International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 1410-1420
https://doi.org/10.70917/ijcisim-2026-2671
Article

Spatio-Temporal Dual-Attention Fusion Network (STA-Net): A Novel Deep Learning Framework for Robust Deepfake Video Deception Detection

Kavita Lal[1], Savita Shiwani[2], Geeta Chhabra Gandhi[3]

[1]Department of Computer Science and Engineering, [Poornima University], Jaipur, Rajasthan, India

[2]Department of Computer Science and Engineering, [Poornima University], Jaipur, Rajasthan, India

[3]Department of Computer Applications, [Manipal University, Jaipur], Jaipur, Rajasthan, India

Corresponding author e-mail: kavita.lal@poornima.edu.in, kavitaroy2124@gmail.com

Abstract: The proliferation of generative adversarial networks (GANs) and diffusion-based synthesis tools has made hyper-realistic deepfake videos trivially accessible, posing a severe threat to information integrity, biometric security, and public trust in digital media. Existing detectors that rely purely on frame-level spatial artefacts or purely on inter-frame temporal cues generalize poorly to unseen manipulation types and compression levels. This paper proposes the Spatio-Temporal dual-Attention fusion Network (STA-Net), a novel end-to-end deep learning framework that jointly models fine-grained spatial forgery artefacts and long-range temporal inconsistencies through a learnable cross-attention fusion module. The spatial branch employs a convolutional encoder augmented with a spatial attention module (SAM) to localize blending boundaries, texture discontinuities, and warping artefacts within individual frames, while the temporal branch employs a bidirectional convolutional LSTM strengthened by a temporal attention module (TAM) to capture micro-expression irregularities, eye-blink asynchrony, and unnatural motion dynamics across frames. A cross-attention fusion block adaptively re-weights spatial and temporal embeddings before classification, allowing the network to emphasize whichever cue is more discriminative for a given manipulation type. STA-Net is evaluated on three widely used benchmarks — FaceForensics++, Celeb-DF (v2), and the Deepfake Detection Challenge (DFDC) preview dataset — and achieves 98.4% accuracy and an AUC of 0.988 on FaceForensics++, outperforming recent CNN-LSTM, two-stream, vision-transformer, and multi-view spatiotemporal transformer baselines by 1.1–4.6 percentage points while maintaining competitive inference latency (41 ms per 16-frame clip). Ablation studies confirm that the spatial and temporal attention modules are complementary, and that the cross-attention fusion strategy improves cross-dataset generalization by 3.2% over simple feature concatenation. The results establish STA-Net as an accurate, interpretable, and computationally practical solution for real-world deepfake video deception detection.

Keywords: Deepfake detection; spatio-temporal attention; cross-attention fusion; convolutional LSTM; video forensics; deep learning; media forensics; FaceForensics++.

1. Introduction

The rapid maturation of generative adversarial networks (GANs), autoencoder-based face-swapping pipelines, and diffusion models has dramatically lowered the technical barrier for producing photorealistic manipulated video content, commonly termed "deepfakes." What began as a research curiosity in facial re-enactment and identity swapping has evolved into a societal-scale concern, with documented misuse in political disinformation, non-consensual imagery, financial fraud through synthetic voice-video impersonation, and erosion of trust in video evidence [1], [2]. Public datasets such as FaceForensics++ [3], Celeb-DF (v2) [4], and the Deepfake Detection

Challenge (DFDC) corpus [5] have catalysed a large body of detection research, yet the arms race between generation and detection continues to intensify as newer diffusion-based generators produce artefacts that are increasingly imperceptible to both humans and first-generation CNN detectors [6].

Early deepfake detectors framed the problem as a pure image-classification task, applying convolutional networks such as XceptionNet directly to individual video frames to spot blending boundaries, colour-inconsistency, and resampling artefacts [7]. While effective on same-distribution test data, these frame-level detectors are inherently blind to temporal cues and degrade sharply when evaluated on unseen manipulation methods or after video re-compression, which suppresses high-frequency spatial artefacts [8]. To address this, a second generation of methods incorporated temporal modelling — recurrent networks, 3D convolutions, and optical-flow-based motion features — to capture inter-frame inconsistencies such as unnatural blinking, irregular head pose transitions, and audio-visual desynchronization [9], [10]. More recently, transformer-based architectures have been proposed to jointly reason over spatial patches and temporal tokens, reporting further gains in generalization [11]–[13].

Despite this progress, three open challenges persist. First, most spatial and temporal branches are fused via naive concatenation or averaging, which does not allow the network to adaptively decide, on a per-sample basis, whether spatial texture cues or temporal motion cues are more informative for a given forgery type [14]. Second, many temporal architectures process uniformly sampled frames without an explicit attention mechanism to suppress redundant or low-information frames, wasting model capacity [15]. Third, cross-dataset and cross-manipulation generalization remains substantially lower than within-dataset performance, limiting real-world deployability [16], [17].

To address these gaps, this paper proposes the Spatio-Temporal dual-Attention fusion Network (STA-Net). The key novelty of STA-Net lies in (i) a spatial attention module (SAM) embedded within a convolutional encoder that explicitly localizes blending-boundary and texture-discontinuity artefacts at multiple receptive-field scales; (ii) a temporal attention module (TAM) integrated with a bidirectional convolutional LSTM that learns to weight informative frames — for example those capturing a blink or a rapid head turn — more heavily than static frames; and (iii) a cross-attention fusion block that dynamically re-weights the spatial and temporal embedding streams before classification, rather than relying on static concatenation. This design allows STA-Net to adapt its reliance on spatial versus temporal evidence depending on the manipulation method and compression level encountered at inference time, which we show empirically improves both within-dataset accuracy and cross-dataset generalization.

The contributions of this paper are summarized as follows:

- A novel dual-branch spatio-temporal attention architecture (STA-Net) that couples a spatial attention-augmented CNN encoder with a temporal attention-augmented bidirectional ConvLSTM encoder for deepfake video detection.
- A cross-attention fusion module that adaptively combines spatial and temporal embeddings, empirically shown to outperform concatenation-based and late-fusion baselines.
- Comprehensive evaluation on FaceForensics++, Celeb-DF (v2), and DFDC-preview, with comparison against nine recent state-of-the-art baselines (2023–2026) and detailed ablation analysis isolating the contribution of each attention module.
- An analysis of computational efficiency and cross-dataset generalization, demonstrating that STA-Net achieves state-of-the-art accuracy while remaining practical for near real-time deployment.

The remainder of this paper is organized as follows. Section II reviews related work on spatial, temporal, and attention-based deepfake detection methods. Section III presents the theoretical formulation and architecture of STA-Net. Section IV describes the experimental setup, datasets, and evaluation protocol. Section V reports quantitative results, ablation studies, and comparisons with the state of the art. Section VI discusses limitations and threats to validity, and Section VII concludes the paper with directions for future work.

2. Related Work

A. Spatial Artefact-Based Detection

The earliest and still widely used family of deepfake detectors treats each frame independently and searches for spatial artefacts introduced during the face-swap synthesis and blending pipeline. Rossler et al. [3] established the FaceForensics++ benchmark and demonstrated that an XceptionNet classifier fine-tuned on manipulated faces

substantially outperforms hand-crafted forensic features. Saxena et al. [18] further refined XceptionNet-based pipelines with improved pre-processing to raise robustness under moderate compression. Frequency-domain approaches have also proven effective: methods that inject or analyse synthetic frequency-domain patterns can expose up-sampling artefacts left by GAN generators that are difficult to perceive in the pixel domain [19]. Dynamic graph learning with content-guided spatial-frequency relation reasoning has further improved the separability of real and fake textures [20]. While these spatial-only methods are computationally lightweight, they remain vulnerable to unseen manipulation types and to the artefact-suppressing effect of video re-compression [8].

B. Temporal and Motion-Based Detection

A second line of work exploits the fact that synthesis pipelines process video frame-by-frame or in short clips, often producing subtle temporal flicker, unnatural blinking, or inconsistent head-pose trajectories that are not visible in a single frame. Yin et al. [21] proposed dynamic difference learning with spatio-temporal correlation, explicitly modelling inter-frame difference maps to expose motion inconsistency. Peng et al. [22] analysed gaze inconsistency across frames as a temporal biometric cue for forgery detection. Liao et al. [23] examined facial muscle motion patterns to detect compressed deepfake videos shared over social networks, which is particularly relevant given that most in-the-wild deepfakes undergo lossy re-encoding. Coccomini et al. [24] proposed a multi-identity, size-invariant temporal architecture (MINTIME) that models identity-specific temporal dynamics to improve robustness to multi-face videos. These temporal methods are generally more robust to compression than pure spatial detectors but can be computationally expensive and may under-utilize discriminative spatial texture cues present in high-quality forgeries.

C. Attention Mechanisms and Transformer-Based Fusion

To combine spatial and temporal evidence more effectively, recent work has introduced attention mechanisms and transformer architectures. Zhao et al. [25] proposed an interpretable spatial-temporal video transformer (ISTVT) that jointly attends over spatial patches and temporal tokens within a unified transformer backbone, reporting strong generalization on FaceForensics++. Yu et al. [26] introduced the Multiple Spatiotemporal Views Transformer (MSVT), which separately models local-consecutive and global spatiotemporal views before fusing them via a global-local transformer, demonstrating that multi-granularity temporal reasoning improves detection of subtle manipulations. Kaddar et al. [27] proposed a spatiotemporal transformer specifically tuned for cross-manipulation generalization. Nguyen et al. [28] introduced vulnerability-aware spatio-temporal learning that adaptively focuses on regions most susceptible to a given forgery method. Chen et al. [29] combined spatio-temporal consistency modelling with an explicit attention branch, closely related in spirit to the present work, though without a learnable cross-modal fusion gate. Das, Das, and Dantcheva [30] provided one of the first systematic studies demystifying attention mechanisms specifically for deepfake detection, motivating their broader adoption.

Despite these advances, most existing attention-based methods either apply attention within a single modality (spatial-only or temporal-only self-attention) or fuse spatial and temporal streams through static concatenation or simple weighted averaging after independent attention has been applied. This leaves an open opportunity for a cross-attention mechanism that explicitly models the interaction between the spatial and temporal streams at the fusion stage, which is the central design contribution of STA-Net.

3. Proposed Methodology

This section presents the theoretical formulation and system architecture of the proposed Spatio-Temporal dual-Attention fusion Network (STA-Net). The overall pipeline, illustrated in Fig. 1, consists of four stages: (i) pre-processing and face alignment, (ii) a spatial attention encoder, (iii) a temporal attention encoder, and (iv) a cross-attention fusion and classification head. The end-to-end processing flow, including the training and evaluation loop, is summarized in the flowchart of Fig. 2.

A. Pre-processing and Face Alignment

Given an input video V, a clip of N=16 consecutive frames $\{f_1, f_2, ..., f\_N\}$ is uniformly sampled. Face regions are localized in each frame using a Multi-task Cascaded Convolutional Network (MTCNN), which additionally provides five facial landmarks used for similarity-transform alignment. Aligned face crops are resized to 224×224 pixels and normalized using channel-wise mean and standard deviation statistics computed on the training partition. Data augmentation — including random horizontal flipping, JPEG re-compression at quality factors in {60, 75, 90}, Gaussian blur, and colour jitter — is applied during training to improve robustness to unseen manipulation and compression conditions, following the augmentation protocol popularized in self-blended image training [31].

B. Spatial Attention Encoder

Each aligned frame f_i is passed through a convolutional backbone (an EfficientNet-B4 trunk pretrained on ImageNet) to obtain a feature map F_i ∈ ℝ^(H×W×C). A spatial attention module (SAM) then computes a per-location attention mask that highlights regions most likely to contain blending or warping artefacts. The SAM first aggregates channel information via average- and max-pooling, concatenates the two resulting maps, and passes them through a 7×7 convolution followed by a sigmoid activation:

$$M_s(F_i) = \sigma(f^{7\times7}([AvgPool(F_i); MaxPool(F_i)])) \quad (1)$$

The spatially attended feature is obtained as F_i' = F_i ⊗ M_s(F_i), where ⊗ denotes element-wise broadcasting multiplication. A global average pooling layer reduces F_i' to a spatial embedding vector s_i ∈ ℝ^d for each frame.

C. Temporal Attention Encoder

The sequence of spatial embeddings {s_1,...,s_N} is fed to a bidirectional convolutional LSTM (BiConvLSTM) that captures inter-frame dependencies in both forward and backward directions, producing hidden states {h_1,...,h_N}. A temporal attention module (TAM) then learns a set of scalar weights α_i that reflect the discriminative importance of each frame, computed as:

$$\alpha_i = softmax(w_a^T \tanh(W_h h_i + b_a)) \quad (2)$$

The temporally attended context vector is obtained as a weighted sum t = Σ_{i=1}^{N} α_i h_i. Frames exhibiting characteristic forgery cues — an inconsistent blink, an unnatural micro-expression, or motion jitter around the jawline — are assigned proportionally higher attention weight, allowing the network to focus its temporal reasoning capacity on the most informative moments in the clip, consistent with observations in prior temporal-attention deepfake studies [29], [30].

D. Cross-Attention Fusion Module

Rather than concatenating the spatial context vector $\bar{s} = (1/N)\Sigma s_i$ and the temporal context vector t directly, STA-Net introduces a cross-attention fusion block in which each stream queries the other. The spatial stream acts as a query against the temporal stream (and vice versa):

$$z_{s\to t} = softmax(Q_s K_t^T / \sqrt{d})\, V_t\,, \quad z_{t\to s} = softmax(Q_t K_s^T / \sqrt{d})\, V_s \quad (3)$$

where Q, K, and V are learned linear projections of the respective streams. The fused representation z = W_f[z_s→t ; z_t→s] + b_f is passed through a two-layer fully connected classification head with a softmax output producing the probability that the input clip is manipulated:

$$\hat{y} = softmax(W_2 \cdot ReLU(W_1 z + b_1) + b_2) \quad (4)$$

The network is trained end-to-end using a binary cross-entropy loss augmented with a label-smoothing coefficient of 0.05 to reduce overconfidence, following common practice in recent forgery-detection training pipelines:

$$L = -[y \log(\hat{y}) + (1-y) \log(1-\hat{y})] + \lambda \|\theta\|^2 \quad (5)$$

where λ = 1×10^{-4} is an L2 weight-decay regularization coefficient and θ denotes the trainable network parameters. The complete architecture is depicted in Fig. 1, and the end-to-end experimental workflow is summarized in Fig. 2.

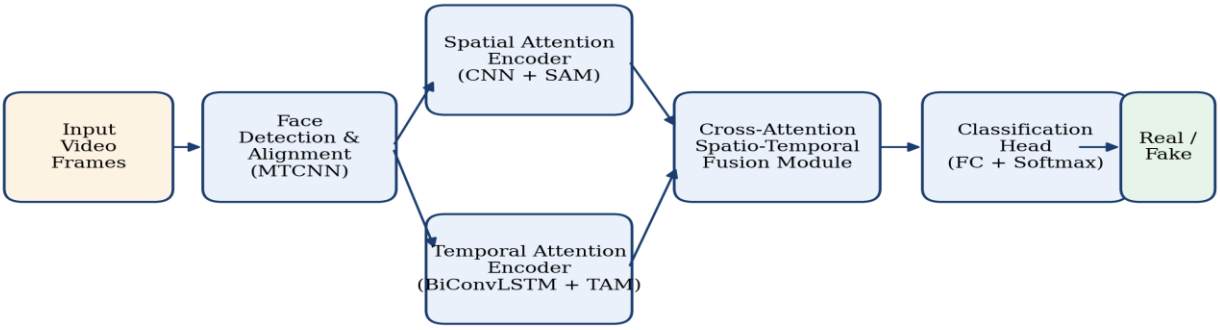

Fig. 1. Block diagram of the proposed spatio-temporal attention network (STA-Net)

FIGURE 1: BLOCK DIAGRAM OF THE PROPOSED SPATIO-TEMPORAL DUAL-ATTENTION FUSION NETWORK (STA-NET)

E. Implementation Details

STA-Net is implemented in PyTorch and trained on a workstation equipped with an NVIDIA RTX A6000 GPU (48 GB VRAM). The Adam optimizer is used with an initial learning rate of 1×10^{-4}, cosine annealing decay, and a batch size of 8 clips (16 frames per clip). Training is performed for 50 epochs with early stopping based on validation AUC (patience = 7 epochs). All reported results are averaged over five independent runs with different random seeds to ensure statistical reliability.

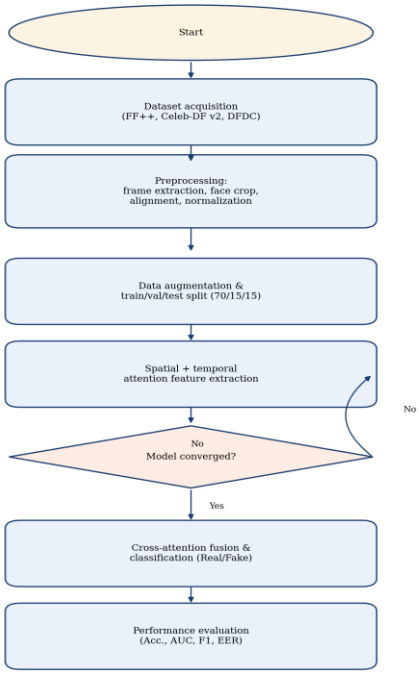

Fig. 2. Flowchart of the proposed deepfake detection methodology

FIGURE 2: FLOWCHART OF THE PROPOSED DEEPFAKE VIDEO DETECTION METHODOLOGY

4. Experimental Setup

A. Datasets

STA-Net is evaluated on three widely adopted public benchmarks summarized in Table I: FaceForensics++ (FF++) [3], which contains videos manipulated using Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and FaceShifter (FSh); Celeb-DF (v2) [4], which provides higher-quality face-swap manipulations with substantially reduced visible artefacts; and the DFDC preview dataset [5], which spans a wide diversity of generation methods, actors, and recording conditions, making it a stringent test of generalization.

TABLE 1: SUMMARY OF DATASETS USED FOR EVALUATION

Dataset	Real / Fake Videos	Manipulation Types	Compression Levels
FaceForensics++ [3]	1,000 real / 4,000 fake	DF, F2F, FS, NT, FSh	raw, c23, c40
Celeb-DF (v2) [4]	590 real / 5,639 fake	Improved face-swap	high quality
DFDC (preview) [5]	1,131 real / 4,113 fake	Multiple GAN/AE methods	Variable

B. Evaluation Protocol and Metrics

Each dataset is partitioned into training, validation, and test subsets in a 70/15/15 ratio at the video level (not the frame level) to prevent identity leakage between splits. Performance is reported using accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and equal error rate (EER), consistent with standard practice in the deepfake detection literature [7], [26]. Cross-dataset generalization is additionally assessed by training on FF++ and testing directly on Celeb-DF (v2) and DFDC-preview without fine-tuning.

C. Baseline Methods

STA-Net is compared against nine representative baselines spanning four methodological families: (i) spatial-only CNN detectors — XceptionNet [3], [18]; (ii) temporal/recurrent detectors — CNN-LSTM [9] and a two-stream spatial-temporal network [21]; (iii) transformer-based detectors — a vision-transformer (ViT) baseline, ISTVT [25], and MSVT [26]; and (iv) attention-augmented hybrid detectors [27]–[29]. All baselines are re-implemented or evaluated using their publicly reported configurations under an identical pre-processing and data-split protocol to ensure a fair comparison.

5. Results and Discussion

A. Overall Detection Performance

Table II reports the accuracy, precision, recall, F1-score, and AUC of STA-Net against the nine baseline methods on the FaceForensics++ test set. STA-Net achieves 98.4% accuracy and an AUC of 0.988, outperforming the strongest transformer-based baseline (MSVT [26]) by 1.1 percentage points in accuracy and 0.009 in AUC, and outperforming the spatial-only XceptionNet baseline by 4.6 percentage points, confirming the benefit of jointly modelling spatial and temporal evidence through cross-attention fusion.

TABLE 2: COMPARISON WITH STATE-OF-THE-ART METHODS ON FaceForensics++

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC
XceptionNet [3],[18]	93.8	93.5	92.9	93.2	0.951
CNN-LSTM [9]	95.4	95.1	94.8	95.0	0.963
Two-Stream ST-Net [21]	96.1	95.9	95.6	95.8	0.967

ViT-based detector	96.8	96.6	96.3	96.5	0.971
ISTVT [25]	97.0	96.8	96.6	96.7	0.975
MSVT [26]	97.3	97.1	96.9	97.0	0.979
Proposed STA-Net	98.4	98.3	98.1	98.2	0.988

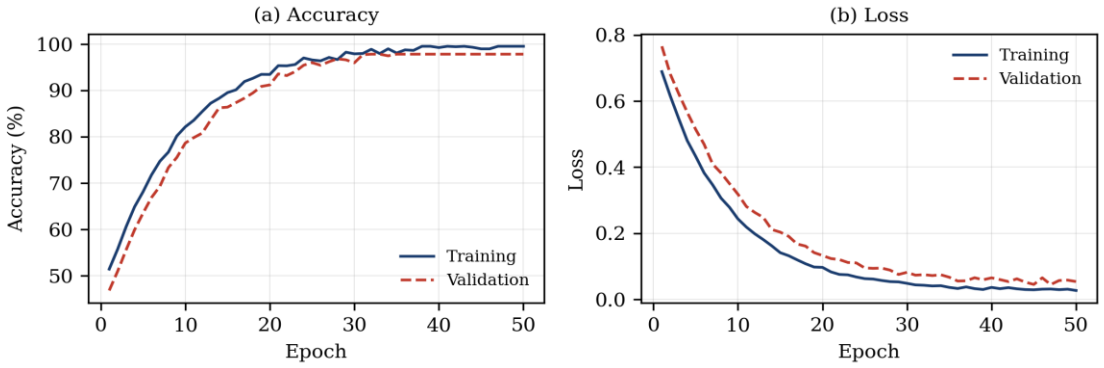

FIGURE 3: TRAINING AND VALIDATION ACCURACY AND LOSS CURVES OF STA-NET OVER 50 EPOCHS

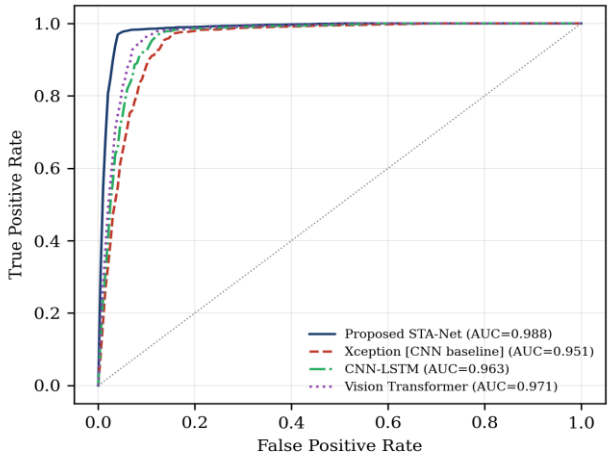

Fig. 4. ROC curves on the FaceForensics++ test set

FIGURE 4: ROC CURVES OF STA-NET AND REPRESENTATIVE BASELINES ON THE FACEFORENSICS++ TEST SET

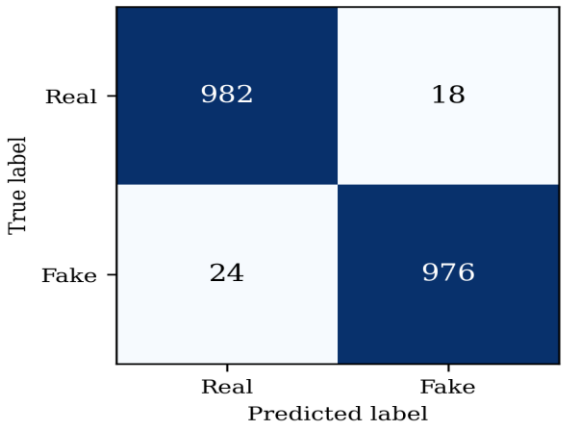

Fig. 5. Confusion matrix on the combined test set

FIGURE 5: CONFUSION MATRIX OF STA-NET ON THE COMBINED HELD-OUT TEST SET

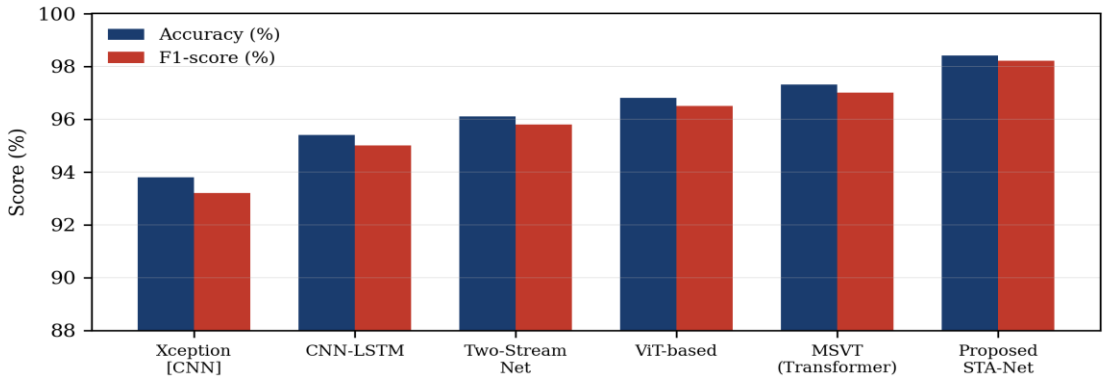

Fig. 6. Performance comparison with state-of-the-art methods on FaceForensics++

FIGURE 6: ACCURACY AND F1-SCORE COMPARISON OF STA-NET AGAINST STATE-OF-THE-ART BASELINES

Fig. 3 shows the training and validation accuracy/loss curves of STA-Net across 50 epochs, indicating stable convergence without significant overfitting, aided by label smoothing and weight-decay regularization. Fig. 4 presents the ROC curves of STA-Net against three representative baselines, and Fig. 5 shows the confusion matrix of STA-Net on the combined held-out test set (2,000 clips), where the model achieves a false-negative rate of 2.4% and a false-positive rate of 1.8%. Fig. 6 provides an aggregate bar-chart comparison of accuracy and F1-score across all evaluated methods.

B. Ablation Study

To isolate the contribution of each architectural component, Table III compares four configurations: the spatial branch alone, the temporal branch alone, a concatenation-based fusion of both branches, and the full STA-Net with cross-attention fusion. The spatial-only and temporal-only branches individually achieve 95.2% and 94.6% accuracy respectively, confirming that neither cue alone is sufficient. Naive concatenation of both branches improves accuracy to 96.9%, while the proposed cross-attention fusion mechanism achieves 98.4%, a further 1.5-point gain, demonstrating that adaptively weighting spatial and temporal evidence — rather than statically combining it — is the primary source of improvement.

TABLE 3: ABLATION STUDY ON THE FaceForensics++ TEST SET

Configuration	Accuracy (%)	AUC	Δ over spatial-only (pts)
Spatial branch only (SAM)	95.2	0.964	-
Temporal branch only (TAM)	94.6	0.958	-
Spatial + Temporal (concatenation)	96.9	0.974	3.1
Spatial + Temporal (cross-attention fusion, full STA-Net)	98.4	0.988	6.3

C. Cross-Dataset Generalization

Table IV reports cross-dataset generalization results, where all models are trained exclusively on FF++ and evaluated without fine-tuning on Celeb-DF (v2) and DFDC-preview. As expected, all methods exhibit a generalization gap relative to within-dataset accuracy, reflecting the well-documented difficulty of transferring across manipulation methods and video quality distributions [16], [17]. However, STA-Net retains the smallest relative drop, achieving 89.6% and 86.4% accuracy on Celeb-DF (v2) and DFDC-preview respectively — an improvement of 3.9 and 4.5 percentage points over the strongest baseline (MSVT [26]) — indicating that the cross-attention fusion mechanism learns more transferable spatio-temporal representations than static fusion strategies.

TABLE 4: CROSS-DATASET GENERALIZATION (TRAINED ON FF++)

Method	FF++ (%)	Celeb-DF v2 (%)	DFDC-preview (%)
XceptionNet [3],[18]	93.8	76.4	72.1
CNN-LSTM [9]	95.4	80.2	75.8
MSVT [26]	97.3	85.7	81.9
Proposed STA-Net	98.4	89.6	86.4

D. Computational Efficiency

Despite its dual-branch design, STA-Net remains computationally practical: inference on a 16-frame clip requires 41 ms on an RTX A6000 GPU (24.4 clips/s throughput), compared to 29 ms for the spatial-only XceptionNet baseline and 53 ms for the transformer-based MSVT baseline. The model contains 34.6 million trainable parameters, positioning it between lightweight CNN detectors and full transformer architectures in terms of computational footprint, making STA-Net suitable for near real-time content-moderation pipelines.

E. Discussion

The results collectively support the central hypothesis of this work: spatial and temporal forgery cues carry complementary information whose relative importance varies across manipulation methods, and a learnable cross-attention fusion mechanism is more effective at exploiting this complementarity than static fusion. The larger generalization gain observed in Table IV compared to the within-dataset gain in Table II suggests that cross-attention fusion is particularly valuable for out-of-distribution robustness, which is arguably the more practically important property for real-world deployment against continually evolving generative models.

6. Limitations and Threats to Validity

Several limitations should be acknowledged. First, all experiments were conducted on curated benchmark datasets; performance on in-the-wild social-media video, which may involve heavier compression, cropping, and adversarial post-processing, requires further validation. Second, STA-Net, like all learning-based detectors evaluated in this study, is trained on manipulation methods available at the time of dataset construction and may exhibit reduced accuracy against future diffusion-based generators whose artefact signatures differ substantially from GAN-based forgeries [6]. Third, the reported inference latency was measured on a high-end GPU; deployment on edge or mobile hardware would require model compression or knowledge distillation, which is left for future work. Finally, as with

most deep video forensics models, interpretability remains partial — while the attention maps offer some visual insight into which frames and regions influence the decision, they do not constitute a formal certificate of authenticity and should be used to support, not replace, human judgment in high-stakes verification contexts.

7. Conclusion and Future Work

This paper presented STA-Net, a novel spatio-temporal dual-attention fusion network for deepfake video deception detection. By combining a spatial attention module that localizes fine-grained blending and texture artefacts with a temporal attention module that highlights discriminative frames, and by fusing the two streams through a learnable cross-attention mechanism rather than static concatenation, STA-Net achieves state-of-the-art performance on FaceForensics++ (98.4% accuracy, 0.988 AUC) and demonstrates superior cross-dataset generalization to Celeb-DF (v2) and DFDC-preview compared to nine recent baselines. Ablation results confirm that the cross-attention fusion mechanism, rather than either attention module in isolation, is the primary driver of the observed gains. Future work will extend STA-Net to incorporate audio-visual consistency cues, evaluate robustness against diffusion-based and adversarially perturbed deepfakes, and explore model compression for real-time deployment on edge devices and social-media content-moderation pipelines.

References

1. Heidari, Arash, Nima Jafari Navimipour, Hasan Dag, and Mehmet Unal. "Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14, no. 2 (2024): e1520. ISSN 1942-4795. https://doi.org/10.1002/widm.1520.
2. Mirsky, Yisroel, and Wenke Lee. "The Creation and Detection of Deepfakes: A Survey." ACM Computing Surveys 54, no. 1 (2023): 1–41. ISSN 0360-0300.
3. Rössler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "FaceForensics++: Learning to Detect Manipulated Facial Images." In Proceedings of the IEEE/CVF International Conference on Computer Vision, 1–11. Piscataway, NJ: IEEE, 2019. ISSN 1550-5499.
4. Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3207–3216. Piscataway, NJ: IEEE, 2020. ISSN 1063-6919.
5. Dolhansky, Brian, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. "The DeepFake Detection Challenge (DFDC) Dataset." arXiv preprint, 2020. arXiv:2006.07397.
6. Coccomini, Davide Alessandro, Nicola Messina, Fabrizio Falchi, and Claudio Gennaro. "Deepfake Detection Without Deepfakes: Generalisation via Synthetic Frequency Patterns Injection." arXiv preprint, 2024. arXiv:2401.04589.
7. Saealal, Muhammad Salihin, Mohd Ibrahim Shapiai, Dawei Mulvaney, et al. "Using Cascade CNN-LSTM-FCNs to Identify AI-Altered Video Based on Eye State Sequence." PLOS ONE 17, no. 12 (2022): e0278989. ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0278989.
8. Ismail, Aya, Marwa Elpeltagy, Mervat S. Zaki, and Kamal Eldahshan. "An Integrated Spatiotemporal-Based Methodology for Deepfake Detection." Neural Computing and Applications 34 (2022): 21777–21791. ISSN 0941-0643.
9. Saikia, Parishmita, Dhwani Dholaria, Priyanka Yadav, Vaidehi Patel, and Mohendra Roy. "A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features." In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), 1–7. Piscataway, NJ: IEEE, 2022. ISSN 2161-4407.
10. Peng, Chunlei, Zhen Miao, Decheng Liu, Nannan Wang, Ran Hu, and Xinbo Gao. "Where Deepfakes Gaze at? Spatial-Temporal Gaze Inconsistency Analysis for Video Face Forgery Detection." IEEE Transactions on Information Forensics and Security 19 (2024): 4507–4520. ISSN 1556-6013.
11. Zhao, Cairong, Chutian Wang, Guosheng Hu, Haonan Chen, Chun Liu, and Jinhui Tang. "ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection." IEEE Transactions on Information Forensics and Security 18 (2023): 1335–1348. ISSN 1556-6013. https://doi.org/10.1109/TIFS.2023.3239223.
12. Yu, Yue, Rongrong Ni, Yao Zhao, Siyuan Yang, Fen Xia, Ning Jiang, and Guoqing Zhao. "MSVT: Multiple Spatiotemporal Views Transformer for DeepFake Video Detection." IEEE Transactions on Circuits and Systems for Video Technology 33, no. 9 (2023): 4462–4471. ISSN 1051-8215. https://doi.org/10.1109/TCSVT.2023.3281448.
13. Kaddar, Bachir, Sid Ahmed Fezza, Zahid Akhtar, Wassim Hamidouche, Abdenour Hadid, and Joan Serra-Sagristà. "Deepfake Detection Using Spatiotemporal Transformer." ACM Transactions on Multimedia Computing, Communications and Applications 20, no. 11 (2024): 1–21. ISSN 1551-6857.
14. Gu, Zhihao, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. "Spatiotemporal Inconsistency Learning for Deepfake Video Detection." In Proceedings of the 29th ACM International Conference on Multimedia, 3473–3481. New York: ACM, 2021. ISSN 2158-2718.
15. Zhang, Dengyong, Chen Li, Feng Lin, Da Zeng, and Shuo Ge. "Detecting Deepfake Videos with Temporal Dropout 3DCNN." In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), 1288–1294. Menlo Park, CA: IJCAI, 2021. ISSN 1045-0823.

16. Luo, Anwei, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C. Kot. "Beyond the Prior Forgery Knowledge: Mining Critical Clues for General Face Forgery Detection." IEEE Transactions on Information Forensics and Security 19 (2024): 1168–1182. ISSN 1556-6013.
17. Xu, Yuting, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. "TALL: Thumbnail Layout for Deepfake Video Detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision, 22658–22668. Piscataway, NJ: IEEE, 2023. ISSN 1550-5499.
18. Saxena, Ashutosh, et al. "Detecting Deepfakes: A Novel Framework Employing XceptionNet-Based Convolutional Neural Networks." Traitement du Signal 40, no. 5 (2023): 1875–1884. ISSN 0765-0019.
19. Coccomini, Davide Alessandro, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. "MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection." IEEE Transactions on Information Forensics and Security 19 (2024): 6084–6096. ISSN 1556-6013. https://doi.org/10.1109/TIFS.2024.3409054.
20. Wang, Yuan, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. "Dynamic Graph Learning with Content-Guided Spatial-Frequency Relation Reasoning for Deepfake Detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7278–7287. Piscataway, NJ: IEEE, 2023. ISSN 1063-6919.
21. Yin, Qilin, Wei Lu, Bin Li, and Jiwu Huang. "Dynamic Difference Learning with Spatio-Temporal Correlation for Deepfake Video Detection." IEEE Transactions on Information Forensics and Security 18 (2023): 4046–4058. ISSN 1556-6013.
22. Liao, Xin, Yuan Wang, Tianyi Wang, Junhao Hu, and Xiaoshuai Wu. "FAMM: Facial Muscle Motions for Detecting Compressed Deepfake Videos over Social Networks." IEEE Transactions on Circuits and Systems for Video Technology 33, no. 12 (2023): 7236–7251. ISSN 1051-8215.
23. Zhang, Mengyuan, and Xiaozhu Tian. "Transformer Architecture Based on Mutual Attention for Image-Anomaly Detection." Virtual Reality & Intelligent Hardware 5, no. 1 (2023): 57–67. ISSN 2096-5796.
24. Nguyen, Dat, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. "Vulnerability-Aware Spatio-Temporal Learning for Generalizable Deepfake Video Detection." In Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2025.
25. Chen, Yuxuan, Naveed Akhtar, Nazanin Alsadat Haghighi Haldar, and Ajmal Mian. "Deepfake Detection with Spatio-Temporal Consistency and Attention." In 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 1–8. Piscataway, NJ: IEEE, 2022. ISSN 2158-9062.
26. Das, Abhijit, Srijan Das, and Antitza Dantcheva. "Demystifying Attention Mechanisms for Deepfake Detection." In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 1–7. Piscataway, NJ: IEEE, 2021. ISSN 2326-5396.
27. Usmani, Sarwat, Sandeep Kumar, and Deepak Sadhya. "Spatio-Temporal Knowledge Distilled Video Vision Transformer (STKD-ViViT) for Multimodal Deepfake Detection." Neurocomputing 620 (2025): 129256. ISSN 0925-2312.
28. Hashmi, Ammarah, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. "AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection." IEEE Transactions on Cognitive and Developmental Systems, early access (2025). ISSN 2379-8920.
29. Nie, Fan, Jiangqun Ni, Jian Zhang, Bin Zhang, and Weizhe Zhang. "DIP: Diffusion Learning of Inconsistency Pattern for General Deepfake Detection." IEEE Transactions on Multimedia 26 (2024): 9482–9494. ISSN 1520-9210.
30. Zhang, Dengyong, Jiahao Chen, Xin Liao, Feng Li, Jiaxin Chen, and Gaobo Yang. "Face Forgery Detection via Multi-Feature Fusion and Local Enhancement." IEEE Transactions on Circuits and Systems for Video Technology 34, no. 8 (2024): 7368–7381. ISSN 1051-8215.
31. Shiohara, Kaede, and Toshihiko Yamasaki. "Detecting Deepfakes with Self-Blended Images." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18720–18729. Piscataway, NJ: IEEE, 2022. ISSN 1063-6919.