International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Comparative Analysis of Frequency-Domain and Spatial-Domain Deep Learning Methods for Detecting Diffusion-Generated Human Faces

Ziad Karim Tabbouche

Abstract

The rapid advancement of diffusion-based image generation models has significantly improved the realism of synthetic human faces, creating major challenges for digital media authentication and forensic analysis. Traditional deepfake detection approaches primarily rely on spatial-domain artifacts, but modern diffusion models generate highly realistic images that minimize visible inconsistencies. Consequently, frequency-domain analysis has emerged as an alternative strategy for identifying subtle spectral anomalies left during the image synthesis process. This paper presents a comparative analysis of frequency-domain and spatial-domain deep learning methods for detecting diffusion-generated human faces. The study evaluates representative convolutional neural networks, transformer-based architectures, and hybrid dual-domain models across publicly available datasets. Experimental findings demonstrate that spatial-domain methods achieve strong performance on known datasets but suffer from reduced generalization when confronted with unseen diffusion models. Frequency-domain approaches exhibit superior robustness against compression and cross-dataset variations due to their ability to capture hidden spectral artifacts. Furthermore, hybrid models integrating both domains outperform single-domain approaches by leveraging complementary discriminative features. The paper also discusses challenges related to dataset bias, adversarial robustness, and model interpretability. The results suggest that future detection systems should adopt collaborative spatial-frequency learning frameworks to improve reliability against increasingly sophisticated generative models.

Keywords: Diffusion Models, Deepfake Detection, Frequency Domain Analysis, Spatial Domain Learning, CNN, Vision Transformers, Synthetic Face Detection, AI-Generated Images, Digital Forensics, Multimedia Security

1. Introduction

Artificial intelligence has transformed image synthesis through the development of deep generative models such as Generative Adversarial Networks (GANs) and diffusion models. Among these technologies, diffusion-based models have recently gained widespread attention because of their ability to generate highly realistic and semantically consistent human faces. Platforms based on Stable Diffusion, DALL·E, Midjourney, and other latent diffusion architectures can now produce synthetic faces that are visually indistinguishable from authentic photographs.

While these technologies offer significant benefits in entertainment, media production, and creative design, they also introduce serious risks associated with misinformation, identity fraud, and digital manipulation. Synthetic human faces generated by diffusion models can be misused in fake social media profiles, impersonation attacks, and deceptive multimedia content. As a result, reliable methods for

distinguishing authentic images from AI-generated images have become increasingly important in digital forensics and cybersecurity.

Existing deepfake detection techniques can generally be categorized into spatial-domain and frequency-domain approaches. Spatial-domain methods analyze visual inconsistencies directly from image pixels, including texture artifacts, facial inconsistencies, and blending anomalies. Deep learning architectures such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are commonly used in this category. However, the quality of modern diffusion-generated images has substantially reduced visible artifacts, limiting the effectiveness of purely spatial analysis.

Frequency-domain methods attempt to overcome these limitations by transforming images into spectral representations using Fourier Transform, Discrete Cosine Transform (DCT), or wavelet analysis. These approaches capture hidden periodic patterns and high-frequency inconsistencies introduced during the image generation process. Recent studies indicate that frequency-domain features may provide better generalization across unseen generative models.

This paper presents a comparative analysis of spatial-domain and frequency-domain deep learning methods for detecting diffusion-generated human faces. The primary objective is to evaluate the strengths, limitations, and practical applicability of each approach under different experimental conditions.

2. Literature Review

Research on deepfake detection has evolved rapidly alongside advances in image generation technologies. Early studies focused primarily on GAN-generated images, while recent investigations increasingly target diffusion-generated content.

Lu and Ebrahimi proposed a benchmark for detecting AI-synthesized human face images and demonstrated that frequency-domain representations improve cross-model generalization for unseen generators.

Liu et al. introduced the Spatial-Phase Shallow Learning (SPSL) framework, which combines spatial information with phase spectrum analysis to capture up-sampling artifacts in forged images. Their study highlighted the importance of phase information in detecting synthetic content.

Suo et al. developed a Dual Domain Attention Mechanism (DDAM) that explicitly models interactions between spatial and frequency representations. Their experiments demonstrated that dual-domain learning improves forgery detection performance across benchmark datasets.

Recent work by Luo and Wang proposed the Frequency-Domain Masking and Spatial Interaction (FMSI) model, which integrates masked image modeling with frequency-domain analysis. Their results showed improved robustness and generalization under compressed and low-quality conditions.

Another significant contribution is the Multi-scale Spatial-Frequency Variance-sensing (MSFV) model, which combines spatial-frequency interactions through bidirectional feature exchange mechanisms. The method achieved state-of-the-art performance on several deepfake benchmarks.

Despite these advances, comparative studies specifically focusing on diffusion-generated human faces remain limited. Most existing works either focus on GAN-generated content or evaluate only a single detection paradigm. Therefore, a comprehensive comparative analysis between spatial-domain and frequency-domain methods is necessary.

3. Methodology

3.1 Dataset Collection

The study considers publicly available datasets containing both authentic and AI-generated human faces. The datasets include images generated using diffusion-based architectures such as Stable Diffusion and DDPM variants. Real face images are collected from benchmark face datasets to ensure balanced evaluation.

The datasets are divided into training, validation, and testing subsets using an 80:10:10 ratio.

3.2 Spatial-Domain Detection Methods

Spatial-domain approaches analyze images directly in the pixel domain. The following models are evaluated:

- ResNet-50
- EfficientNet-B4
- Vision Transformer (ViT)
- Swin Transformer

These models are trained using RGB face images resized to a fixed resolution. Data augmentation techniques including horizontal flipping, rotation, and brightness adjustment are applied to improve generalization.

3.3 Frequency-Domain Detection Methods

Frequency-domain methods transform images into spectral representations before feature extraction. The following transformations are used:

- Fast Fourier Transform (FFT)
- Discrete Cosine Transform (DCT)
- Wavelet Transform

CNN-based architectures are then trained on the transformed representations to learn discriminative frequency patterns associated with synthetic images.

The Fourier Transform equation used in the analysis is:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=}$$

0
N
−
1
f
(
x
,
y
)
e
−
j
2
π
(
u
x
M
+
v
y
N
)
F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}
F(u,v)=∑x=0M−1 ∑y=0N−1 f(x,y)e−j2π(Mux +Nvy)
where
f
(
x
,
y
)
f(x,y)
f(x,y) represents the spatial image and
F
(
u
,
v
)
F(u,v)
F(u,v) denotes its frequency representation.

3.4 Hybrid Spatial-Frequency Models

Hybrid models combine both spatial and spectral information using feature fusion mechanisms. The architectures evaluated include:

- Dual-stream CNN frameworks
- Attention-based fusion networks
- Spatial-frequency transformer models

Feature maps from both domains are concatenated before classification.

3.5 Evaluation Metrics

Performance is measured using:

- Accuracy
- Precision
- Recall
- F1-Score
- Area Under Curve (AUC)
- Equal Error Rate (EER)

Cross-dataset testing is also performed to assess generalization capability.

4. Experimental Results

4.1 Spatial-Domain Performance

Spatial-domain models achieved high accuracy on datasets containing known diffusion generators. Vision Transformers demonstrated superior feature learning capability due to their global attention mechanisms. However, performance degraded significantly under heavy image compression and unseen generation models.

ResNet-50 achieved strong baseline performance but showed sensitivity to low-quality synthetic images. Transformer-based models improved robustness but required higher computational resources.

4.2 Frequency-Domain Performance

Frequency-domain methods demonstrated better robustness against compression and adversarial perturbations. FFT-based models successfully identified hidden spectral inconsistencies that were not visible in the spatial domain.

DCT-based representations produced stable detection accuracy across multiple datasets. Frequency-domain detectors also generalized more effectively to unseen diffusion generators compared with purely spatial models.

The findings align with recent studies showing that diffusion models introduce characteristic spectral distributions detectable through frequency analysis.

4.3 Hybrid Model Performance

Hybrid spatial-frequency models consistently outperformed single-domain approaches. The integration of complementary features improved robustness and reduced false positives.

The FMSI and MSFV architectures demonstrated strong cross-dataset performance by combining local spatial textures with global spectral inconsistencies.

Experimental observations suggest that hybrid learning frameworks provide the most reliable strategy for future diffusion-generated face detection systems.

5. Discussion

The comparative analysis reveals that both spatial-domain and frequency-domain methods possess distinct strengths and limitations.

Spatial-domain approaches excel at capturing semantic inconsistencies and local texture artifacts. Their effectiveness, however, depends heavily on the visibility of forgery traces. As diffusion models continue improving visual realism, these artifacts become increasingly difficult to detect.

Frequency-domain approaches provide improved robustness because generative models often introduce subtle spectral irregularities during image synthesis. These hidden patterns remain detectable even when visual artifacts disappear. Nevertheless, frequency-based methods may lose sensitivity under extreme post-processing operations.

Hybrid approaches represent a promising direction because they leverage complementary information from both domains. By integrating spatial semantics with spectral analysis, hybrid frameworks improve generalization and robustness against unseen attacks.

Another important challenge involves adversarial robustness. Attackers may deliberately manipulate spectral signatures to evade frequency-domain detectors. Therefore, future research should investigate adversarially robust multimodal detection systems.

Model interpretability also remains an open issue. Many deep learning detectors operate as black-box systems, making forensic validation difficult in legal or security-sensitive environments.

6. Conclusion

This paper presented a comparative analysis of frequency-domain and spatial-domain deep learning methods for detecting diffusion-generated human faces. Experimental findings indicate that spatial-domain models achieve strong performance on familiar datasets but struggle with generalization to unseen diffusion generators. Frequency-domain methods provide improved robustness and cross-dataset transferability by capturing hidden spectral artifacts introduced during image synthesis.

Hybrid spatial-frequency frameworks achieved the best overall performance by combining complementary features from both domains. These results suggest that future detection systems should move toward collaborative multimodal architectures capable of adapting to evolving generative technologies.

As diffusion models continue advancing, the development of reliable, interpretable, and generalizable detection systems will remain essential for maintaining trust in digital media ecosystems.

References

1. Lu, Y., & Ebrahimi, T. “Towards the Detection of AI-Synthesized Human Face Images.” arXiv, 2024.
2. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., & Yu, N. “Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain.” arXiv, 2021.
3. Suo, Y., Zhao, X., Guo, Y., Li, Y., & Wang, Y. “A Dual Domain Attention Mechanism for Face Forgery Detection.” IEEE IJCB, 2023.
4. Luo, X., & Wang, Y. “Frequency-Domain Masking and Spatial Interaction for Generalizable Deepfake Detection.” Electronics, 2025.
5. Li, Z., Tang, W., Gao, S., Wang, Y., & Wang, S. “Multiple Contexts and Frequencies Aggregation Network for Deepfake Detection.” PLOS ONE, 2026.

6. Qiao, M., Tian, R., & Wang, Y. “Towards Generalizable Deepfake Detection with Spatial-Frequency Collaborative Learning and Hierarchical Cross-Modal Fusion.” arXiv, 2025.
7. Wang, X., Guo, H., Hu, S., Chang, M.-C., & Lyu, S. “GAN-generated Faces Detection: A Survey and New Perspectives.” arXiv, 2022.
8. Awrahman, B. J., Awrahman, Z. J., & Aziz, C. F. “Enhanced Detection of Diffusion Model-Generated Deepfakes Using CNN Feature Maps and Forgery Traces.” Journal of Al-Qadisiyah for Computer Science and Mathematics, 2024.
9. Sato, Y., Yamashita, T., Fujiyoshi, H., & Hirakawa, T. “Frequency-Domain Detection of AI-Generated Images via Radial Spectrum Templates.” 2026.
10. Wang, Z., Yu, Z., Zhao, C., Zhu, X., & others. “Deep Spatial Gradient and Temporal Depth Learning for Face Anti-Spoofing.” CVPR, 2020.