ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection

Cairong Zhao, Chutian Wang, Guosheng Hu, *Senior Member, IEEE*, Haonan Chen, Chun Liu, *Member, IEEE*, and Jinhui Tang, *Senior Member, IEEE*

***Abstract*— With the rapid development of Deepfake synthesis technology, our information security and personal privacy have been severely threatened in recent years. To achieve a robust Deepfake detection, researchers attempt to exploit the joint spatial-temporal information in the videos, like using recurrent networks and 3D convolutional networks. However, these spatial-temporal models remain room to improve. Another general challenge for spatial-temporal models is that people do not clearly understand what these spatial-temporal models really learn. To address these two challenges, in this paper, we propose an Interpretable Spatial-Temporal Video Transformer (ISTVT), which consists of a novel decomposed spatial-temporal self-attention and a self-subtract mechanism to capture spatial artifacts and temporal inconsistency for robust Deepfake detection. Thanks to this decomposition, we propose to interpret ISTVT by visualizing the discriminative regions for both spatial and temporal dimensions via the relevance (the pixel-wise importance on the input) propagation algorithm. We conduct extensive experiments on large-scale datasets, including FaceForensics++, FaceShifter, DeeperForensics, Celeb-DF, and DFDC datasets. Our strong performance of intra-dataset and cross-dataset Deepfake detection demonstrates the effectiveness and robustness of our method, and our visualization-based interpretability offers people insights into our model.**

***Index Terms*— Deepfake detection, video transformer, deep learning interpretability.**

Manuscript received 3 July 2022; revised 18 October 2022 and 5 December 2022; accepted 2 January 2023. Date of publication 23 January 2023; date of current version 31 January 2023. This work was supported in part by the National Natural Science Fund of China under Grant 62076184, Grant 61976158, Grant 61976160, and Grant 62076182; in part by the Shanghai Innovation Action Project of Science and Technology under Grant 20511100700; in part by the Shanghai Natural Science Foundation under Grant 22ZR1466700; and in part by the Fundamental Research Funds for the Central Universities and the State Key Laboratory of Integrated Services Networks, Xidian University. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Bin Li. *(Cairong Zhao and Chutian Wang contributed equally to this work.) (Corresponding author: Cairong Zhao.)*

Cairong Zhao and Chutian Wang are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: zhaocairong@tongji.edu.cn).

Guosheng Hu is with Oosto, 38330 Belfast, U.K. (e-mail: huguosheng100@gmail.com).

Haonan Chen is with Alibaba Group, Hangzhou 310052, China (e-mail: haolan.chn@alibaba-inc.com).

Chun Liu is with the College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China (e-mail: liuchun@tongji.edu.cn).

Jinhui Tang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jinhuitang@njust.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2023.3239223, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2023.3239223

I. Introduction

IN RECENT years, videos created by face manipulation methods, i.e. Deepfake videos, have raised great threats to our privacy security. These fake videos can be created easily via the open source software (e.g. Deepfakes [1], DeepFaceLab [2]). Therefore, it is important to develop an effective and robust Deepfake detection method. Fortunately, a lot of Deepfake detectors [3], [4], [5], [6], [7], [8] have been devised and they have achieved success on the large-scale Deepfake datasets [9], [10], [11].

Generally speaking, existing Deepfake detection methods can be divided into frame-based and video-based methods. The frame-based methods [4], [12], [13], [14] take a single frame as the input and mainly focus on the spatial artifacts generated by the forgery process (e.g. blurred edge, inconsistent textures, etc.). The video-based methods take a frame sequence as the input and try to extract temporal artifacts (e.g. inconsistent textures and structures between the frames) in the fake videos. Although the frame-based methods have achieved great success on various datasets and competitions like Deepfake Detection Challenge [11], these methods have the following critical drawbacks: (1) Frame-based methods do not perform well on the latest Deepfake video forgery manipulations (e.g. NeuralTextures [15]) because these manipulations have made great progress in improving visual effects and imitating realistic facial structures. (2) Frame-based methods have a great performance drop on the low quality videos [9], where the spatial details are blurred. (3) Frame-based methods may overfit to the spatial features generated by a specific Deepfake synthesis method, leading to the weak generalization capacity [5]. To overcome these drawbacks, in recent years, researchers start to investigate the video-based methods, aiming to exploit the joint spatial-temporal information [5], [6], [16], [17], [18]. Another motivation for the use of video-based methods is that current Deepfake synthesis methods are all frame-based, the consistency and harmony between the frames are not well-preserved during the forgery. Therefore, the video-based methods can potentially capture this temporal inconsistency, leading to more robust Deepfake detection.

However, current video-based Deepfake detection methods, like directly using the C3D [19] or LSTM [20] do not achieve promising results (experimental results and comparison see in Section IV-C.1). Concurrently, video-based transformers have made great progress in other computer vision tasks,

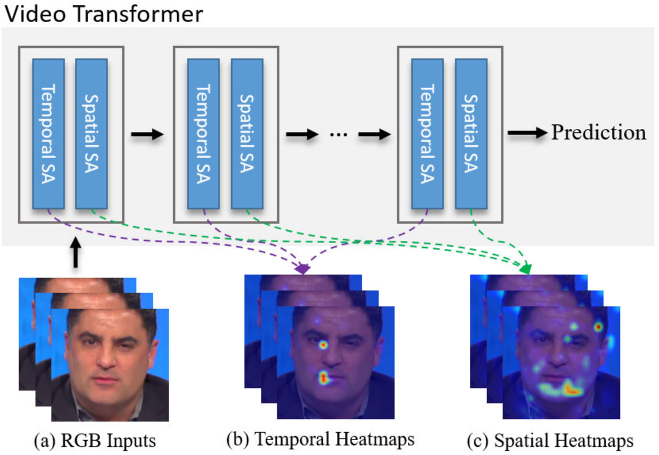


Fig. 1. A overview of our work. Our Interpretable Spatial-Temporal Video Transformer (ISTVT) decomposes the self-attention along two dimensions (spatial and temporal). Our ISTVT can also interpret the model decision-making process by visualizing the discriminative areas separately from spatial and temporal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

like action recognition [21], [22], [23], [24], [25], video object detection [26], [27], [28], [29]. Not surprisingly, the attempt of employing the video transformers for Deepfake detection has been done by Khan and Dai [6], but they use the vanilla form of the vision transformer [30], leading to the high computational complexity and the lack of interpretability. Motivated by these video transformers and addressing Deepfake challenges, we propose a novel Interpretable Spatial-Temporal Video Transformer (ISTVT) architecture and its corresponding interpretability method, illustrated in Fig. 1. Our ISTVT can (1) effectively conduct the Deepfake detection and (2) interpret the ways of self-attention detecting Deepfake. Specifically, for (1), we systematically analyzed different variants of spatial-temporal self-attention, and propose to use a decomposed spatial-temporal self-attention mechanism in each transformer block of ISTVT. Based on this decomposition, a novel self-subtract mechanism is proposed to force the model to capture the inter-frame temporal inconsistency of Deepfake videos. With the decomposition and self-subtract mechanism, ISTVT can effectively capture the subtle hints of Deepfake videos from both spatial (e.g. blurred edge) and temporal (cross-frame inconsistency) dimensions. For (2), motivated by relevance propagation algorithm [31], we propose a model interpretation method to visualize the temporal and spatial heatmaps (i.e. (b) and (c) in Fig. 1) separately, offering the interpretability in both temporal and spatial dimensions within a transformer.

Our contributions can be summarized as:

- We propose a novel Interpretable Spatial-Temporal Video Transformer model (ISTVT) for Deepfake detection. ISTVT decomposes a self-attention module into spatial and temporal ones and then an innovative self-subtract mechanism is proposed to guide the decomposed temporal self-attention to effectively detect temporal artifacts. Thus, our ISTVT is very accurate and robust.
- Motivated by the relevance propagation algorithm [31], we propose a new interpretation method that can visualize the class-discriminative heatmaps for the decomposed spatial and temporal self-attentions. To our knowledge, this is the first work to investigate transformer interpretability along spatial and temporal dimensions separately for not only Deepfake detection but other computer vision tasks in general.
- We conduct extensive experiments on various Deepfake datasets including FaceForensics++ [9], FaceShifter [32], DeeperForensics [33], Celeb-DF [10] and DFDC [11] datasets. The comparison and visualization results demonstrate the effectiveness and robustness of our method.

II. Related Work

A. Video-Based Deepfake Detection

With the rapid development of Deepfake synthesis methods, fake videos have become difficult to discriminate by using only spatial information. Also, frame-based methods have a great performance drop on the cross-dataset tests. To address these challenges, researchers start to investigate video-based methods for more accurate and more robust Deepfake detection. Early works, like [17] and [34] propose hand-crafted features based on facial landmarks for Deepfake detection. However, limited by the accuracy of landmark detection, their performance and robustness are not promising. Another type of video-based Deepfake detection solution is based on general video analysis models like C3D [19], I3D [35], and LSTM [20]. But these models do not work well on the Deepfake detection task and their performance is even worse than the frame-based methods [5]. Recently, some specialized spatial-temporal methods have been proposed for video-based Deepfake detection. As an early attempt, Sabir et al. [36] combine a convolution network (CNN) and a recurrent network (RNN) for Deepfake detection. However, the combination of CNN and RNN is proved not effective in Deepfake detection in the Deepfake Detection Challenge (DFDC) in 2020. Therefore, this scheme is gradually abandoned by the researchers. Li et al. [18] regard faces in each frame as instances and propose a multi-instance learning mechanism. This method achieves very promising performance and is robust to video compression. Yang et al. [37] propose to exploit lips movement to prevent Deepfake attack on speaker authentication systems. Khan and Dai [6] first employ the video transformer to detect Deepfake videos. They also introduce an incremental learning strategy to generalize various manipulation methods. However, their transformer architecture is not optimized for Deepfake detection and thus has no significant performance advantage over traditional frame-based solutions. Gu [5] propose to use a well-designed module to learn the spatial-temporal inconsistency for Deepfake detection. FTCN [38] uses a 3D fully-convolutional network to extract spatial-temporal information and a temporal transformer to extract long-range temporal information. Although some of these video-based solutions achieve promising performance improvement compared with the frame-based ones. The separate interpretations of temporal and spatial information are rarely mentioned. In [39], Pino et al. find that current explainers (e.g. GradCAM [40], SHAP [41]) can produce explanations of Deepfake with promising inter-frame consistency and centredness. In addition, Peng et al. [42] propose a method to obtain a more human-friendly interpretation by enhancing original suspect images. However, the spatial and temporal information is not handled separately in current methods. Thus, it is difficult to separately explain or visualize what spatial or temporal information is extracted by the models. Therefore, the interpretability of these methods is limited, which is not conducive to the development of Deepfake detection.

B. Interpretability of Video Transformers

Although vision transformers have made great progress in vision tasks recently, the visualization of vision transformers, especially video transformers, is rarely discussed. A trivial solution is to directly regard the self-attention scores as the visualization results. But this attempt losses most of the meaningful information in the transformer components, thus the visualization results are not ideal. Another approach is using the GradCAM [40] and its variants via regrading the token sequence as the feature maps in convolutions networks. Chefer et al. [31] propose an advanced visualization method for transformer networks based on LRP [43] and Deep Taylor Decomposition [44]. They derive the relevance propagation formula of self-attention, skip-connection, and layer-normalization blocks in the transformers based on the theory in [44]. The visualization quality of this method is significantly better than raw-attention methods and GradCAM-based methods. For video transformers, the specialized visualization mechanism is rarely mentioned. In [22], Zhang et al. use rollout [45] to visualize the spatial attention and spatial-temporal attention, but the separate temporal attention has not been discussed. Because of the increased dimension, we have to consider the spatial and temporal attention respectively to better interpretability. A trivial solution is applying the GradCAM-based methods or rollout [45], as mentioned, directly to the video transformers. But these methods are not suitable for the visualization of transformers according to previous works [31]. Since the proposed decomposition of spatial and temporal self-attention make the separate interpretation of spatial and temporal information becomes possible, the generalization of the well-performed LRP-based method to our video transformer is worth discussing.

III. Methodology

For the video-based Deepfake detection, the inputs are frame sequences with the shape of $T \times C_i \times H_i \times W_i$, where T, C_i, H_i, W_i denote the length of the sequences, number of the image channels, frame height, and frame width respectively. To our knowledge, current public released Deepfake synthesis methods are frame-based, thus the inter-frame relation is not considered in the Deepfake videos, potentially introducing temporal artifacts. According to previous works [5], [16], [34], learning these artifacts can significantly improve the performance and robustness. In contrast to using complex structures (e.g. SIL and TIL blocks proposed in [5]) or hand-craft features in previous works (e.g. rPPG used in [16]), we devise a novel Interpretable Spatial-Temporal Video Transformer (ISTVT) structure for general, effective, and interpretable learning of the spatial-temporal information. The proposed ISTVT is composed of a feature extractor based on Xception [46], a video transformer network with decomposed spatial-temporal self-attention, and an MLP network for prediction. Finally, a visualization method is proposed to visualize and interpret the temporal and spatial self-attention respectively. This work can help us learn more about the internal principle of video-based Deepfake detection, and thus is meaningful to the research community.

A. Network Architecture

The overall structure of ISTVT is shown in Figure 2. Since Deepfake detection mainly focuses on the low-level texture information [3], [14], we use a tiny convolutional network composed of several Xception blocks [46] (i.e. the entry flow of Xception network) to extract textural features. Specifically, given a frame sequence $\mathbf{S} \in \mathbb{R}^{T \times C_i \times H_i \times W_i}$, we feed all the frames to the Xception blocks and the generated feature maps are of shape $\mathbf{F} \in \mathbb{R}^{T \times C \times H \times W}$, where C, H, W denote the number of the feature map channels, feature map height, and feature map width. The feature maps $\mathbf{F}$ are then split into 1×1 patches and flattened into vectors, each vector represents C feature patterns at different spatial positions. We directly use them as the input tokens without linear projection. 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. Following the previous works [23], [30], we add a learnable position embedding to the tokens. The preprocessed token tensor $\mathbf{I} \in \mathbb{R}^{(T+1) \times (HW+1) \times C}$ then flows into M spatial-temporal transformer blocks. The output of the transformer blocks $\mathbf{O}$ keeps the same shape with $\mathbf{I}$, and the classification token at the first temporal position and first spatial position of the output tokens (i.e. $\mathbf{O}_{(0,0,:)}$) is used for final prediction via the prediction head (i.e. an MLP network). The proposed ISTVT is trained by the binary cross-entropy (BCE) classification loss since Deepfake detection is a binary classification task. Next, we introduce the details of the spatial-temporal transformer block.

B. Spatial-Temporal Transformer Block

According to the procedure of Deepfake manipulation algorithms, the frames are forged separately, thus the inter-frame relation is not considered during the generation of fake videos. Therefore, the spatial artifacts (e.g. blurred edges and textures, abnormal facial structures) between the frames are independent of each other. This fact inspires us to decompose the spatial and temporal self-attentions in the transformer. Hence we propose a decomposed spatial-temporal self-attention to construct the transformer blocks. In detail, the input tensor of a self-attention block is projected to the query, key, and value features $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ respectively via the linear projection layers. These features are then split into different heads at the latest dimension, thus its shape becomes

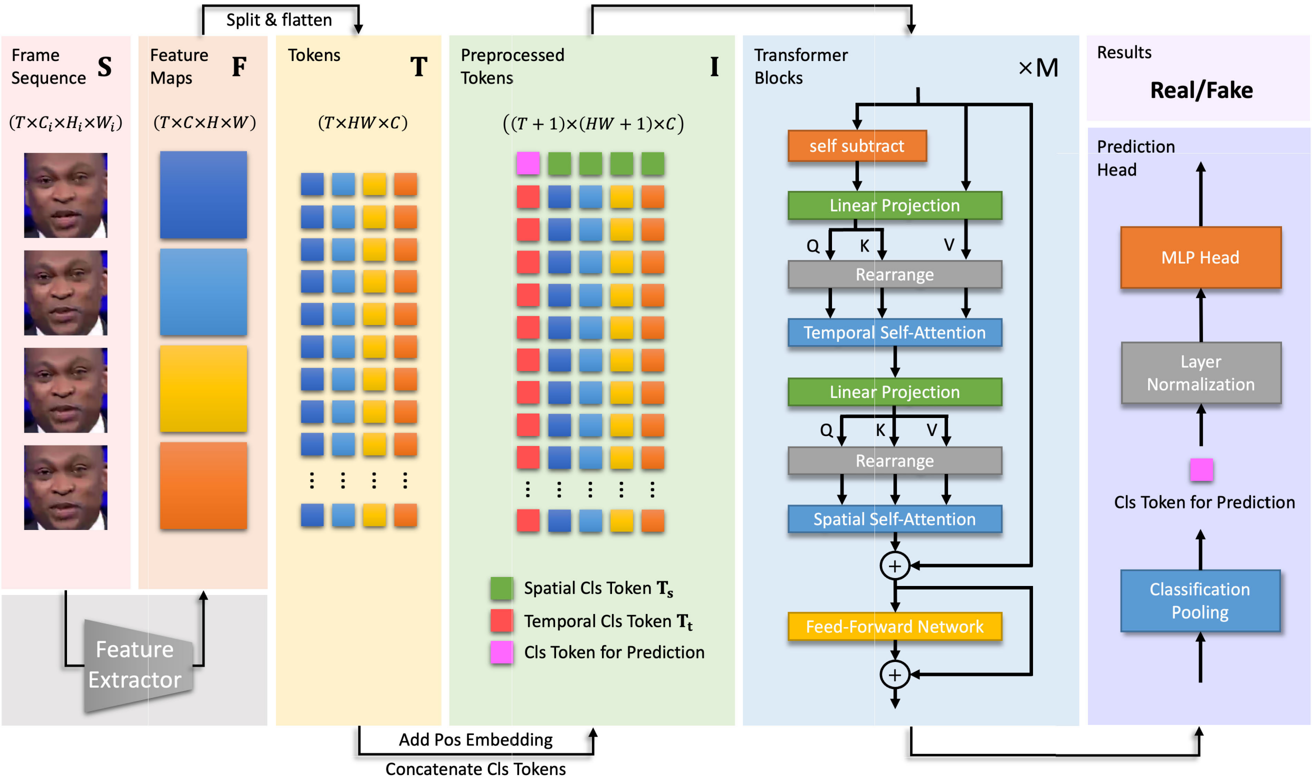


Fig. 2. The architecture of the proposed ISTVT. The feature extractor extracts the texture features from the input sequence and generates the feature maps. We split the feature maps into patches and flatten them to form the token sequence. The token sequence is then concatenated to the classification tokens and added with a position embedding. A spatial-temporal video transformer, consisting of several decomposed spatial-temporal blocks, takes the preprocessed tokens as inputs and outputs the results.

$(T+1) \times (HW+1) \times N \times D$, where N is the number of heads and $D = \frac{C}{N}$. We then conduct self-attention on the spatial and temporal dimensions respectively. Specifically, as shown in Fig. 3, for the temporal self-attention, computation is conducted on each spatial position j (i.e. patches of the same location among all the frames) simultaneously, formulated as:

$$\mathbf{O}^t_{(:,j,:,:)} = softmax(\frac{\mathbf{Q}_{(:,j,:,:)} \cdot \mathbf{K}^\top_{(:,j,:,:)}}{\sqrt{D}}) \cdot \mathbf{V}_{(:,j,:,:)} \tag{1}$$

where $\mathbf{O}^t_{(:,j,:,:)}$ denotes the output of the temporal self-attention at spatial index j.

For the spatial self-attention, computation is conducted on each temporal position k (i.e. all the patches in each frame) simultaneously, formulated as:

$$\mathbf{O}^s_{(k,:,:,:)} = softmax(\frac{\mathbf{Q}_{(k,:,:,:)} \cdot \mathbf{K}^\top_{(k,:,:,:)}}{\sqrt{D}}) \cdot \mathbf{V}_{(k,:,:,:)} \tag{2}$$

Similarly, $\mathbf{O}^s_{(k,:,:,:)}$ denotes the output of the spatial self-attention at temporal index k. In the actual implementation shown in Fig. 2, the query, key, and value features are rearranged to the shape of $N \times (HW+1) \times (T+1) \times D$ and $N \times (T+1) \times (HW+1) \times D$ for the temporal and spatial self-attentions respectively. Then we simply conduct matrix multiplications at the last two dimensions simultaneously. Compared with the vanilla self-attention, the proposed spatial-temporal self-attention reduces the computational complexity of matrix multiplication from $\mathcal{O}(T^2H^2W^2)$ to $\mathcal{O}(T^2+H^2W^2)$.

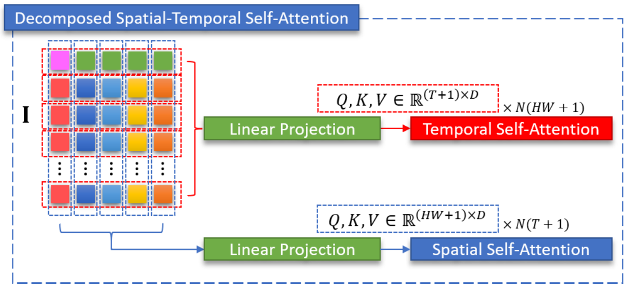

Decomposed Spatial-Temporal Self-Attention
I
$Q, K, V \in \mathbb{R}^{(T+1)×D}$
$× N(HW+1)$
Linear Projection
Temporal Self-Attention
$Q, K, V \in \mathbb{R}^{(HW+1)×D}$
$× N(T+1)$
Linear Projection
Spatial Self-Attention

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

C. Self-Subtract Mechanism

To encourage the temporal self-attention to focus more on the inter-frame distortion and reduce the redundant changeless features, we apply a self-subtract mechanism to the input tokens before the projection to queries and keys for temporal self-attention. This operation generates feature residuals to replace the original features and can help the network to ignore the redundant information and focus on discriminative inconsistent temporal features. Specifically, as shown in Fig. 4, for an input tensor $\mathbf{I} \in \mathbb{R}^{(T+1)\times(HW+1)\times C}$ of the temporal self-attention module, we subtract the tokens that are adjacent at the temporal dimension (i.e. have an adjacent index on dimension 0) except for the classification token, formulated as:

$$\mathbf{I}' = cat((\mathbf{I}_{(0:2,:,:)}, \mathbf{I}_{(2:,:,:)} - \mathbf{I}_{(1:-1,:,:)}), dim = 0) \tag{3}$$

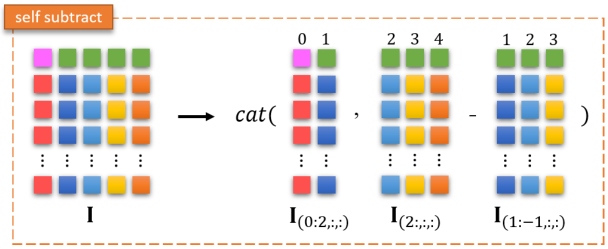


Fig. 4. Details of the proposed self-subtract mechanism.

where $cat((\mathbf{A},\mathbf{B}), dim = d)$ denotes concatenating tensors $\mathbf{A}$ and $\mathbf{B}$ at d-th dimension. To preserve the important spatial information in the tokens fed to the spatial self-attention, we project $\mathbf{I}'$ to the queries and keys, while values are the projections of the original $\mathbf{I}$. By this way, it is possible to study more discriminative temporal information (e.g. inter-frame inconsistency) while keeping important spatial artifacts. This simple but effective mechanism can significantly improve the Deepfake detection performance and robustness. We demonstrate this by experimental results (Sec. IV-E) and visualization results (Sec. V).

At last, residual connections, layer normalization, and a feed-forward network are also employed to construct the transformer block following the previous works [23], [30]. The complete spatial-temporal transformer module in ISTVT is a combination of M spatial-temporal transformer blocks.

D. Model Interpretability

Model interpretation is generally important for understanding the decision-making of deep models. For Deepfake detection, the synthetic frames are very realistic, and humans cannot even distinguish that. Thus it is important to investigate the interpretability to understand how our model makes decisions.

Benefitting from the decomposed self-attention mechanism, we can interpret our model from two dimensions (temporal and spatial) via visualizing the discriminative and salient areas. In [31], the visualization method based on LRP [43] and Deep Taylor Decomposition [44] has achieved success on the image-based vision transformer. Motivated by this, we generalize this from image-based transformer to the video transformer. Following the relevance propagation rule of linear, residual connection, add, layer normalization and self-attention layers proposed in [31] and [44], the relevance $\mathbf{R}_t^{(m)}, \mathbf{R}_s^{(m)}$ of the temporal and spatial self-attention modules in each m-th transformer block ($m = 1, 2, \ldots, M$) of shape $N \times (HW + 1) \times (T+1) \times (T+1)$ and $N \times (T+1) \times (HW+1) \times (HW+1)$ are computed firstly. Since Deepfake detection is a binary classification task, only the relevance of class 0 (indicates the Fake class) is considered. The outputs $\mathbf{U}_d$ ($d \in \{t, s\}$, denotes temporal or spatial) for the visualization are defined by:

$$\bar{\mathbf{A}}_{d(i,:,:)}^{(m)} = \mathbf{I} + max\left(\mathbb{E}_h(\mathbf{R}_{d(:,i,:,:)}^{(m)} \circ \nabla\mathbf{A}_{d(:,i,:,:)}^{(m)}), 0\right) \quad (4)$$

$$\mathbf{U}_d^{(i,:,:)} = \bar{\mathbf{A}}_{d(i,:,:)}^{(1)} \cdot \bar{\mathbf{A}}_{d(i,:,:)}^{(2)} \cdot \ldots \cdot \bar{\mathbf{A}}_{d(i,:,:)}^{(M)} \quad (5)$$

where $\mathbf{I}$ denotes the identity matrix, $\circ$ indicates the Hadamard product, $\cdot$ indicates the matrix multiplication, $\mathbb{E}_h$ indicates the mean across the head dimension. The subscripts $(i, :, :)$ and $(:, i, :, :)$ indicate that $\bar{\mathbf{A}}_d^{(m)}$ and $\mathbf{U}_d$ are computed on each spatial position and temporal position i simultaneously for the temporal and spatial self-attentions respectively. $\nabla\mathbf{A}_d^{(m)}$ is the gradient of the temporal or spatial attention of the m-th transformer block, which can be formulated as:

Algorithm 1 Visualization Method for ISTVT

Input:
Trained model, $f(x)$
Input frame sequence, X
Output:
Temporal heatmaps corresponding to the input, U_t
Spatial heatmaps corresponding to the input, U_s

```
Compute $f(X)$ via the forward propagation
Conduct the backward propagation, save the gradients of each layer
Compute and save the relevance of each layer via the relevance propagation rules
Initialize $U_t$, $U_s$
for $m = 1$ to $M$ do
    Load the gradients $\nabla A_t^{(m)}, \nabla A_s^{(m)}$ and relevance $R_t^{(m)}, R_s^{(m)}$ of the temporal and spatial self-attention layers in the $m$-th transformer block
    for $i = 1$ to $HW + 1$ do
        Compute $\bar{A}_{t(i,:,:)}^{(m)}$ by Formula 4
        Update $U_t^{(i,:,:)} \leftarrow U_t^{(i,:,:)} \cdot \bar{A}_{t(i,:,:)}^{(m)}$
    end for
    for $i = 1$ to $T + 1$ do
        Compute $\bar{A}_{s(i,:,:)}^{(m)}$ by Formula 4
        Update $U_s^{(i,:,:)} \leftarrow U_s^{(i,:,:)} \cdot \bar{A}_{s(i,:,:)}^{(m)}$
    end for
end for
$U_t \leftarrow U_t^{(1:,0,1:)}, U_s \leftarrow U_s^{(1:,0,1:)}$
Rearrange $U_t, U_s$ to shape $(T, H, W)$
Upscale $U_t, U_s$ to shape $(T, H_i, W_i)$ via the bilinear interpolation
return $U_t, U_s$
```

$$\nabla\mathbf{A}_d^{(m)} = \nabla softmax\left(\frac{\mathbf{q}_d^{(m)} \cdot \mathbf{k}_d^{(m)\top}}{\sqrt{D}}\right) \quad (6)$$

The results of previous steps are the tensors $\mathbf{U}_t$ and $\mathbf{U}_s$ of shape $(HW + 1) \times (T + 1) \times (T + 1)$ and $(T + 1) \times (HW + 1) \times (HW + 1)$, where $\mathbf{U}_t$ can be regarded as a sequence of $HW + 1$ matrices of each spatial position, and $\mathbf{U}_s$ can be regarded as a sequence of $T + 1$ matrices of each temporal position. Each row in each matrix corresponds to a relevance map of each token given the other tokens. To obtain a class-discriminative heatmap, we only consider the relevance maps corresponding to the classification tokens (first row in the matrices), except for the relevance to themselves. Besides, the matrices of irrespective classification token (first matrix of each sequence) are discarded from the matrix sequences. After the aforementioned operations, the shapes of $\mathbf{U}_t$ and $\mathbf{U}_s$ become $HW \times 1 \times T$ and $T \times 1 \times HW$ respectively. We then rearrange the results to the shape of $T \times H \times W$, and upscale them to the original input size $T \times H_i \times W_i$ via a bilinear interpolation step to achieve the final visualization heatmaps. The detailed steps of the proposed visualization method for ISTVT are shown in Algorithm 1.

IV. Experiments

A. Datasets

We use FaceForensics++ [9], FaceShifter [32], DeeperForensics [33], Celeb-DF [10], and DFDC [11] datasets for the experiments in this section. FaceForensics++ contains 1000 real videos collected from the internet and their corresponding forged videos manipulated by Deepfakes [1], FaceSwap [56], Face2Face [57], and NeuralTextures [15]. FaceShifter and DeeperForensics synthesis high-quality fake videos based on the real videos in FaceForensics++ dataset. Celeb-DF is a new dataset which contains 590 real and 5639 fake videos of high visual quality created by advanced manipulation methods. DFDC [11] is the preview dataset of the Deepfake Detection Challenge in 2020 synthesised by two unknown manipulation methods.

B. Implementation Details

1) Data Preprocessing: Following the previous works [5], [18], we use MTCNN [58] to detect the faces in the frames. The facial region in each video frame is cropped and resized to 300 × 300 as the facial image. We use sequences that consist of 6 continuous facial images to train and test the models. To ensure that the relative position of the face in each image is stable, we use the facial landmarks to align the faces and crop boxes in each frame. Specifically, for each frame, we use the tip of the nose as the center of the bounding box, and 1.25 times the maximum height and width of the face as the height and width of the box. For a fair comparison, we only use 270 frames in each video for the experiments on the FaceForensics++ dataset following the previous work [9]. For the other datasets, all the detected facial images are employed.

2) Settings: We use the entry flow of the Xception network [46] as the feature extractor. For each $300 \times 300 \times 3$ RGB facial image in the input sequence, the extractor generates a $19 \times 19 \times 728$ feature map. We split and flatten the feature maps to the 1×1 patches. For the transformer blocks, M is set to 12, N is set to 8. We use the SGD optimizer to train the model, the initial learning rate is set to 0.0005 and adjusted dynamically by a warm-up strategy. We train our models on 4 Tesla V100 GPUs for up to 100 epochs and select the best model based on the validation accuracy.

C. Detection Performance

1) Intra-Dataset Comparisons: In this section, we conduct experiments on FaceForensics++, Celeb-DF, and DFDC datasets. In each experiment, all the models are trained and tested on the same dataset. We report and compare the video-level detection accuracy of the models (the video-level predictions are the average scores of the video sequence predictions for our video-transformer based methods). This protocol allows us to fairly compare the capacity of capturing the forgery artifacts of the models.

Since the FaceForensics++ (FF++) dataset has four manipulation methods with different qualities (i.e. HQ and LQ), we conduct experiments on each manipulation and quality, with a total of eight settings. The results are shown in Table I. Apart from XN-avg (average score of each frame predicted by a Xception network [46]), all the other methods in the table are video-based. To compare with the existing video transformers, we re-implement VTN [6], VidTr [22], and ViViT [21] and conduct experiments on the Deepfake datasets. We use the public code of ViViT and VTN, and implement VidTr by modifying the self-attention of ViViT. Specifically, VTN proposes to conduct self-attention on all the tokens, like in [23], and uses a complete Xception network to extract features. Compared with VTN, our ISTVT makes great progress on accuracy. VidTr proposes a temporal down-sampled spatial-temporal method to reduce the redundant information in the videos. But this mechanism may erase the discriminative temporal artifacts in the fake videos, thus its performance is even worse than frame-based XN-avg. ViViT introduces four different variants of self-attention video transformers. We choose the best-performed factorized-encoder variant based on their experimental results. The results show this kind of video transformer does not work well on the Deepfake detection task and the performance is bad. Compared with the existing video transformers and other state-of-the-art video-based methods, our proposed ISTVT achieves significantly better performance on FaceForensics++ dataset. We note that STIL [5] focuses on extracting the temporal and spatial inconsistency and also achieves promising performance. This also proves the importance and effectiveness of capturing temporal and spatial artifacts in Deepfake detection.

For Celeb and DFDC datasets, we use the raw videos for the experiments. As the results shown in Table I, we also achieve the best results compared with the state-of-the-art methods. We notice that on DFDC dataset, video transformer-based methods perform significantly better than the CNN-based methods (VTN, VidTr, and our ISTVT achieve accuracy higher than 90%, while CNN-based is lower than 90%). It might result from that the DFDC dataset has various artifacts in complex scenarios generated by different manipulation methods. Compared with the CNN-based models, transformer-based models have a stronger representation learning capacity to handle them.

2) Cross-Dataset Comparisons: Current Deepfake videos are usually forged by various synthesis methods. Therefore, the generalization capacity is an important indicator to evaluate the Deepfake detection methods. To test the generalization capacity of ISTVT and compare it with the previous methods, we train the models on the FaceForensics++ dataset and test them on Celeb-DF, DFDC, FaceShifter (FSh), and DeeperForensics (DFo) datasets. Following the previous works [4], [5], [55], we employ the AUROC (Area Under Receiver Operating Characteristic) as the metric. The results are shown in Table II. We achieve very promising results compared to the previous methods,

TABLE I

Deepfake Detection Accuracy (%) on FaceForensics++ (FF++), Celeb-DF, and DFDC Datasets. DF, F2F, FS, and NT Denote the Deepfakes, Face2Face2, FaceSwap, and NeuralTextures Manipulation Methods. Best Results Are Highlighted. * Indicates Our Re-Implementation of the Existing Transformer-Based Methods. Results of Other Methods Are Collected From Their Published Papers

Methods	FF++ (HQ)				FF++ (LQ)				Celeb-DF	DFDC
	DF	F2F	FS	NT	DF	F2F	FS	NT		
XN-avg [46]	98.9	98.9	99.6	95.0	96.8	91.1	94.6	87.1	99.4	84.6
C3D [19]	92.9	88.6	91.8	89.6	89.3	82.9	87.9	87.1	-	-
I3D [35]	92.9	92.9	96.4	90.4	91.1	86.4	91.4	78.6	99.2	80.8
LSTM [20]	**99.6**	99.3	98.2	93.9	96.4	88.2	94.3	88.2	95.7	79.0
TEI [47]	97.9	97.1	97.5	94.3	95.0	91.1	94.6	90.4	99.1	87.0
ADDNet-3d [48]	92.1	83.9	92.5	78.2	90.4	78.2	80.0	69.3	95.2	79.7
S-MIL [18]	98.6	99.3	99.3	95.7	96.8	91.4	94.6	88.6	99.2	83.8
S-MIL-T [18]	**99.6**	**99.6**	**100.0**	94.3	97.1	91.1	96.1	86.8	98.8	85.1
SlowFast [49]	-	-	-	-	97.5	94.9	95.0	82.5	-	-
F^3-Net [50]	-	-	-	-	98.6	95.8	97.2	86.0	-	-
STIL [5]	**99.6**	99.3	**100.0**	95.4	98.2	92.1	97.1	91.8	**99.8**	89.8
FTCN [38]	97.9	97.1	98.2	96.1	-	-	-	-	-	-
VTN* [6]	**99.6**	99.3	99.6	95.4	97.9	92.1	95.7	90.4	99.3	91.7
VidTr* [22]	99.3	97.5	99.3	93.6	95.7	90.7	93.9	88.2	99.3	90.4
ViViT* [21]	82.9	82.9	83.9	80.7	81.4	78.6	80.0	72.1	98.8	83.4
ISTVT (ours)	**99.6**	**99.6**	**100.0**	**96.8**	**98.9**	**96.1**	**97.5**	**92.1**	**99.8**	**92.1**

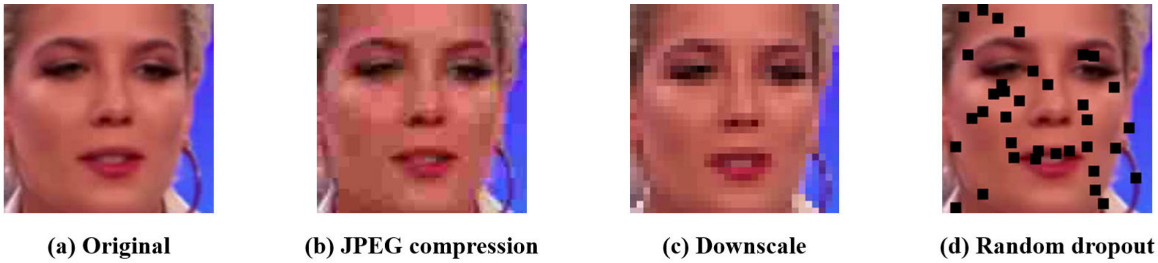

(a) Original (b) JPEG compression (c) Downscale (d) Random dropout

Fig. 5. Examples of the perturbation. For JPEG compression, quality factor is set to 10; for downscale, image scale is set to 0.2; for random dropout, number of dropout regions is set to 36.

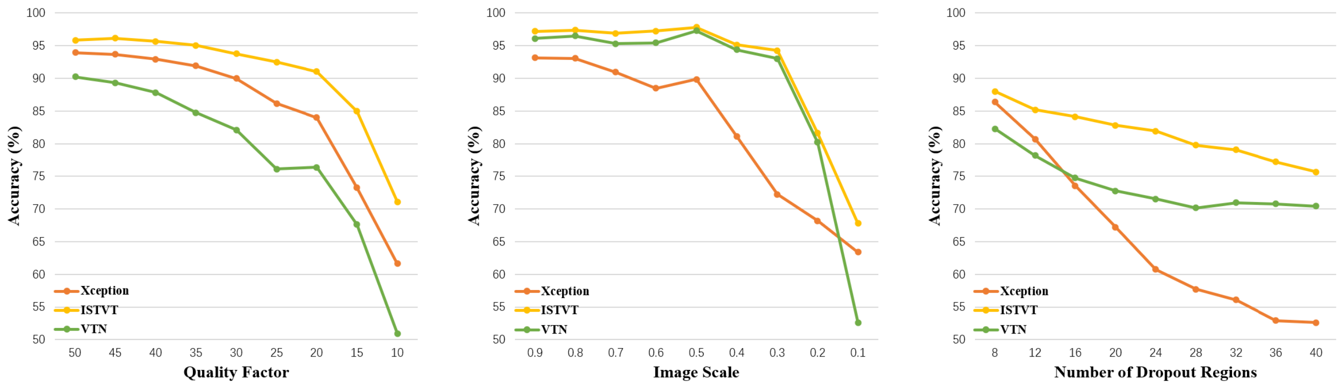
TABLE II

AUROC(%) of the Models Trained on FaceForensics++ Dataset and Tested on Unseen Datasets. Best Results Are Highlighted. * Indicates Our Re-Implementation of the Existing Transformer-Based Methods. Results of Other Methods Are From [51]

Methods	Celeb-DF	DFDC	FSh	DFo	Avg
Xception [46]	73.7	70.9	72.0	84.5	75.3
CNN-Aug [52]	75.6	72.1	65.7	74.4	72.0
PatchForensics [53]	69.6	65.6	57.8	81.8	68.7
CNN-GRU[36]	69.8	68.9	80.8	74.1	73.4
Multi-task[54]	75.7	68.1	66.0	77.7	71.9
FWA-DSP [55]	69.5	67.3	65.5	50.2	63.1
Face X-ray [4]	79.5	65.5	92.8	86.8	81.2
LipForensics [51]	82.4	73.5	97.1	97.6	87.7
FTCN [38]	**86.9**	74.0	98.8	**98.8**	**89.6**
VTN* [6]	83.2	73.5	98.7	97.7	88.3
VidTr* [22]	83.3	73.3	98.0	97.9	88.1
ViViT* [21]	80.5	72.8	96.8	95.3	86.4
ISTVT (ours)	84.1	**74.2**	**99.3**	98.6	89.1

which show the robustness improvement achieved by our method. Speficically, since fake videos in FaceShifter and DeeperForensics dataset are synthesised from the real videos in FaceForensics++ dataset, the generalization to these two dataset is very successful. And we find that both video-based methods and frame-based methods can achieve promising performance. However, for the completely unseen Celeb-DF and DFDC dataset, the generalization can be much more difficult. And the video-based methods (i.e. LipForensics [51], FTCN [38], three baseline video-transformer based methods, and ISTVT) obviously outperforms the frame-based state-of-the-art methods (i.e. Face X-ray [4]). This experimental result strongly demonstrate the importance of exploiting temporal information in improving the generalization capacity of Deepfake detection models.

We notice that FTCN also attempts to exploit the temporal inconsistency in videos to improve the generalization capacity of Deepfake detection. The difference is that FTCN uses a 3DCNN to extract spatial-temporal information and a temporal transformer to enhance the learning of long-term relationships. But our model focus on learning short-term inter-frame inconsistency via completely separated spatial-temporal self-attention. Thus, since FTCN has a longer input (32 frames) than ours (6 frames). It can perform better on the datasets like Celeb-DF, where the light conditions and head poses are consistent among the video frames. However, our method still has the following advantages: (1) Better cross-dataset performance on difficult datasets: for the datasets with more complex conditions like DFDC, it is difficult to discriminate whether the inconsistency is brought



(a) JPEG compression **(b) Downscale** **(c) Random dropout**

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

by the Deepfake artifacts or the changes of head pose or light condition in a long sequence. Hence, our method has better performance on the DFDC dataset compared with FTCN since we focus on the local inconsistency. (2) Better interpretability: since we separately handle the spatial and temporal information in ISTVT, we can know what spatial and temporal artifacts the model learns from the videos. But for FTCN, the visualization is conducted on the whole model and is hard to discriminate the spatial and temporal clues. (3) Better intra-dataset performance: due to the learning of fine temporal inconsistency and better learning capacity of the video transformer, our model has better intra-dataset performance on the FaceForensics++ dataset.

D. Robustness to Perturbations

In this section, we investigate the robustness of our method and compare with two baseline methods (i.e. Xception and VTN). All the models are trained and tested on the Celeb-DF dataset. We introduce three common noise and perturbations, i.e. JPEG compression, downscale, and random dropout, and apply them respectively to the testing set (training set is not augmented). Some examples of the perturbation are shown in Fig. 5. We test the proposed ISTVT and two baseline methods on the frame sequences that are perturbed to vary degrees and report the performance in Fig. 6. Next, we compare the performance of each model on different types of perturbations in detail.

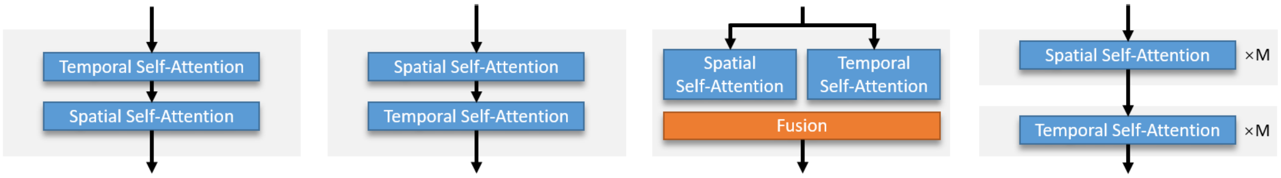
1) JPEG Compression: JPEG compression is widely used in network image transmission. We control the degree of perturbation by adjusting the quality factor of JPEG compression. The results in Fig. 6(a) demonstrate that our ISTVT is more robust than the two baseline methods. In addition, our performance advantages over VTN demonstrate that the proposed decomposed spatial-temporal self-attention and the self-subtract mechanism is very effective in improving the robustness of video transformers.

2) Downscale: This perturbation downsampling all the frame images in the sequences. We control the degree of perturbation by adjusting the scale of downsampling. In this scenario, the spatial information is heavily corrupted, while the inter-frame relationship (i.e. temporal information) is relatively well preserved. Therefore, as the results in Fig. 6(b) show, the video-based methods (i.e. ISTVT and VTN) is significantly more robust than the frame-based Xception network. In addition, since our ISTVT employs the self-subtract mechanism to learn the inter-frame inconsistency, it can detect the temporal artifacts better than the traditional VTN, especially when the image scale is very low (e.g. setting to 0.2 or 0.1).

3) Random Dropout: This perturbation randomly dropout 16×16 square regions in the original frame image. To generate a consistent frame sequence, the dropout mask keeps the same across all the frames. We control the degree of perturbation by adjusting the number of dropout regions. This perturbation randomly erases spatial and temporal information in the frame sequences and generates sharp texture warping artifacts which may confuse the model. Thus, as the results in Fig. 6(c) show, the frame-based Xception can not work when the number of the dropout regions exceeds 36. The reason is that important spatial artifacts are erased in this situation, hence the frame-based methods do not have enough information to make a correct prediction, thus predicting all the frames as real. However, the video-based methods can still work since some temporal artifacts are still preserved.

E. Ablation Study

1) Self-Attention: In the video transformer, the architecture of the spatial-temporal self-attention plays an important role in performance [21]. Thus, we investigate different variants of spatial-temporal self-attentions in this section. We test four self-attentions, illustrated in Fig. 7. Specifically, Decomposed attention with temporal or spatial first is the proposed version in Section III-B; Parallel Attention conducts the spatial and temporal self-attention simultaneously and merges the results via a fusion layer; for Decomposed Block, all the spatial self-attention layers are before the temporal self-attention layers, like the form in ViViT [21]. The results of the vanilla self-attention (i.e. results of VTN) are also compared as a baseline. We employ these variants to the proposed ISTVT respectively and compare the performance on FaceForensics++ and Celeb-DF datasets.



(a) Decomposed Attention (temporal first)   **(b) Decomposed Attention (spatial first)**   **(c) Parallel Attention**   **(d) Decomposed Block**

Fig. 7. Four variants of the spatial-temporal self-attention.

TABLE III

DEEPFAKE DETECTION ACCURACY (%) AND AUROC (%) OF ISTVT WITH DIFFERENT SELF-ATTENTION VARIANTS. TEMPORAL FIRST AND SPATIAL FIRST INDICATE THE DECOMPOSED ATTENTION WITH TEMPORAL OR SPATIAL FIRST

Methods	FF++		Celeb-DF	
	Acc	AUROC	Acc	AUROC
Vanilla	93.3	94.0	99.0	99.5
Temporal First	94.9	96.3	99.6	99.8
+ Self-subtract	**96.0**	**96.7**	**99.8**	**99.9**
Spatial First	93.4	95.8	99.6	99.6
+ Self-subtract	93.8	96.0	99.6	99.8
Parallel Attention	90.9	92.6	98.8	99.3
+ Self-subtract	91.7	93.2	99.0	99.5
Decomposed Block	78.2	80.5	98.3	99.1
+ Self-subtract	79.6	81.9	98.7	99.2
Spatial Only	92.8	93.5	98.5	99.0
Temporal Only	92.5	93.3	98.0	98.4
+ Self-subtract	92.7	93.8	98.3	98.9

For FaceForensics++ dataset, experiments are conducted on the LQ version of the complete FaceForensics++ (FF++) dataset. The results reported in Table III demonstrate that the decomposed attention mechanism outperforms other variants of self-attention, and conducting temporal self-attention first is a better choice. We also separately remove the spatial and temporal self-attention to test the effectiveness of our self-attention mechanism and the importance of the spatial and temporal information. Specifically, since the temporal or spatial classification tokens are discarded when temporal or spatial self-attention is removed, the original classification token for prediction (i.e. $\mathbf{O}_{(0,0,:)}$) is not available in this ablation study. Therefore, we employ the mean of all the spatial classification tokens or temporal classification tokens as the input of the prediction head to address this problem.

As expected, since frame-based methods are able to have promising intra-dataset performance, interesting results can be also achieved by using only spatial self-attention. Besides, using only the temporal information also achieves promising accuracy. This demonstrates that temporal information is a sufficient basis for Deepfake detection.

We also conduct ablation studies on the proposed self-subtract mechanism by applying it on all the variants of spatial-temporal self-attention respectively. The results show that our method is effective on all five variants, thus is a general method to improve the spatial-temporal attention.

2) Transformer Settings: The hyperparameter settings of the transformer are also important factors affecting model performance and efficiency. We test and compare the detection performance and computational budget of different input frame sequence length T and model depth M (i.e. number of transformer blocks) settings on the FaceForensics++ dataset in this section. The results are shown in Table IV and Table V respectively. The results show that (1) simply increasing the sequence length can not improve the performance effectively. In contrast, employing a short sequence (e.g. $T = 2$) can achieve an interesting performance with a low computational budget. The reason is that the discriminative temporal information is mainly the inter-frame inconsistency (i.e. short-term temporal information) for Deepfake detection [5]. Thus, the long-term information in long frame sequences is useless for Deepfake detection and therefore brings no performance improvement. (2) Employing a deeper model with more transformer blocks can not effectively improve the Deepfake detection performance either. Many previous works have proved that Deepfake detection mainly relies on the low-level mesoscopic features [3], [14], [59]. Thus,

TABLE IV

COMPARISON BETWEEN DIFFERENT SEQUENCE LENGTH T

T	2	4	6	8	12	16
AUC (%)	96.1	96.5	96.7	96.7	96.6	96.6
GFLOPs	99.6	167.9	236.1	304.4	440.9	577.4

TABLE V

COMPARISON BETWEEN DIFFERENT MODEL DEPTH M

M	4	6	8	12	16	24
AUC (%)	95.4	96.2	96.5	96.7	96.7	96.8
GFLOPs	89.7	126.3	162.9	236.1	309.4	455.8

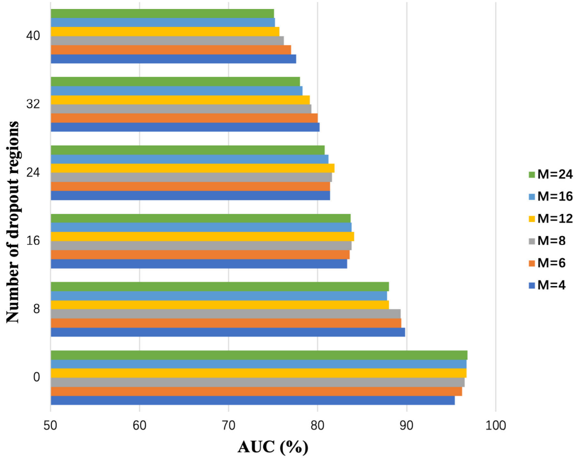


Fig. 8. Ablation study on different model depth and augmentation degree.

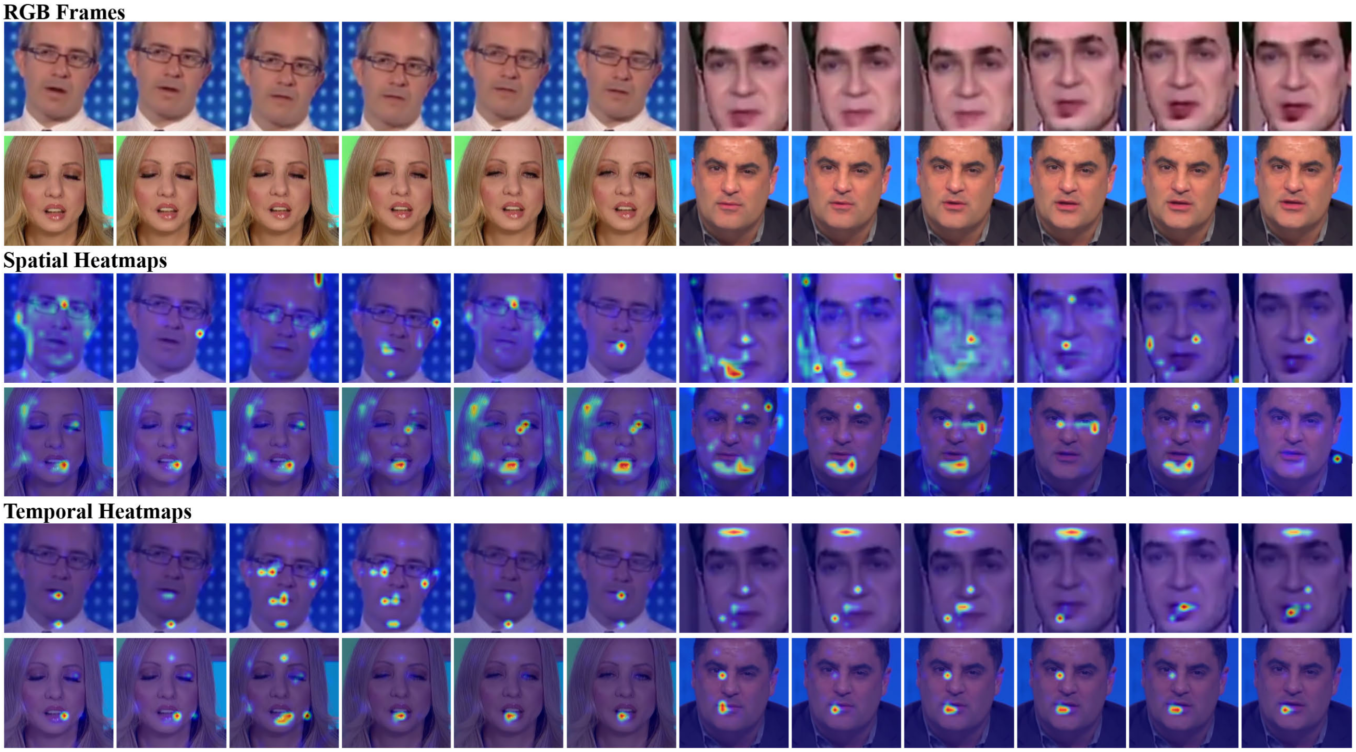


Fig. 9. Examples of the visualization results. We sample the frame sequences from the test set of FaceForensics++ dataset.

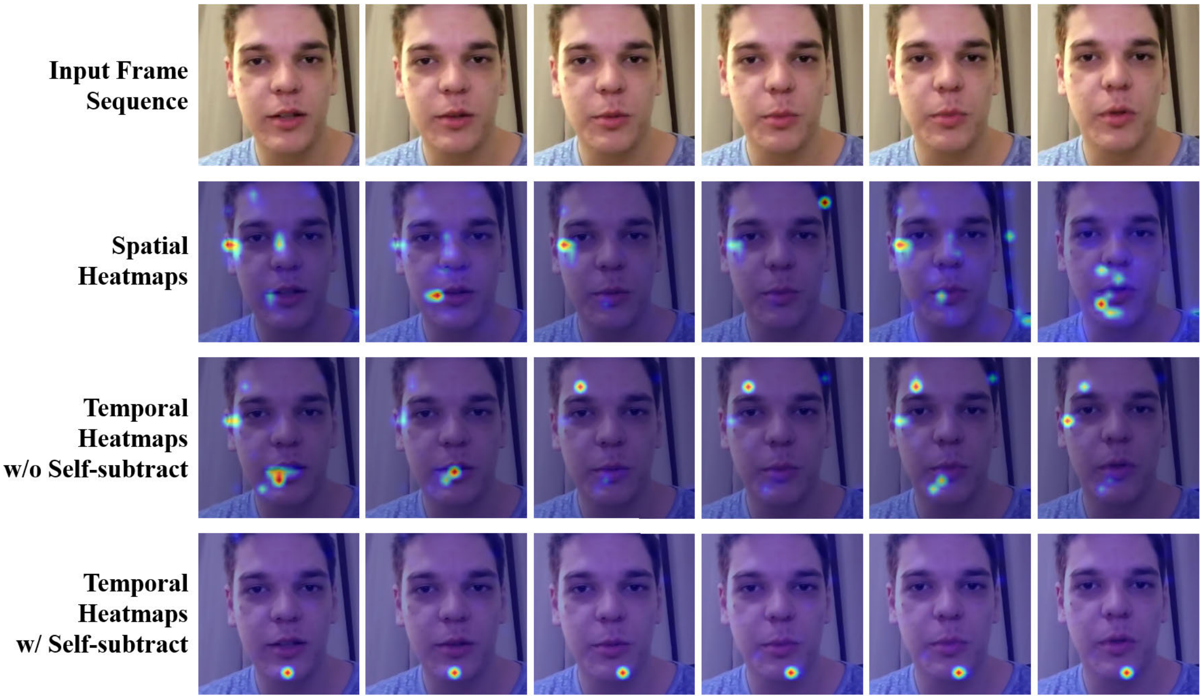

Input Frame Sequence
Spatial Heatmaps
Temporal Heatmaps w/o Self-subtract
Temporal Heatmaps w/ Self-subtract

Fig. 10. Example of a challenging case. The input frame sequence is sampled from a real video in FaceForensics++ dataset, the model without self-subtract predicts it as fake, while the model with self-subtract predicts it as real.

it is difficult to improve performance by learning semantic information through deeper models. Also, a larger model may cause overfitting, and thus reduce the robustness. To further investigate the relationship between the model depth and its robustness, we conduct experiments on the augmented data and compare the performance of different models. We use the random dropout augmentation to corrupt the semantic information in the frames. We report the AUROC (AUC) in Fig. 8. The results show that though deeper models have slightly better baseline performance than the smaller models, they may overfit the semantic information and have worse robustness than the smaller ones. Taking performance, robustness, and efficiency into account, we finally set T to 6 and M to 12.

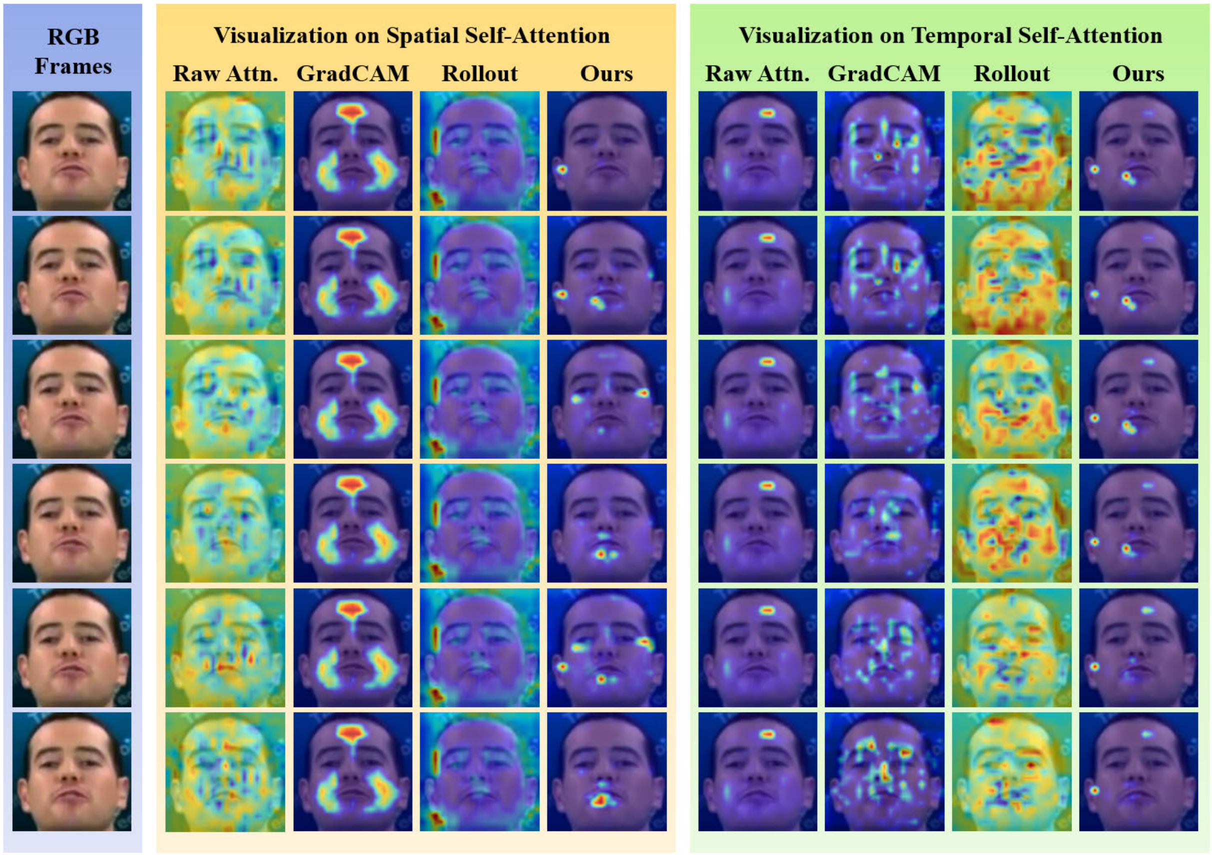

RGB Frames
Visualization on Spatial Self-Attention
Raw Attn.
GradCAM
Rollout
Ours
Visualization on Temporal Self-Attention
Raw Attn.
GradCAM
Rollout
Ours

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

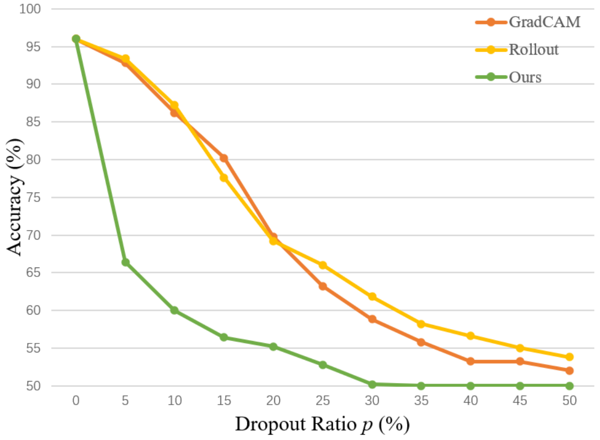


Fig. 12. Results of the quantitative test of different explanation methods.

V. Interpretability by Visualization

In this section, we use the proposed visualization method for ISTVT (i.e. Algorithm 1) to visualize the temporal and spatial heatmaps separately. We sample the fake frame sequences from the test set of FaceForensics++ dataset, and compute the heatmaps of each frame, the results are shown in Fig. 9. From the visualization results, it is obvious that spatial self-attention focuses on the blending edge (i.e. boundaries of the manipulated areas and backgrounds) and abnormal borders of the facial features. It helps people to understand how ISTVT makes decisions in the spatial dimension. It can also be observed that the discriminative regions of different spatial heatmaps are very different, showing that spatial attention can capture frame-independent information as claimed in Section III-B. Compared to the spatial self-attention, it is observed that (1) the temporal one ignores the unchanged areas and focuses on the moving areas (e.g. lips, jaws); (2) the discriminative regions of temporal heatmaps are very consistent. It shares the insights that (1) our temporal module can capture the temporal inconsistency caused by movement. It matches our intuition that it is hard for frame-based manipulation methods to generate consistent and realistic details around the moving area. (2) our temporal module captures the inter-frame discriminative areas while the spatial module mainly captures the within-frame artifacts.

To explain the high robustness of the proposed ISTVT, we collect some challenging cases and try to exploit our visualization method to explain why the models work or make mistakes on these situations. As shown in Fig. 10, the input frame sequence is sampled from a real video in the test set of FaceForensics++ dataset. This sequence has abnormal facial textures (i.e. the sun reflections in the upper left corner). Thus, the spatial self-attention may falsely focus on these areas. Conversely, the temporal self-attention tries to ignore these abnormal textures. However, without the self-subtract mechanism, the temporal self-attention may also focus on these textures (e.g. the first, fifth and sixth frames). Thus, the model incorrectly classified it as fake. While the model with self-subtract mechanism can completely ignore them since these textures are relatively consistent between the frames. Thus it can correctly predict the input sequence as real. This example clearly demonstrates the effectiveness of our self-subtract mechanism and the usefulness of our visualization method.

We also try to utilize other common visualization methods for vision transformers, including raw attention (Raw Attn.), GradCAM [40], and Rollout [45], to interpret our ISTVT. Specifically, raw attention means using the attention map of the last transformer layer. GradCAM is a common visualization method based on gradients. We backprop the gradients of the spatial self-attention layer or temporal self-attention layer in the last transformer block respectively for the generation of spatial or temporal heatmaps. Rollout [45] is obtained by a recursive matrix multiplication of all the attention maps. Similar to Algorithm 1, Rollout is conducted on the spatial and temporal self-attention in ISTVT respectively. We use the aforementioned visualization methods to interpret a same ISTVT trained on FaceForensics++ dataset. The results are shown and compared with our proposed method in Fig. 11. By observing the visualization results, we find that the raw attention can not interpret the spatial self-attention, and can not accurately interpret the temporal self-attention either (it indicates that the forehead region contains temporal artifacts, however, this region stays constant between the frames). The visualization of spatial self-attention given by GradCAM considers the model is focusing on the cheeks and forehead, and the visualization of temporal self-attention seems fragmented. Using Rollout can roughly interpret the spatial self-attention, but it does not work on the temporal self-attention. Compared with these three existing visualization methods, our improved algorithm based on [31] generates more concentrated and reasonable results on both spatial and temporal self-attention. Thus, our visualization algorithm is very suitable to interpret video transformers.

To further verify our method and compare it with other baseline methods. We conduct an experiment based on perturbation to quantitatively analyze the explanation capacity. Specifically, we dropout the top p% pixels with the largest scores in the heatmaps generated by different visualization methods (spatial and temporal heatmaps are added up in this experiment). The experiment is conducted on the FaceForensics++ dataset. We record and compare the changes in accuracy under different degrees of perturbation in Fig. 12. The results show that dropping 10% pixels guided by visualization results generated by our method can almost completely confuse the model (i.e. accuracy drops to less than 60%). As a comparison, if the perturbation is guided by the visualization results generated by Rollout or GradCAM, since these two methods are less effective in explaining where the model is focusing on, we need to erase more than 30% of the pixels to confuse the model. This experiment strongly demonstrates that our interpretation method is much more accurate than the baseline methods.

Clearly, our interpretability by visualization strategy can be also applied to other spatial-temporal downstream tasks, helping people to understand how the spatial-temporal makes decisions.

VI. Conclusion and Future Work

In this paper, we propose the Interpretable Spatial-Temporal Video Transformer (ISTVT) for Deepfake detection. We employ the decomposed spatial-temporal self-attention and the self-subtract mechanism in ISTVT. Our method significantly improves the performance and robustness compared with the previous video-based Deepfake detection methods, including existing video transformers. Then, motivated by [31], we propose a visualization method for video transformers to interpret what temporal and spatial features are captured by our model and contribute to the prediction respectively. This separate visualization is novel for not only the Deepfake detection, but also for the video transformers. Also, our visualization method is general, and it can be applied to other video-based transformers as well.

In future work, we will further investigate the visualization results and try to improve the design of Deepfake detection models through it. We will also further optimize our interpretability by visualization strategy so that we can have a deeper understanding of how the Deepfake videos are detected by the deep learning models.

References

[1] *Deepfakes Github*. Accessed: Oct. 24, 2021. [Online]. Available: https://github.com/deepfakes/faceswap

[2] *Deepfacelab*. Accessed: Oct. 24, 2021. [Online]. Available: https://github.com/iperov/DeepFaceLab

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[4] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5001–5010.

[5] Z. Gu et al., "Spatiotemporal inconsistency learning for DeepFake video detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3473–3481.

[6] S. A. Khan and H. Dai, "Video transformer for deepfake detection with incremental learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1821–1828.

[7] P. Yu, J. Fei, Z. Xia, Z. Zhou, and J. Weng, "Improving generalization by commonality learning in face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 547–558, 2022.

[8] Y. Wang, C. Peng, D. Liu, N. Wang, and X. Gao, "ForgeryNIR: Deep face forgery and detection in near-infrared scenario," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 500–515, 2022.

[9] A. Rossler, 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)*, Oct. 2019, pp. 1–11.

[10] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," 2019, *arXiv:1909.12962*.

[11] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[12] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," 2019, *arXiv:1910.12467*.

[13] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5781–5790.

[14] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1–10.

[15] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, 2019.

[16] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 15, 2020, doi: 10.1109/TPAMI.2020.3009287.

[17] M. Li, B. Liu, Y. Hu, L. Zhang, and S. Wang, "Deepfake detection using robust spatial and temporal features from facial landmarks," in *Proc. IEEE Int. Workshop Biometrics Forensics (IWBF)*, May 2021, pp. 1–6.

[18] X. Li et al., "Sharp multiple instance learning for DeepFake video detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1864–1872.

[19] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: Generic features for video analysis," 2014, *arXiv:1412.0767*.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," 2021, *arXiv:2103.15691*.

[22] Y. Zhang et al., "VidTr: Video transformer without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13577–13587.

[23] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," 2021, *arXiv:2102.00719*.

[24] R. Girdhar, J. Joao Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 244–253.

[25] Z. Liu et al., "Video swin transformer," 2021, *arXiv:2106.13230*.

[26] L. He et al., "End-to-end video object detection with spatial–temporal transformers," 2021, *arXiv:2105.10920*.

[27] J. Yin, J. Shen, C. Guan, D. Zhou, and R. Yang, "LiDAR-based online 3D video object detection with graph-based message passing and spatiotemporal transformer attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11495–11504.

[28] X. Yang, H. Wang, D. Xie, C. Deng, and D. Tao, "Object-agnostic transformers for video referring segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 2839–2849, 2022.

[29] Z. Xu, D. Chen, K. Wei, C. Deng, and H. Xue, "HiSA: Hierarchically semantic associating for video temporal grounding," *IEEE Trans. Image Process.*, vol. 31, pp. 5178–5188, 2022.

[30] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[31] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791.

[32] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5074–5083.

[33] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2889–2898.

[34] G. Wang, J. Zhou, and Y. Wu, "Exposing deep-faked videos by anomalous co-motion pattern detection," 2020, *arXiv:2008.04848*.

[35] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[36] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[37] C.-Z. Yang, J. Ma, S. Wang, and A. W.-C. Liew, "Preventing DeepFake attacks on speaker authentication by dynamic lip movement analysis," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1841–1854, 2021.

[38] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15044–15054.

[39] S. Pino, M. J. Carman, and P. Bestagini, "What's wrong with this video? Comparing explainers for deepfake detection," 2021, *arXiv:2105.05902*.

[40] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," 2019, *arXiv:1905.00780*.

[41] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds. Dec. 2017, pp. 4765–4774.

[42] B. Peng, S. Lyu, W. Wang, and J. Dong, "Counterfactual image enhancement for explanation of face swap deepfakes," in *Pattern Recognition and Computer Vision* (Lecture Notes in Computer Science), vol. 13535, S. Yu, Z. Zhang, P. C. Yuen, J. Han, T. Tan, Y. Guo, J. Lai, and J. Zhang, Eds. Shenzhen, China: Springer, Nov. 2022, pp. 492–508.

[43] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2016, pp. 63–71.

[44] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.

[45] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," 2020, *arXiv:2005.00928*.

[46] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[47] Z. Liu et al., "TEINet: Towards an efficient architecture for video recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11669–11676.

[48] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2382–2390.

[49] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[50] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 86–103.

[51] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5039–5049.

[52] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot.. for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8695–8704.

[53] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 103–120.

[54] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2019, pp. 1–8.

[55] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," 2018, *arXiv:1811.00656*.

[56] *Faceswap Github*. Accessed: Oct. 24, 2021. [Online]. Available: https://github.com/MarekKowalski/FaceSwap/

[57] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[58] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[59] C. Wang, C. Zhao, and G. Hu, "Multi-definition video deepfake detection via semantics reduction and cross-domain training," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6.

Cairong Zhao received the B.S. degree from Jilin University in 2003, the M.S. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2006, and the Ph.D. degree from the Nanjing University of Science and Technology in 2011. He is currently a Professor with the College of Electronic and Information Engineering, Tongji University. He works on visual and intelligent learning, including computer vision, pattern recognition, and visual surveillance. He has authored more than 40 journals and conference papers in these areas.

Chutian Wang received the B.Sc. degree from the Department of Computer Science and Technology, Tongji University, in 2020, where he is currently pursuing the master's degree. His research interests include computer vision and pattern recognition.

Guosheng Hu (Senior Member, IEEE) received the Ph.D. degree from the Centre for Vision, Speech and Signal Processing, University of Surrey, U.K., in June 2015. He was a Post-Doctoral Researcher at the LEAR Team, INRIA Grenoble Rhone-Alpes, France, from May 2015 to May 2016. He is currently a Senior Researcher of Oosto, Belfast, U.K. His research interests include deep learning, pattern recognition, and biometrics (mainly face recognition).

Haonan Chen received the B.Sci. and Ph.D. degrees in instrument science and engineering from Zhejiang University, Hangzhou, China, in 2014 and 2020, respectively. He is currently an Engineer-Algorithm-Applied Algorithm with Alibaba Group, Alibaba Cloud Intelligence Business Group, Tmall Genie-AI. His research interests include deep learning, pattern recognition, biometrics (mainly face recognition), and multimodal visual and audio algorithm.

Chun Liu (Member, IEEE) received the Ph.D. degree in geodesy and surveying from Tongji University, Shanghai, China, in 2001. From 2001 to 2003, he was a Visiting Scholar with the Development of Infrastructure for Cyber Hong Kong, The Hong Kong Polytechnic University, Hong Kong. From 2007 to 2008, he was a Senior Visiting Scholar with the Laboratory of Mapping and Geographic Information Systems, The Ohio State University, Columbus, OH, USA. In 2011, he was a Senior Visiting Scholar with the Institute of Spatial Geodesy and Engineering Survey, University of Nottingham, Nottingham, U.K. He is currently a Full Professor of cartography and geography information engineering with the College of Surveying and Geo-Informatics, Tongji University. His research interests include multisource point cloud coupled semantic cognition and spatial–temporal smart sensing.

Jinhui Tang (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China (USTC) in July 2003 and July 2008, respectively. From July 2008 to December 2010, he was a Research Fellow with the School of Computing, National University of Singapore. During that period, he visited the School of Information and Computer Science, UC Irvine, from January 2010 to April 2010, as a Visiting Research Scientist. From September 2011 to March 2012, he visited Microsoft Research Asia, as a Visiting Researcher. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include large-scale multimedia search, social media mining, and computer vision. He has authored more than 80 journals and conference papers in these areas. He is a member of the ACM. He serves as an Editorial Board Member for *Pattern Analysis and Applications*, *Multimedia Tools and Applications*, *Information Sciences*, and *Neurocomputing*, a technical committee member for about 30 international conferences, and a reviewer for about 30 prestigious international journals. He co-received the Best Paper Award in ACM Multimedia 2007, PCM 2011, and ICIMCS 2011.