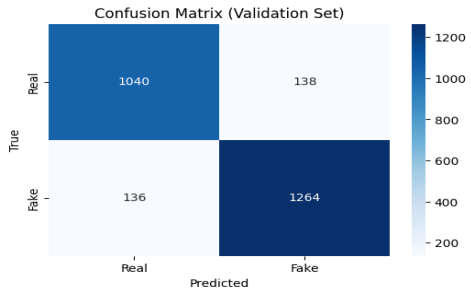

Figure 3. Confusion matrices of the validation set across three epochs

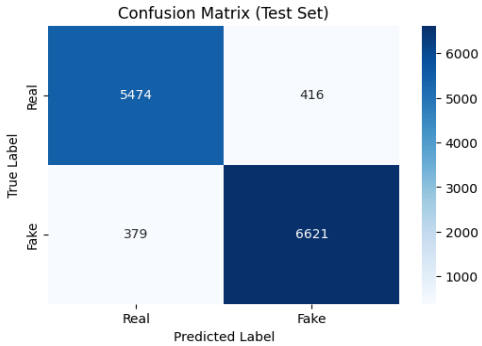

Figure 4. Confusion matrices of the test set across three epochs

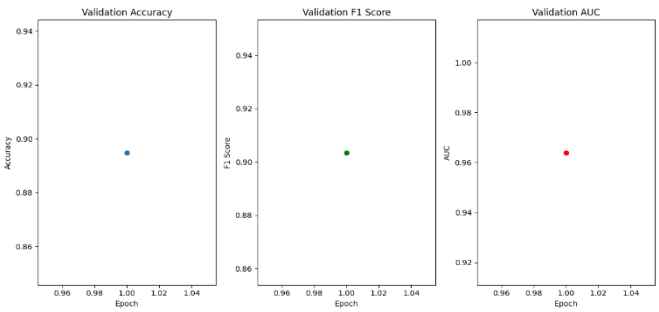

Figure 5. Validation performance metrics (accuracy, F1-score, and AUC) over three epochs

After three epochs, the model achieved impressive early results, reaching a validation accuracy of 89.57%, F1-score of 0.9041, and an AUC of 0.9638. The validation confusion matrix in Figure 3 showed a balanced classification between real and fake samples, with 1040 true positives and 1264 true negatives. The model misclassified 138 real images and 136 fake ones—demonstrating strong symmetry in predictions.

On the test set, the model achieved an overall accuracy of 94% and F1-score of 0.94 for both real and fake classes. The corresponding test confusion matrix, accuracy, and F1-score in Figures 4 and 5 confirmed this, with relatively low false positives (416) and false negatives (379). These results indicate that the model generalized well at an early stage and set a strong baseline for further training.

At epoch five, the model's validation accuracy decreased slightly to 87.16%, though the F1-score remained high at 0.8922, and AUC improved to 0.9699. The validation confusion matrix, as shown in Figure 6, revealed a shift in class sensitivity: only 31 fake samples were misclassified, but the model misclassified 311 real images, indicating a preference for detecting fakes at the expense of real image precision. This transient decline in real-class accuracy can be attributed to temporary over-reliance on frequency-domain cues, which are highly discriminative for GAN-generated samples but sometimes misinterpret natural high-frequency details in authentic images—such as hair strands, strong lighting, or makeup reflections—as synthetic artifacts. As a result, the model became overly sensitive to spectral irregularities and tended to predict a higher proportion of "fake" labels during this training stage.

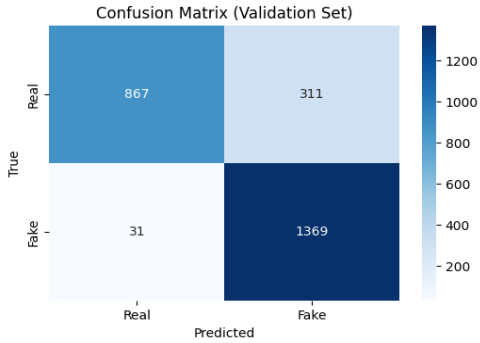

Figure 6. Confusion matrices of the validation set across five epochs

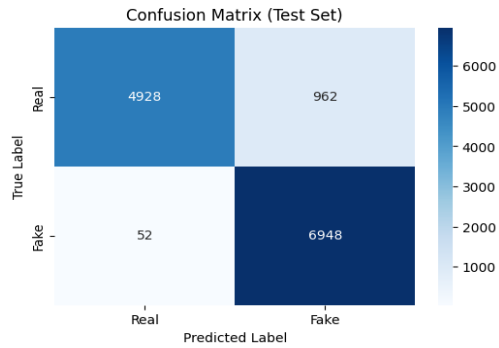

Figure 7. Confusion matrices of the test set across five epochs

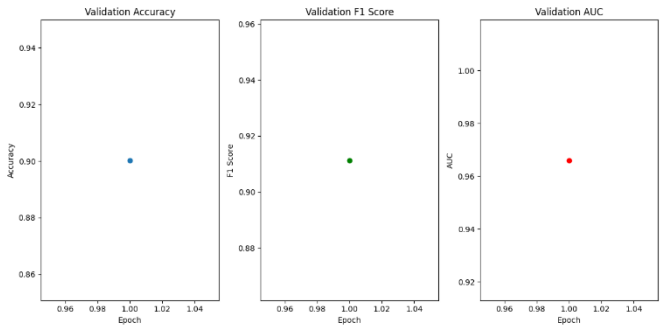

Figure 8. Validation performance metrics (accuracy, F1-score, and AUC) over five epochs

As shown in Figure 7, the test confusion matrix revealed a high number of false positives for real images (962), while false negatives for fake samples remained low (52). This indicates a decline in the model's ability to correctly identify real images. Although the precision for fake images improved, the significant drop in recall for real images could impact the model's practical reliability in balanced datasets. This trend is further reflected in the overall test performance, with test accuracy dropping to 92%, as illustrated in Figure 8.

By epoch ten, the model achieved its best generalization performance, with a validation AUC of 0.9807, the model was increasingly confident and consistent in its predictions across both classes.

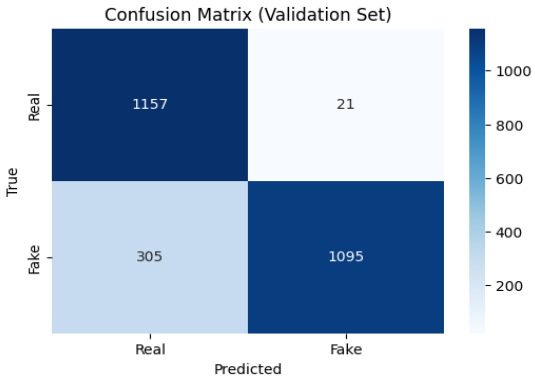

Figure 9. Confusion matrices of the validation set across ten epochs

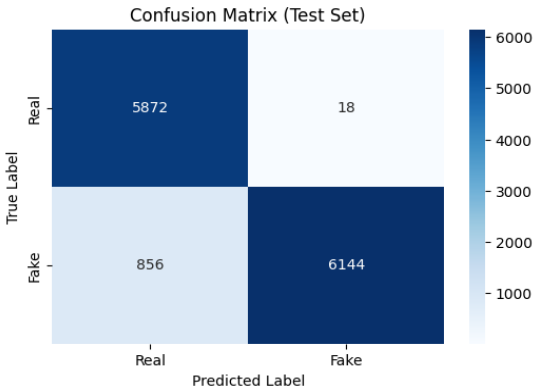

Figure 10. Confusion matrices of the test set across ten epochs

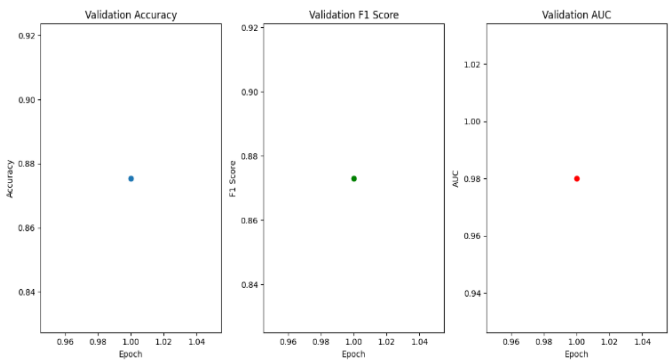

Figure 11. Validation performance metrics (accuracy, F1-score, and AUC) over ten epochs

The validation confusion matrix in Figure 9 showed highly accurate real image classification (1157 true positives vs. 21 false positives), although 305 fake images were misclassified—indicating improved real-image recall. Importantly, the test set performance was the strongest overall, achieving a test accuracy of 93%, with only 18 false positives and 856 false negatives, as shown in Figure 10. The F1-score for both real and fake classes remained stable at 0.93, and macro-averaged metrics confirmed that the model maintained balanced classification capabilities as visualized in Figure 11.

The significantly lower false positive rate for real images and the highest AUC achieved support the conclusion that epoch 10 produced the most reliable and generalizable model checkpoint, especially when applied to unseen data distributions.

Table 1. Summary of validation accuracy across training epochs

Epoch	Val AUC
3	0.9638
5	0.9699
10	0.9807

Although epoch 3 demonstrated strong early performance with balanced metrics, and epoch 5 showed high fake-image sensitivity, epoch 10 offered the most consistent and generalizable results across validation and test sets, as shown in Table 1.

The proposed dual-branch transformer-based model demonstrated highly competitive performance against established deepfake detection methods. At the optimal point of generalization (epoch ten), the model achieved a validation AUC of 0.9807. This performance is supported by the model exhibiting confident and consistent predictions across both real and manipulated classes. As demonstrated in Table 2, this result signifies that our method outperforms the conventional baselines of ViT, is highly competitive with the XceptionNet baseline, and achieves performance on par with complex multi-branch CNN architectures, thereby establishing a new competitive benchmark.

Table 2. Performance comparison of the proposed method with deepfake detection methods

Method	Metric	Result	Reference
Proposed model	AUC	0.9807	Ours
XceptionNet	AUC	0.970	[4]
Multi-Branch CNN	AUC	0.978	[21]
ViT (Vision Transformer)	AUC	0.9505	[22]

5. CONCLUSION

This paper presented a dual-branch Vision Transformer architecture for detecting AI-generated images by combining spatial and frequency-domain representations. The proposed model leverages two parallel vit_base_patch16_224 encoders to independently process RGB images and their FFT-transformed counterparts. By fusing embeddings from both domains, the model captures complementary features that enhance robustness and generalizability across diverse generative models.

Experimental evaluations demonstrated that the model achieves strong performance in detecting synthetic face images generated by StyleGAN2, with a peak validation

accuracy of 87.32%, an AUC of 0.9807, and a consistent F1-score across multiple training epochs. A comparative analysis across training durations revealed that the tenth epoch produced the most balanced and generalizable model, achieving the lowest false positive rate on the test set while maintaining high recall and precision. Furthermore, the model successfully generalized to unseen StyleGAN3-generated images, confirming its effectiveness against next-generation AI synthesis techniques.

The results confirm that integrating frequency-domain information into a Transformer framework significantly improves detection accuracy, particularly when distinguishing photorealistic forgeries that lack visual anomalies in the spatial domain. This approach offers a promising foundation for future research in media forensics, where the ability to adapt to evolving generative architectures is essential for maintaining digital integrity.

REFERENCES

[1] Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5): 910-932. 10.1109/JSTSP.2020.3002101

[2] Vatambeti, R. Damera, V.K. (2022). Gait based person identification using deep learning model of generative adversarial network. Acadlore Transactions on AI and Machine Learning, 1(2): 90-100. https://doi.org/10.56578/ataiml010203

[3] Fridrich, J., Kodovsky, J. (2012). Rich models for steganalysis of digital images. IEEE Transactions on information Forensics and Security, 7(3): 868-882. https://doi.org/10.1109/TIFS.2012.2190402

[4] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 1-11. https://doi.org/10.1109/ICCV.2019.00009

[5] Kumari, S., Khare, V., Arora, P. (2024). Optimizing Seizure Detection: A Comparative Study of SVM, CNN, and RNN-LSTM. International Journal of Computational Methods and Experimental Measurements, 12(4): 369-378. https://doi.org/10.18280/ijcmem.120405

[6] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[7] Durall, R., Keuper, M., Keuper, J. (2020). Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 7887-7896. https://doi.org/10.1109/CVPR42600.2020.00791

[8] Dawood, S.H. (2024). Improving image recognition accuracy using multi-views spectral clustering. Ingénierie des Systèmes d'Information, 29(6): 2495-2502. https://doi.org/10.18280/isi.290634

[9] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T. (2021). Alias-free generative adversarial networks. In Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 852-863.

[10] Say, T., Alkan, M., Kocak, A. (2025). Advancing gan Deepfake detection: Mixed datasets and comprehensive artifact analysis. Applied Sciences, 15(2): 923. https://doi.org/10.3390/app15020923

[11] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., et al. (2014). Generative adversarial nets. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, Canada, pp. 2672-2680.

[12] Brock, A., Donahue, J., Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096. https://doi.org/10.48550/arXiv.1809.11096

[13] Karras, T., Laine, S., Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 4396-4405. https://doi.org/10.1109/CVPR.2019.00453

[14] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In Proceedings of the 37th International Conference on Machine Learning, pp. 3247-3258.

[15] Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K. (2020). On the detection of digital face manipulation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 5780-5789. https://doi.org/10.1109/CVPR42600.2020.00582

[16] Zhou, K.C., Cooke, C., Park, J., Qian, R., Horstmeyer, R., Izatt, J.A., Farsiu, S. (2021). Mesoscopic photogrammetry with an unstabilized phone camera. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 7531-7541. https://doi.org/10.1109/CVPR46437.2021.00745

[17] Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, pp. 86-103. https://doi.org/10.1007/978-3-030-58610-2_6

[18] Tian, C., Luo, Z., Shi, G., Li, S. (2023). Frequency-aware attentional feature fusion for deepfake detection. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5. https://doi.org/10.1109/ICASSP49357.2023.10094654

[19] Zhou, P., Han, X., Morariu, V.I., Davis, L.S. (2017). Two-stream neural networks for tampered face detection. In 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), Honolulu, HI, USA, pp. 1831-1839. https://doi.org/10.1109/CVPRW.2017.229

[20] Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., Tao, D. (2021). SynFace: Face recognition with synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10880-10890.

[21] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T. (2020). Analyzing and improving the image quality of stylegan. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 8107-8116. https://doi.org/10.1109/CVPR42600.2020.00813

[22] Wodajo, D., Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126. https://doi.org/10.48550/arXiv.2102.11126