Detection of Deepfake Videos Using Long-Distance Attention

Wei Lu, *Member, IEEE*, Lingyi Liu, Bolin Zhang, Junwei Luo, Xianfeng Zhao, *Senior Member, IEEE*, Yicong Zhou, *Senior Member, IEEE*, and Jiwu Huang, *Fellow, IEEE*

***Abstract*— With the rapid progress of deepfake techniques in recent years, facial video forgery can generate highly deceptive video content and bring severe security threats. And detection of such forgery videos is much more urgent and challenging. Most existing detection methods treat the problem as a vanilla binary classification problem. In this article, the problem is treated as a special fine-grained classification problem since the differences between fake and real faces are very subtle. It is observed that most existing face forgery methods left some common artifacts in the spatial domain and time domain, including generative defects in the spatial domain and interframe inconsistencies in the time domain. And a spatial-temporal model is proposed which has two components for capturing spatial and temporal forgery traces from a global perspective, respectively. The two components are designed using a novel long-distance attention mechanism. One component of the spatial domain is used to capture artifacts in a single frame, and the other component of the time domain is used to capture artifacts in consecutive frames. They generate attention maps in the form of patches. The attention method has a broader vision which contributes to better assembling global information and extracting local statistic information. Finally, the attention maps are used to guide the network to focus on pivotal parts of the face, just like other fine-grained classification methods. The experimental results on different public datasets demonstrate that the proposed method achieves state-of-the-art performance, and the proposed long-distance attention method can effectively capture pivotal parts for face forgery.**

Manuscript received 8 June 2021; revised 25 January 2022, 30 May 2022, and 25 October 2022; accepted 27 December 2022. Date of publication 6 January 2023; date of current version 9 July 2024. This work was supported by the National Natural Science Foundation of China under Grant U2001202, Grant 62072480, Grant U19B2022, and Grant U1636202. *(Corresponding author: Wei Lu.)*

Wei Lu, Lingyi Liu, Bolin Zhang, and Junwei Luo are with the School of Computer Science and Engineering, Guangdong Province Key Laboratory of Information Security Technology, Ministry of Education Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University, Guangzhou 510006, China (e-mail: luwei3@mail.sysu.edu.cn; liuly83@mail2.sysu.edu.cn; zhangblin3@mail2.sysu.edu.cn; luojw8@mail2.sysu.edu.cn).

Xianfeng Zhao is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China (e-mail: zhaoxianfeng@iie.ac.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

Jiwu Huang is with the Guangdong Key Laboratory of Intelligent Information Processing and the Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China, and also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518055, China (e-mail: jwhuang@szu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2022.3233063

***Index Terms*— Attention mechanism, deepfake detection, face manipulation, spatial and temporal artifacts.**

I. Introduction

THE deepfake videos are designed to replace the face of one person with another's. The advancement of generative models [1], [2], [3] makes deepfake videos become very realistic. In the meantime, the emergence of some face forgery applications [4], [5] enables everyone to produce highly deceptive forged videos. Now, deepfake videos are flooding the Internet. In the internet era, such technology can be easily used to spread rumors and hatred, which brings great harm to society. Thus the high-quality deepfake videos that cannot be distinguished by human eyes directly have aroused interest among researchers. An effective detection method is urgently needed.

The general process of generating deepfake videos is shown in Fig. 1. Firstly, the video is divided into frames and the face in each frame is located and cropped. Then, the original face is converted into the target face by using a generative model and spliced into the corresponding frame. Finally, all frames are serialized to compose the deepfake video. In these processes, two kinds of defects are inevitably introduced. In the process of generating forged faces, the visual artifacts in the spatial domain are introduced by the imperfect generation model. In the process of combining frame sequences into videos, the inconsistencies between frames are caused by the lack of global constraints.

Many detection methods are proposed [6], [7], [8] based on the defects in the spatial domain. Some of the methods take advantage of the defects of face semantics in deepfake videos, because the generative models lack global constraints in the process of fake face generation, which introduces some abnormal face parts and mismatched details in the face from a global perspective. For example, face parts with abnormal positions [8], asymmetric faces [9], and eyes with different colors [6]. However, it's fragile to rely entirely on these semantics. Once the deepfake videos do not contain the specific semantic defects that the method depends on, the performance will be significantly degraded.

There are also some "deep" approaches [7], [10], which attempt to excavate spatial defects according to the characteristics of the deepfake generators. However, compared with image contents, the forgery traces in the spatial domain are very weak, and the convolutional networks tend to extract

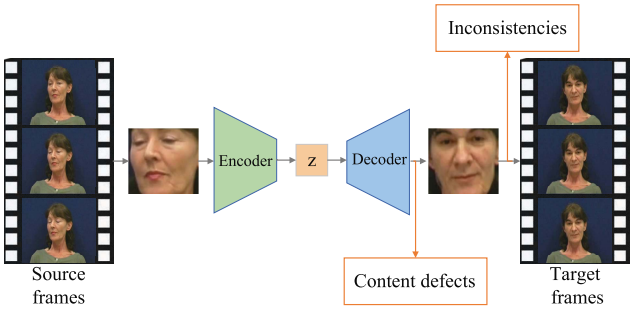

Fig. 1. Generation process of deepfake videos. The original video is divided into frames and cropped out of the faces. The target face is generated by an encoder-decoder which introduces content defects. Then the target face is spliced back to the original frame and inconsistencies are introduced. Finally, all the frames are synthesized into a fake video.

image content features rather than the traces [11]. So blindly utilizing deep learning is not very effective in catching fake content [12].

Since the deepfake video is synthesized frame by frame, and there is no precise constraint between the frame sequences, the inconsistencies in the time domain will be introduced. Some methods exploit these defects of the time domain. The movements of eyes are exploited in [13]. Li et al. [14] use the human blink frequency in the video to detect the deepfake videos. The movement of lip [15] and the heart rate [16] are also exploited as the identification basis between authentic videos and deepfake videos in the time domain. The optical flows and the movement patterns of the real face and fake face are classified in [17] and [18], respectively.

All of the methods mentioned above take deepfake detection as a vanilla binary classification problem. However, as the counterfeits become more and more realistic, the differences between real and fake ones will become more and more subtle and local which makes such global feature-based vanilla solutions work not well [19].

Similar problems have been studied in the field of fine-grained classification. Fine-grained classification aims to classify very similar categories, such as species of the bird, models of the car, and types of the aircraft [20]. Since deepfake detection and fine-grained classification share the same spirit, that learning subtle and discriminative features, in [19], the deepfake detection is reformulated as a fine-grained classification task. And a convolutional attention module with 1×1 is adopted to make a network focus on the subtle but critical regions.

However, combining global semantics is just as important as focusing on local areas. Because some defects are normal from a local or isolated perspective, but abnormal from a global perspective. For example, uncoordinated head postures [21], mismatched facial expressions and head movements [22], and mismatched eye details [23]. These kinds of defects exist between different parts of the face at a long-distance. In other words, the local areas of focus should be determined according to the global semantics [24], and modeling long-distance dependencies in both the spatial domain and time domain are important. But it is not directly for the convolutional attention mechanism, especially when the kernel is small. The global pooling may be a choice for assembling global information,

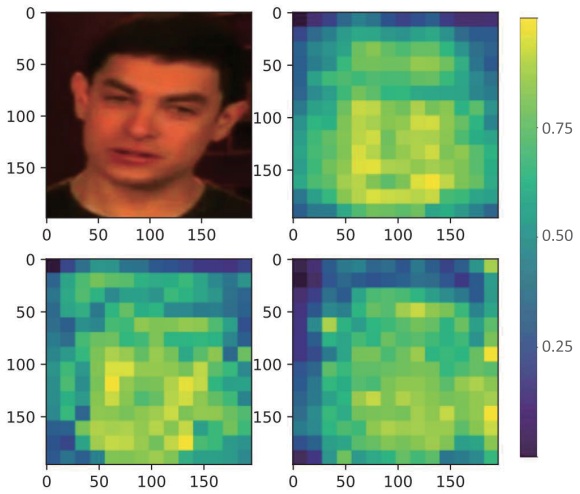

Fig. 2. Attention maps generated by the novel long-distance attention mechanism. Pivotal facial regions are emphasized in patches by these maps. The brighter and yellower the color, the more important it means to the network.

however, the weak forgery clues will be averaged by this operation, resulting in a loss of distinguishability [19].

Vision Transformer (ViT) [25] is a widely used model, which can draw global dependencies and assemble global information relying entirely on a self-attention mechanism. However, in [25] the deficiencies are also presented. Firstly, unlike convolutional networks, it does not intuitively feel the relationships between adjacent pixels, and the position relationship only exists between patches. Secondly, when trained on mid-sized datasets, such as ImageNet, ViT yields modest accuracies of a few percentage points below the SOTA convolutional neural networks of comparable size [25]. Thus, some methods [26], [27], [28], [29] modify the ViT to apply to their own framework. In [26] and [27], the image is processed by the convolutional network before being input into the ViT. In [29], dual cross-modality attention is proposed to model the interaction between the low-frequency textures and the high-frequency noises. In [28], the transformer architecture is used as the backbone to extract multiscale information. Because we want to combine the advantages of convolutional networks in feature extraction with the advantages of ViT in constructing long-range dependencies, we draw lessons from the fine-grained classification and propose a novel long-distance attention mechanism according to the characteristics of deepfake videos. The long-distance attention mechanism is designed to determine the pivotal parts of forgery by assembling information from a global perspective. Long-distance attention is adopted in our spatial-temporal model to exploit the defects in the spatial domain and time domain. The spatial-temporal model is used to generate attention maps in the form of patches and guides the network to focus on pivotal local parts of the face. An example of our attention maps is shown in Fig. 2, different colors represent different levels of importance to neural networks. From the color, we can see that the brighter and yellower the color, the more important it means to the network. More detailed attention maps are shown in Section VI.

The contributions of this article are summarized as follows.

1) The experience of the fine-grained classification field is introduced, and a novel long-distance attention mechanism is proposed that can generate guidance by assembling global information.
2) The attention mechanism with a longer attention span is more effective for deepfake detection tasks. This combination form of a convolutional network and ViT is beneficial to make full use of the advantages of the two network architectures. And in the process of generating attention maps, the nonconvolution module is also feasible.
3) A spatial-temporal model is proposed to capture the defects in the spatial domain and time domain, according to the characteristics of deepfake videos, the model adopts the long-distance attention as the main mechanism to construct multilevel semantic guidance. The experimental results show that it achieves state-of-the-art performance.

The remainder of this article is organized as follows. In Section II, we first discuss the related work in the field of fine-grained classification. Then, the classical ViT is introduced briefly. In Section III, we analyze the defect characteristics of deepfake videos. In Section IV, the proposed method is introduced in detail. Section V discusses the experimental results. The ablation analysis is given in Section VI. The conclusion is presented in Section VII.

II. Related Works

A. Fine-Grained Classification

In the past few years, the performance of general image classification tasks has been significantly improved. From the amazing start of Alexnet [30] in Imagenet [31], the method based on deep learning almost dominates the Imagenet competition. However, for fine-grained object recognition [32], [33], [34], [35], there are still great challenges. The main reason is that the two objects are almost the same from the global and apparent point of visual. Therefore, how to recognize the subtle differences in some key parts is a central theme for fine-grained recognition. Earlier works [36], [37] leverage human-annotated bounding box of key parts and achieve good results. But the disadvantage is that it needs expensive manual annotation, and the location of manual annotation is not always the best distinguishing area [38], which completely depends on the cognitive level of the annotator.

Since the key step of fine-grained classification is focusing on more discriminative local areas [39], many weakly supervised learning methods [20], [38], [40] have been proposed. Most of them use kinds of convolutional attention mechanisms to find the pivotal parts for detection. Fu et al. [40] use a recurrent attention convolutional neural network (RA-CNN) to learn discriminative region attention. Hu et al. [41] propose a channel-wise attention method to model interdependencies between channels. In [38], a multiattention convolutional neural network is adopted and more fine-grained features can be learned. Hu et al. [20] propose a weakly supervised data augmentation network using attention cropping and attention dropping.

Deepfake detection and fine-grained classification are similar, that attempt to classify very similar things. Thus we learn from the experience in this field and leverage the attention maps generated with long-range information to make the networks focus on pivotal regions.

B. Vision Transformer

Transformer [42], a kind of self-attention architecture, is initially applied in natural language processing (NLP) and shows excellent performance. Its variant in the field of computer vision, ViT [25], is first proposed by the Google team in 2020 and attracts a lot of attention. In the vision, attention is usually used as a component of convolutional networks while keeping the overall structure. ViT shows that reliance on CNNs is not necessary. To apply the transformer to images directly, they first split the image into patches and project the patches to linear embedding. As a classification model, it generates a final discriminant vector through several stacking layers of self-attention modules. The self-attention modules are used to integrate the features of each patch with the self-attention mechanism. The self-attention mechanism is a stunning mechanism, which draws global dependencies and assembles global information.

ViT is a good choice for tasks that require building long-distance dependencies. For example, in temporal action localization task [43], [44], [45], it is important to obtain information from an action's context. In [43], the dependencies are modeled by a variant of ViT. It may be a promising candidate to deal with the detection of deepfake videos since the deepfake videos need to be considered from a global perspective and focused on the critical regions. However, we find that it is not effective to apply ViT to deepfake detection directly. Therefore, we learn from ViT and propose a novel long-distance attention mechanism. It is used to guide the backbone network to focus on critical regions by assembling global information.

III. Analysis of Deepfake

The deepfake videos, generated by GANs [1] and VAEs [2], are formidably realistic and difficult for human eyes to discriminate.

Since the differences between authentic videos and deepfake videos are subtle, detectors that blindly utilize deep learning are not effective in catching fake content [12]. Similar problems have been studied in the field of fine-grained classification. A crucial experience is that using an attention mechanism to make the network focus on pivotal local regions can greatly improve classification performance.

The generative models also have some inherent defects, which make deepfake detection possible. Whether it's GANs or VAEs, the generative networks will have an up-sampling process in the generation process to generate high-resolution images from latent coding [1], [2]. This allows the network to fill in detail in the rough image. Deconvolution allows the model to draw a larger square from a point in the small graph. However, deconvolution is prone to uneven overlap, especially when the kernel size cannot be divided by the step size.

In theory, the neural network can learn the weight parameters carefully to avoid this kind of defect, but in fact, the neural network cannot completely avoid this kind of defect [46]. This overlapping style is reflected in two dimensions. The uneven overlapping multiplication of two coordinate axes results in the image block similar to chessboard [46], and resulting in a loss of facial texture details. Liu et al. [47] observe that up-sampling is a necessary step of most face forgery techniques and utilizes phase spectrum to capture the up-sampling defects of face forgery. Since the up-sampling occurs between adjacent pixels, it is advantageous to capture the local information and collect statistics by using small blocks of appropriate size [48]. On the other hand, deepfake often generates abnormal face semantics. For example, unconvincing specular reflections in the eyes, either missing or represented as white blobs, or roughly modeled teeth, which appear as a single white blob [23]. The semantics and textures of the human face also appear in the form of the region [49]. Therefore, the processing of facial features in the form of patches is conducive to extracting local statistical information and capturing forgery traces. In long-distance attention, the input image is divided into many nonoverlapping small patches to collect local information.

However, some face semantics are normal from the local perspective but abnormal from the global perspective. That's because the GANs lack global constraints which introduce abnormal facial parts and mismatched details. It is observed that the density distributions of normalized face landmark locations on real and GAN-synthesized fake faces are different [8] because there is no coordination mechanism in the generation process of face components. This also leads to the asymmetry of the face [9]. In addition to the global structure as clues, the difference in details between facial components is also a key to the detection. For example, human eyes are always separated by a certain distance and have the same color, but the eyes of the fake face sometimes show a different color [23]. An example of defects in the spatial domain is shown in Fig. 3. The first row reflects defects in a local region, and the next two rows reflect defects from a wider vision. It is also observed that biological signals are not coherently preserved in different synthetic facial parts [12]. Therefore, assembling global semantic information and considering the location relationship between facial components will help to find these generative defects.

In addition to the generative defects in the spatial domain, temporal defects also exist in deepfake videos. In [14], the temporal inconsistencies are caught by the frequency of eye blinking. The inconsistency is also reflected in the face motion. The face motion patterns of real videos and deepfake videos have some differences and can be used for classification [18]. Furthermore, there is a strong correlation between facial expression and head movement [22]. Changing the former without modifying the latter may expose a manipulation. It is also observed that temporal consistencies of human biological signals are not well preserved in GAN-erated content [12]. Thus, it is beneficial to modeling the continuity of the face in the videos for deepfake detection. We exploit these inconsistencies in the time domain with a temporal model.

Fig. 3. Some typical defects of deepfake videos in the spatial domain. The images in the first row reflect some local defects, i.e., obvious forgery clues in the mouth of the left and middle pictures, and a strange facula near the hair in the right picture. The second row contains faces with weird eyes. The third row contains faces with abnormal face structures. (a) Local defects. (b) Mismatched eyes. (c) Abnormal structures.

IV. Proposed Method

A. Overview

In this section, the motivation to use long-distance attention is given first and then the proposed model is described briefly.

As aforementioned, there is no precise global constraint in the deepfake generation model, which always introduces disharmony between local regions in the face of forgery from a global perspective. In addition to the artifacts that exist in each forgery frame itself, there are also inconsistencies (e.g., unsmooth lip movement) between frame sequences because the deepfake videos are generated frame by frame. To capture these defects, a spatial-temporal model is proposed, which has two components for capturing spatial and temporal defects respectively. Each component has a novel long-distance attention mechanism that can be used to assembling the global information to highlight local regions.

Based on the observation [50] that the artifacts caused by the generation model are mainly preserved in textural information of shallow features, the attention maps generated by the spatial component are adopted to recalibrate the shallow features maps which are generated by the first several convolutional layers. As inconsistency occurs in relative high-level semantic features, the attention maps generated by the temporal attention component are used to guide the relative high-level semantic features.

The framework of the spatial-temporal model is shown in Fig. 4. Two essential components are integrated into the backbone network: 1) a spatial attention component for capturing spatial disharmony and focusing on shallow features and 2) a temporal attention component for capturing temporal inconsistencies and focusing on mid-level features. Both the attention maps are used to recalibrate the feature maps and make the network focusing on pivotal local regions.

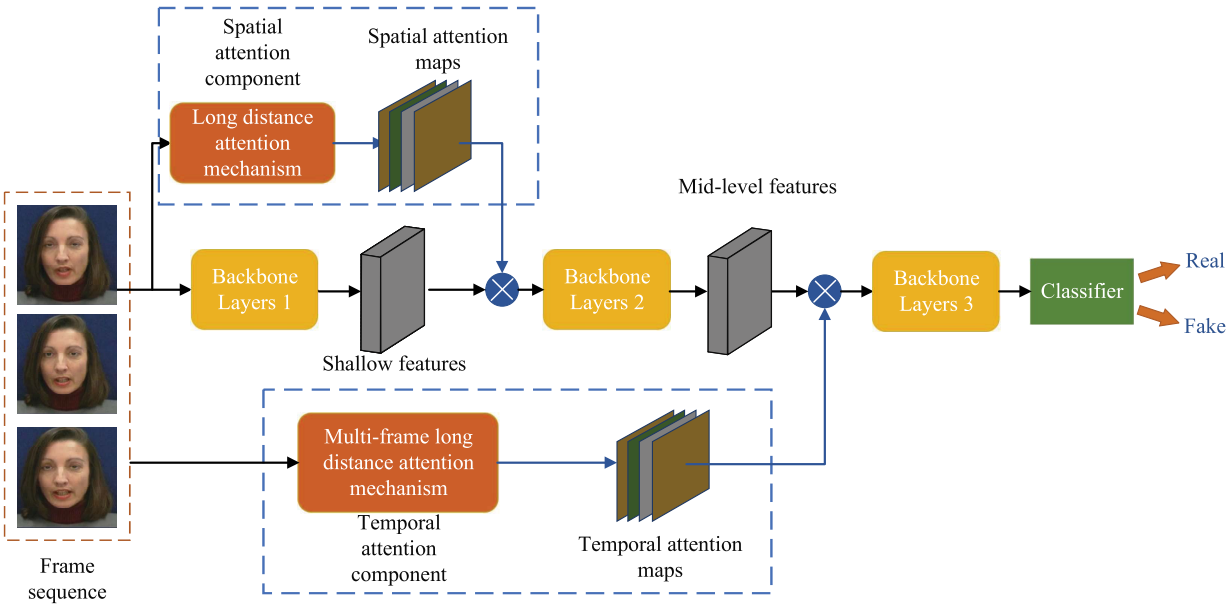

Fig. 4. Framework of the proposed method. There are two essential components in the framework. A spatial attention module for capturing the spatial defects in a single frame, and a temporal attention module for capturing the temporal inconsistencies between consecutive frames. The components are used to generate guidance to make the backbone network focus on pivotal local regions.

The backbone adopted is the Xception [51] which performs well in the vision field.

About the loss function, considering that most of the public datasets do not have a ground-truth of marked defect areas, at the same time, in the real deepfake scenario, the highly realistic forgery makes it difficult to mark valuable suspicious areas even for a human. Thus we absorb the experience in the fine-grained classification field to adopt a weakly supervised learning way. This also brings some benefits. The proposed attention module can be easily applied to more models as a pluggable module, and the performance impact brought by the attention module itself can be compared more fairly, rather than influenced by the additional prior benefits brought by annotation. In general, the loss function of the whole model is a cross-entropy loss of the output of the network and the label of the input.

B. Long-Distance Attention

In faces, a semantic region is a small area with rich information, like human eyes. Based on the observation we have mentioned, the long-distance attention mechanism is proposed to model the interdependencies between the semantic regions to perform feature recalibration mediately. It contributes to using global information to selectively emphasize informative regions and suppress regions that are useless for forgery detection.

As the key parts of face forgery can be regarded as many small areas with abnormal clues, the image is divided into many nonoverlapping small patches. These patches contain the local statistical information that might imply potential forgery clues. And then the weight of fake confidence for a small area corresponding to each patch is obtained which is achieved by long-distance attention.

Denote the input image as $I \in \mathbb{R}^{H\times W\times C}$, and the resolution is $H \times W$, C is the number of channels. The image is divided into a sequence of small patches $P = [p_1, p_2, \ldots, p_N]$. Therefore, there will be $N = HW/s$ patches, each one has C channels and the resolution $s \times s$. Then each patch is flattened and mapped to a D dimension vector with a trainable linear projection $f_z(P)$, which transfers patches to embedding $Z = [z_1, z_2, \ldots, z_N]$ for ease of processing [52]. Considering that the position of each patch reflects the spatial relationship between them, in order to reserve positional information, position embedding is added to the patch embedding to compose patch features [53]. The position embedding is shaped in a learning way. To model the internal relationship between the patch features, a necessary global forgery template t is utilized [24]. The template t is used to model the global association of a latent forgery property space. In order to intuitively understand the so-called latent forgery property space, an inaccurate example is the optical flow space of the patches, the optical flow sometimes reflects an irregular variation of the deepfake videos. Since there may be more than one forgery property space, multiple templates are adopted. At the same time, the patch features will be mapped to the representations $X = [x_1, x_2, \ldots, x_N]$ in each latent space, which is implemented with a learnable transformation matrix U. Both the matrix and template are shaped in a learning way. After that, the template in each latent forgery property space is used to consult each representation to get the forgery property activation [54]. The activation is treated as the attention weight, and adopted to guide the feature maps.

As shown in Fig. 5, long-distance attention consists of three main steps: 1) the patches are flattened to patch embedding Z and added the position embedding to compose the patch features; 2) the patch features are mapped to the representations $X = [x_1, x_2, \ldots, x_N]$ of a latent forgery property space, by a learnable transformation matrix U; and 3) finally, the global forgery template t is used to consult each representation to obtain the activation rate of each representation.

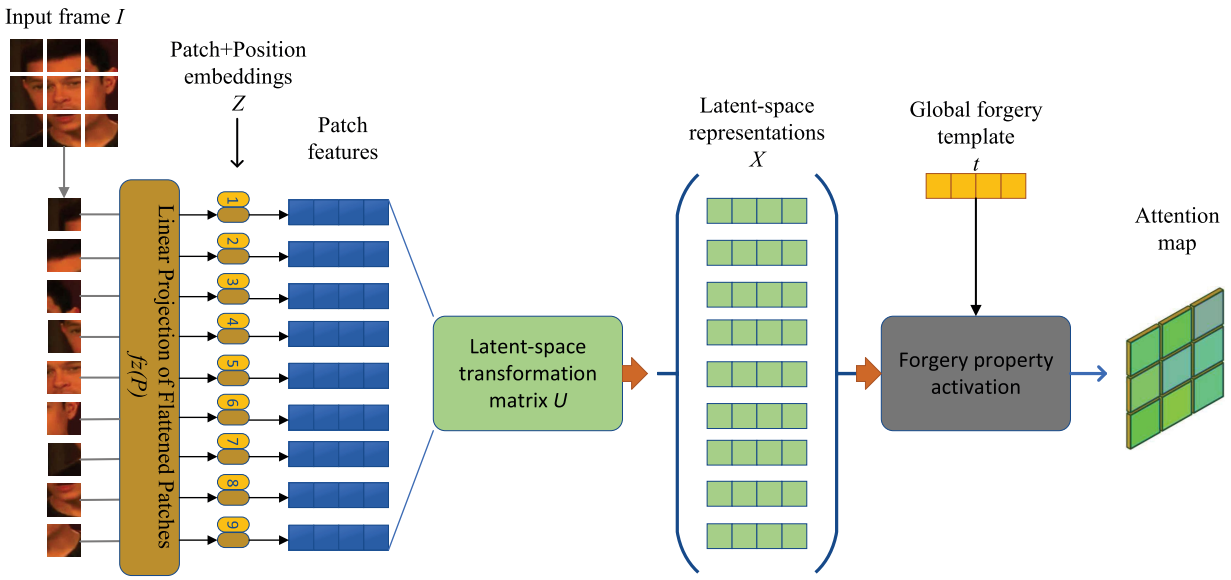

Fig. 5. Proposed long-distance attention mechanism. The image is split into small patches. The patches are linearly projected to patch embeddings and the position embedding is added. Then, the embedding is transformed into representations in a latent space by a matrix. Lastly, a global forgery template rectified by learning is used to activate the forgery property of each representation to generate attention maps.

Since the activation rates represent the confidence level of each patch with the suspicious region, they are reshaped to an attention map with the same resolution as the feature maps of the backbone, and applied by elementwise multiplication to emphasize pivotal regions.

As aforementioned, there are not only one forgery property space, in fact, we adopt 12 such attention module to produce different attention maps of different latent space, and linearly combined into m final attention maps for a robust and efficient reason [19], more discussion is given in VI-B.

C. Spatial Attention Model

In this section, we introduce the overall spatial attention model in detail. The spatial attention model is designed to capture the artifacts that existed in the spatial domain with a single frame. As aforementioned, since there is no precise global constraint between face parts which will introduce disharmonious facial structures and mismatched texture details [23], it is beneficial to generate guidance from a global aspect. Most of the existing methods use pooling to deal with the problem [41], such as global pooling, channel pooling, and so on. However, compared with the local association, the global association information is very weak and difficult to be established [12]. On the other hand, defects such as oversampling and insufficient texture appear in the local area, so an appropriate size of the local receptive field is benefiting for the collection of this statistical information. With the long-distance attention mechanism, these problems can be well balanced.

As we want to use the long-distance attention mechanism to capture the defects of the spatial domain in a global perspective, a single frame of the tested video is used as the input. And to recalibrate the importance between regions, the attention maps generated by a single frame are adopted to the feature maps of the backbone network. As textural features exist in shallow features [19], we make the attention works with the first several layers of the backbone. More specifically, the input image I which is used for the backbone and the spatial attention model is reshaped to the resolution 398×398 and 224×224 respectively. Then the convolutional feature maps are extracted by the first several layers of the backbone. And the spatial attention module receives a relatively small image which is tackled by the attention mechanism we have described above. Finally, the attention maps generated by the mechanism are element-wise multiplied by shallow feature maps to get the emphasized feature maps.

D. Temporal Attention Model

The movement of human faces is a complex and delicate process. For example, the facial expressions and head movements are strongly correlated [55] and changing the former without modifying the latter may expose a manipulation [22]. However, since the deepfake videos are synthesized frame by frame and do not precisely model the correlation between frames, it almost inevitably introduces inconsistencies. In order to capture these temporal inconsistencies, consecutive frames of video are required. For the frame to be detected, the next n frames are also utilized in the temporal attention model. The number n is determined by experiments in Section VI-C. For temporal attention, we care more about the variation of videos in the time dimension, so we calculate the motion residuals between adjacent frames as the inputs. The calculation of the motion residual r^t can be formulated as

$$r^t = I^{t+1} - I^t.$$

The temporal difference operation is simple and does not introduce any extra parameters but is capable of modeling the temporal inconsistency efficiently [56]. As shown in Fig. 6, the frame I and the motion residuals are split into patches,

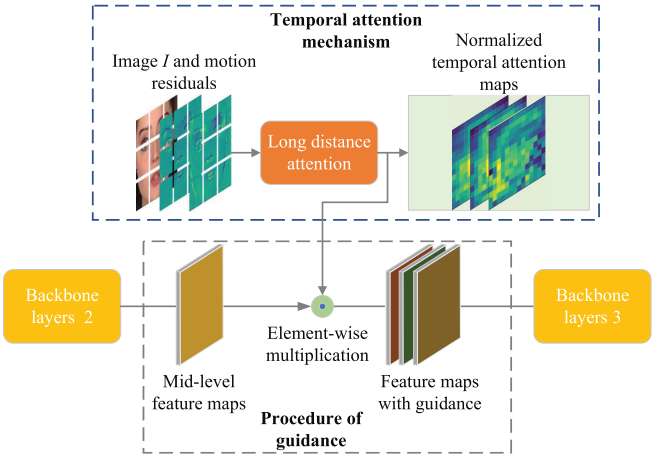

Fig. 6. Process of temporal attention generation. The image of the tested video and its following motion residuals are used as inputs. Then the attention maps are generated by the long-distance attention and adopted to guide the mid-level feature maps.

and all of the patches are composed of a sequence to be the input of the model. In this way, the template t of the temporal attention model is used to model the inconsistency between frames and to obtain the activation rate of each region in a latent inconsistency space. The activation rates are used as the attention weights and represent the confidence of inconsistency in each region. Since the inconsistency in the time domain is relatively high-level semantics compared with the features in the spatial domain, the temporal attention maps are applied to relatively high-level feature maps. In the same way, the attention maps are reshaped to the same size with relatively high-level feature maps and element-wise multiplied.

V. Experiments

In this section, the experiment setups are introduced first and then we present extensive experimental results to demonstrate the superiority of our method.

A. Datasets and Implementation Details

Two mainstream deepfake datasets are used in our experiment, including FaceForensics++ (FF++) [57] and Celeb-DF [58]. FaceForensics++ and Celeb-DF are both large-scale datasets that are widely used in face forgery detection. FaceForensics++ dataset consists of four kinds of face forgery videos, which are generated by four state-of-the-art methods, i.e., DeepFake (DF) [4], FaceSwap (FS) [5], Face2Face (F2F) [59] and NeuralTexture (NT) [60]. For each video of FF++, it has two compression versions (i.e., HQ, LQ), which are compressed by H.264 with constant rate quantization parameters set by 23 and 40. Celeb-DF is a great challenge to the current detection methods. It consists of more than 5000 deepfake videos, and the real videos are gathered from social media. Benefiting from an elaborate generation model, the generated videos are very realistic. For all video frames, we use Dlib [61] to detect and crop faces. The aligned facial images are resized to 398 $\times$ 398 for the backbone network and 224 $\times$ 224 for attention modules respectively. And the size of all patches is set to 16 $\times$ 16.

TABLE I

Quantitative Comparisons Among Recent Methods and the Proposed on FaceForensics++ Datasets With Low-Quality (Heavy Compression) and High-Quality (Light Compression). ACC(%) and AUC(%) Are Adopted, and the Best Performances Are Marked as Bold

Methods	LQ		HQ	
	ACC	AUC	ACC	AUC
Steg. Features [62]	55.98	–	70.97	–
Cozzolino et al. [63]	58.69	–	78.45	–
Bayar et al. [11]	66.84	–	82.97	–
MesoNet [7]	70.47	–	83.10	–
Face X-ray [64]	–	61.60	–	87.40
Two Branch [65]	–	86.59	–	98.70
Xception [51]	86.86	89.30	95.73	96.30
EfficientNet-B4 [66]	86.67	88.20	96.63	99.18
Multi-attentional [19]	86.95	87.26	96.37	98.97
F^3-Net [67]	90.43	93.30	97.52	98.10
M2TR [28]	92.35	94.22	98.23	99.48
Ours	**95.81**	**98.49**	**99.51**	**99.88**

TABLE II

Quantitative Comparisons on Celeb-DF Datasets. ACC(%) and AUC(%) Are Adopted

Methods	ACC	AUC
MesoNet [7]	–	53.6
I3D [68]	76.08	83.00
C3D [69]	78.67	84.00
FaceNetLSTM [70]	79.83	–
Hu et al. [71]	80.74	87.00
Xception [51]	89.55	89.91
FakeCatcher [12]	91.50	–
XcepTemporal [72]	97.83	–
Ours	**99.13**	**99.87**

Xception [51] is the backbone we adopted which has 12 main blocks and some feature extraction layers at the beginning. Our framework is implemented by PyTorch. During the training phase, the backbone network is initialized randomly and the ViT network in our long-distance attention mechanism is initialized by its pre-trained weights. The batch size is set to 32. In our experiments, the learning rate is set as 0.0003, which is determined by experiments. The networks are optimized by SGD with momentum $=$ 0.9. The total number of training epochs is set to 20, and the learning rate is reduced to half every five epochs. The quantity of attention maps is set by experiments, and the default number is 4, more discussion is given in Section VI-B.

Face forgery detection is a binary classification task, that is, gives a judgment of the tested video whether it is fake or real. Two evaluation metrics are adopted in our experiments, Accurency rate (ACC) is the most intuitive evaluation metric. AUC is another metric we adopted.

B. Comparison Experiments

Comparisons are conducted between current state-of-the-art deepfake detection methods and the proposed method. For deepfake video detection, there are frame-level and video-level detection methods. The frame-level detection method can give a decision for each frame of the video whether the frame has been tampered while the video-level detection method can give

TABLE III

QUANTITATIVE COMPARISON [FRAME-LEVEL ACC (%) AND AUC (%)] ON FACEFORENSICS++ WITH FOUR DIFFERENT MANIPULATION METHODS, I.E., DEEPFAKES(DF) [4], FACE2FACE(F2F) [59], FACESWAP(FS) [5], NEURALTEXTURES(NT) [60]. THE PROPOSED METHOD IS CAPABLE OF DEALING WITH DIFFERENT MANIPULATION METHODS

Methods	DF [4]		F2F [59]		FS [5]		NT [60]	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Steg. Features [62]	73.64	–	73.72	–	68.93	–	63.33	–
Cozzolino et al. [63]	85.45	–	67.88	–	73.79	–	78.00	–
Rahmouni et al. [73]	85.45	–	64.23	–	56.31	–	60.07	–
Bayar et al. [11]	84.55	–	73.72	–	82.52	–	70.67	–
C3D [69]	85.10	91.00	73.12	88.00	72.11	87.00	60.30	59.00
Hu et al. [71]	94.64	98.00	86.48	94.00	85.27	94.00	80.05	90.00
MesoNet [7]	87.27	–	56.20	–	61.17	–	40.67	–
Xception [51]	95.15	99.08	83.48	93.77	92.09	97.42	77.89	84.23
Spatial-phase [47]	93.48	98.50	86.02	94.62	92.26	98.10	76.78	80.49
FakeCatcher [12]	94.87	–	96.00	–	95.75	–	89.12	–
Ours	**99.47**	**99.79**	**99.98**	**100.00**	**98.27**	**99.46**	**93.25**	**98.61**

TABLE IV

QUANTITATIVE COMPARISON [VIDEO-LEVEL ACC (%) AND AUC (%)] ON FACEFORENSICS++ WITH FOUR DIFFERENT MANIPULATION METHODS, I.E., DEEPFAKES(DF) [4], FACE2FACE(F2F) [59], FACESWAP(FS) [5], NEURALTEXTURES(NT) [60]. THE PROPOSED METHOD IS CAPABLE OF DEALING WITH DIFFERENT MANIPULATION METHODS

Methods	DF [4]		F2F [59]		FS [5]		NT [60]	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Xception [51]	96.79	97.64	91.79	93.17	96.07	97.05	81.79	82.36
F^3-Net [67]	97.14	97.85	94.64	95.60	97.14	97.48	83.57	84.64
High-frequency [29]	97.86	98.20	94.64	95.54	96.79	97.64	81.79	83.89
Multi-attentional [19]	98.57	99.07	93.93	95.46	97.86	99.05	85.00	87.94
Ours	**99.29**	**99.64**	**99.64**	**99.75**	**98.58**	**99.81**	**94.29**	**96.31**

a decision whether the video has been tampered [23], [74]. As our method is capable of making decision for each frame, it is a frame-level method. We first perform our experiments on FaceForensics++ [57], which has been widely tested in this field. As aforementioned, there are two compression versions in FaceForensics++ [57], HQ represents the low-level compression version, and LQ represents the high-level compression version. The performances demonstrated in Table I are tested on both HQ (c23) and LQ (c40) versions with ACC and AUC metrics. The experimental results indicate that the proposed method achieves state-of-the-art performance on both versions of FaceForensics++ [57]. In general, the performance of most methods in high-compressed video is not as good as that in low-compressed video. This is because the video will lose a lot of texture details after high compression, which is one of the main pieces of information that networks need to pay attention to. Since the proposed method takes into account the inconsistencies of multiple frames in the time domain, compared with the other methods, the performance degradation is relatively small. Another noteworthy point is that, compared with the backbone Xception [51], the proposed method has a significant improvement as shown in Fig. 7, which is benefiting from the spatial-temporal guidance.

The Celeb-DF [58] datasets are also adopted. As shown in Table II, although the Celeb-DF is very realistic, the proposed model can effectively capture the defects and achieve better performance than the other methods.

To evaluate the spatial-temporal model's ability to capture defects introduced by different manipulation methods, the model is trained and tested on different manipulation methods

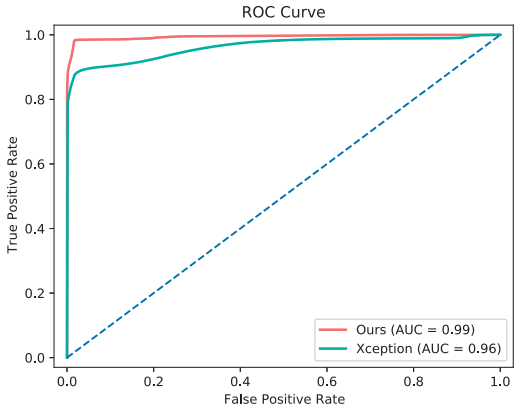

Fig. 7. ROC curves for the Xception and spatial-temporal model on HQ of FaceForensics++.

TABLE V

COMPARISON OF SOME ATTENTION MODELS AND OUR PROPOSED ON FACEFORENSICS++

Models	ACC(%)
SENet [41]	78.22
Look closer [40]	87.47
MMAL [75]	92.33
WSDAN [20]	93.81
Ours	**99.51**

in FaceForensics++ [57]. As the results shown in Table III, for different manipulation methods, our method achieves better performances than the other methods. It confirms that the proposed spatial-temporal model is capable of capturing various kinds of defects introduced by different manipulation

TABLE VI

CROSS-DATASET EVALUATION WITH AUC(%). TRAINED ON HQ AND LQ OF FACEFORENSICS++ AND TESTED ON CELEB-DF. OUR METHOD OUTPERFORMS MOST DEEPFAKE DETECTION METHODS

Methods	FF++	Celeb-DF
Two-stream [10]	70.10	53.80
MesoNet [7]	84.70	54.80
FWA [76]	80.10	56.90
Xception [51]	99.70	48.20
Multi-task [77]	76.30	54.30
Capsule [78]	96.60	57.50
DSP-FWA [76]	93.00	64.60
Two Branch [65]	93.18	73.41
F^3-Net [67]	98.10	65.17
M2TR [28]	99.50	65.70
Spatial-phase [47]	96.91	76.88
High-frequency [29]	–	**79.40**
Multi-attentional [19]	99.80	67.44
Ours	**99.97**	70.33

methods. This may be because the defects introduced by these operation methods have some common characteristics, and the attention mechanism helps to excavate these characteristics. We also conduct experiments for video-level prediction on LQ(c40) versions of FaceForensics++ to decide whether a video is fake or not. From Table IV we can see that due to our spatial-temporal attention which takes inconsistency between consecutive frames into account, our method achieves the best performance on four different manipulation datasets.

As we treat the problem as a fine-grained classification problem, we compare our model with some state-of-the-art fine-grained classification models. The models are reproduced and migrated to deepfake detection. As shown in Table V, the performance is acceptable but not satisfactory compared with the SOTA deepfake detection method.

C. Cross-Dataset Performance

In this part, the transferability of our framework is evaluated. The cross-dataset result is shown in Table VI. To compare with other methods, we train our model on both HQ and LQ of FaceForensics++ [57] and tested on Celeb-DF [58]. Since there are many differences between datasets, such as different video compression methods, common scenes, camera angles, and so on, it is a challenging task for most detection methods. All the methods have different degrees of decline in the cross-dataset task. As the experimental results show, although our method is not specially designed for cross-dataset performance, it still has a better performance than most methods. Two-Branch [65] is elaborately designed for transferability and achieves better results. However, our in-dataset performance is better than theirs. The comparison with the backbone network also confirms that our spatial-temporal model can effectively emphasize local regions, thus improving transferability.

D. Ability of Capturing Defects

Since the defects in deepfake videos are subtle, it may not be an initiative for human eyes to discriminate the differences between the attention maps generated by real and fake faces. To intuitively understand how long-distance attention works, we manually add the defects that we mentioned in Fig. 3 to the frames of authentic videos and examine the differences between attention maps of the real and tampered faces. Although deepfake videos generally do not produce such obvious traces of forgery, the use of these obviously tampered faces helps to intuitively understand how long-distance attention can capture these local and global defects. As shown in Fig. 8, the first column consists of real faces, the second column consists of attention maps of the real faces, the third column is the tampered version of the first column by a certain manipulation, and the last column consists of attention maps of the tampered faces. The areas highlighted in the fourth column but not highlighted in the second column are marked with red boxes, and it can be seen that they are highly coincident with the tampering position in the face. The first row is an example of local defects. The tampered image is Gaussian blurred to simulate the texture defects in deepfake videos. The mouth area of the real face is blurred, and it can be seen that the attention map generated by the tampered face highlights the corresponding area. The second row is an example of mismatched eyes. The face's right eye pupil is painted red. Obviously, the area of the attention map corresponding to the abnormal eye is highlighted. The last row is an example of abnormal face structures. The mouth of the face is distorted. Therefore, the generated attention map highlights the corresponding area of this abnormal structure. These results indicate that a long-distance attention mechanism can capture the defects in local and global perspectives. Thus, a long-distance attention mechanism is useful to generate guidance from the local and global perspectives, and make the backbone network focus on the pivotal regions.

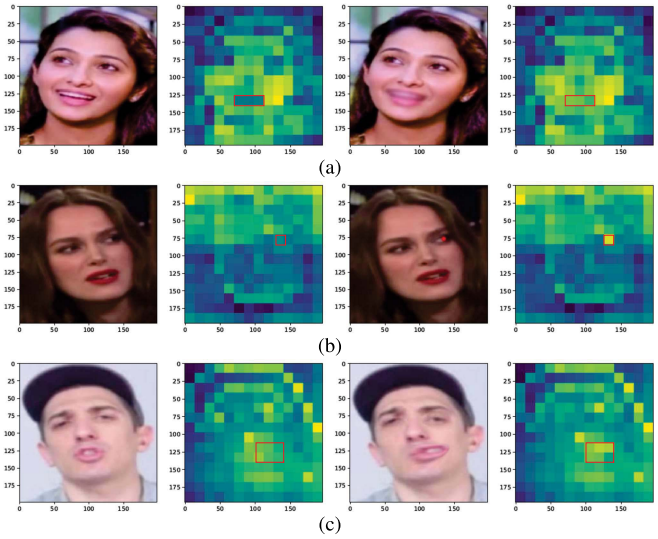

Fig. 8. Attention maps effectively emphasize the tampered region. The highlighted area in the attention map is highly coincident with the tampered area of the real face. (a) Capturing local defects. (b) Capturing mismatched eyes. (c) Capturing abnormal structures.

E. Analysis on Failure Cases

Although it is difficult to pinpoint the logic of the neural network and how it makes bad decisions, we did find some typical errors in the detection process. As shown in Fig. 9, they can be roughly classified into two categories, one is abnormal inputs, such as incomplete faces, and sideways face,

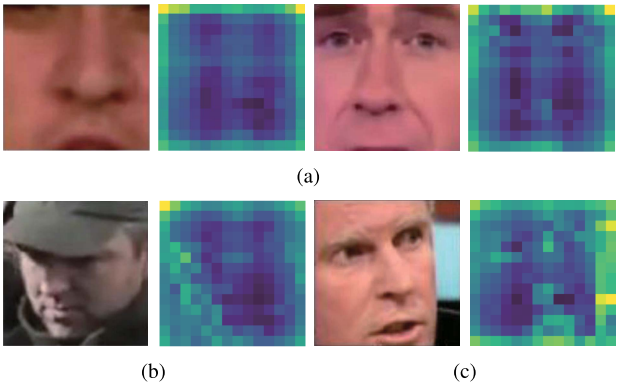

Fig. 9. Failure cases of our model and their corresponding attention maps, which can be roughly classified into abnormal input (incomplete face, excessive head angle) and abnormal attention. The model misclassifies these real images as fake. (a) Incomplete face. (b) Excessive head angle. (c) Abnormal attention.

and the other is abnormal processing of the model, such as generating abnormal attentions. Attention maps generated by our model assign lower weights to facial area and higher weights at the frontier, which can hardly help the network focus on important regions and make wrong predictions. To deal with the abnormal inputs, as we know, the neural network can handle the data is has "seen" very well, so it will be helpful to manually eliminate the abnormal data to improve the performance, or through data enhancement means to increase the quantity of such data. To deal with the errors of the model itself, it is helpful to introduce some additional guidance and confrontation ideas to improve the robustness of the model and improve cross-dataset performance, which is also one of our future work directions.

F. Comparison on Computation Complexity

As a deep neural network model, computational complexity is an important measure. The flops (floating-point operations per second) represent the amount of computing needed for the network. The total params are also an import measure, reflecting the size of the model. We compare our models with Xception [51] and multiattentional [19] in flops, total params, and accuracy (ACC) on HQ of FaceForensics++. The results are shown in Table VII. The Xception is the backbone we adopted. The multiattentional [19] is another attention model for the deepfake detection task which shares the same perspective with us from fine-grained classification field. Compared with the backbone, our models have a significant performance improvement with an acceptable increment of computational complexity. Compared to multiattentional, although our models require a little extra computation, it has a significant performance improvement which presents the effectiveness of the novel long-distance attention mechanism.

VI. Ablation Analysis

In this section, we discuss the effectiveness of the temporal attention model and the spatial attention model respectively, and further discuss the influence of model parameters on performance.

TABLE VII

Comparison on Computation Cost and ACC on HQ of FaceForensics++

Models	Flops (G)	Total params (M)	ACC (%)
Xception [51]	14.60	22.38	95.73
Multi-attentional [19]	14.61	21.00	96.37
spatial	14.92	23.17	99.11
temporal	15.41	24.74	98.80
spatial-temporal	15.82	25.57	99.51

TABLE VIII

Models Are Trained and Tested on FaceForensics++ HQ, Both Spatial Model and Temporal Model Are Effective and Have a Significant Improvement, the Best Performance Is Achieved by the Spatial-Temporal Model

Models	Xception [51]	spatial	temporal	spatial-temporal
ACC(%)	95.73	99.11	98.80	**99.51**
AUC(%)	96.30	99.76	99.85	**99.88**

A. Effectiveness of Spatial-Temporal Model

To evaluate the effectiveness of the spatial model and temporal model, we separately use the spatial model, the temporal model, and the combination of the two models to compare the performance with the backbone. All of the models are trained and tested on the FaceForensics++ [57] with ACC and AUC metrics. The comparison results are shown in Table VIII. It can be clearly seen that the proposed temporal model and spatial model both have significant performance improvement compared with the backbone. The best performance is present by the spatial-temporal model, confirming that both the spatial model and temporal model are effective. More specifically, compared with the backbone network, each model has at least 3 percent performance improvement in ACC and AUC metrics, and the combination of the two will have a better effect. At the same time, it can be observed that the spatial model is slightly better than the temporal model. We think this may be because the defects in the spatial domain are more common in deepfake videos.

In order to understand the guiding role of attention maps intuitively, the attention maps produced by the model are visualized. The spatial attention maps are shown in Fig. 10. The first two columns of attention maps are generated from real video frames, while the last two columns of attention maps are generated from forged video frames. Although all attention maps successfully capture the semantic regions of the human face, the slight difference is that the highlight regions of spatial attention maps from fake videos are more concentrated. This phenomenon is also reflected in the temporal attention maps. As shown in Fig. 11, the weight of temporal attention maps generated by real video frames is more uniform, while the temporal attention map generated by fake video focuses on a few areas. We think it's caused by irregular, tiny jitters that often occur in deepfake videos, especially near the mouth and the edge of the face. It is consistent with the highlight of the temporal attention maps of fake.

B. Quantity of Attention Maps

In order to enhance the diversity of the guidance generated by the spatial model and temporal model, and avoid

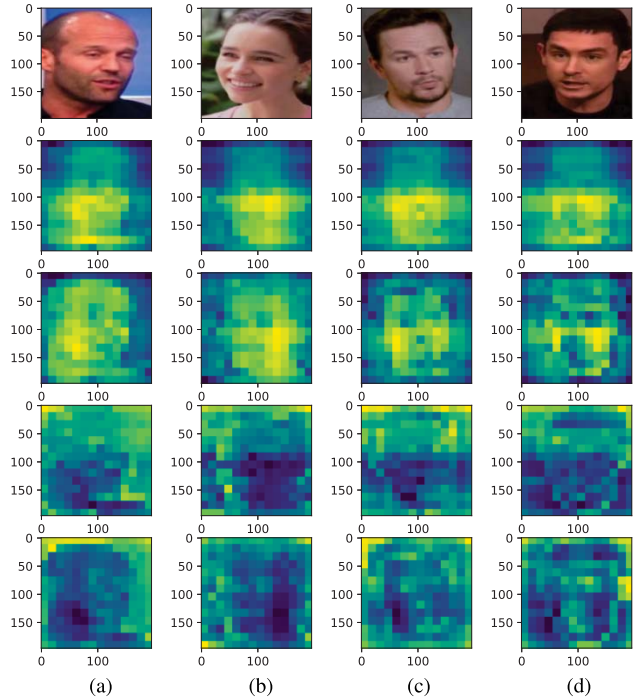

Fig. 10. Spatial attention maps generated by real videos and fake videos. The highlighted regions represent potential artifacts in spatial domain and fake videos are more concentrated. (a) real1. (b) real2. (c) fake1. (d) fake2.

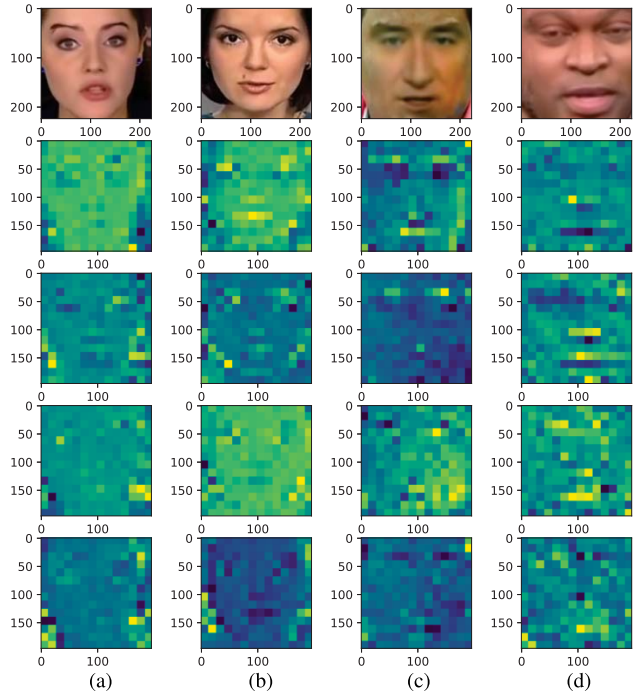

Fig. 11. Temporal attention maps generated by real videos and fake videos. The highlighted regions represent degree of inconsistency between consecutive frames within that area in temporal domain and fake videos are more concentrated. (a) real1. (b) real2. (c) fake1. (d) fake2.

TABLE IX
COMPARISON OF MODELS WITH DIFFERENT NUMBER m OF ATTENTION MAPS

m	1	2	3	4	5
ACC(%)	97.79	99.35	99.47	**99.51**	99.38
AUC(%)	99.73	99.62	**99.94**	99.88	99.91

TABLE X
COMPARISON OF MODELS WITH DIFFERENT NUMBER n OF CONSECUTIVE FRAMES

n	2	3	4	5
ACC(%)	99.33	**99.51**	99.48	99.12
AUC(%)	99.85	99.88	**99.93**	99.76

generating guidance limited from a single latent space, multiple long-distance attention modules are used in each model. At the same time, in order to enhance the robustness and stability of the guidance, 1×1 convolution kernel is used to combine the guidance and generate m final attention maps. To verify the effectiveness of the multiattention maps and explore the optimal quantity of attention maps, experiments are conducted on the influence of the quantity of attention maps on the performance of the model. The models are trained on FaceForensics++ [57] with the same hyper-parameters except for the quantity of attention maps. As the result shown in Table IX, since multiple attention maps provide more diversity of guidance, the model using multiple attention maps has a better performance than the model using a single attention map, and the best ACC is obtained with $m = 4$, and the best AUC is obtained with $m = 3$. When the number of maps increases to a certain number, blindly increasing the number cannot bring obvious performance improvement.

C. Quantity of Consecutive Frames

In the temporal model, multiple consecutive frames are used to mine the inconsistencies. Although more consecutive frames carry more temporal information, too many sequences will make it difficult for the model to establish information association. In order to explore how many consecutive frames can provide enough temporal information for the proposed model, the temporal models with different numbers of consecutive frames are used to explore the optimal number of consecutive frames. The experimental results of different numbers of consecutive frames are shown in Table X. It can be seen that three consecutive frames are enough for the proposed temporal model to build the information association of the patches in the time domain.

D. Resolution of Inputs

The image resolution is a factor that can affect the performance of the neural network. The input of the long-distance attention module is 224×224, which is the input requirement of ViT since we need to load the pre-trained weights of ViT partly, so the input size of this part is fixed. Meanwhile, the attention map produced by the attention module represents the importance of each patch. This also allows the network to focus on the key areas, which is applied by element-wise multiplication between feature maps and corresponding attention maps. We choose 398×398 resolution because of the need for size correspondence when multiplying. With this resolution, the size of the shallow feature map is exactly an integer multiple of the size of the attention map, so we can expand the attention maps linearly and project the attention map onto the feature map, which is why our feature map emphasizes the key areas in the form of patches. Without this

TABLE XI
QUANTITATIVE COMPARISONS AMONG RECENT METHODS AND THE PROPOSED ON CELEB-DF DATASETS WITH DIFFERENT RESOLUTION

Methods	Resolution			
	256×256	299×299	398×398	448×448
High-Frequency	89.97	90.02	90.36	91.51
RFM	**96.68**	**97.18**	97.65	**97.31**
Multi-attentional	93.36	93.5	94.9	94.26
Xception	90.61	90.82	90.87	90.46
Ours	94.87	95.4	**99.13**	96.91

correspondence, we need to use adaptive averaging pooling to match the sizes of the attention maps and the feature maps, which will bring the loss of information and impact the correspondence between the attention map and the feature map. Furthermore, it will hurt the performance as well. Besides, for fine-grained classification tasks, in order to make the features after convolution still have sufficient identification, so as to distinguish important local areas, the image resolution will be higher. Higher resolution helps achieve higher performance on fine-grained visual categorization datasets [79]. Therefore, we absorbed their experience and used a relatively large image resolution to ensure sufficient differentiation of the region and granularity of the attention guidance.

In order to explore the influence of image resolution on the current deepfake detection methods, experiments are conducted on Celeb-DF with several SOTA methods at different resolutions. For our method, since the size of the shallow feature map obtained with other resolutions is not an integer multiple of the size of the attention map, we use adaptive averaging pooling to match the sizes of the attention maps and the feature maps. The other methods are reproduced based on their published code and modified as less as possible to adapt to the inputs of different resolutions. It is worth pointing out that, all of the methods use Xception [51] as the backbone to make it easier to compare the effectiveness of modules. As shown in Table XI, for most of them the effect of high resolution is positive as more than one percent performance improvement between large and small resolutions. The proposed method achieves the best performance at 398×398 resolution and is more affected by resolution. This is because, at resolution 398×398, the patch-based activation weights can affine into patch-based attention maps. At the same time, the global average pooling makes patch-based weak supervision less feasible. Thus in order to play a better role in the proposed module, the size of the shallow feature map should be exactly an integer multiple of the size of the attention map.

VII. CONCLUSION

In this article, we detect deepfake video from the perspective of fine-grained classification since the difference between fake and real faces is very subtle. According to the generation defects of the deepfake generation model in the spatial domain and the inconsistencies in the time domain, a spatial-temporal attention model is designed to make the network focus on the pivotal local regions. And a novel long-distance attention mechanism is proposed to capture the global semantic inconsistency in deepfake. In order to better extract the texture information and statistical information of the image, we divide the image into small patches and recalibrate the importance between them. Extensive experiments have been performed to demonstrate that our method achieves state-of-the-art performance, showing that the proposed long-distance attention mechanism is capable of generating guidance from a global perspective. Apart from the spatial-temporal model and the long-distance attention mechanism, we think the main contribution of this article is that we confirm not only focusing on pivotal areas is important, but combining global semantics is also critical. This is a noteworthy point, which can be a strategy to improve current models.

REFERENCES

[1] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[3] Q. Duan and L. Zhang, "Look more into occlusion: Realistic face frontalization and recognition with BoostGAN," *IEEE Trans. Netw. Learn. Syst.*, vol. 32, no. 1, pp. 214–228, Jan. 2021.

[4] *Deepfake*. Accessed: Sep. 18, 2019. [Online]. Available: http://www.github.com/deepfakes/

[5] *Faceswap*. Accessed: Sep. 30, 2019. [Online]. Available: http://www.github.com/MarekKowalski/

[6] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa, HI, USA, Jan. 2019, pp. 83–92.

[7] 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)*, Hong Kong, Dec. 2018, pp. 1–7.

[8] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing GAN-synthesized faces using landmark locations," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, Paris, France, Jul. 2019, pp. 113–118.

[9] D.-T. Dang-Nguyen, G. Boato, and F. G. De Natale, "Discrimination between computer generated and natural human faces based on asymmetry information," in *Proc. 20th Eur. Signal Process. Conf.*, Bucharest, Romania, Aug. 2012, pp. 1234–1238.

[10] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1831–1839.

[11] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, Vigo, Spain, Jun. 2016, pp. 5–10.

[12] 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.

[13] M. Li, B. Liu, Y. Hu, and Y. Wang, "Exposing deepfake videos by tracking eye movements," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, Jan. 2021, pp. 5184–5189.

[14] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Hong Kong, Dec. 2018, pp. 1–7.

[15] 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, doi: 10.1109/TIFS.2020.3045937.

[16] S. Fernandes et al., "Predicting heart rate variations of deepfake videos using neural ODE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, Seoul, South Korea, Oct. 2019, pp. 1721–1729.

[17] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 1205–1207.

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

[19] 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. 2185–2194.

[20] T. Hu, H. Qi, Q. Huang, and Y. Lu, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," 2019, *arXiv:1901.09891*.

[21] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 8261–8265.

[22] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Los Angeles, CA, USA, Jun. 2019, pp. 1–8.

[23] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.

[24] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and local structure preservation for feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1083–1095, Jun. 2013.

[25] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

[26] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," 2021, *arXiv:2102.11126*.

[27] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Deepfake detection scheme based on vision transformer and distillation," 2021, *arXiv:2104.01353*.

[28] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2TR: Multi-modal multi-scale transformers for deepfake detection," in *Proc. Int. Conf. Multimedia Retr.*, 2022, pp. 615–623.

[29] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16317–16326.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[32] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2846–2854.

[33] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, Jun. 2016.

[34] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Los Alamitos, CA, USA, Jun. 2015, pp. 842–850.

[35] H. Wang, L. Jiao, S. Yang, L. Li, and Z. Wang, "Simple and effective: Spatial rescaling for person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 145–156, Jan. 2022, doi: 10.1109/TNNLS.2020.3027589.

[36] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1173–1182.

[37] H. Zhang et al., "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 1143–1152.

[38] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5219–5227.

[39] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognit.*, vol. 76, pp. 704–714, Apr. 2018.

[40] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4476–4484.

[41] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Apr. 2020.

[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, vol. 30, 2017, pp. 1–11.

[43] P. Tirupattur, K. Duarte, Y. S. Rawat, and M. Shah, "Modeling multi-label action dependencies for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1460–1470.

[44] T. Zhao, J. Han, L. Yang, B. Wang, and D. Zhang, "SODA: Weakly supervised temporal action localization based on astute background response and self-distillation learning," *Int. J. Comput. Vis.*, vol. 129, no. 8, pp. 2474–2498, Aug. 2021.

[45] L. Yang, H. Peng, D. Zhang, J. Fu, and J. Han, "Revisiting anchor mechanisms for temporal action localization," *IEEE Trans. Image Process.*, vol. 29, pp. 8535–8548, 2020.

[46] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016, *arXiv:1603.07285*.

[47] H. Liu et al., "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.

[48] M. Yang, L. Zhang, S. C.-K. Shiu, and D. Zhang, "Robust kernel representation with statistical local features for face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 900–912, Jun. 2013.

[49] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2201–2215, Nov. 2016.

[50] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?" in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, San Jose, CA, USA, Mar. 2019, pp. 506–511.

[51] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.

[52] Y. Liu, Q. Sun, X. He, A.-A. Liu, Y. Su, and T.-S. Chua, "Generating face images with attributes for free," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2733–2743, Jun. 2021.

[53] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, Oct. 2015.

[54] X. Fang, S. Teng, Z. Lai, Z. He, S. Xie, and W. K. Wong, "Robust latent subspace learning for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2502–2515, Jun. 2018.

[55] M. Shao, Y. Zhang, and Y. Fu, "Collaborative random faces-guided encoders for pose-invariant face representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1019–1032, Apr. 2018.

[56] Z. Gu et al., "Spatiotemporal inconsistency learning for deepfake video detection," in *Proc. 29th ACM Int. Conf. Multimedia*, New York, NY, USA, 2021, pp. 3473–3481.

[57] 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)*, Seoul, South Korea, Oct. 2019, pp. 1–11.

[58] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3204–3213.

[59] 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)*, Las Vegas, NV, USA, Jun. 2016, pp. 2387–2395.

[60] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, p. 66, Jul. 2019.

[61] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[62] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.

[63] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur.*, Philadelphia, PA, USA, 2017, pp. 159–164.

[64] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 5000–5009.

[65] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–(ECCV)*. Cham, Switzerland: Springer, 2020, pp. 667–684.

[66] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, vol. 97, 2019, pp. 6105–6114.

[67] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Computer Vision–(ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 86–103.

[68] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 4724–4733.

[69] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497.
[70] S. J. Sohrawardi et al., "Poster: Towards robust open-world detection of deepfakes," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, London, U.K., 2019, pp. 2613–2615.
[71] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1089–1102, Mar. 2022, doi: 10.1109/TCSVT.2021.3074259.
[72] A. Chintha et al., "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1024–1037, Aug. 2020.
[73] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proc. IEEE Workshop Inf. Forensics Secur. (WIFS)*, Rennes, France, Dec. 2017, pp. 1–6.
[74] X. Li et al., "Sharp multiple instance learning for deepfake video detection," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, 2020, pp. 1864–1872, doi: 10.1145/3394171.3414034.
[75] F. Zhang, M. Li, G. Zhai, and Y. Liu, "Multi-branch and multi-scale attention learning for fine-grained visual categorization," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2021, pp. 136–147.
[76] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Los Angeles, CA, USA, Jun. 2019, pp. 1–7.
[77] 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)*, Tampa, FL, USA, Sep. 2019, pp. 1–8.
[78] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 2307–2311.
[79] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4109–4118.

Wei Lu (Member, IEEE) received the B.S. degree in automation from Northeast University, Shenyang, China, in 2002, and the M.S. and Ph.D. degrees in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2005 and 2007, respectively.

He was a Research Assistant with Hong Kong Polytechnic University, Hong Kong, from 2006 to 2007. He is currently a Professor with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include multimedia forensics and security, data hiding and watermarking, and AI security.

Dr. Lu is an Associate Editor for the *Signal Processing* and the *Journal of Visual Communication and Image Representation*.

Lingyi Liu received the B.S. degree in information security from Jinan University, Guangzhou, China, 2020, and the M.S. degree in computer science and technology from Sun Yat-sen University, Guangzhou, in 2022.

His research interests include multimedia forensics and security.

Bolin Zhang received the B.S. degree in computer science and technology from Sun Yat-sen University, Guangzhou, China, in 2021, where he is currently pursuing the M.S. degree with the School of Computer Science and Engineering.

His research interests include multimedia forensics and security.

Junwei Luo received the B.S. degree in information security and the M.S. degree in computer science and technology from Sun Yat-sen University, Guangzhou, China, in 2020 and 2022, respectively.

His research interests include multimedia forensics and security.

Xianfeng Zhao (Senior Member, IEEE) received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2003.

Since 2012, he has been a Professor with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences (CAS), Beijing, China, where he is the Leader of Multimedia Security Group. He has also been a Professor with the University of CAS, since 2015. His research interests are the fields in multimedia security including information hiding, multimedia forensics, and the related AI technologies. In these fields, he has published more than 100 articles.

Dr. Zhao organized International Workshop on Digital-forensics and Watermarking (IWDW) five times as co-chairs. He is an Editorial Board Member of the journals on forensics including *IJDCF* and *FSI-R*.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, in 1992, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA, in 2008 and 2010, respectively.

He is currently a Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Prof. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE) and was recognized as the Highly Cited Researcher in Web of Science in 2020. He received the Third Price of the Macao Natural Science Award as a Sole Winner in 2020 and a co-recipient in 2014 and was a recipient of the Best Editor Award for his contributions to *Journal of Visual Communication and Image Representation* in 2020. He has been a leading Co-Chair of the Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society since 2015. He serves as an Associate Editor for the IEEE TRANSACTIONS ON NEUTRAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and four other journals.

Jiwu Huang (Fellow, IEEE) received the B.S. degree from Xidian University, Xian, China, in 1982, the M.S. degree from Tsinghua University, Beijing, China, in 1987, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 1998.

He is currently a Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include multimedia forensics and security.

Dr. Huang is a member of the Signal Processing Society Information Forensics and Security Technical Committee. He was a General Co-Chair of the IEEE Workshop on Information Forensics and Security in 2013 and a TPC Co-Chair of the IEEE Workshop on Information Forensics and Security in 2018. He is an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.