

Generalizable deepfake video detection for multimedia security via temporal–spatial artifact localization and adaptive sequence modelling

Kavita Lal *

Savita Shiwani [†]

Department of Computer Science and Engineering

Poornima University

Jaipur 302022

Rajasthan

India

Geeta Chhabra Gandhi [§]

Department of Computer Applications

Manipal University Jaipur

Jaipur 303007

Rajasthan

India

Abstract

Due to the rise in face manipulation algorithms, it is difficult to distinguish real videos from computer-generated videos. For this, the paper proposes HABSNet, a model that performs artifact localization and temporal consistency analysis. The facial regions identified by MTCNN are analysed by ResNeXt-50 backbone model with the help of two-stream attention mechanism for detecting blending and frequency artifacts, as well as a BiLSTM based inter-frame self-attention gate for detecting unnatural eye movements and head pose changes. Trained on the FaceForensics++ (C23) dataset, the HABSNet achieves an accuracy of 94.88% and generalizes well on Celeb-DF v2 (0.950) and DFDC (0.920). For multimedia security the proposed framework contributes by allowing identification of manipulated videos across varied deepfake generation techniques.

Subject Classification: Primary 68T07, Secondary 68M25, 68T45.

Keywords: Deepfake detection, Bidirectional LSTM, ResNeXt-50, Temporal self-attention, Cross-dataset generalization, FaceForensics++, Multimedia security.

* E-mail: kavitaroy@gmail.com (Corresponding Author)

[†] E-mail: savita.shiwani@poornima.edu.in

[§] E-mail: geeta.gandhi@jaipur.manipal.edu

1. Introduction

The progression of deep neural networks that can produce highly realistic human faces has transitioned from being a research novelty to a commonly accessible tool for everyday users. Modern techniques for altering facial features largely rely on Generative Adversarial Networks (GANs) [1], which enable realistic identity swaps by employing an adversarial optimization process involving a generator and a discriminator. Hence, developing a robust deepfake detection system has become a research challenge for multimedia security and cyber forensic.

The main technical limitation is that features within the artworks- spatial attributes of manipulation artifacts occurring in semantically meaningful face areas (such as regions surrounding eyes and mouth, changes from skin to hair) have not been fused with temporal characteristics of stabilities during syntheses, consisting of frame-to-frame inconsistency of blinking patterns, luminance-shift and unrealistic head rotation [3].

Most previous architectures focus on one dimension at the expense of the other: spatial-only detectors disregard inter-frame context, whereas temporal models aggregate information across frames in a manner that obscures detailed spatial features.

We present HABSNet, for multimedia security, which incorporates a dual-headed attention module for spatially selective feature refinement and a temporal self-attention-enhanced BiLSTM to model coherence across frame sequences, addressing these limitations. This work presents several key contributions:

- (1) We then propose a fundamental dual-branch attention module that combines channel modulation with spatial saliency weighting, allowing the model to identify which feature types and which spatial regions are forensic.
- (2) We improve the temporal processing branch by incorporating a self-attention gating mechanism applied to the BiLSTM output, enabling adaptive re-weighting of inter-frame relationships that surpasses the structural constraints imposed by fixed recurrence in conventional sequence models.
- (3) We perform an ablation analysis to measure the individual contributions of and the synergies among the various architectural modules.

2. Background

2.1 *Early heuristic and signal-based methods*

Initially, detection was systematically approached using low-level physiological and signal processing cues. The authors in [1] showed that faces generated using GANs were likely to exhibit unusual dynamics in the eye regions, as the training datasets predominantly consisted of images with open eyes. [2] Demonstrate 3D lighting consistency, with the direction of estimated illumination remaining uniform across both facial and non-facial areas.

Although these techniques were significant during their era, they have essentially become outdated due to modern synthesis approaches.

2.2 Deep Learning-Based Detection Frameworks

Most contemporary approaches are derived from deep networks that are trained in an end-to-end manner. FaceForensics++ [5], exhibited significant failure in cross-manipulation generalization, which was attributed to overfitting the statistical properties of method-specific artifacts.

Additional architectural refinements have extended accuracy boundaries in more recent studies. [6] Proposed the MSATT as a multi-scale synthetic artifacts trace tracker based on dilated convolution hierarchies and chromatic irregularities. Together, these advances constitute the design principles guiding HABSNet: a structured methodology for aggregating spatial, spectral and temporal features into a single cohesive multiscale paradigm, enabling generalisation across data sets.

3. Datasets and Experimental Setups

3.1 Dataset Descriptions

FaceForensics++ (FF++) with Compression Level C23: FF++ [5] comprises 1,000 real video sequences, and 4000 manipulated sequences using Face2Face, FaceSwap and Neural Textures.

Celeb-DF v2 — This dataset [6] includes 590 real celebrity video clips, and 5,639 decent quality fake videos (6.229 in total).

Deepfake Detection Challenge (DFDC) The DFDC dataset [7] contains 23,654 real videos and 104,500 fake videos generated from 398 actors using eight different manipulation techniques.

3.2 Training Protocol and Hyperparameters

All the experiments are performed on an NVIDIA RTX 3090 GPU (24 GB VRAM). The model was trained for 101 epochs using the Adam optimizer with learning rate 1×10^{-3} , which is reduced if validation does not improve, yielding a batch size of 32 and forty gradient steps per epoch. We used the FF++ C23 training split and five-fold cross-validation, with all metrics reported as mean $\pm$ standard deviation across folds. We trained only on the complete FF++ C23 corpus and tested without any fine-tuning or domain adaptation on held-out Celeb-DF v2 and DFDC test sets, appropriate for cross-dataset evaluation. All performance improvements were statistically significant with p-values < 0.05 using paired t-tests. Data augmentation included random horizontal flipping, brightness and contrast jitter ($\pm 20\%$) and Gaussian blur ($\sigma \in [0.1, 1.0]$).

4. Proposed Methodology

4.1 Video Preprocessing and Face Extraction

For the detection of face bounding boxes in a video, we used MTCNN since it gets the job done regardless of the distance (scale) and lighting/occlusion conditions. Frame Processing: Finally, it makes a lot of computational sense to

process multiple frames together. If not equal to one, the bounding box aspect ratio is employed in padding that maintains the composition of the original images till they resize to $112 \times 112 \times 3$ pixels.

1. Frame sampling

$$F_S = \{I_t \mid t = 0, \Delta, 2\Delta, \dots, t < T\}$$

Here, I_t denotes the video frame at time index t , T is the total number of frames, and Δ is the sampling interval.

2. Face bounding-box detection

$$B_t = \text{MTCNN}(I_t) = (x_t, y_t, w_t, h_t)$$

MTCNN returns the detected face bounding box, with top-left coordinates (x_t, y_t) , width w_t , and height h_t .

3. Bounding-box margin expansion

$$w'_t = (1 + 2m)w_t, \quad h'_t = (1 + 2m)h_t$$

The margin ratio m expands the detected region symmetrically to retain surrounding facial context.

4. Aspect ratio and target padding

$$r_t = \frac{w'_t}{h'_t},$$

$$p_x = \frac{\max(0, h'_t - w'_t)}{2}$$

$$p_y = \frac{\max(0, w'_t - h'_t)}{2}$$

Symmetric padding creates a square crop without distorting the original face geometry. Padding is applied after clipping the expanded crop to image boundaries.

5. Face resizing and normalization

$$X_t = \frac{\text{Resize}(\text{Pad}(\text{Crop}(I_t, B'_t), p_x, p_y), 112, 112)}{255}$$

The padded face crop is resized to 112×112 pixels and pixel intensities are normalized to $[0, 1]$; X_t has dimensions $112 \times 112 \times 3$.

6. Mini-batch construction

$$\mathcal{B}_k = \text{Stack}(X_{\{kN\}}, X_{\{kN+1\}}, \dots, X_{\{\min((k+1)N-1, M-1)\}})$$

For batch size N and M extracted faces, the k -th batch contains at most N normalized face tensors, enabling efficient parallel processing.

4.2 Feature Extraction using ResNeXt-50

ResNeXt-50 [5], an enhanced version of the residual network, is used to extract spatial features from $112 \times 112 \times 3$ face crops. ResNeXt-50 produces scale-invariant hierarchical feature maps. In lower layers, the feature maps correspond to primitive features like edges, colour and texture; mid-layers represent local configurations while deeper layers describe semantically coherent facial parts. Global average pooling + batch norm then flat layer

[B,2048,4 x 4] use residual affine transform $C_j = (W_a \cdot B_j + b_a) + B_j$ where B_j is the backbone normalized representation at frame j extracted and W_a and b_a are learnable affine parameters.

4.3 Dual-Branch Hierarchical Attention Module

Figure 1 depicts the dual branch attention module.

Channel Attention Branch: The channel attention module takes a global average pool over the spatial dimension and prints to two layers of a bottleneck mlp (bottleneck ratio $r=16$) with a sigmoid at the top to produce the per-channel significance vector like in [11].

Spatial Attention Branch: Each position in the k grid of spatial (i.e., x, y) positions $\tau\{j\}$ has a single (learned) parameter $\alpha_{j,t}$, which is obtained through a 1×1 convolutional projection projecting high-dimensional (2048-dimension channel space) feature map into a scalar field. The output of this scalar field is then fed into a spatial softmax which assigns probability-like weights to each small 4×4 grid of labels. P_j is the attended context, for frame j the output is:

$$F_j^{attn} = \sum_t \alpha_{j,t} B_{j,b} C_{j,t}.$$

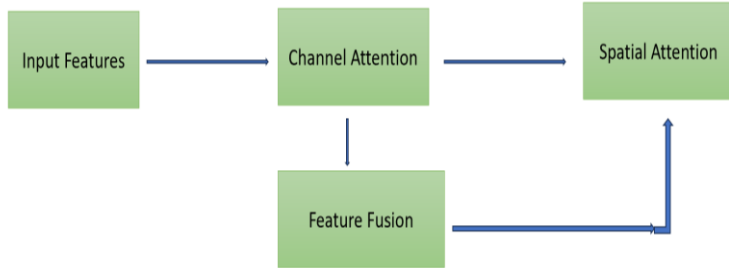


Figure 1
Dual Branch attention module with channel and spatial branches

4.4 Temporal Self-attention on Bidirectional LSTM

Next, the sequence of attended frame descriptors $F^{attn1} \dots F^{attnN}$ is fed into a Bidirectional LSTM (BiLSTM) [8] to model the temporal relationships. Following this, the input forget and output gates are applied individually for each direction, resulting in the final hidden state of the BiLSTM with a dimension of 2048. At each timestep, the forward and backward hidden states are combined to form $H = \{h_1 \dots h_N\}$, a 4096-dimensional representation.

4.5 Classification Head

The temporal aggregated representation output is fed into a two-layer fully connected classifier, with Dropout(0.5) layer between them used for regularization by avoiding overfitting. We also used a SoftMax activation function across two classes (real/fake) for the output layer and derive predictive label in an argmax fashion over the class probability vector. This loss is

leveraged for regularization and end-to-end training of the network with the binary cross-entropy loss applying label smoothing ($\epsilon=0.1$).

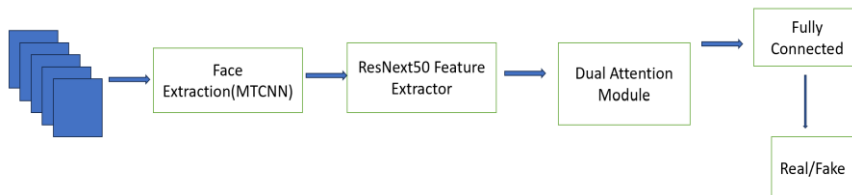


Figure 2

Architecture of HABSNet framework for deepfake video detection

5. Experimental Results and Discussion

5.1 Benchmark Comparison on FF++ C23

HABSNet shown in Figure 2 was compared with the eight baseline and state-of-the-art architectures only on FaceForensics++ C23 in Table 1 based on five-fold cross-validation. All of those are unique and valuable pieces of this architecture, having much higher representation power than a backbone alone.

Table 1

Performance Comparison on FF++ C23 (Mean $\pm$ Std, 5-Fold CV)

Model Used	Accuracy	Precision	Recall	F1- Score	Area Under Curve	EER (%)
HABSNet (Proposed)	0.9488 $\pm$ 0.013	0.921 $\pm$ 0.009	0.936 $\pm$ 0.008	0.928 $\pm$ 0.007	0.967 $\pm$ 0.006	3.8 $\pm$ 0.4
ResNeXt-50 (backbone only)	0.883 $\pm$ 0.016	0.874 $\pm$ 0.019	0.861 $\pm$ 0.015	0.867 $\pm$ 0.013	0.908 $\pm$ 0.011	9.1 $\pm$ 1.1
LSTM (temporal only)	0.887 $\pm$ 0.015	0.882 $\pm$ 0.016	0.874 $\pm$ 0.014	0.878 $\pm$ 0.012	0.915 $\pm$ 0.010	8.3 $\pm$ 1.0
EfficientNet-B4 [7]	0.924 $\pm$ 0.013	0.94 $\pm$ 0.01	0.908 $\pm$ 0.011	0.924 $\pm$ 0.009	0.950 $\pm$ 0.007	5.1 $\pm$ 0.6
MobileNetV2	0.900 $\pm$ 0.016	0.93 $\pm$ 0.018	0.888 $\pm$ 0.015	0.904 $\pm$ 0.013	0.928 $\pm$ 0.011	6.9 $\pm$ 0.8
ViT-B/16 [8]	0.915 $\pm$ 0.014	0.92 $\pm$ 0.012	0.910 $\pm$ 0.010	0.915 $\pm$ 0.009	0.945 $\pm$ 0.008	5.5 $\pm$ 0.7
BMNet [9]	0.920 $\pm$ 0.013	0.93 $\pm$ 0.011	0.905 $\pm$ 0.012	0.918 $\pm$ 0.010	0.952 $\pm$ 0.007	4.8 $\pm$ 0.6
sFace X-ray [10]	0.905 $\pm$ 0.015	0.91 $\pm$ 0.014	0.895 $\pm$ 0.013	0.902 $\pm$ 0.011	0.935 $\pm$ 0.009	6.2 $\pm$ 0.8
Multi-task GAN-det. [11]	0.912 $\pm$ 0.014	0.92 $\pm$ 0.013	0.900 $\pm$ 0.011	0.910 $\pm$ 0.010	0.940 $\pm$ 0.008	5.8 $\pm$ 0.7

5.2 Robustness to Quality Variations of Compression

Table 2 displays detection accuracy for three different compression levels from FaceForensics++ (Low Quality or LQ, high level of compression; High Quality or HQ, low level of compression; RAW, uncompressed), as well as per-

frame inference latency on the RTX 3090. This reduction in inference time (~ 45 ms) is significantly lower than the other ViT based architectures which are recording overall time complexities within range of 60–80 ms.

Table 2
Accuracy vs. Compression Quality Level and Inference Speed on FF++

Method	LQ (%)	HQ (%)	RAW (%)	Inference (ms/frame)
SVM	69.43 $\pm$ 2.7	72.81 $\pm$ 2.5	75.16 $\pm$ 2.3	18 $\pm$ 3
CNN	85.37 $\pm$ 1.9	89.62 $\pm$ 1.7	90.74 $\pm$ 1.6	33 $\pm$ 4
RNN	86.92 $\pm$ 1.8	88.47 $\pm$ 1.6	89.53 $\pm$ 1.5	38 $\pm$ 5
LSTM [8]	87.64 $\pm$ 1.7	89.08 $\pm$ 1.5	89.91 $\pm$ 1.4	44 $\pm$ 4
ConvLSTM	88.73 $\pm$ 1.6	90.34 $\pm$ 1.4	90.87 $\pm$ 1.3	50 $\pm$ 5
HABSNet (Proposed)	91.51$\pm$1.2	92.87$\pm$1.0	94.88$\pm$0.9	46$\pm$4

5.3 Ablation Study

Next, we ablate the impact of each architectural component in Table 3.

Table 3
Ablation Study done on FaceForensics++ C23.

Model Variant	AC	AUC	EER (%)
Full HABSNet	0.9488$\pm$0.011	0.971$\pm$0.007	3.6$\pm$0.4
w/o Dual Attention Module	0.881 $\pm$ 0.016	0.912 $\pm$ 0.011	9.2 $\pm$ 1.2
w/o BiLSTM (CNN-only pathway)	0.897 $\pm$ 0.015	0.929 $\pm$ 0.010	7.1 $\pm$ 0.9
w/o Temporal Self-Attention Gate	0.912 $\pm$ 0.013	0.943 $\pm$ 0.009	5.7 $\pm$ 0.7

5.4 Cross-Dataset Generalization

Table 4 assesses cross-dataset generalization by training exclusively on FF++ C23 and testing on Celeb-DF v2 and DFDC, without employing any domain adaptation techniques. HABSNet obtains an AUC of 0.950 ± 0.009 on Celeb-DF v2 and 0.920 ± 0.010 on DFDC. This generalization benefit indicates that the dual attention mechanism successfully focuses on manipulation-invariant forensic signatures key artifacts revealed by the GAN synthesis process rather than method-specific texture patterns that do not generalize across different datasets.

Table 4

Cross-Dataset Generalization (Train: FF++ C23, Test: Held-Out Datasets; AUC $\pm$ Std). pp = percentage points.

Test Dataset	HABSNet AUC	ViT-B/16 AUC	Absolute Gain
Celeb-DF v2 [3]	0.950 $\pm$ 0.009	0.760 $\pm$ 0.012	+19.0 pp (p<0.001)
DFDC [4]	0.920 $\pm$ 0.010	0.680 $\pm$ 0.015	+24.0 pp (p<0.001)

6. Limitations and Future Directions

Although HABSNet achieved impressive results, there are several limitations that need to be addressed. In cases of high compression (quantization parameter > C40), the AUC decreases to approximately 0.85 across all data, as aggressive quantization eliminates crucial residual high-frequency signals that the attention module relies on. Adversarial disturbances specifically designed to deceive the network can lower accuracy by approximately 8%, a limitation common to all gradient-based deep learning classifiers. New manipulation techniques that generate artifacts with a qualitatively different distribution than those in the training corpus are also undetectable. Looking ahead, we plan to investigate several new directions: (i) incorporating differentiable JPEG simulation layers into the training process to enhance low-quality performance (ii) adversarial training that dynamically generates perturbations during training, making our model resilient to adaptive attacks (iii) multimodal audio-visual synchrony analysis aimed at identifying manipulated academic voices, along with paired visual face swapping, as a common real-world example of deepfakes.

7. Conclusion

In this paper, we introduce HABSNet, a hierarchical attention-enhanced bidirectional sequence network designed for multimedia security and detection of deepfake videos. HABSNet consists of a dual-branch, multi-attentive framework that combines channel recalibration with the saliency weighting of spatial features derived from deep 2D-convolutional expressions symbolic to the intermediate stage ResNeXt-50 backbone networks, along with a BiLSTM mechanism for encoding temporal dynamics through self-attention gating, which contributes to its superior performance on both in-distribution and transductive dataset testing scenarios. The suggested architecture attains 94.88% accuracy and 0.967 AUC on FaceForensics++ C23, while demonstrating strong generalization with 0.950 and 0.920 AUC on Celeb-DF v2 and DFDC, respectively.

References

- [1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1–11 (2019), doi: 10.1109/ICCV.2019.00009.
- [2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3207–3216 (2020), doi: 10.1109/CVPR42600.2020.00327.
- [3] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton-Ferrer, “The DeepFake Detection Challenge (DFDC) dataset,” arXiv preprint arXiv:2006.07397 (2020), doi: 10.48550/arXiv.2006.07397.
- [4] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5987–5995 (2017), doi: 10.1109/CVPR.2017.634.
- [5] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face: Real-time face capture and reenactment of RGB videos,” *Commun. ACM*, vol. 62, no. 1, pp. 96–104 (Jan. 2019), doi: 10.1145/3292039.
- [6] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face X-ray for more general face forgery detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5001–5010 (2020), doi: 10.1109/CVPR42600.2020.00505.
- [7] S. Kalra, C. H. S. Kumar, Y. Sharma, and G. S. Chauhan, “FakeExpose: Uncovering the falsity of news by targeting the multimodality via transfer learning,” *J. Inf. Optim. Sci.*, vol. 44, no. 3, pp. 301–314 (2023).
- [8] S. Kalra, Y. Bansal, Y. Sharma, and G. S. Chauhan, “FakeSpotter: A blockchain-based trustworthy idea for fake news detection in social media,” *J. Inf. Optim. Sci.*, vol. 44, no. 3, pp. 515–527 (2023), doi: 10.47974/JIOS-1411.

- [9] V. Sapkal, P. Gupta, and A. B. Khan, "Deep learning-based phishing website identification system using CNN-LSTM classifier," *J. Inf. Optim. Sci.*, vol. 44, no. 3, pp. 315–330 (2023).
- [10] S. Sharma and B. Sharma, "Optimized data analysis through hybrid approach over trusted security environment," *J. Inf. Optim. Sci.*, vol. 44, no. 3, pp. 331–340 (2023).
- [11] A. Singh and P. K. Roy, "Blockchain for Secure Computing: Technologies, Challenges, and Future Scope," *Journal of Information and Optimization Sciences*, vol. 47, no. 7, pp. 2535–2582 (Jul. 2026), doi: 10.47974/JIOS-1980.

Received March, 2026