



Spatial-Frequency Fusion for Robust Detection of AI-Generated Images across Unseen Generators

Yateesh Kumar.V¹, Dr. Densy John Vadakkan^{2*}

¹Department of Computer Applications, CMR University, Bengaluru, Karnataka, India.

^{2*}School of Science and Computer Studies, CMR University, Bengaluru, Karnataka, India.

*Corresponding Author.

How to cite this paper:

Yateesh Kumar.V¹, Dr. Densy John Vadakkan², "Spatial-Frequency Fusion for Robust Detection of AI-Generated Images across Unseen Generators", IJIRE-V7I5-01-04.



Copyright © 2026
by author(s) and
Fifth Dimension
Research

Publication. This work is licensed under the
Creative Commons Attribution International
License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>

Abstract: The rapid development of generative artificial intelligence has made it increasingly difficult to distinguish AI-generated images from authentic photographs. Generative adversarial networks, diffusion models, and text-to-image systems can produce highly realistic visual content, creating challenges for digital forensics, misinformation control, and content authentication. Existing detectors frequently depend on generator-specific artifacts and may lose effectiveness when images originate from unseen generators or undergo resizing, compression, or blur. This study proposes a spatial-frequency feature-fusion framework for binary detection of real and AI-generated images. A convolutional neural network extracts spatial and structural representations from RGB images, while a second branch analyzes frequency-domain characteristics obtained through a two-dimensional Fourier transform. The resulting representations are fused before classification. Data augmentation is incorporated to improve robustness against common image transformations. The framework is designed for evaluation using the GenImage benchmark, which contains more than one million real/fake image pairs and explicitly supports cross-generator and degraded-image evaluation [10]. Performance is assessed using accuracy, precision, recall, F1-score, and AUROC. The experimental design emphasizes generalization rather than relying only on in-distribution accuracy.

Key Words: AI-generated images, image forensics, deep learning, convolutional neural networks, Fourier transform, frequency-domain analysis, generative AI, cross-generator detection.

I. INTRODUCTION

Generative artificial intelligence has significantly advanced the automatic synthesis of visual content. Modern generative adversarial networks and diffusion-based systems can produce images with a high degree of visual realism, reducing the reliability of human inspection and conventional forensic techniques. This development has practical implications for misinformation, identity manipulation, fraud, copyright disputes, and digital-media authentication. The need for automated detection has therefore increased as generated content becomes easier to create and distribute.

Earlier studies demonstrated that generated images may contain statistical fingerprints and architectural artifacts that can be learned by machine-learning detectors [1], [2], [5]. CNN-based approaches subsequently showed that detectors can learn generation artifacts directly from image data [6], while later work investigated the possibility of generalizing detection across generators [9]. However, the emergence of diffusion-based image synthesis has changed the detection landscape. Reconstruction-based approaches such as DIRE exploit differences between an input image and its reconstruction under a diffusion model [11], while other methods investigate reconstruction and inversion features for detecting unseen text-to-image generators [13], [14].

Large-scale benchmarks have also highlighted the difficulty of robust detection. GenImage provides more than one million real/generated image pairs and includes cross-generator and degraded-image tasks involving low-resolution, blur, and compression conditions [10]. More recent research continues to show that difficult generated images can challenge existing detectors. These observations motivate a detector that does not depend exclusively on visible spatial artifacts but combines complementary spatial and spectral information.

This work therefore proposes a two-branch spatial-frequency architecture. The spatial branch learns semantic and structural information from RGB images, whereas the frequency branch analyzes spectral patterns using a two-dimensional Fourier transform. The two representations are fused before binary classification. The central hypothesis is that complementary information from the spatial and frequency domains can provide a more robust representation for distinguishing real images from AI-generated images, particularly under cross-generator and degraded-image conditions.

II. RELATED WORK

AI-generated image detection research can be broadly grouped into artifact-based CNN approaches, fingerprint-based methods, reconstruction-based approaches, and foundation-model-based methods. Early fingerprint studies reported generator-specific traces in GAN-produced images [1], [2]. Co-occurrence statistics and deep learning were also used to

identify GAN-generated images [5]. CNN-based detectors subsequently demonstrated that generated images could be distinguished using learned visual artifacts [6].

Generalization has become a central research issue. Universal detector studies investigated whether representations learned from one set of generators can transfer to unseen generators [9]. CLIP-based representations have also been explored for cross-generator detection, providing a semantic feature space that can complement low-level forensic cues [9].

Diffusion-oriented approaches introduced reconstruction-based signals. DIRE uses diffusion reconstruction error as a detection representation [11], while DRCT uses reconstruction and contrastive learning to improve detection of diffusion-generated images [13]. FakeInversion uses Stable Diffusion inversion features to detect images generated by unseen text-to-image systems [14]. Recent work has increasingly combined semantic and low-level or frequency-related representations, reinforcing the value of complementary feature streams.

Frequency-domain analysis is particularly relevant because image synthesis pipelines can produce spectral inconsistencies that are not always obvious in RGB space. A recent review notes that Fourier, wavelet, and cosine transforms can expose periodic artifacts, noise distributions, and frequency-component differences associated with synthetic image generation. [15] this supports the use of a frequency branch alongside a conventional spatial CNN.

III. RESEARCH GAP

Despite substantial progress, several limitations remain. First, a detector trained on a limited set of generators may perform poorly on images produced by unseen systems. Second, continual changes in diffusion and text-to-image models can reduce the stability of generator-specific artifacts. Third, common transformations such as resizing, JPEG compression, and blur may weaken forensic evidence. GenImage explicitly identifies cross-generator and degraded-image evaluation as important tests of real-world robustness [10]. Fourth, a single-domain representation may not capture all discriminative information available in an image.

The research gap addressed in this study is therefore the need for a relatively simple and reproducible detector that combines spatial representations with frequency-domain evidence and evaluates performance under both generator shifts and image degradation.

IV. PROBLEM STATEMENT

The objective is to develop a deep learning system capable of distinguishing authentic images from AI-generated images while maintaining useful performance when the test images originate from generators not represented in training. The system should also remain reasonably robust when images are resized, blurred, or compressed.

V. OBJECTIVES

- Develop a binary classifier for real versus AI-generated images.
- Extract spatial and structural information using a convolutional backbone.
- Extract complementary spectral information using Fourier analysis.
- Fuse spatial and frequency representations before classification.
- Use data augmentation to improve robustness to common image transformations.
- Evaluate cross-generator generalization using a benchmark designed for this purpose.
- Compare the proposed fusion model with spatial-only and frequency-only baselines.

VI. PROPOSED METHODOLOGY

The proposed framework consists of five major stages: preprocessing, spatial feature extraction, frequency feature extraction, feature fusion, and classification. Figure 1 summarizes the complete processing pipeline.

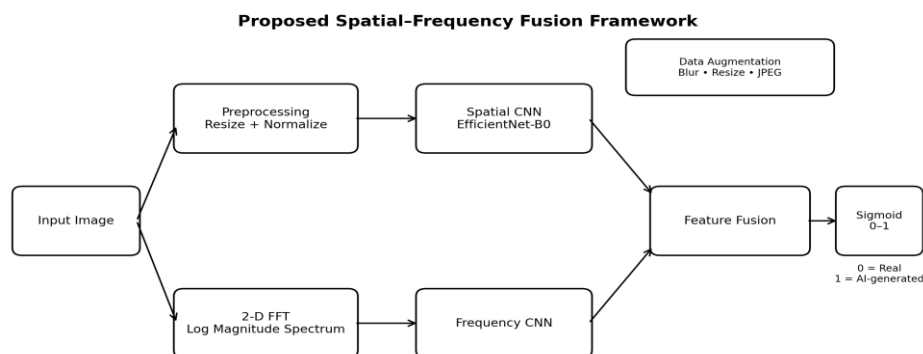


Fig. 1. Proposed spatial–frequency fusion framework for AI-generated image detection.

A. Preprocessing

Input images are resized to a fixed resolution and normalized before being passed to the two feature branches. During training, augmentation operations such as resizing, blur, and JPEG-compression simulation are introduced to expose the model to common forms of image degradation. The purpose is to reduce over-reliance on a narrow set of pixel-level artifacts.

B. Spatial Feature Extraction

The spatial branch processes the RGB image using an EfficientNet-B0 or comparable convolutional backbone. The backbone learns hierarchical representations ranging from local edges and textures to higher-level structural patterns. These features provide information about object structure, texture consistency, and other spatial characteristics that may differ between natural and synthesized images.

C. Frequency Feature Extraction

The frequency branch first applies a two-dimensional Fast Fourier Transform (FFT) to the input image. The magnitude spectrum is converted to a logarithmic representation to improve the visibility of low- and high-frequency components. The resulting spectral representation is then processed by a dedicated CNN. Frequency-domain analysis provides a complementary view of the image because some synthesis artifacts are more readily represented as spectral irregularities than as directly visible RGB patterns.

D. Feature Fusion

Let z_s denote the spatial feature vector and z_f denote the frequency feature vector. The fused representation is obtained by concatenation, $z = [z_s ; z_f]$, followed by fully connected layers. This late-fusion strategy allows the two branches to learn domain-specific representations independently before combining their complementary evidence.

E. Classification

The fused representation is passed to a fully connected classification layer followed by a sigmoid activation. The output $p \in [0,1]$ represents the estimated probability that an input image belongs to the AI-generated class. A threshold, typically 0.5, can be used to convert the probability into a binary decision.

VII. SYSTEM ARCHITECTURE

The complete architecture begins with a common input image. One path performs preprocessing and spatial feature extraction using a CNN, while the second path transforms the image into the frequency domain and extracts spectral features. The two feature vectors are concatenated and passed to a classification head. This design is intended to reduce dependence on a single type of forensic clue.

VIII. DATASET AND EXPERIMENTAL SETUP

GenImage is selected as the principal benchmark because it contains more than one million pairs of AI-generated and real images and includes images produced by multiple modern generators. The benchmark also defines cross-generator classification and degraded-image classification tasks, making it suitable for evaluating generalization and robustness rather than only standard in-distribution accuracy [10]. □ cite □ turn0search0 □ turn0search24 □

The dataset should be separated into training, validation, and test subsets while avoiding unnecessary overlap of generator conditions between training and evaluation. The experimental design should include the following comparisons: (1) RGB-only CNN, (2) frequency-only CNN, (3) the proposed spatial–frequency fusion model, (4) cross-generator testing, and (5) degraded-image testing.

IX. IMPLEMENTATION

The proposed implementation uses Python and PyTorch for model development. OpenCV and PIL can be used for image preprocessing, NumPy for Fourier transformations, and scikit-learn for metric computation. Binary cross-entropy is used as the principal classification loss, while Adam can be used for optimization. Early stopping and model checkpointing are recommended to reduce overfitting and preserve the best validation model.

For reproducibility, the final experimental report should specify the input resolution, batch size, learning rate, number of epochs, optimizer parameters, augmentation probabilities, train/validation/test counts, hardware configuration, random seed, and exact generator partitions. These details are essential before reporting numerical performance.

X. PERFORMANCE EVALUATION

Performance is evaluated using accuracy, precision, recall, F1-score, and AUROC. Accuracy is defined as $(TP+TN)/(TP+TN+FP+FN)$, precision as $TP/(TP+FP)$, recall as $TP/(TP+FN)$, and F1-score as $2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$. AUROC is used to assess ranking performance across classification thresholds. The evaluation should report results separately for in-distribution, cross-generator, and degraded-image conditions. Confidence intervals or repeated-run statistics should be included where feasible so that improvements are not attributed to random variation.

XI. RESULTS AND DISCUSSION

The supplied manuscript does not contain completed experimental measurements; therefore, numerical accuracy, precision, recall, F1-score, and AUROC values should not be invented. The final results section should be populated after the implementation is executed on the selected GenImage splits.

The primary comparison should be between the RGB-only baseline, frequency-only baseline, and proposed spatial–frequency fusion model. The most important evidence should come from cross-generator and degraded-image tests because

these settings directly address the generalization limitations identified in prior research. GenImage was designed specifically to examine these conditions [10].

The expected analytical question is whether combining spatial and spectral information improves robustness relative to either branch alone. If the fusion model improves cross-generator performance while retaining competitive in-distribution accuracy, this would support the hypothesis that complementary representations reduce dependence on generator-specific artifacts. If performance declines under compression or blur, the analysis should identify which transformations cause the largest degradation and whether augmentation mitigates that effect.

XII. LIMITATIONS

The proposed detector may remain sensitive to previously unseen generation mechanisms, adversarial perturbations, severe compression, and changes in the distribution of natural images. A binary detector also cannot, by itself, reliably identify which generator produced an image. Furthermore, performance on a benchmark does not guarantee effectiveness on future commercial or open-source generators. The framework should therefore be regarded as a detection model rather than a definitive proof of image provenance.

XIII. CONCLUSION

This study presents a spatial–frequency feature-fusion framework for detecting AI-generated images. The approach combines spatial representations learned from RGB images with complementary spectral representations obtained through Fourier analysis. The two feature streams are fused before binary classification, while data augmentation is used to improve robustness to common image transformations.

The principal contribution of the proposed design is its emphasis on complementary evidence and cross-generator evaluation rather than reliance on a single generator-specific artifact. The GenImage benchmark provides an appropriate basis for this evaluation because it contains large-scale real and generated image pairs and explicitly supports cross-generator and degraded-image testing [10].

Future work can investigate transformer-based backbones, adaptive frequency representations, explainable localization of forensic evidence, generator attribution, continual learning, and robustness against adversarial manipulation. These directions are important because image-generation technology continues to evolve and detection systems must generalize beyond the conditions used during training.

References

1. F. Marra, D. Gagnaniello, L. Verdoliva, and G. Poggi, “Do GANs leave artificial fingerprints?” MIPR, 2019.
2. N. Yu, L. S. Davis, and M. Fritz, “Attributing fake images to GANs: Learning and analyzing GAN fingerprints,” ICCV, 2019.
3. Y. Li and S. Lyu, “Exposing DeepFake videos by detecting face warping artifacts,” CVPRW, 2019.
4. A. Rössler et al., “FaceForensics++: Learning to detect manipulated facial images,” ICCV, 2019.
5. L. Nataraj et al., “Detecting GAN generated fake images using co-occurrence matrices,” 2019.
6. S.-Y. Wang et al., “CNN-generated images are surprisingly easy to spot...for now,” CVPR, 2020.
7. J. Frank and T. Holz, “[RE] CNN-generated images are surprisingly easy to spot...for now,” 2021.
8. Y. Ju et al., “Fusing global and local features for generalized AI-synthesized image detection,” ICIP, 2022.
9. U. Ojha, Y. Li, and Y. J. Lee, “Towards universal fake image detectors that generalize across generative models,” CVPR, 2023.
10. M. Zhu et al., “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image,” NeurIPS, 2023. DOI: 10.52202/075280-3398.
11. Z. Wang et al., “DIRE for diffusion-generated image detection,” ICCV, 2023.
12. J. Ricker, D. Lukovnikov, and A. Fischer, “AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error,” CVPR, 2024.
13. B. Chen et al., “DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images,” ICML, 2024.
14. G. Cazenavette et al., “FakeInversion: Learning to detect images from unseen text-to-image models by inverting Stable Diffusion,” CVPR, 2024.
15. S. Yan et al., “A sanity check for AI-generated image detection,” ICLR, 2025.
16. C. Tan et al., “C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection,” AAAI, 2025.
17. F. Marra et al., “Incremental learning for the detection and classification of GAN-generated images,” WIFS, 2019.
18. H. Lin et al., “Revisiting DIRE: Towards universal AI-generated image detection,” Neural Networks, 2026.