

CascadeNet: A Two-Stage Hybrid Learning Framework for Explainable Deepfake Forensics

Jatin Yadav

21bcs104@iiitdmj.ac.in

Indian Institute of Information Technology Design and Manufacturing Jabalpur

Divyanshu Sharma

Indian Institute of Information Technology Design and Manufacturing Jabalpur

Pritee Khanna

Indian Institute of Information Technology Design and Manufacturing Jabalpur

Neha Gour

Indian Institute of Information Technology Design and Manufacturing Jabalpur

Research Article

Keywords: Deepfake Detection, Image Manipulation, Deep Learning, EfficientNet, UNet, Semantic Segmentation, Facial Forgery, Media Forensics, Two-Stage Detection, FF++ Dataset, Facial Manipulation Localization, Image Authentication

Posted Date: April 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8823156/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

CascadeNet: A Two-Stage Hybrid Learning Framework for Explainable Deepfake Forensics

Jatin Yadav^{1*}, Divyanshu Sharma¹, Pritee Khanna¹, Neha Gour¹

^{1*}Department of Computer Science and Engineering, PDPM Indian Institute of Information Technology Design and Manufacturing Jabalpur, Jabalpur, Madhya Pradesh, 482005, India.

*Corresponding author(s). E-mail(s): 21bcs104@iiitdmj.ac.in;
Contributing authors: 21bcs079@iiitdmj.ac.in; pkhanna@iiitdmj.ac.in;
g.neha@iiitdmj.ac.in;

Abstract

This paper presents a novel two-stage deepfake detection framework that helps in finding the exact locations of the region of the image where a deepfake has happened. In the first stage, a binary classification approach is used to find out whether an image or video frame has undergone any manipulation or not. This classifier achieves a high accuracy of 99.9% in classifying whether the image is deepfake or not while maintaining computational efficiency with the help of transfer learning and data augmentation methods. After the image is successfully classified as a deepfake, the second stage uses a U-Net segmentation model with a ResNet encoder to find out the exact regions within each frame where a deepfake has happened, providing pixel-level manipulation boundaries. Extensive experiments on the challenging FF++ dataset demonstrate 99.9% accuracy in detection and 99.8% Receiver Operating Characteristic - Area Under Curve (ROC-AUC) Score in localization tasks, with minimal false positives. The proposed architecture significantly decreases computational costs and time by specifically choosing only those frames that are deepfake and using a pre-trained model to help reduce time.

Keywords: Deepfake Detection, Image Manipulation, Deep Learning, EfficientNet, UNet, Semantic Segmentation, Facial Forgery, Media Forensics, Two-Stage Detection, FF++ Dataset, Facial Manipulation Localization, Image Authentication

1 Introduction

Deepfakes are a crucial challenge to the integrity of digital media [1, 2]. Using advanced techniques such as GANs, deepfakes are capable of producing fake media that cannot be distinguished from real media, which is a serious threat to national security, public trust, and business reputation. As deepfakes spread across the internet, these technologies are becoming increasingly available and their misuse potential is increasing as well.

Existing methods for deepfake detection are based on binary classification [3, 4]. Although existing methods achieve higher accuracy, they have various drawbacks, including being black-box methods without localization, poor generalization across different generation methods, and high computational cost.

In this paper, we propose a two-stage method that combines binary classification with semantic segmentation. EfficientNetB0 is used for the initial filtering of potential deepfakes, followed by UNet for the localization of manipulated areas at the pixel level. Our method provides state-of-the-art accuracy and localization precision with fast inference speed.

The key contributions are:

- **A Computationally Efficient Cascade Architecture:** The proposed system serves as a high-speed filter, preventing the segmentation model from wasting GPU cycles on frames that are not deepfakes.
- **Explainable Forensic Evidence:** It provides pixel-level masks of manipulated areas, moving beyond simple binary labels, which enhances interpretability.
- **State-of-the-Art Performance:** Rigorous evaluations on the challenging datasets show superior performance

in both detection accuracy and localization precision compared to the existing available techniques.

The paper is organized as follows: Section 2 reviews related work, Section 3 presents the methodology, Section 4 describes experimental setup, Section 5 analyzes results, Section 6 presents ablation studies, and Section 7 concludes the work.

2 Related Work

The rapid development of deepfake technology has attracted a wide range of researchers to develop its detection methods. Rana et al. [5] conducted a detailed literature review and classified deepfake detection approaches, demonstrating the evolution from traditional deepfake detection techniques to modern deep learning-based methods. Similarly, Sunil et al. [6] gathered various techniques and evaluation metrics, presenting a wide range of performance.

Binary classification has been a prominent strategy in deepfake detection, with multiple neural networks achieving high accuracy on benchmark datasets. The landmark paper by Rössler et al. [7] introduced the FaceForensics++ dataset, a comprehensive benchmark for CNN-based detection techniques that set crucial baselines in the field of study. More recent techniques have focused on improving generalization across various types of deepfakes. Reiss et al. [8] introduced a new approach that can detect deepfakes without viewing any examples during training, a major leap forward in zero-shot detection. Haliassos et al. [9] proposed a detection method emphasizing the inconsistencies in lip activity, which performed well on multiple datasets.

Beyond image-based approaches, multi-modal methods leveraging audio-visual inconsistencies have shown promise for deepfake detection [10–13]. These techniques exploit lip-sync discrepancies and audio-visual mismatches to identify manipulated content, though they require additional modalities beyond the visual channel that our framework focuses on.

Self-supervised learning has been identified as a promising approach for deepfake detection, which requires less dependency on large amounts of labeled data. Haliassos et al. [14] showed that real talking faces can be leveraged using self-supervised learning for robust forgery detection. This approach allows models to learn the inherent characteristics of real videos without the need for explicit labeling of forged videos. Temporal coherence analysis has been shown to be very effective for video-based deepfake detection. Zheng et al. [15] analyzed the use of temporal incoherencies between frames for forgery detection, even when the frames look realistic individually. Their model generalized very well to unseen forgery methods by focusing on the basic temporal characteristics that are difficult to emulate realistically.

3 Proposed Method

A two-stage cascaded system is designed for effective and explainable deepfake detection and localization. Figure 1 illustrates the complete architecture of the proposed framework. The system works sequentially: first, it performs binary classification to identify potentially manipulated content, and then only the detected deepfakes are passed to the more computationally intensive second stage for pixel-level localization. This two-stage

cascade design offers a significant computational advantage. The U-Net segmentation model is used only when the image is classified as a deepfake, reducing overall computational load while maintaining high detection performance. This approach yields substantial efficiency gains when authentic content is larger than manipulated content.

3.1 Preprocessing Pipeline

The preprocessing pipeline for the proposed deepfake detection system consists of two main components. For the binary classification model, input images are resized to 256×256 pixels and normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], standard deviation = [0.229, 0.224, 0.225]) to ensure consistent input distribution. For the segmentation model, a frame differencing approach is used, where frames are extracted from both original and manipulated videos. These frames are then processed by computing the absolute difference between corresponding frames, converting the result to grayscale, and applying a binary threshold (threshold = 30) to generate precise segmentation masks. This dual preprocessing strategy allows our system to effectively learn both global and localized manipulation patterns while maintaining computational efficiency.

3.2 Binary Classification with EfficientNetB0

EfficientNetB0 [16] is selected as the binary classifier because it provides the best trade-off between accuracy and computational complexity. The model is based on compound scaling and features a series of mobile inverted bottleneck convolutional (MBCConv) layers

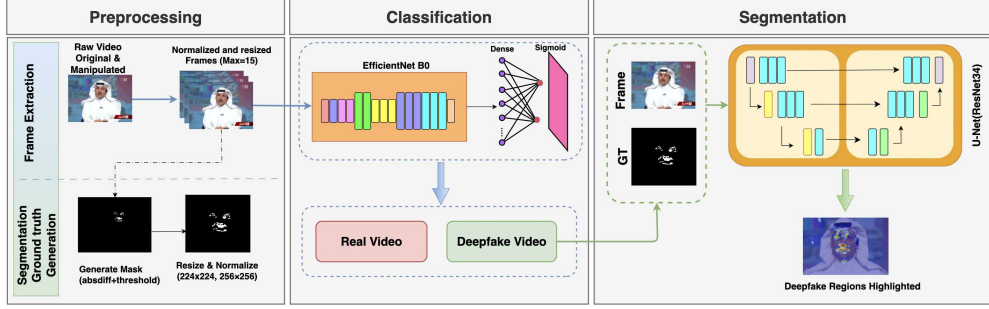


Fig. 1: Overview of the two-staged deepfake detection and localization framework.

with squeeze-and-excitation (SE) optimization. The SE component is capable of adaptively recalculating channel-wise feature responses by learning to focus on more informative channels and downweighting less informative ones. The model is pre-trained on ImageNet, and the classification layer is swapped with a custom head that includes global average pooling, dropout ($p = 0.5$), and a fully connected layer ($1280 \rightarrow 1$). The model is trained with binary cross-entropy loss; the training parameters are listed in Section 4.2.

3.3 Region Localization with UNet

For pixel-level localization of manipulated regions, a UNet architecture with a ResNet34 encoder backbone is used. This encoder-decoder model with skip connections is particularly effective for semantic segmentation tasks, as it combines high-level semantic features with fine-grained spatial information to generate accurate masks. The detailed tensor dimensions for each stage of the ResNet34-based U-Net architecture are provided in Table 1.

The encoder progressively extracts hierarchical features through a series of convolutional blocks, while the decoder reconstructs the segmentation mask

Table 1: ResNet34-based U-Net tensor dimensions.

Stage	Component	Output Shape
<i>Encoder</i>		
Input	–	$3 \times 224 \times 224$
Enc1	Conv1 ($7 \times 7, s=2$)	$64 \times 112 \times 112$
Enc2	MaxPool + Layer1	$64 \times 56 \times 56$
Enc3	Layer2	$128 \times 28 \times 28$
Enc4	Layer3	$256 \times 14 \times 14$
Enc5	Layer4	$512 \times 7 \times 7$
Center	Two Conv-BN-ReLU	$512 \times 7 \times 7$
<i>Decoder</i>		
Dec5	UpConv + Skip(Enc4)	$256 \times 14 \times 14$
Dec4	UpConv + Skip(Enc3)	$128 \times 28 \times 28$
Dec3	UpConv + Skip(Enc2)	$64 \times 56 \times 56$
Dec2	UpConv + Skip(Enc1)	$64 \times 112 \times 112$
Dec1	UpConv	$64 \times 224 \times 224$
Output	1×1 Conv + Sigmoid	$1 \times 224 \times 224$

by combining upsampled features with encoder skip connections.

These skip connections preserve fine-grained spatial details that would otherwise be lost during downsampling. The UNet architecture consists of a ResNet34 backbone pre-trained on ImageNet as the encoder, two convolutional blocks with batch normalization in the center, a series of up-convolutional blocks with skip connections as the decoder, and a one-channel convolutional layer with sigmoid activation as the output. The final segmentation mask is produced by applying

a 1×1 convolution followed by sigmoid activation.

For training the segmentation model, ground truth masks that indicate manipulated regions are required. For the FF++ dataset, which lacks explicit manipulation masks, a difference-based approach is employed. The manipulation mask is generated through frame differencing:

$$M(x, y) = \begin{cases} 1, & |I_f(x, y) - I_r(x, y)| > \tau \\ 0, & \text{else} \end{cases} \quad (1)$$

where I_{fake} and I_{real} are corresponding frames from manipulated and original videos, respectively, and $\tau = 30$ is the binarization threshold. Morphological operations are subsequently applied to refine the mask boundaries. The segmentation model is trained using binary cross-entropy loss.

The model is optimized using the AdamW optimizer with a learning rate of 1×10^{-4} , a batch size of 16, and trained for 15 epochs with a validation split of 80/20. The model with the best validation performance is saved for inference.

3.4 Inference Pipeline

During inference, preprocessed frames are passed to the EfficientNetB0 classifier, and only those classified as deepfakes (probability > 0.5) are forwarded to the UNet segmentation model for pixel-level localization. The system generates heatmaps by overlaying predicted masks onto original frames and computes frame-level and video-level statistics for interpretable forensic analysis.

4 Experimental Setup

4.1 Datasets

Experiments were performed on a variety of benchmark datasets to ensure a comprehensive evaluation of the proposed method:

FaceForensics++ (FF++) is used in both high-quality (HQ) and heavily compressed low-quality (LQ) versions, containing manipulated videos from four methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures, split into training (70%), validation (15%), and test (15%) sets. **Celeb-DF(v2)** contains 590 original and 5,639 deepfake videos with minimal visual artifacts, making detection challenging. **DFDC**, released by Facebook AI, provides approximately 100,000 videos created using diverse generation techniques across various subjects, settings, and compression levels, serving as a large-scale benchmark.

Table 2: Summary of the datasets used for experimentation.

Dataset	Split	Real	Fake	Total
Celeb-DF(v2)	Train	350	350	700
	Val	70	70	140
	Test	280	280	560
DFDC	Train	300	300	600
	Val	50	50	100
	Test	250	250	500
FF++ (HQ)	Train	400	400	800
	Val	80	80	160
	Test	320	320	640
FF++ (LQ)	Train	500	500	1000
	Val	100	100	200
	Test	400	400	800

Table 2 summarizes the dataset splits and the number of real and fake samples used for training, validation, and testing across Celeb-DF, DFDC, and FF++ (HQ/LQ). Collectively, these datasets

form a strong experimental foundation for benchmarking deepfake detection models across diverse manipulation techniques, visual qualities, and compression settings. In particular, the availability of both high- and low-quality variants of FF++ enables systematic examination of how detection performance changes with increasing compression. This evaluation is especially important for real-world scenarios, where videos are often subjected to substantial quality degradation during capture, storage, or transmission.

4.2 Implementation Details

For each video, 15 frames are extracted to retain sufficient temporal information while maintaining efficiency. Frames are preprocessed according to the pipeline described in Section 3.1.

The EfficientNetB0 classifier was initialized with ImageNet pre-trained weights and fine-tuned on the preprocessed dataset. The model was trained using Binary Cross-Entropy loss, optimized with Adam at an initial learning rate of 1×10^{-4} , and a cosine annealing scheduler reducing the learning rate to a minimum of 1×10^{-6} . Training was conducted with a batch size of 32 for up to 30 epochs, with early stopping after 5 epochs. Regularization was applied through a weight decay of 1×10^{-5} and a dropout rate of 0.5 to mitigate overfitting.

For pixel-level manipulation localization, a U-Net-based segmentation network incorporating a ResNet34 encoder was initialized with ImageNet pretraining. The model was trained using Binary Cross-Entropy with Logits as the loss function and optimized with AdamW, with an initial learning rate of 1×10^{-4} . Training was performed with a batch size of 16 for 15 epochs. All experiments were

carried out using PyTorch 2.7.0 on an Apple M3 Pro processor with 36 GB of unified memory.

5 Results and Discussion

5.1 Stage 1: Binary Classification Results

The classification stage is evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. The EfficientNetB0 classifier consistently showed strong performance in separating real videos from manipulated ones across all datasets. As illustrated in Figure 2, both training and validation curves exhibit steady convergence, indicating effective learning with little evidence of overfitting.

To provide interpretability and visual explanation of the classification decisions, Grad-CAM (Gradient-weighted Class Activation Mapping) visualizations are presented in Figure 3. The heatmaps highlight the regions of the input images that most strongly influence the model’s classification decision. For authentic images, the model focuses on natural facial features, while for deepfake images, the attention is concentrated on manipulated regions, particularly around the face and facial boundaries. This demonstrates that the EfficientNetB0 classifier has learned to identify manipulation artifacts rather than relying on spurious correlations.

5.2 Stage 2: Manipulation Localization Results

The localization stage is evaluated using Intersection over Union (IoU) and AUC-ROC. The UNet-based segmentation model achieved precise localization of manipulated regions, as evidenced by both quantitative metrics and qualitative

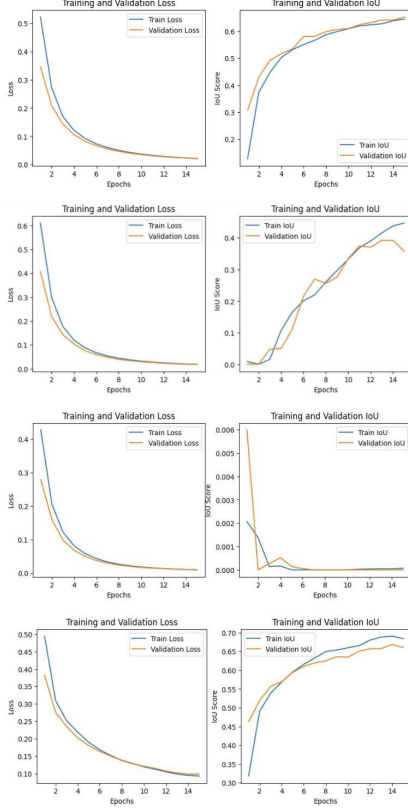


Fig. 2: Training and validation metrics for Stage 1 classification across four datasets (top to bottom): FF++ HQ, FF++ LQ, Celeb-DF, and DFDC.

visualizations. The ROC curves in Figure 5 demonstrate the model’s strong discriminative capability, with AUC scores exceeding 0.99 across all datasets. Table 3 provides detailed performance metrics, highlighting the model’s consistent performance under varying quality conditions. Figure 4 illustrates the mask generation process with example frames from the FF++ dataset, while Figure 6 presents sample segmentation outputs, demonstrating the model’s ability to precisely identify manipulated regions while maintaining clean boundaries around authentic facial features.



Fig. 3: Grad-CAM visualizations for EfficientNetB0 classification (left to right): Original image, Deepfake image, Grad-CAM overlay on original, and Grad-CAM overlay on deepfake.



Fig. 4: Visualization of the localization process (top to bottom): Original frame, Manipulated frame, and Localization heatmap highlighting the forged regions.

5.3 Comparative Analysis

Classification Performance

Table 4 compares the obtained classification accuracy against state-of-the-art methods. The proposed approach demonstrates consistent superiority across all

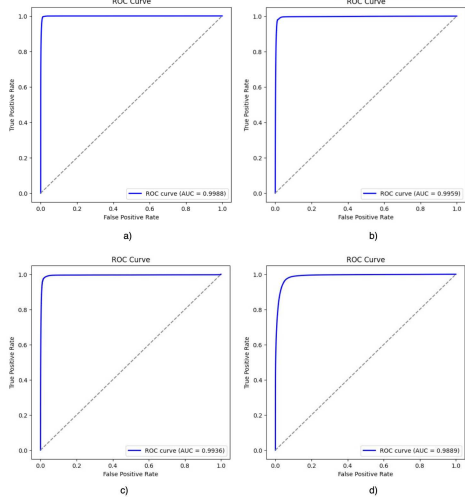


Fig. 5: ROC curves for segmentation-based localization on four datasets: (a) FF++ HQ, (b) FF++ LQ, (c) Celeb-DF, and (d) DFDC.

Table 3: Segmentation performance.

Dataset	AUC	IoU
FF++ (HQ)	0.9988	0.892
FF++ (LQ)	0.9921	0.842
Celeb-DF	0.9875	0.821
DFDC	0.9832	0.803



Fig. 6: Segmentation Qualitative Analysis: Predicted binary masks for spatial localization of manipulated regions.

evaluated datasets, particularly in challenging low-quality scenarios.

Table 4: Comparative classification accuracy (%) against state-of-the-art methods.

Method	HQ	LQ	CDF	DFDC
XceptionNet (2019) [7]	95.7	89.2	86.4	84.1
Mesonet (2018)	93.1	87.5	84.7	82.3
Multi-attention (2020)	96.8	91.7	88.5	86.3
CNN-LSTM (2021)	96.2	90.1	87.3	85.2
KAN (2025) [17]	95.0	—	88.3	—
CViT (2024) [18]	97.0	—	—	—
DFS (2024) [19]	98.0	—	96.3	—
Ours (2025)	99.9	95.2	92.1	90.7

* Dashes (—) indicate results were not reported.

6 Ablation Studies

Ablation experiments were conducted to evaluate key design decisions. Table 5 compares backbone architectures for binary classification.

Table 5: Performance comparison of different backbone architectures for the classification model.

Backbone	Acc. (%)	F1 (%)	AUC	Params (M)	Infer. (ms)
EfficientNetB0	94.3	94.7	0.985	5.3	18.5
ResNet18	92.1	92.4	0.967	11.7	15.2
ResNet50	93.6	93.8	0.978	25.6	29.7
MobileNetV3	91.8	92.0	0.963	5.4	14.3
VGG16	92.9	93.1	0.972	138.4	35.8

EfficientNetB0 achieved the best accuracy-efficiency trade-off. MobileNetV3 offered faster inference but with a 2.5% accuracy drop. VGG16 and ResNet50 had higher parameter counts without proportional gains.

Preprocessing Impact:

Face detection and alignment improved accuracy by 3.2% and IoU by 0.053. ImageNet normalization added 1.8% accuracy. Focusing on facial regions improved accuracy by 4.7% while reducing computation by 32%.

Table 6: Ablation study on segmentation model and loss functions.

Encoder	Loss	IoU	Boundary F1
ResNet18	BCE	0.723	0.692
ResNet18	Dice	0.731	0.711
ResNet18	BCE+Dice	0.742	0.725
ResNet34	BCE	0.761	0.734
ResNet34	Dice	0.770	0.748
ResNet34	BCE+Dice	0.784	0.763
ResNet50	BCE+Dice	0.791	0.772

Segmentation Model:

Table 6 shows encoder and loss function comparisons. Combined BCE+Dice loss outperformed individual losses. ResNet34 was selected over ResNet50 (+0.007 IoU) due to better speed-accuracy trade-off. The proposed cascade architecture significantly optimizes computational efficiency by selectively invoking the segmentation stage only for frames identified as manipulated.

7 Conclusion

This work proposes a two-stage cascade framework for deepfake detection and localization that combines an EfficientNetB0-based classifier with a UNet segmentation model to balance detection accuracy, interpretability, and computational efficiency. The classifier achieves 99.9% accuracy on high-quality FF++ data and more than 90% on various other datasets like Celeb-DF, DFDC. The segmentation stage attains an AUC of 0.9988 for manipulation localization, enabling selective application of computationally intensive processing. Experimental analysis confirms the effectiveness of key design choices, including face alignment, data augmentation, and hybrid loss optimization. By providing both reliable detection and spatial localization of manipulated regions, the proposed

framework supports explainable forensic analysis and contributes toward practical and trustworthy deepfake detection systems. Despite remaining challenges in generalization, temporal modeling, and efficiency on constrained devices, the approach establishes a scalable foundation for future research in real-time media authentication.

Statements and Declarations

Funding: The authors did not receive support from any organization for the submitted work.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

- [1] Verdoliva, L.: Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing* **14**(5), 910–932 (2020)
- [2] Chesney, R., Citron, D.K.C.: Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review* **107**(6), 1753–1820 (2019)
- [3] Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. *ACM Computing Surveys* **54**(1), 1–41 (2021)
- [4] Tolosana, R., Vera-Rodríguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion* **64**, 131–148 (2020)
- [5] Rana, M.S., Nobi, M.N., Murali, B., Sung, A.H.: Deepfake detection: A systematic literature review. *IEEE Access* **10**, 25494–25513 (2022)

- [6] Sunil, R., Mer, P., Diwan, A., Mahadeva, R., Sharma, A.: Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation. *Heliyon* **11**(3), 42273 (2025)
- [7] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: Learning to detect manipulated facial images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019)
- [8] Reiss, T., Cavia, B., Hoshen, Y.: Detecting deepfakes without seeing any. In: *arXiv Preprint arXiv:2311.01458* (2023)
- [9] Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021)
- [10] Zhou, Y., Lim, S.-N.: Joint audio-visual deepfake detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14780–14789 (2021)
- [11] Feng, C., Chen, Z., Owens, A.: Self-supervised video forensics by audio-visual anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10491–10503 (2023)
- [12] Shahzad, S.A., Hashmi, A., Khan, S., Peng, Y.-T., Tsao, Y., Wang, H.-M.: Lip sync matters: A novel multi-modal forgery detector. In: *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1885–1892 (2022)
- [13] Hashmi, A., Shahzad, S.A., Lin, C.-W., Tsao, Y., Wang, H.-M.: Avtenet: Audio-visual transformer-based ensemble network exploiting multiple experts for video deepfake detection. In: *arXiv Preprint arXiv:2310.13103* (2023)
- [14] Haliassos, A., Mira, R., Petridis, S., Pantic, M.: Leveraging real talking faces via self-supervision for robust forgery detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14950–14962 (2022)
- [15] Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15044–15054 (2021)
- [16] Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114 (2019)
- [17] Jabbar, N., Bhatti, S.M., Rashid, M., Jaffar, A., Akram, S.: Single-layer kan for deepfake classification: Balancing efficiency and performance in resource constrained environments. *PLoS ONE* **20**(7), 0326565 (2025)
- [18] Soudy, A.H., Sayed, O., Tag-Elser, H., Hussein, S., Ali, A.: Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications* **36**(31), 19759–19775 (2024)
- [19] Naskar, G., Mohiuddin, S., Malakar, S., Cuevas, E., Sarkar, R.: Deepfake detection using deep feature stacking and meta-learning. *Heliyon* **10**(4), 25933 (2024)