A Dual–Stream Spatial–Frequency Feature Learning Framework for Generalized Deepfake Detection

Notebook for the ImageCLEF Lab at CLEF 2026

Jayasree R*, P Mirunalini and K A Muthukumar

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India

Abstract

Deepfake generation techniques have rapidly evolved in recent years, producing highly realistic manipulated media that poses significant risks to digital trust, media authenticity, and information security. Detecting deepfakes generated by unseen methods remains a major challenge, as many existing approaches tend to learn generator-specific artifacts rather than generalized representations. In this work, we propose a dual-stream spatial–frequency feature learning framework for generalized deepfake detection. The proposed system simultaneously learns spatial representations from RGB images and spectral representations from Fast Fourier Transform (FFT)-based frequency images using two EfficientNet-B0 backbones. A consistency regularization strategy is introduced to encourage alignment between spatial and frequency-domain feature representations, improving robustness toward unseen manipulations. The model was trained using multiple manipulation categories from the FaceForensics++ dataset with extensive augmentation techniques to improve generalization. Experimental evaluation on the ImageCLEF 2026 Deepfake Detection task achieved a final score of 0.5562, demonstrating the framework's capability to detect both organizer-generated and participant-generated deepfakes under generalized evaluation settings.

Keywords

Deepfake Detection, Generalized Deepfake Detection, Spatial–Frequency Learning, FFT-based Feature Extraction, Dual-Stream Learning, Feature Consistency Learning, Image Forensics

1. Introduction

The rapid advancement of generative artificial intelligence has significantly improved the realism of synthetic media, leading to the widespread emergence of deepfakes. Modern deepfake generation methods based on Generative Adversarial Networks (GANs) and diffusion models are capable of producing highly convincing manipulated images with minimal visible artifacts. While these technologies provide benefits in multimedia generation and entertainment applications, they also introduce serious concerns related to misinformation, digital forgery, identity manipulation, and media authenticity. As the visual quality of generated content continues to improve, reliable deepfake detection has become an important and challenging research problem.

A major limitation of many existing deepfake detection approaches is poor generalization toward unseen manipulations. Several methods tend to learn generator-specific artifacts and therefore experience performance degradation when evaluated on unknown deepfake generation techniques. The ImageCLEF 2026 evaluation campaign [1] introduced the Deepfake Detection Task [2], focusing on generalized multimodal deepfake detection under unseen evaluation settings. The image deepfake detection subtask involves binary classification of images into real and manipulated categories without providing labeled training data, encouraging the development of robust and generalized detection strategies. Deepfake manipulations may introduce inconsistencies not only in the spatial appearance of images but also in their frequency-domain representations, motivating the use of complementary multi-domain feature learning for improved robustness against diverse deepfake generation techniques.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

jayasree2370002@ssn.edu.in (J. R); miruna@ssn.edu.in (P. Mirunalini); muthukumarka@ssn.edu.in (K. A. Muthukumar)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Related Work

Early deepfake detection methods primarily operated in the spatial domain using RGB images and deep convolutional neural network backbones to learn forgery-specific artifacts. Several approaches explored reconstruction-based objectives for face forgery detection [3], attention-guided feature refinement for long-distance temporal dependencies [4], and EfficientNet-based classifiers for robust real-and-fake media classification [5]. Although these methods achieved strong performance on known manipulations, many suffered from limited generalization and performance degradation when evaluated on unseen deepfake generators [6, 7].

To improve robustness against unseen manipulations, recent studies incorporated spectral representations alongside spatial information. Frequency-aware methods based on Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT) were introduced to capture hidden inconsistencies produced during image synthesis and upsampling operations [8, 9]. More recent dual-stream architectures jointly learn spatial and spectral representations using multi-domain feature extraction and cross-modal fusion strategies to capture complementary forgery patterns [10, 11]. Despite recent progress, many existing methods continue to rely heavily on generator-specific artifacts and often struggle to generalize to highly realistic unseen manipulations. Furthermore, explicit consistency learning between spatial and frequency-domain representations remains relatively underexplored in generalized deepfake detection settings. Motivated by these limitations, the proposed framework incorporates consistency regularization between spatial and frequency representations using dual EfficientNet-B0 backbones operating on RGB and FFT-based image representations.

3. System Overview

3.1. Dataset Overview

The proposed framework was developed using the FaceForensics++ dataset [12], which is one of the widely used benchmark datasets for deepfake detection research. The dataset contains both authentic and manipulated facial videos generated using multiple face manipulation techniques. For this work, frames were extracted from real videos and multiple manipulation categories, including Deepfakes, FaceSwap, NeuralTextures, Face2Face, FaceShifter, and DeepFakeDetection. The use of diverse manipulation methods enabled the model to learn generalized forgery characteristics instead of relying solely on generator-specific artifacts.

The evaluation was performed using the ImageCLEF 2026 Deepfake Detection dataset, which consists of organizer-generated deepfakes, participant-generated deepfakes, and authentic real images. The image deepfake detection subtask contains $24,404$ unlabeled test images that must be classified into real or deepfake categories. Since labeled training data was not provided as part of the challenge, the task emphasized generalization toward unseen manipulations and diverse deepfake generation techniques.

3.2. Data Preprocessing

The FaceForensics++ videos were converted into image frames for model training. Frames were extracted at evenly spaced intervals to ensure temporal diversity and reduce redundancy among consecutive frames. The extracted frames were resized to 224×224 resolution to match the input requirements of the backbone networks.

Two complementary representations were generated for each image. The first representation consisted of the original RGB image used for spatial feature learning. The second representation was obtained by applying Fast Fourier Transform (FFT) to the grayscale version of the image to capture spectral characteristics and frequency-domain inconsistencies introduced during image synthesis. The FFT magnitude spectrum was normalized and converted into a three-channel image for processing through the frequency branch of the network.

To improve robustness and generalization toward unseen manipulations, multiple augmentation techniques were applied during training, including random resized cropping, horizontal flipping, rotation, color jittering, and Gaussian blur. These augmentations help reduce overfitting to specific artifacts and improve the ability of the model to detect deepfakes under varying visual conditions.

3.3. Methodology

The proposed framework, as depicted in Fig. 1, follows a dual-stream architecture designed to learn complementary spatial and frequency-domain representations for generalized deepfake detection. Given an input image, two parallel feature extraction branches are employed. The first branch processes the RGB image to capture spatial characteristics such as texture inconsistencies, blending artifacts, and visual manipulation patterns. The second branch processes the frequency-domain representation generated using Fast Fourier Transform (FFT) to capture hidden spectral irregularities introduced during image synthesis and upsampling operations.

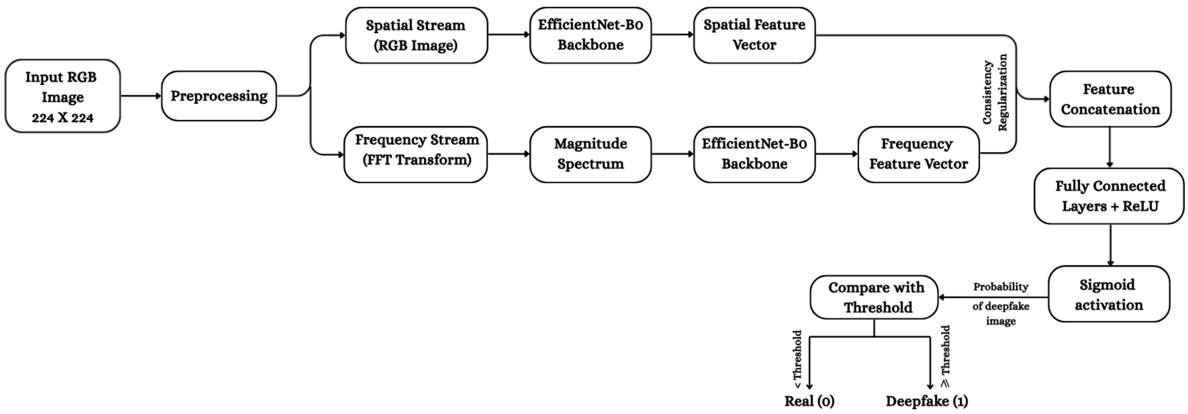

Figure 1: Proposed dual-stream spatial–frequency framework for generalized deepfake detection.

Both spatial and frequency representations are processed using EfficientNet-B0 backbones for feature extraction. The spatial branch receives the resized RGB image, while the frequency branch receives the normalized FFT magnitude spectrum converted into a three-channel image. The extracted feature representations from both branches are concatenated and passed through fully connected classification layers for binary prediction of real or deepfake images.

To improve robustness toward unseen manipulations, a consistency regularization strategy is incorporated between the spatial and frequency feature representations. This term encourages alignment between the feature representations learned by the two branches. Real images generally exhibit coherent spatial and spectral characteristics, whereas manipulated images may introduce discrepancies between visual appearance and underlying frequency patterns due to synthesis and upsampling operations. By minimizing the feature divergence between the spatial and frequency representations, the framework promotes complementary multi-domain learning and improves robustness toward unseen deepfake manipulations. The overall training objective combines binary cross-entropy classification loss with consistency regularization loss, defined as:

$$L_{\text{total}} = L_{\text{BCE}} + \lambda L_{\text{consistency}} \tag{1}$$

where L_{BCE} represents the binary cross-entropy loss for deepfake classification, $L_{\text{consistency}}$ denotes the feature consistency regularization term, and λ controls the contribution of the consistency constraint during training.

The framework was trained using the Adam optimizer with a learning rate of 1×10^{-4}, batch size of 16, and image resolution of 224×224. The training process was performed for five epochs using augmented image samples generated from multiple manipulation categories of the FaceForensics++ dataset.

During inference, the same preprocessing pipeline was applied to the test images, including image resizing and FFT generation. The trained model produced prediction probabilities using a sigmoid activation function, followed by threshold-based binary classification to identify real and deepfake images. Multiple decision thresholds were experimentally evaluated during validation-based calibration analysis, and a threshold value of 0.45 was selected for the final submission based on balanced prediction performance under generalized evaluation settings.

4. Results and Discussion

The proposed dual-stream spatial–frequency framework was evaluated on the unseen test dataset provided as part of the ImageCLEF 2026 Deepfake Detection challenge. Table 1 presents the official evaluation results obtained on the image deepfake detection subtask. The framework achieved a final weighted score of 0.5562 under generalized evaluation settings. Comparatively higher performance was observed on organizer-generated deepfakes and secret real ground-truth data, indicating the ability of the model to capture both spatial and spectral inconsistencies associated with manipulated content.

Table 1
Official ImageCLEF 2026 Image Deepfake Detection Results

Evaluation Metric	Score
Baseline Deepfakes (Images)	0.6454
Organizers Real Data (Images)	0.5309
Organizers Real GT Data (Images)	0.6287
Participant Deepfake Data (Images)	0.5557
Participant Deepfake Data Weighted (Images)	0.4677
Final Score (Images)	**0.5562**

The framework demonstrated stable prediction behavior without severe overfitting during training, suggesting reasonable generalization capability across multiple manipulation categories. The higher performance on baseline deepfakes compared to participant-generated manipulations further indicates that the framework is more effective at detecting traditional forgery patterns than highly sophisticated unseen generation methods. The moderate performance on organizer real data also highlights the difficulty of maintaining strong discrimination between authentic and manipulated content under generalized evaluation settings. Overall, the experimental results demonstrate the potential of combining spatial and frequency-domain feature learning for generalized deepfake detection while also highlighting the challenges associated with highly realistic in-the-wild manipulations.

5. Conclusion and Future Work

A dual-stream spatial–frequency framework for generalized deepfake detection was presented for the ImageCLEF 2026 Deepfake Detection Task. By jointly learning RGB and FFT-based representations with consistency regularization, the framework demonstrated the capability to capture complementary forgery characteristics under generalized evaluation settings involving unseen manipulations.

Despite promising results, generalized deepfake detection remains challenging for highly realistic manipulations generated using advanced synthesis pipelines. Future work may focus on incorporating explicit face localization and face-aware feature extraction to better capture subtle facial inconsistencies. Transformer-based architectures and self-supervised representation learning may further improve robustness toward unseen deepfake generators. In addition, temporal consistency modeling for video-level analysis and multi-scale spectral feature learning could enhance the detection of complex manipulations that suppress visible spatial and frequency-domain artifacts. Further investigation into calibration strategies and domain generalization techniques may also improve adaptability under real-world deepfake scenarios involving diverse image qualities, compression levels, and post-processing operations.

Acknowledgments

We extend our gratitude to the developers of the CLEF2026 team and the organizers of ImageCLEF for hosting the competition.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: check grammar and spelling, paraphrase, and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[2] D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, B. Ionescu, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, Overview of imageclef 2026 deepfake task: Multimodal detection and generation of deepfakes, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[3] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, X. Yang, End-to-end reconstruction-classification learning for face forgery detection, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4103–4112. doi:10.1109/CVPR52688.2022.00408.

[4] W. Lu, L. Liu, B. Zhang, J. Luo, X. Zhao, Y. Zhou, J. Huang, Detection of deepfake videos using long-distance attention, IEEE Transactions on Neural Networks and Learning Systems 35 (2024) 9366–9379. doi:10.1109/TNNLS.2022.3233063.

[5] E. Jain, D. Kundra, Efficientnet-based deepfake detection: A robust approach for real and fake media classification, in: 2024 Global Conference on Communications and Information Technologies (GCCIT), 2024, pp. 1–6. doi:10.1109/GCCIT63234.2024.10862025.

[6] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, L. Ma, Spatiotemporal inconsistency learning for deepfake video detection, 2021. URL: https://arxiv.org/abs/2109.01860. arXiv:2109.01860.

[7] K. Sun, S. Chen, T. Yao, H. Liu, X. Sun, S. Ding, R. Ji, Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion, 2024. URL: https://arxiv.org/abs/2410.04372. arXiv:2410.04372.

[8] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, 2020. URL: https://arxiv.org/abs/2007.09355. arXiv:2007.09355.

[9] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, Y. Wei, Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024. URL: https://arxiv.org/abs/2403.07240. arXiv:2403.07240.

[10] M. Qiao, R. Tian, Y. Wang, Towards generalizable deepfake detection with spatial-frequency collaborative learning and hierarchical cross-modal fusion, 2025. URL: https://arxiv.org/abs/2504.17223. arXiv:2504.17223.

[11] C. Peng, T. Chen, D. Liu, Y. Zheng, N. Wang, Spatial-frequency dual-stream reconstruction for deepfake detection, in: Pattern Recognition and Computer Vision: 7th Chinese Conference, PRCV 2024, Urumqi, China, October 18–20, 2024, Proceedings, Part XI, Springer-Verlag, Berlin, Heidelberg, 2024, p. 473–487. URL: https://doi.org/10.1007/978-981-97-8795-1_32. doi:10.1007/978-981-97-8795-1_32.

[12] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, 2019. URL: https://arxiv.org/abs/1901.08971. arXiv:1901.08971.