Signal, Image and Video Processing (2026) 20:147
https://doi.org/10.1007/s11760-026-05191-y

ORIGINAL PAPER

STGT: Spatio-temporal graph transformer for robust deepfake detection using dynamic facial features

Ayushi Pandey[1] · **Arun Solanki**[1]

Received: 23 July 2025 / Revised: 8 January 2026 / Accepted: 7 February 2026
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2026

Abstract

With the rising number of cases of financial fraud and political manipulation around, deepfake creation is getting sophisticated and exponentially increasing, researchers need to come up with remarkable innovations rapidly to detect this synthetic media. As traditional methods like CNN and ViTs fail to capture the minute difference between facial regions. There are optical inconsistencies as they ignore motion, and they only analyse rigid regions, hence struggles with pose variations. To overcome these problems, Spatio-Temporal Graph Transformer (STGT) is proposed using spatio-temporal graphs. For graph conversion, facial features are extracted dynamically along with optical flow features using Farneback algorithm. Facial region features are represented with nodes which are connected by temporal edges. The graph is sent to a 4X graph transformer, which uses dynamic attention pooling instead of traditional pooling for final prediction. The experiment is conducted using FaceForensics++ dataset consisting 1000 real and 1000 fake videos. STGT achieves 99% accuracy, which is a boost in accuracy compared to other state-of-the-art algorithm for detecting deepfake videos. This approach achieves 0.99 AUC by simultaneously analysing geometric, motion and hierarchical features.

Keywords Deepfake · Graph transformer · Attention pooling · Gelu · FaceForensics++ · Optical flow

1 Introduction

Deepfake technology has evolved from a research curiosity to a societal threat, enabled by rapid advances in generative AI. Modern creation tools using GANs [6], VAEs [14], and diffusion models [11] can now produce synthetic media nearly indistinguishable from reality, with applications ranging from entertainment to malicious disinformation. Recent techniques like StyleGAN3 [13] achieve photorealistic face swaps, while GAN-based Neural Radiance Fields (GNeRF) [21] enable full facial reenactments without perfect lighting consistency.

These developments have serious implications. Deepfakes have been used in financial scams, political manipulation [23], and personal harassment, with detection becoming increasingly difficult [19]. Where early forgeries showed obvious flaws like blurred edges [1] or missing physiological signals [18], newer methods maintain consistent artifacts across videos [24] and preserve high-frequency details [2]. This escalation demands detection systems that combine technical sophistication with real-world applicability.

Existing detection approaches struggle with three key challenges. First, traditional methods relying on 68 facial landmarks [20] miss subtle texture anomalies and micro-expressions that reveal state-of-the-art forgeries [19]. Second, frame-by-frame CNN analysis [1, 18] fails to capture temporal inconsistencies like unnatural eye blinking [18] or implausible facial dynamics [8]. Third, current systems poorly integrate different detection clues - for example, when facial symmetry [25] appears normal but GAN fingerprints [16] or optical flow patterns [8] reveal manipulation. These limitations become acute with real-world complications: video compression [4], intentional evasion tactics [12], and emerging techniques like SimSwap [2]. Even robust baselines like FakeSpotter [27], which relies on layer-wise neuron behavior analysis. This fails to model spatial-temporal relationships between facial regions a limitation addressed by our framework.

✉ Arun Solanki
asolanki@gbu.ac.in

Ayushi Pandey
ayushi.pandey26@gmail.com

[1] Department of Computer Science, Gautam Buddha University, Greater Noida, Uttar Pradesh, India

Published online: 10 March 2026

Our Spatial-Temporal Graph Transformer (STGT) solves these problems through three innovations: (1) high-resolution 468 point facial graphs [20], (2) multi-scale feature extraction using EfficientNet-B3 [25], and (3) edge-aware graph attention [26] that learns relationships between facial regions. By unifying spatial and temporal analysis through transformer networks [7, 15, 28], STGT outperforms both landmark-based [22] and hybrid CNN-RNN [24] approaches across diverse deepfake types.

In summary, this paper makes three key contributions:

i. A dynamic facial region extraction pipeline that leverages 468-point MediaPipe landmarks and adaptive multi-scale cropping for enhanced artifact detection.
ii. A hybrid feature representation combining EfficientNet-B3 embeddings, optical flow-based motion features, and region-type embeddings for comprehensive deepfake analysis.
iii. An edge-aware Spatial-Temporal Graph Transformer that models both intra-frame spatial relationships and inter-frame temporal dependencies using transformer-based attention.

2 Literature review

The field of deepfake detection was systematically organized with the introduction of the FaceForensics++ benchmark by Rössler et al. (2019) [23], which established standardized evaluation protocols and demonstrated the significant impact of video compression on detection performance. This foundational work provided the community with critical insights into the challenges of real-world deployment by including multiple manipulation methods and realistic compression scenarios. The benchmark's multilevel dataset design enabled researchers to systematically evaluate detector robustness under varying quality conditions.

Early approaches focused on temporal analysis, beginning with Afchar et al. (2018) [1] introducing MesoNet, a compact CNN architecture specifically designed for facial video forgery detection. Their shallow network design focused on mesoscopic features that proved particularly effective for identifying artifacts in early deepfake generations. Güera et al. (2018) [8] developed their novel two-stream architecture combining convolutional networks for spatial feature extraction with recurrent networks for temporal sequence modeling, highlighting the importance of motion cues in distinguishing authentic from manipulated content. Their approach demonstrated how optical flow patterns could reveal inconsistencies in synthetic facial movements that static frame analysis might miss.

Sabir et al. (2019) [24] introduced a recurrent convolutional network that effectively captured manipulation artifacts persisting across video frames. Their architecture demonstrated how combining spatial and temporal processing could improve detection of subtle inconsistencies in synthetic media by analyzing both individual frame anomalies and their temporal evolution. The method's hierarchical feature aggregation showed particular effectiveness in identifying blending artifacts at facial boundaries. The field underwent a significant transformation with the introduction of transformer architectures by Dang et al. (2020) [3]. Their approach leveraged self-attention mechanisms to identify long-range dependencies and inconsistencies in videos, representing a paradigm shift from convolutional to attention-based processing in media forensics. The transformer's ability to model relationships between distant facial regions proved especially valuable for detecting global inconsistencies in high-quality deepfakes.

Recent advances have explored multimodal detection strategies. Liu et al. (2021) [19] developed a frequency domain analysis approach that complemented traditional spatial analysis by examining spectral artifacts in manipulated content. Their method revealed how GAN-generated faces often contain distinctive high frequency patterns that differ from natural images. Hussain et al. (2021) [12] conducted crucial research on adversarial robustness, systematically evaluating vulnerabilities in contemporary detection systems by developing targeted attack methods that could bypass state of the art detectors.

The current state of the field is represented by innovative hybrid architectures. Gu et al. (2021) [7] proposed a spatiotemporal inconsistency learning framework that analyzed both spatial artifacts and temporal anomalies in videos. While not using graph networks, their method effectively combined spatial and temporal attention mechanisms. Most recently, Ramadhani et al. (2024) [22] advanced video vision transformers by incorporating facial landmark guidance and self-attention mechanisms, creating a framework that maintained detection accuracy even under heavy compression by focusing on anatomically consistent facial dynamics.

3 Proposed architecture

The Spatial-Temporal Graph Transformer (STGT) introduces a novel deepfake detection framework that combines dynamic facial modelling with attention-based graph reasoning and hierarchical graph attention. Unlike traditional methods that rely on static facial landmarks or frame-by-frame analysis, the STGT constructs a spatio-temporal graph G where, $G = (V, E)$ where nodes $v_i \in V$ represent dynamically cropped regions (eyes, mouth, etc.) and edges $e_{ij} \in E$

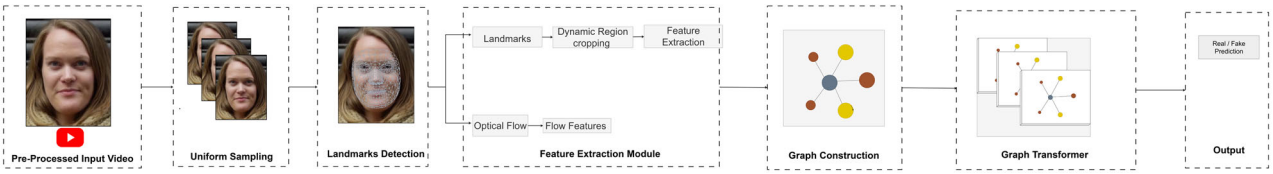

Fig. 1 Architecture of proposed model

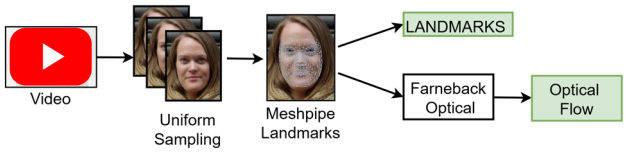

Fig. 2 Input module for facial region extraction

encode both anatomical relationships and temporal motion patterns that is spatial-temporal relationships.

The Proposed STGT architecture as illustrated in figure 1 can be broadly divided into three modules: 1) Input module 2) Processing module and 3) Output module. In input module video is received and cropped to extracted the facial region. The video extracted is put into the processing module where facial region is extracted, graph from the image is made and the graph made is put into graph transformer [28] for deepfake prediction. The facial region is extracted using landmark detection via mediapipe [20], dynamic region cropping and multi-scale feature extraction. From each video 30 frames are taken via uniform sampling which avoids bias towards high-motion segments. Landmarks and optical flow are extracted from frames as shown in figure 2. Each facial region is processed through a multi-scale feature extractor with EfficientNet-B3 to capture both high-level semantics (e.g., facial symmetry) and low-level artifacts (e.g., GAN-induced noise). EfficientNet-B3 was selected as the feature backbone due to its proven effectiveness in balancing model capacity and computational efficiency. This enables extraction of both semantic features and texture patterns essential for deepfake detection.

MediaPipe uses a lightweight CNN with a heatmap-based regression loss to detect 468 3D facial landmarks.

$$\mathcal{L}_{\text{landmark}} = \sum_{i=1}^{468} \|\hat{\mathbf{p}}_i - \mathbf{p}_i^*\|_2^2 + \lambda \|\Theta\|_2 \tag{1}$$

Where p_i^* are ground truth landmarks and θ are model parameters.

Landmarks extracted are further cropped using dynamic region cropping. This multi-scale cropping strategy ensures robustness to resolution variations and facial poses, allowing the model to capture artifacts at different granularities. From

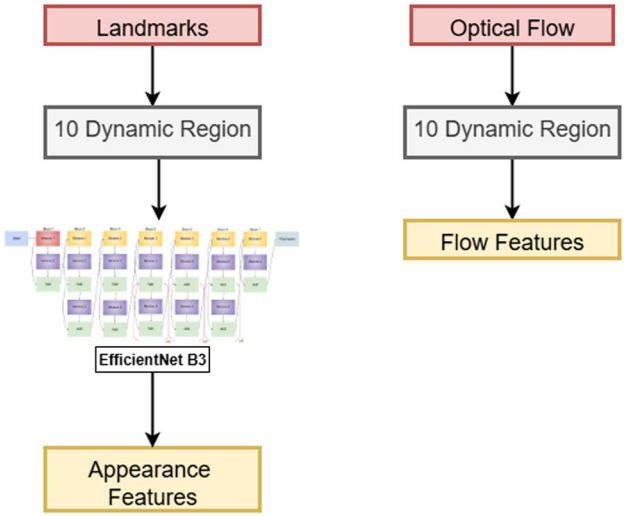

Fig. 3 Feature extraction process

fine-texture anomalies at smaller scales to structural inconsistencies at larger scales.This adaptive cropping mechanism implements a content-aware sampling strategy grounded in geometric measure theory. For each of the 10 facial regions R_k (eyes, mouth, etc.), BB_k is computed with boundary checks:

$$\text{BB}_k = (\mu_x \pm \alpha\sigma_x, \mu_y \pm \alpha\sigma_y) \tag{2}$$

where μ_x, μ_y are the mean coordinates of landmarks in R_k, σ_x, σ_ytheir standard deviations, and $\alpha = 1.8$ a scaling factor (empirically optimized). Further multi-scale feature extraction is done on each cropped region R_k which is resized to scales $0.8s, s, 1.2s$ (default s=224×224), then processed by EfficientNet-B3 [25]. The Scale invariance captures artifacts at different frequencies, low resolution $(0.8x)$ focuses on structural anomalies and high resolution $(1.2x)$ detects high-frequency GAN artifacts. Higher landmark density captures subtle artifacts (e.g., lip-sync errors) missed by [1, 16]. Adaptive cropping handles large pose variations better than fixed RoIs in [3, 22]. The multi-scale feature extraction further enhances stability, with the feature difference between scales bounded by

$$\|f_{s_1} - f_{s_2}\| \leq L_\phi(0.2s) + \epsilon_{\text{quant}} \tag{3}$$

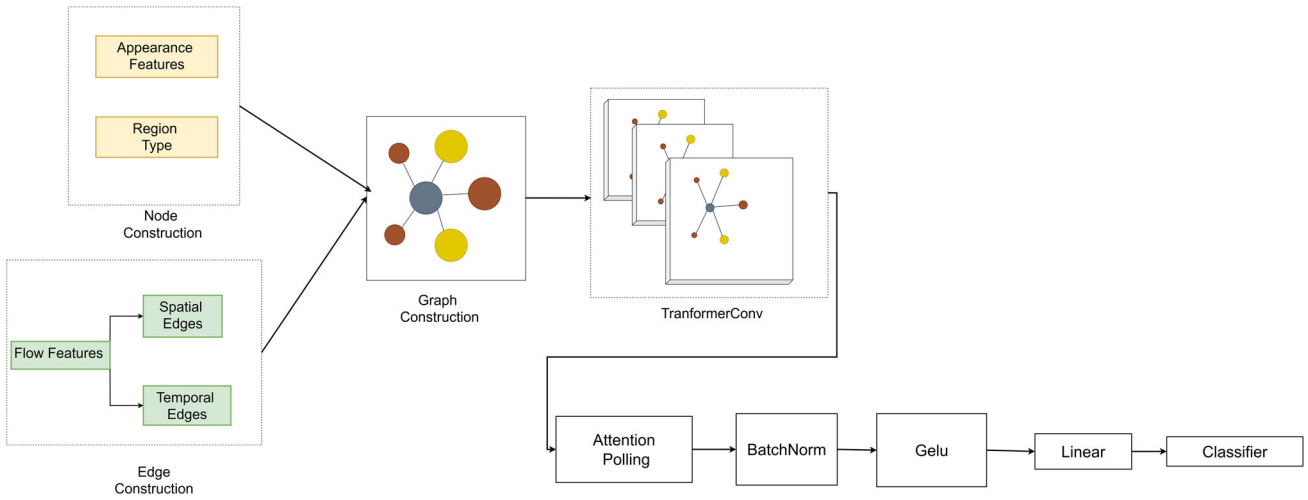

Fig. 4 Graph construction and classification

where L_ϕ is the Lipschitz constant of EfficientNet-B3, f_{s_1}, f_{s_2} are features extracted from the same facial region at different scales (0.8s, s,1.2s where s=224x224) and ϵ_{quant} is the quantization error from resizing/rescaling operations. Stability bound features ensures that the region extraction process is both precise and resilient to variations in facial pose and expression.

Optical flow using Farneback optical flow [5] is used to further enhances motion awareness by tracking unnatural facial movements between frames obtained via uniform sampling. It provides robust, efficient motion estimates by modeling pixel neighborhoods with polynomial expansions. This also balances accuracy and speed. Figure 3 explains how landmarks are converted to appearance features and optical flow into flow features which is further used to make graph.

Spatial-Temporal graph construction uses edge -conditioned attention to dynamically weigh relationships between facial regions. Figure 4 shows how graph is constructed and classification is done for the available data. A video is represented as a graph sequence G_t which is constructed to model both anatomical relationships and temporal dynamics. Video sequence $\{\mathcal{G}_t = (V_t, E_t)\}_{t=1}^{T}$ where nodes $V_t = \{v_i^t\}$ with $v_i^t = [\mathbf{f}_i^t \| \mathbf{m}_i^t \| \mathbf{r}_i]$ where f_i^t is appearance features and m_i^t is optical flow and r_i One hot region type (eye, mouth, etc.) and Edges $\epsilon_t = \epsilon_t^{\text{spatial}} \cup \epsilon_t^{\text{temporal}}$

$$\mathbf{m}_i^t = \left[\|F(\mathbf{p}_i)\|_2, \angle F(\mathbf{p}_i)\right] \in \mathbb{R}^2 \tag{4}$$

The edges are formed based on spatial proximity and temporal consistency. The spatial edges connect nodes within G_t via k-NN based on landmark distances and temporal edges are links between corresponding nodes across consecutive frames. In comparison to traditional fully connected graph attention networks [26], the proposed graph construction uses anatomically constrained edges that follow the natural connectivity of facial landmarks. This helps eliminate noise from irrelevant connections while preserving meaningful spatial relationships between facial regions, leading to more accurate representation of facial structure. The above graph construction is surpassed beyond motion-agnostic methods like the RNN based [8] by incorporating optical flow features into its node representations. This helps to capture dynamic facial movements and temporal inconsistencies. The expected node degree $\mathbb{E}[d]$ as a function of landmark spatial density λ and temporal correlation, providing insights into the graph's connectivity and robustness to occlusions. The framework ensures that the graph

construction process captures both static anatomical relationships and dynamic facial movements, which are critical for detecting subtle deepfake artifacts.

$$\mathbb{E}[d] = \pi r_{\text{spatial}}^2 + T \cdot P(\cos\theta > T_{\text{flow}}) \tag{5}$$

The proposed Spatial-Temporal Graph Transformer (STGT) architecture builds upon the constructed facial region graphs to detect deepfake artifacts through edge-conditioned attention and temporal message passing. Each transformer layer updates node [14] where attention weight a_{ij} are computed as:

$$a_{ij} = \text{softmax}_j\left(\frac{(\mathbf{W}_Q \mathbf{v}_i)^T(\mathbf{W}_K \mathbf{v}_j + \mathbf{W}_A \mathbf{a}_{ij})}{\sqrt{d}}\right) \tag{6}$$

where W_Q, W_K, W_V are learned projections and W_A is edge attribute projection. The TransformerConv layer provides three key benefits for deepfake [26]:

- Structural Awareness: By incorporating edge features E_{ij}, it can model both spatial (within-frame) and temporal (between-frame) relationships between facial regions.

- Dynamic Attention: The attention mechanism automatically learns which facial regions are most relevant for detecting manipulated artifacts, which may vary across different videos.
- Multi-head Processing: Using multiple attention heads allows the model to jointly attend to information from different representation subspaces, capturing diverse types of manipulation clues.

This architecture uses attention pooling [17] which offers several key advantages for deepfake detection. First, it enables adaptive importance weighting, allowing different facial regions to contribute dynamically to the final decision based on their detected relevance. This is particularly useful in deepfake analysis, where manipulated artifacts may appear inconsistently across different facial regions. Second, the attention weights (α_i) provide interpretability, as they indicate which regions (e.g., eyes, mouth, or jawline) were most influential in the classification decision, helping to identify telltale signs of manipulation. If α_i is high, node i strongly influences the deepfake detection decision and if α_i is low, the node contributes minimally. Finally, the pooling operation is permutation invariant, meaning it is unaffected by the ordering of nodes in the facial graph a crucial property for processing unstructured facial region data.

$$s_i = \sigma\left(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{h}_i + \mathbf{b}_1) + \mathbf{b}_2\right) \tag{7}$$

Where h_i is feature vector of node i, W_1, b_1 are weights and bias for the first transformation, $\tanh(\cdot)$ is hyperbolic tangent activation which reduces scores to $[-1, 1]$, W_2, b_2 are weights and bias for the second transformation, $\sigma(\cdot)$ sigmoid activation which maps score to [0, 1] and s_i is raw importance score for node i.

$$\alpha_i = \frac{\exp(s_i)}{\sum_j \exp(s_j)} \tag{8}$$

Where softmax normalizes score across all nodes to sum to 1 and α_i is final attention weight for node i.

$$\mathbf{h}_G = \sum_i \alpha_i \mathbf{h}_i \tag{9}$$

Where h_G is the final pool for the entire graph. Figure 4 also shows that the model uses Gaussian Error Linear Unit (GELU) [10] which is an activation function further enhances the model's capability compared to traditional ReLU [9]. Its smooth gradient facilitates better optimization, which is particularly important given the complex relationships in facial graphs. Unlike ReLU's hard thresholding, GELU provides a probabilistic interpretation, effectively acting as a stochastic gating mechanism that smoothly transitions between identity and zero suppression based on input magnitude. Empirically, GELU has been shown to outperform ReLU in transformer-based architectures, making it a natural choice for enhancing feature learning in deepfake detection tasks where subtle artifacts must be accurately modeled.

The final prediction is computed via hierarchical reasoning pipeline that combines spatial-temporal graph analysis with attention-based pooling.

$$\mathbf{y} = \text{MLP}(\text{READOUT}\{\mathbf{v}_i^t\}) \tag{10}$$

$$\text{READOUT} = \sum_{i,t} \beta_i \mathbf{v}_i^t \tag{11}$$

Where $\beta_i = \sigma(w^T v_i)$ is a learned attention weight which dynamically highlights manipulated regions (eg. Mouth for lip-sync errors) across frames, unlike standard mean/max pooling [17]. This is critical for interpretability, as β_i reveals which nodes contributed most to detection. The READOUT operates on spatial-temporal nodes v_i^t combining appearance, optical flow and region type which lacks in prior work [8, 26]. The proposed model is able to process graph in chunks, enabling 4x longer videos than [22].

Table 1 summarises how the STGT method has some key advantages over the traditional approaches. STGT offers more accurate, robust and interpretable deepfake detection.

4 Process flow of proposed work

The proposed Spatial-Temporal Graph Transformer (STGT) framework as shown in figure 5 follows a meticulously designed pipeline that systematically addresses the limitations of existing deepfake detection methods. The following steps are followed the multistage pipeline:

- **Step 1**: Video is input where frame by frame processing is done where face frame of video is extracted and uniform sampling of 30 frames is taken.
- **Step 2**: High resolution landmark extraction is done using mediapipe face mesh which generates 468 points per frames. This is done to capture subtle geometric distortions like asymmetric eye movements, unnatural lip contours.
- **Step 3**: The face is dynamically segmented into 10 regions like eyes, nose, cheeks etc with multiscale cropping (0.8s, s,1.2s) which ensures robustness among different resolutions and facial poses.
- **Step 4**: Using EfficientNet-B3 features are extracted from cropped face region which captures fine appearance like skin texture.
- **Step 5**:Farneback optical flow looks for motion anomalies like irregular blinking, mouth movement etc.

Table 1 Comparison between traditional methods and STGT

Feature	Traditional Methods	STGT (Proposed)
Landmark granularity	68 points (coarse)	468 points (fine-grained)
Motion modelling	Frame difference	Optical flow + temporal edges
Feature Fusion	Late fusion (post processing)	Early fusion (graph nodes)
Explainability	Black-box decisions	Attention maps for localization

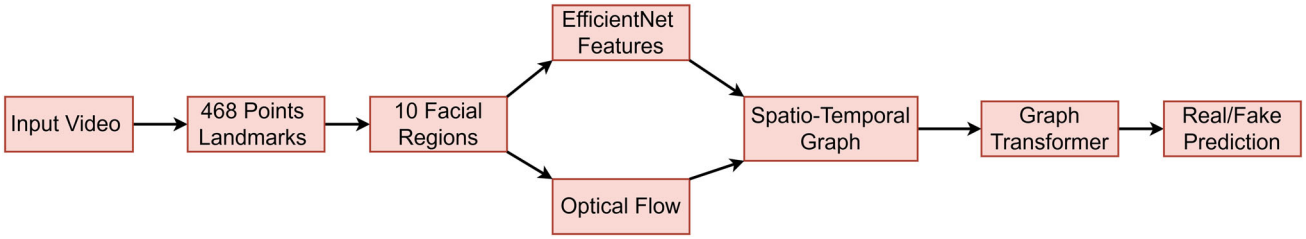

Fig. 5 Process flow of STGT framework

- **Step 6**:Graph is constructed using nodes and edges. Nodes are made from regions' features extracted and edges are made using spatial connection (proximity between eyes to nose) and temporal connection taken from optical flow which look for connection across frames.
- **Step 7**:The novel edge-conditioned graph transformer is used to analyse both intra-frame spatial relationships and inter-frame temporal dependencies which helps to find inconsistent facial lighting or unnatural micro expressing. Attention pooling is used instead of traditional pooling as it helps with dynamically finding important facial region. On other hand GELU is used instead of ReLU.
- **Step 8**:Classification is done which predicts real/fake prediction.

5 Result and analysis

The experiment is performed on FaceForensics++ [23] dataset which is a benchmark dataset used for deepfake detection. It contains 1000 real videos and 4000 manipulated videos generated with 4 different methods. For the experiment 1000 real and 1000 fake videos were taken but after preprocessing and graph generation 992 real and 996 fake videos were converted into graphs for graph transformer. The experiment is set up on Intel i5 processor, 32 GB ram, and NVIDIA GTX 1650 graphic cards. Dataset is divided into 70% train, 15% test and 15% validation and model is run for 100 epoch with early stopping at 48 epochs. Early stopping is a regularization technique used during training to prevent overfitting and optimize model performance.

Figure 6 illustrates that the proposed model demonstrates high performance, with a test accuracy of 99%. Additionally, the model achieves a recall of 100% and achived an F1-score of 99.01%, confirming its effectiveness in the provided dataset.

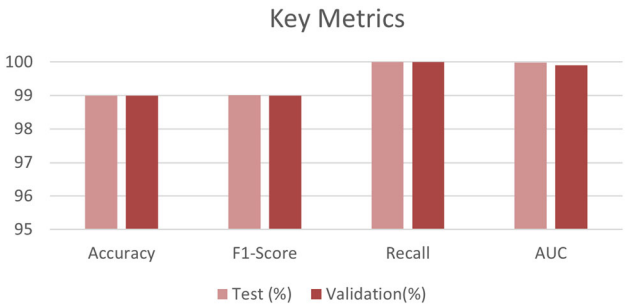

Fig. 6 Key metrics of the proposed model

Table 2 Key metrics of the proposed model for Test and Validation

Metric	Test (%)	Validation (%)
Accuracy	99.0	98.99
F1-Score	99.01	99.0
Recall	100	100
AUC	99.98	99.9

Table 2 shows the performance of the proposed model which gave exceptional performance in deepfake detection as evidenced above by validation and test metrics. With AUC being 0.99, it indicates an excellent classification performance on and the model is able to separate real and false cases. Model shows accuracy of 99% as compared to 90.27% of ViViT inspired model [22]. The near identical metrics across both sets confirm robust generalization without overfitting. The model demonstrated high recall across test scenarios, leaving no false negatives.

The ROC is used to plot True Positive Rate (Recall) against False Positive Rate at different classification threshold. The

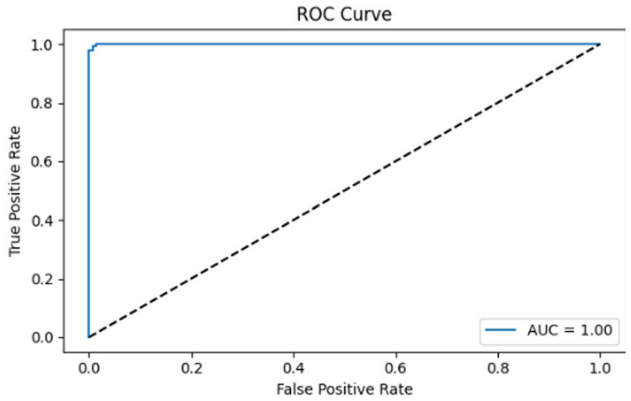

Fig. 7 ROC curve of proposed model

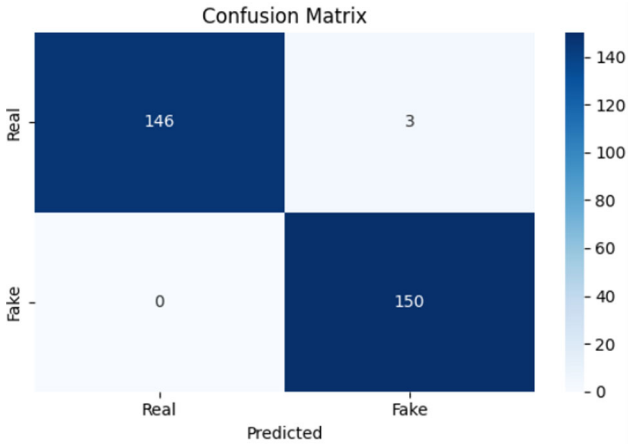

Fig. 8 Confusion matrix of proposed model

ROC curve shown in figure 7 validates the model and it also shows optimal balance between sensitivity and specificity. This near-perfect separability suggests that the model's attention mechanisms and GELU activated features effectively distinguish even subtle manipulation artifacts in facial regions. The confusion matrix as shown in figure 8 imply that true positives (deepfakes correctly flagged) and true negatives (real videos correctly classified) dominate, with false positives (real videos misclassified as fake) accounting for only 1% of cases. Due to 100% recall, there is no false negative cases. From confusion matrix it can be seen there are 146 true positives, 0 false negative, 3 false positives and 150 true negatives. This performance suggests that the model's architecture particularly the TransformerConv layers and attention pooling excels at capturing spatial-temporal inconsistencies unique to deepfake.

As seen from table 3 our STGT achieves superior performance compared to CNN based Xception by 3.9%, DenseNet by 5.5% and MesoNet by 15% on FaceForensics++ dataset. STGT model outperforms recent graph-based deepfake detector Ramadhani et al.'s landmark-guided ViT [22] by 8.73%. Unlike these works, STGT jointly optimizes

Table 3 Model accuracy comparison

Model	Accuracy (%)
XceptionNet [22]	95.10
DenseNet-121 [22]	93.50
MesoNet [22]	84.00
Ramadhani et al. [22]	90.27
STGT (Proposed)	99.00

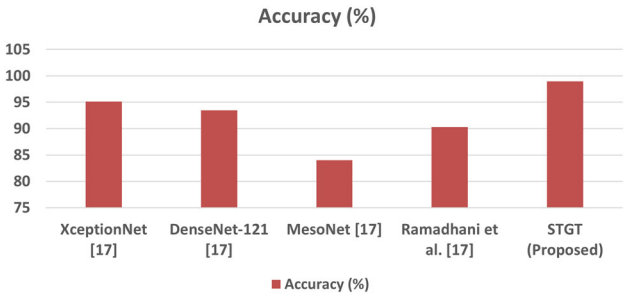

Fig. 9 Comparison between different models

spatial-temporal attention over facial regions, achieving near perfect separability. The gains highlight the importance of graph structured reasoning for capturing manipulation artifacts.

In figure 9 the graph visually confirms proposed model's superiority achieving the highest accuracy of 99% on faceforensics++ compared to CNN based models like Xception 95.1%, DenseNet 93.5%, MesoNet 84% and ViT based approach [22] 90.27%. The attention mechanism's interpretability allows practitioners to trace decisions back to influential facial regions (e.g., unnatural eye movements or mouth distortions), while GELU's smooth gradients likely stabilize training. In practice, such high performance positions the model as a viable tool for platforms needing reliable deepfake screening.

6 Conclusion

In conclusion, the Spatial-Temporal Graph Transformer establishes a new state-of-the-art in deepfake detection, as evidenced by its 99.00% accuracy and complementary metrics on the FaceForensics++ benchmark. The model's superior performance relative to both frequency-domain and transformer based alternatives stems from its unique integration of edge conditioned attention for anatomical consistency modeling, temporal graph reasoning for inter-frame anomaly detection, and multiscale feature fusion for comprehensive artifact capture. Beyond raw detection accuracy, the STGT offers practical advantages including computational efficiency from graph sparsity and improved interpretabil-

ity through anatomically grounded attention weights. These advances position the STGT as a robust solution for real world deepfake detection while providing a flexible framework for future extensions to emerging manipulation techniques and broader multimedia forensic applications.

7 Limitations and future work

While the STGT framework achieves state-of-the-art performance on the FaceForensics++ benchmark, several limitations must be acknowledged. The computational complexity of multiscale EfficientNetB3 feature extraction and graph transformer operations necessitates requires substantial resources, potentially limiting real-time deployment on constrained devices.Our evaluation primarily utilized FaceForensics++, which may not fully represent emerging deepfake techniques. Diffusion-based approaches and generalization across demographic groups require further investigation. Additionally, vulnerability to adversarial attacks targeting graph-based detection, Diffusion based approach and limited cross-dataset validation highlight needs for broader testing. Future work will develop lightweight STGT variants, extend evaluation to diverse datasets (DFDC, Celeb-DF) and emerging techniques, incorporate adversarial training for enhanced robustness, and explore real-time implementation strategies for practical applications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11760-026-05191-y.

Author Contributions A.P. conceptualization, methodology, Data curation, Analysis and visualisation, Writing - original draft , Writing-review and editing. A.S. Supervision, conceptualization, Data curation Writing- review and editing. All authors have read and agreed to the published version of the manuscript.

Funding This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Data Availability Data that support the finding of this study is openly available at https://github.com/ondyari/FaceForensics . Also research paper regarding the data is cited in the manuscript as "Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11). "

Declarations

Competing interests The authors declare no competing interests.

References

1. Afchar, D., Nozick, V., Yamagishi, J., et al.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS), IEEE, pp 1–7 (2018)
2. Chen, R., Chen, X., Ni, B., et al.: Simswap: An efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM international conference on multimedia, pp 2003–2011 (2020)
3. Dang, H., Liu, F., Stehouwer, J., et al.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, pp 5781–5790 (2020)
4. Dolhansky, B., Bitton, J., Pflaum, B., et al.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
5. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis, Springer, pp 363–370 (2003)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
7. Gu, Z., Chen, Y., Yao, T., et al.: Spatiotemporal inconsistency learning for deepfake video detection. In: Proceedings of the 29th ACM international conference on multimedia, pp 3473–3481 (2021)
8. Güera, D., Delp, EJ.: Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, pp 1–6 (2018)
9. He, J, Li, L., Xu, J., et al.: Relu deep neural networks and linear finite elements. arXiv preprint arXiv:1807.03973 (2018)
10. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
12. Hussain, S., Neekhara, P., Jere, M., et al.: Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3348–3357 (2021)
13. Karras, T., Aittala, M., Laine, S., et al.: Alias-free generative adversarial networks. Adv. Neural. Inf. Process. Syst. **34**, 852–863 (2021)
14. Kingma, DP., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Kipf, T.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
16. Korshunov, P., Marcel, S.: Deep-fakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
17. Lee, J., Lee, I., Kang, J.: Self-attention graph pooling. In: International conference on machine learning, pmlr, pp 3734–3743 (2019)
18. Li, Y., Chang, MC., Lyu, S.: In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877 (2018)
19. Liu, H., Li, X., Zhou, W., et al.: Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 772–781 (2021)
20. Lugaresi, C., Tang, J., Nash, H., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
21. Meng, Q., Chen, A., Luo, H., et al.: Gnerf: Gan-based neural radiance field without posed camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6351–6361 (2021)
22. Ramadhani, K.N., Munir, R., Utama, N.P.: Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self attention. IEEE Access **12**, 8932–8939 (2024)
23. Rossler, A., Cozzolino, D., Verdoliva, L., et al.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1–11 (2019)

24. Sabir, E., Cheng, J., Jaiswal, A., et al.: Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) **3**(1), 80–87 (2019)
25. Tan, M.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 pp 6105–6114 (2019)
26. Veličković, P., Cucurull, G., Casanova, A., et al.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
27. Wang, R., Juefei-Xu, F., Ma, L., et al.: Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. arXiv preprint arXiv:1909.06122 (2019)
28. Yun, S., Jeong, M., Kim, R., et al.: Graph transformer networks. Adv. Neural. Inf. Process. Syst. **32**, (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.