ISTVT:

Interpretable Spatial-Temporal Video Transformer for Deepfake Detection

Group Members

Jaykant-230506

Abhinay Teja-231044

Raghubir-230823

JLS Laharii-230510

Overview

The Importance of Deepfake Detection

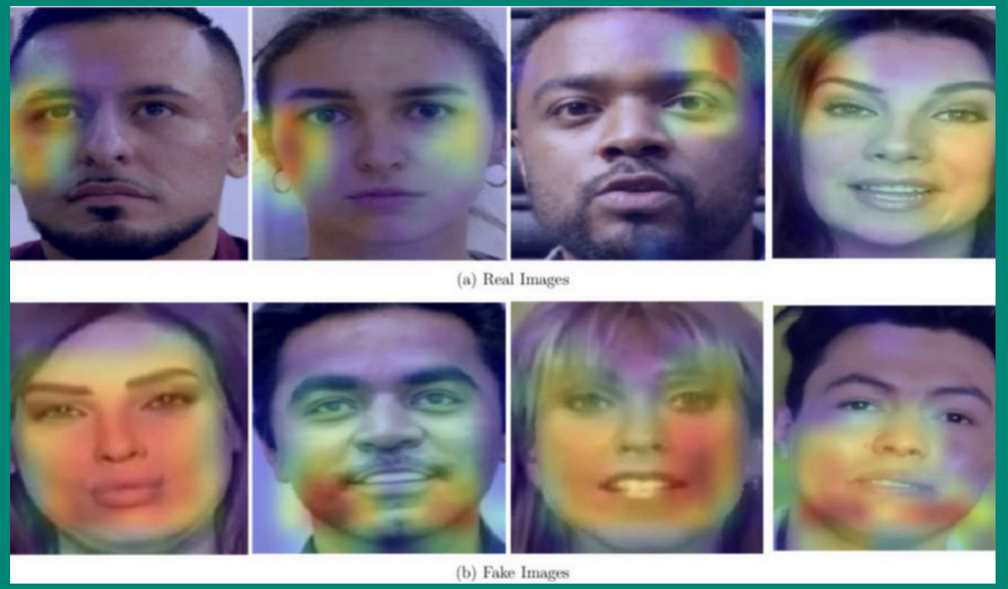
- In the recent years, the development of deepfake synthesis has been increasing at a rapid rate.
- Deepfakes can mimic real people with high accuracy, increasing concern in cybersecurity, finance and many other sectors.
- Can be used to falsely represent leaders, causing a serious threat to national security.

Deepfake Detection Methods

- **Frame-based :** These methods take a single frame as input and mainly focus on spatial artifacts generated by the forgery process (e.g. blurred edge).
- **Video-based :** The video-based methods take a frame sequence as the input and try to external temporal artifacts(e.g. inconsistant structure between frames).
- Mainly Deepfake synthesis methods are frame-based, therefore the video-based methods can potentially capture the temporal inconsistansis.

Why ISTVT?

- Frame-based methods lack temporal awareness and perform poorly in cross-dataset tests.
- Sabir et al.’s CNN + RNN Combination approach was shown to be ineffective in DFDC 2020.
- Li et al.’s regard faces in each frame as instances and proposed multi instance learning mechanisms.
- And many more methods are introduced but interpretability of these models is limited.



As shown above, Even real images receive strong attention, showing that existing models struggle to cleanly interpret manipulation cues highlighting the need for better spatial-temporal disentanglement as done in ISTVT.

What is ISTVT?

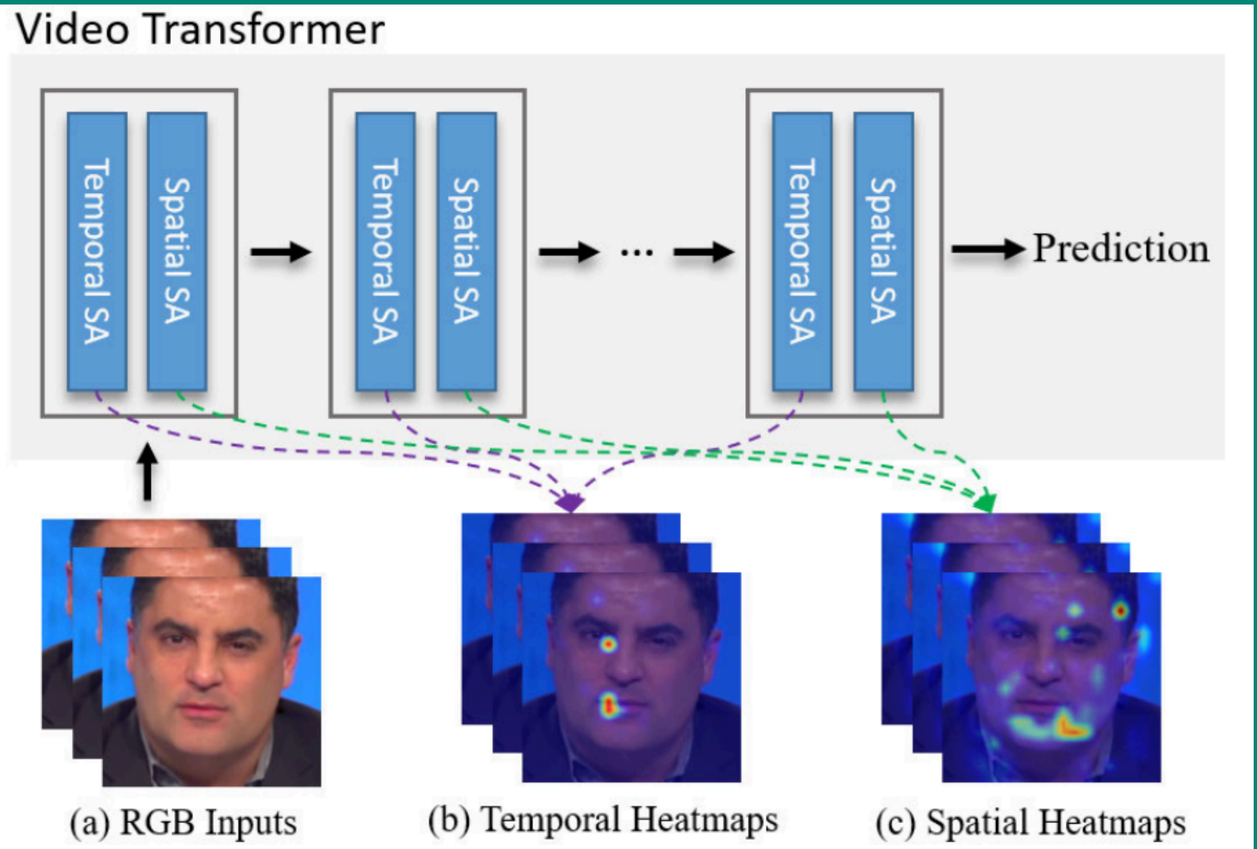


Fig. 1. A overview of our work. Our Interpretable Spatial-Temporal Video Transformer (ISTVT) decomposes the self-attention along two dimensions (spatial and temporal). Our ISTVT can also interpret the model decision-making process by visualizing the discriminative areas separately from spatial and temporal dimensions.

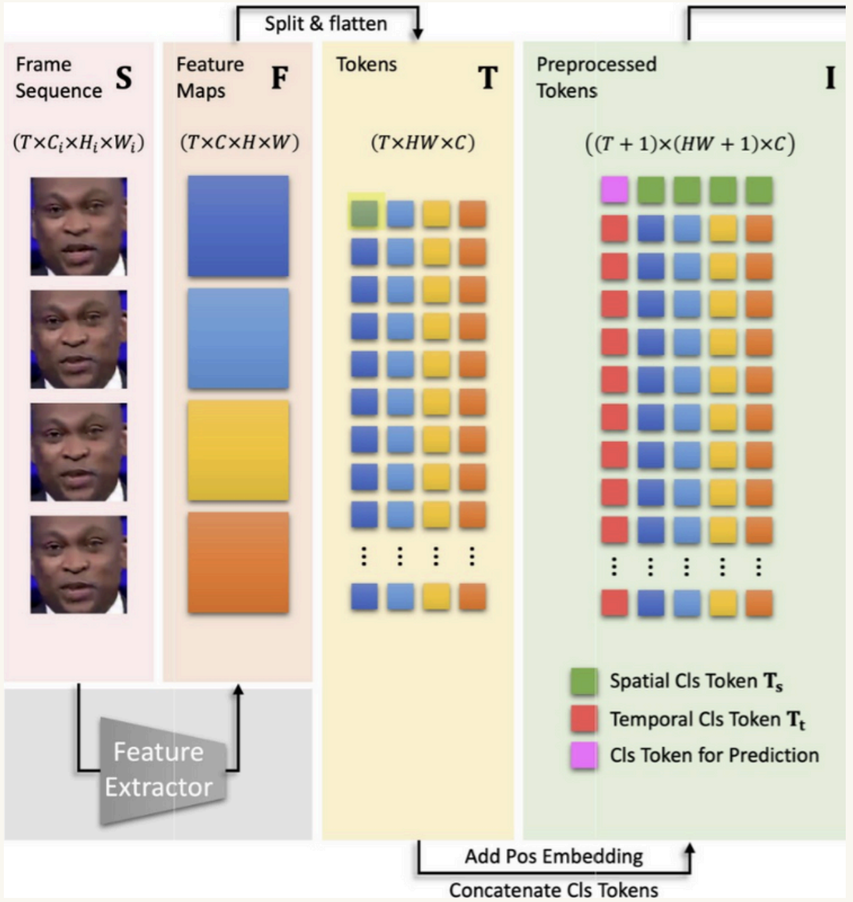
- ISTVT is a novel model specifically designed to detect Deepfake videos by capturing both spatial and temporal inconsistencies.
- Unlike frame-by-frame approaches, ISTVT takes in video sequences ($T \times C \times H \times W$) to learn inter-frame relationships(where T, C, H, W denote the length of the sequences, number of the image channels, frame height, and frame width respectively).
- It provides separate visualizations for spatial and temporal attention, offering insights into how and why the model makes its predictions.

Methodology

Feature Extraction with Xception blocks:

The proposed ISTVT is composed of a feature extractor based on Xception (a CNN architecture).

- Input: Frame sequence $\mathbf{S} \in \mathbb{R}^{T \times C_i \times H_i \times W_i}$
- Output: Feature maps $\mathbf{F} \in \mathbb{R}^{T \times C \times H \times W}$,where C, H, W denote the number of the feature map channels, feature map height, and feature map width.
- Feature maps are split into *1×1 patches* (tokens).



Tokenization and Positional Encoding:

The tokens $\mathbf{T} \in \mathbb{R}^{T \times HW \times C}$ are then concatenated with spatial classification tokens $\mathbf{T_s} \in \mathbb{R}^{T \times 1 \times C}$ and temporal classification tokens $\mathbf{T_t} \in \mathbb{R}^{1 \times (HW+1) \times C}$ successively, and a learnable position embedding is added.

The preprocessed token tensor $\mathbf{I} \in \mathbb{R}^{(T+1) \times (HW+1) \times C}$ then flows into M spatial temporal transformer blocks.

Decomposed Spatial Temporal Transformer:

- Instead of vanilla attention, ISTVT splits attention into spatial and temporal parts that helps in increasing robustness and accuracy.
- The input tensor of a self-attention block is projected to the query, key, and value features Q, K, and V respectively via the linear projection layers. These features are then split into different heads at the latest dimension, thus its shape becomes (T + 1) × (HW + 1) × N × D, where N is the number of heads and D = C/N.
- Temporal Self-Attention: Operates across frames, per spatial location j, Captures inter-frame inconsistencies (e.g., blinking artifacts).
- Spatial Self-Attention: Operates within each frame, across all patches, Captures spatial manipulation cues (e.g., texture blur, edge inconsistencies).

$$\mathbf{O}^{t}_{(:,j,:,:)} = softmax\left(\frac{\mathbf{Q}_{(:,j,:,:)} \cdot \mathbf{K}^{\top}_{(:,j,:,:)}}{\sqrt{D}}\right) \cdot \mathbf{V}_{(:,j,:,:)} \quad (1)$$

$$\mathbf{O}^{s}_{(k,:,:,:)} = softmax\left(\frac{\mathbf{Q}_{(k,:,:,:)} \cdot \mathbf{K}^{\top}_{(k,:,:,:)}}{\sqrt{D}}\right) \cdot \mathbf{V}_{(k,:,:,:)} \quad (2)$$

where equation 1 and 2 represent the outputs of the temporal and spatial self attention at spatial index j and temporal index k respectively.

Compared with the vanilla self-attention, the proposed spatialtemporal self-attention reduces the computational complexity of matrix multiplication from $\mathcal{O}(T^2H^2W^2)$ to $\mathcal{O}(T^2+H^2W^2)$

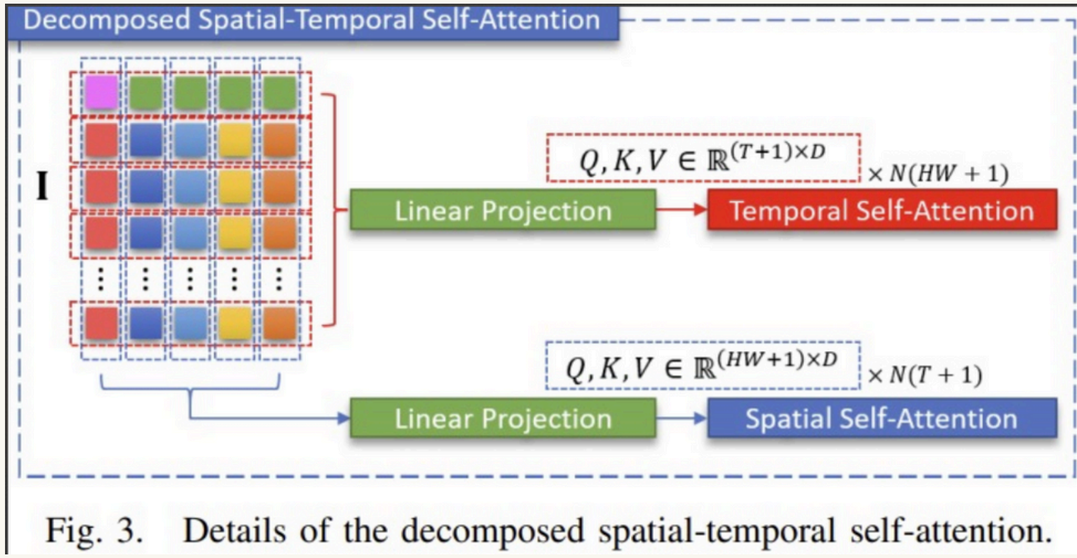


Fig. 3. Details of the decomposed spatial-temporal self-attention.

Self Subtraction Mechanism:

To encourage the temporal self-attention to focus more on the inter-frame distortion and reduce the redundant changeless features, self-subtract mechanism is applied to the input tokens before the projection to queries and keys for temporal self-attention.

- Computes frame differences to highlight dynamic changes:

$$\mathbf{I}' = cat((\mathbf{I}_{(0:2,:,:)}, \mathbf{I}_{(2:,:,:)} - \mathbf{I}_{(1:-1,:,:)}), dim = 0)$$

- To preserve the important spatial information in the tokens fed to the spatial self-attention, we project I' to the queries and keys, while values are the projections of the original I.
- By this way, it is possible to study more discriminative temporal information (e.g. inter-frame inconsistency) while keeping important spatial artifacts.

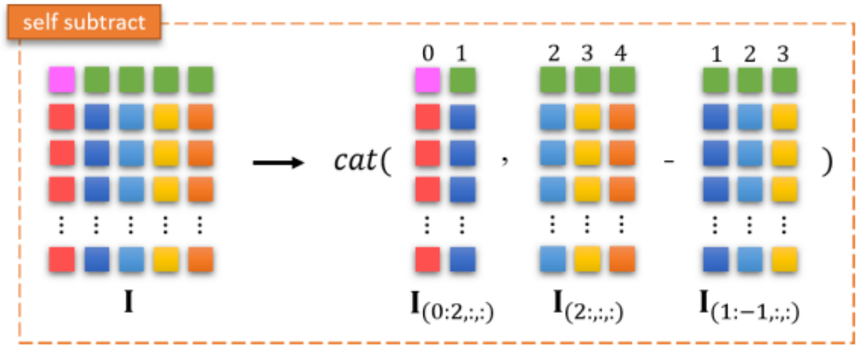
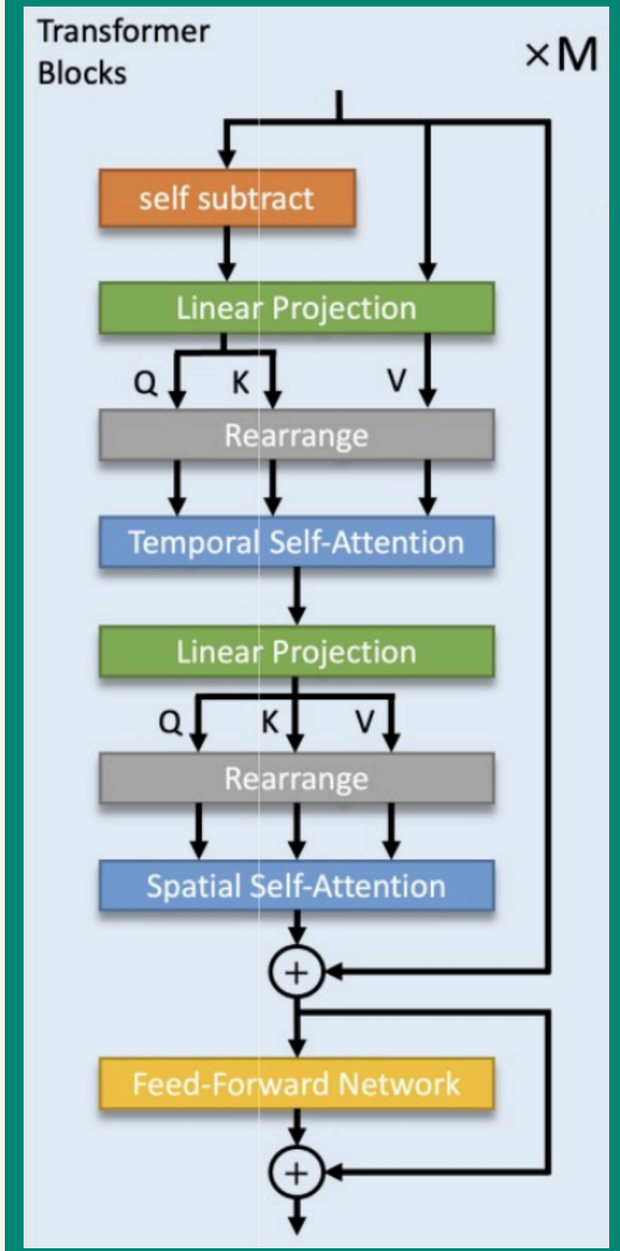


Fig. 4. Details of the proposed self-subtract mechanism.

- The complete spatial-temporal transformer module in ISTVT is a combination of M spatial-temporal transformer blocks.



Model Interpretability:

- Deepfakes are highly realistic, even humans struggle to detect them. Hence, Interpretability helps us understand the model's decision-making.
- ISTVT enables separate interpretation of spatial and temporal attention due to decomposed self-attention.
- Uses Layer-wise Relevance Propagation (LRP) with Deep Taylor Decomposition.
- The relevance $R_t^{(m)}$, $R_s^{(m)}$ of the temporal and spatial self-attention modules in each m^{th} transformer block (m = 1, 2, . . . , M) of shape N × (HW + 1)×(T +1)×(T +1) and N ×(T +1)×(HW +1)×(HW +1) are computed firstly.
- The outputs U_d (d ∈ {t,s}, denotes temporal or spatial) for the visualization are defined by:

$$\bar{\mathbf{A}}_{d(i,:,:)}^{(m)} = \mathbf{I} + max\left(\mathbb{E}_h(\mathbf{R}_{d(:,i,:,:)}^{(m)} \circ \nabla\mathbf{A}_{d(:,i,:,:)}^{(m)}), 0\right)$$
$$\mathbf{U}_d^{(i,:,:)} = \bar{\mathbf{A}}_{d(i,:,:)}^{(1)} \cdot \bar{\mathbf{A}}_{d(i,:,:)}^{(2)} \cdot \ldots \cdot \bar{\mathbf{A}}_{d(i,:,:)}^{(M)}$$

$$\nabla\mathbf{A}_d^{(m)} = \nabla softmax\left(\frac{\mathbf{q}_d^{(m)} \cdot \mathbf{k}_d^{(m)\top}}{\sqrt{D}}\right)$$

Where, I denotes the identity matrix, ◦ indicates the Hadamard product, · indicates the matrix multiplication, E_h indicates the mean across the head dimension.

- U_d is computed on each spatial position and temporal position i simultaneously for the temporal and spatial self-attentions respectively.
- At last considering only the updated attention weights of classification tokens the shapes of Ut and Us become HW × 1 × T and T × 1 × HW respectively, then rearranged to the shape of T × H × W, and upscale them to the original input size T × Hi × Wi via a bilinear interpolation step to achieve the final visualization heatmaps.

Results and Visualizations:

1. Datasets: The experiments use five major deepfake datasets: FaceForensics++, FaceShifter, DeeperForensics, Celeb-DF, and DFDC. These datasets contain real and high-quality fake videos generated by various known and unknown manipulation techniques.
2. Implementation:

- The preprocessing involves detecting and aligning faces using MTCNN and facial landmarks, cropping frames to 300×300, and using 6-frame sequences for training.
- Features are extracted using the Xception network, producing 19×19×728 maps split into patches. The model is trained with SGD using a warm-up learning rate strategy over 100 epochs on 4 Tesla V100 GPUs.

3. Detection Performance:

- The proposed ISTVT model outperforms existing video and CNN-based methods in intra-dataset Deepfake detection by effectively capturing temporal and spatial artifacts, especially on challenging datasets like DFDC.
- ISTVT demonstrates strong cross-dataset generalization, clearer interpretability, and superior performance on unseen deepfakes by focusing on short-term temporal inconsistencies and leveraging video-based temporal modeling.

4. Robustness to Perturbations:
   - ISTVT shows greater robustness to JPEG compression than baseline methods, thanks to its decomposed spatial-temporal self-attention and self-subtract mechanism.
   - It outperforms VTN and frame-based methods under heavy downsampling by effectively capturing temporal artifacts through its self-subtract mechanism.
   - ISTVT remains effective under spatial dropout perturbations by leveraging preserved temporal artifacts, while frame-based methods like Xception fail when key spatial cues are lost.
5. Ablation Study:
   - Ablation studies show that ISTVT performs best with decomposed attention (temporal first), short input sequences (T=6), and moderate depth (M=12), balancing accuracy, robustness, and efficiency for Deepfake detection.

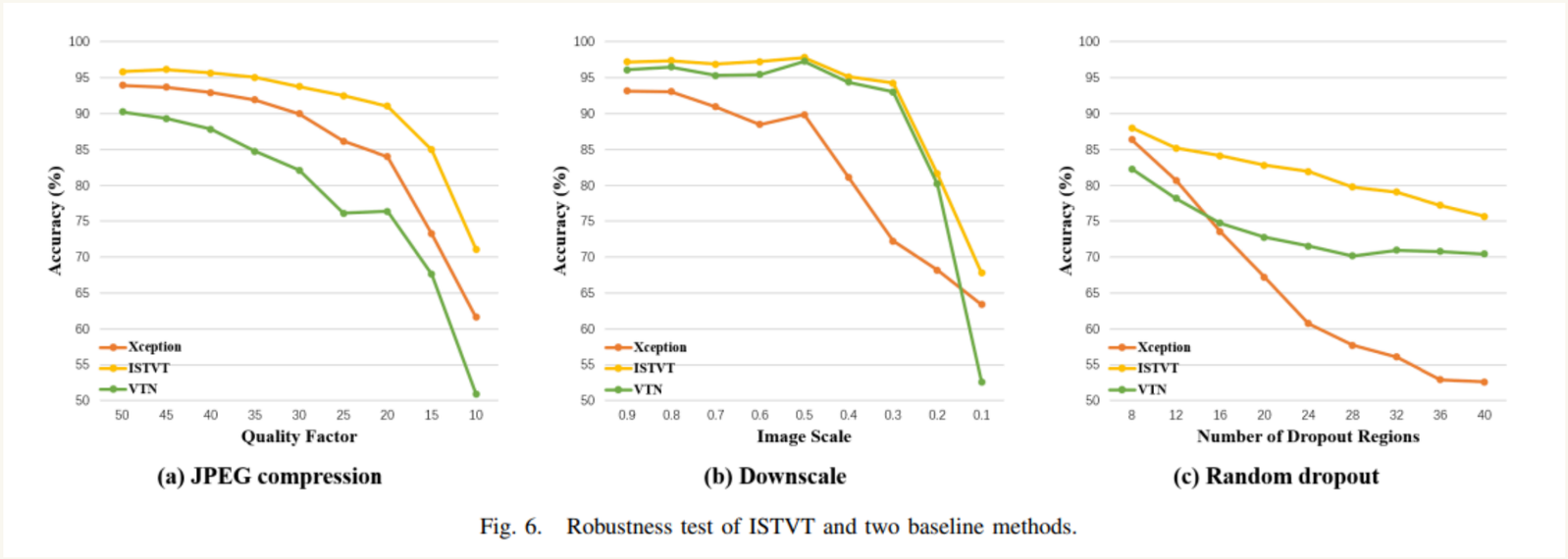
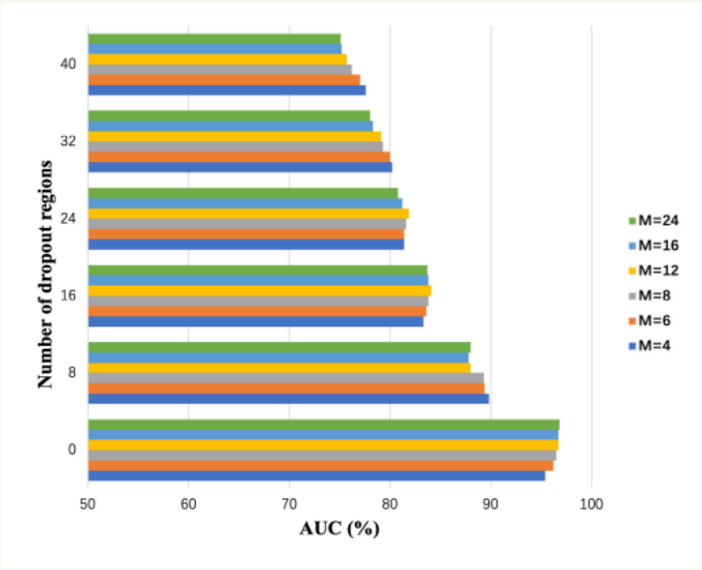


Fig. 6. Robustness test of ISTVT and two baseline methods.



Interpretability by Visualization:

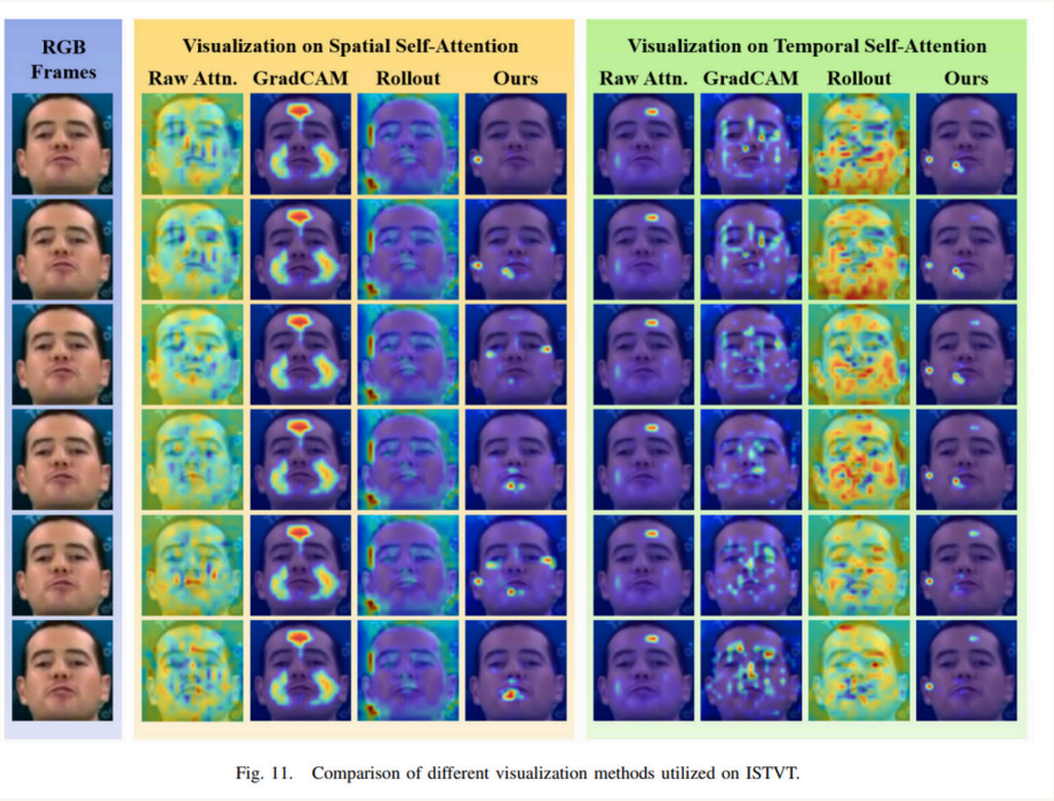


Fig. 11. Comparison of different visualization methods utilized on ISTVT.

- ISTVT's visualization highlights that spatial attention focuses on artifacts like blending edges, while temporal attention targets moving areas (e.g., lips, jaw), capturing inter-frame inconsistencies.
- The self-subtract mechanism helps ignore consistent irrelevant patterns (e.g., lighting artifacts), improving robustness and preventing misclassification.
- Common visualization methods (Raw Attention, GradCAM, Rollout) fail to clearly interpret both spatial and temporal attention, whereas the proposed method provides more accurate and focused results.

- Dropping just 10% of top-importance pixels (as identified by ISTVT's method) reduces accuracy significantly, proving its superior explanation quality compared to GradCAM or Rollout.
- The proposed visualization method offers meaningful insight into ISTVT's decision-making and surpasses existing techniques in both clarity and relevance.
- This visualization strategy can be extended to other spatial-temporal tasks to improve interpretability and understanding of model behavior.

Conclusion

This paper introduces ISTVT, an interpretable spatial-temporal video transformer for Deepfake detection, using decomposed self-attention and a self-subtract mechanism to enhance performance and robustness. A novel visualization method is proposed to separately interpret spatial and temporal features, improving model transparency. The method is generalizable to other video transformers.

Thank You...