The Visual Computer (2024) 40:1083–1094
https://doi.org/10.1007/s00371-023-02833-x

ORIGINAL ARTICLE

Local attention and long-distance interaction of rPPG for deepfake detection

Jiahui Wu[1] · Yu Zhu[1,2] · Xiaoben Jiang[1] · Yatong Liu[1] · Jiajun Lin[1]

Accepted: 6 March 2023 / Published online: 29 March 2023

Abstract
With the development of generative models, abused Deepfakes have aroused public concerns. As a defense mechanism, face forgery detection methods have been intensively studied. Remote photoplethysmography (rPPG) technology extract heartbeat signal from recorded videos by examining the subtle changes in skin color caused by cardiac activity. Since the face forgery process inevitably disrupts the periodic changes in facial color, rPPG signal proves to be a powerful biological indicator for Deepfake detection. Motivated by the key observation that rPPG signals produce unique rhythmic patterns in terms of different manipulation methods, we regard Deepfake detection also as a source detection task. The Multi-scale Spatial–Temporal PPG map is adopted to further exploit heartbeat signal from multiple facial regions. Moreover, to capture both spatial and temporal inconsistencies, we propose a two-stage network consisting of a Mask-Guided Local Attention module (MLA) to capture unique local patterns of PPG maps, and a Temporal Transformer to interact features of adjacent PPG maps in long distance. Abundant experiments on FaceForensics + + and Celeb-DF datasets prove the superiority of our method over all other rPPG-based approaches. Visualization also demonstrates the effectiveness of the proposed method.

Keywords Digital video forensics · Deepfake · PPG · CNN

1 Introduction

Deepfake refers to a type of face manipulation or replacement methods based on deep learning. With the development of generative models [1–3], the technical barrier for face forgery is getting lower and lower, and anyone can easily create realistic face forged contents by ready-made models or tools. Deepfakes may also be exploited by malicious users to create false political information and spread pornographic content.

As a defense mechanism, face forgery detection has been proposed to respond to the challenge brought by Deepfake. The task of face forgery detection is commonly defined as a real-fake binary classification problem. According to the face forgery generation procedure, two helpful conclusions can be drawn to detect Deepfakes, including (1) pixel modification occurs only in local regions of the face, which inevitably leads to spatial inconsistencies such as blending boundaries, and (2) since the forged video is generated frame by frame, temporal inconsistency across frames like facial position jittering cannot be eliminated. Frame-level detection methods mainly focus on the first observation, while video-level approaches based on the second.

Face forgery detection via biological signals provides another way of thinking. Heartbeat signal is a typical biological signal. Photoplethysmography (PPG) is a heart rate monitoring technology used in biomedicine [4]. As the hemoglobin level changes due to periodic heartbeats, the skin's absorption rate of light changes accordingly. The development of remote photoplethysmography (rPPG) [5] technology makes it possible to capture the subtle changes in skin color from recorded videos. Since facial pixel modifications and inter-frame discrepancies inevitably disrupt periodic changes in skin color, previous work [6, 7] has proved that the rPPG signal is a powerful biological indicator for face forgery detection.

In this paper, the Multi-scale Spatial–Temporal representation of PPG is adopted to further exploit heartbeat signal

✉ Yu Zhu
zhuyu@ecust.edu.cn

1 School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

2 Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai 200032, China

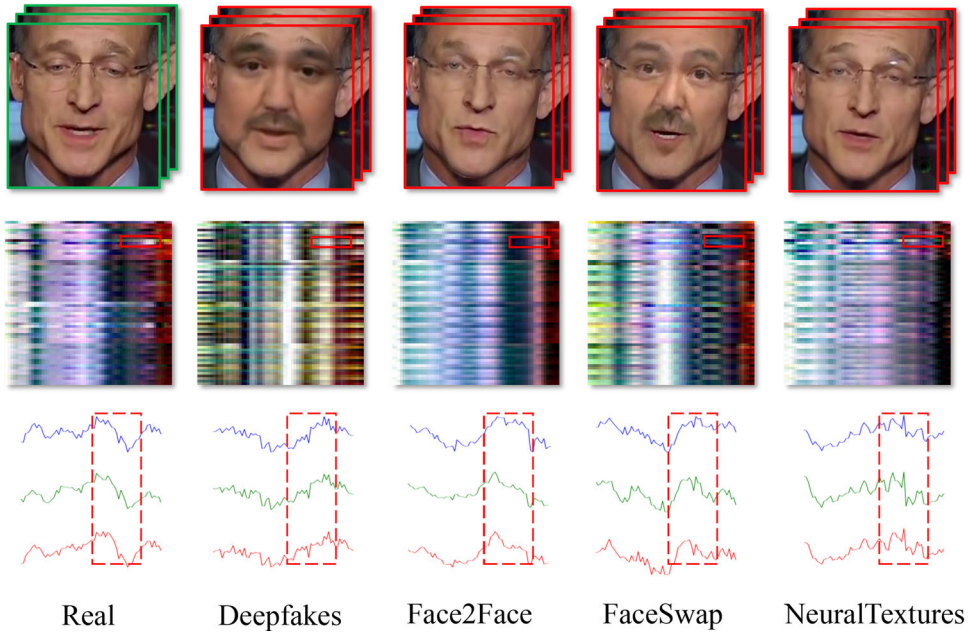

Fig. 1 An example of Multi-scale PPG maps (second row) and rPPG signals (third row) generated from real videos and various manipulations, i.e., Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Each forgery method presents unique rhythmic patterns of rPPG signals

calculated from different facial regions. As shown in Fig. 1, the key observation is that consistent rPPG signals are not yet preserved in Deepfakes, and pseudo signals produce unique rhythmic patterns in terms of different generation methods. Therefore, not only do we regard face forgery detection as a binary classification problem, but also as a source detection task for the recognition of different generation methods behind fake videos. To utilize both spatial and temporal information, a two-stage network is designed for face forgery detection and categorization. The Mask-Guided Local Attention module (MLA) is proposed to highlight the modified regions of PPG maps and guide the network to better detect the unique rhythmic patterns of different manipulation methods. Moreover, Transformer [8] is introduced to fully interact high-level temporal features between adjacent video clips in long distance. Abundant experiments prove the superiority of the proposed method, which outperforms all other rPPG-based methods in both face forgery detection and categorization. Extension experiment demonstrates the strong generalization ability of the proposed methods against newly added manipulation sources. To show the effectiveness of each component, we also conduct detailed ablation study on various setups for comparison.

In summary, the contributions of this paper are listed as three-fold:

(1) A two-stage network is designed to detect both spatial and temporal inconsistencies, which consists of a Mask-Guided Local Attention module (MLA) to highlight local regions of PPG maps and a Transformer to interact temporal adjacent features in long distance.

(2) We utilize the Multi-scale Spatial–Temporal representation of PPG which contains information of multiple facial regions. The visualization shows that unique patterns of PPG maps can be identified in terms of different generation methods with the help of Mask-Guided Local Attention module (MLA).

(3) Extensive experiments on various datasets are conducted to demonstrate the effectiveness and extension capability of the proposed method, which outperforms all other rPPG-based methods in terms of both forgery detection and categorization tasks.

2 Related work

2.1 Deepfake generation

Deepfake has been receiving more and more attention over the past decades. Variational Autoencoders (VAE) [9] and Generative Adversarial Networks (GAN) [1] are mainly adopted to generate Deepfakes in recent approaches. The existing facial manipulation methods can be divided into two major categories: facial identity manipulation and facial expression manipulation. Deepfakes (DF) refers to a type of facial identity manipulation method that was spread via online forums such as FakeApp which adopts two VAE models and a shared encoder to train and reconstructs the source and target faces. Poisson blending [10] and color transform algorithm [11, 12] are also used to mix the source face image with the background, which also applies to popular Deepfake open-source tools such as DeepFaceLab [13]. FaceSwap (FS)

is a graphics-based approach to transfer the face region from a source video to a target video based on sparse detected facial landmarks and 3D template model. FaceShifter (FSH) [14] is a novel two-stage framework designed for high fidelity and occlusion aware facial identity manipulation. On the other hand, Face2Face (F2F) [15] and NeuralTextures (NT) [16] are two typical facial expression manipulation methods. F2F is a facial reenactment system that manipulates the target video with the expressions of the source video while keeping the target person's facial identity unchanged. NT adopts the rich signal stored in learned neural textures of the target person and performs facial reenactment by deferred neural renderer. However, regardless of the manipulation method, there will be spatial texture inconsistencies in each frame since the video background is constant. Meanwhile, since the video tampering is operated frame by frame, Deepfake inevitably contains temporal discrepancies.

2.2 Deepfake detection based on CNN

Early researches [17–20] mainly use hand-crafted features to distinguish real and fake videos. With the rise of deep learning and the continuous upgrading of face forgery technology, Convolutional Neural Network (CNN) has become the mainstream of Deepfake detection. A number of well-designed backbone networks are used to extract crucial features, such as Mesonet [21], CapsuleNet [22], and Xception [23]. In addition to replacing the backbone, another kind of methods focuses more on the local spatial inconsistency within the forged frame. For example, FaceXray [24] detects forgery by revealing the blending boundaries of Deepfakes. Dang et al. [25] propose a plug-in local attention module to highlight features in the modified regions. PRRNet [26] fuses pixel-wise similarity and region-level similarity to learn local differences by spatial attention mechanism. Chen et al. [27] measure the similarity between different local areas by calculating Multi-scale Patch Similarity, and fuse frequency information with RGB channels to obtain a more comprehensive representation of local features. However, these methods only focus on frame-level forgery traces and tend to ignore cross-frame information at the video level.

On the other hand, many works use 3D CNN or Recurrent Neural Networks (RNN) to explore the temporal inconsistency of Deepfakes. Lima et al. [28] transfer the 3D network pre-trained on the action recognition task for video classification. Montserrat et al. [29] propose a weighting mechanism that automatically selects relevant frames and combine CNN with GRU [30] to extract both spatial and temporal features. With the success of ViT [8] in the field of computer vision, Transformer [31] has also been introduced to detect Deepfakes. Zheng et al. [32] propose a hybrid network combining a fully temporal convolution network with a Temporal Transformer. Xu et al. [33] fuse the visual semantic sequence with the contexture feature sequence extracted by Transformer. Khan et al. [34] utilize both RGB image and UV texture map as two-stream inputs for Transformer to learn the fused features. These methods demonstrate the effectiveness of the long-distance self-attention mechanism of the Transformer architecture in Deepfake detection. The proposed method utilizes spatiotemporal representation of PPG, combines CNN with Transformer, and adopts the local attention mechanism. In other words, we take both the spatial and temporal inconsistency of Deepfake into account.

2.3 Deepfake detection based on explainable methods

In addition to using pure CNN, another kind of method explores a variety of explainable methods. Malolan et al. [35] make use of explainable AI (XAI) techniques including Local Interpretable Model-Agnostic Explanations (LIME) and Layer-Wise Relevance Propagation (LRP) to provide clear visualizations of the salient regions of the image focused on by the model. Jayakumar et al. [36] propose a model-agnostic high precision explainer named "Anchors" XAI to visually explain the predictions of a deepfake detector and obtain better performance than LIME.

Benefit from their clear physical meanings, biological signals provide another scope of the explainable approaches. Early attempts adopt biological signals such as eye blinking [37], head posture [38], and lip movement [39]. The development of rPPG makes it possible to estimate heart rate from recorded face videos, and rPPG signals are also used in Deepfake detection. FakeCatcher [7] first introduces heartbeat signal into Deepfake detection and proposed a spatial–temporal map of chromatic-based PPG and its power spectral density (PSD). Ciftci et al. [6] adopt the same form of PPG maps and demonstrate that different manipulation methods generate their own unique heartbeat rhythms. Boccignone et al. [40] calculate path-wise rPPG signals and spectrums, and measures both intra-patch and inter-patch coherence of rPPG. DeepRhythm [41] uses Motion-Magnified Spatial–Temporal Representation (MMSTR) of PPG to enhance facial color changes and amplify heartbeat signals. This work also adopts prior predictions from face-based network to weight the input features. Liang et al. [42] further study the interaction between adjacent PPG maps. All these PPG-based methods prove that Deepfakes are not yet capable of maintaining consistent heartbeat signals, which is a strong and explainable evidence to detect forgery videos. However, these methods do not focus on the detailed local discrepancy between PPG maps generated by different manipulation sources. Without any prior knowledge, we adopt the Multi-scale Spatial–Temporal representation of PPG to comprehensively represent facial skin color changes caused by heartbeat activity. In addition, a two-stage network is proposed. On the one hand,

the Mask-Guided Local Attention module (MLA) is used to focus on the spatial local regions of PPG maps. On the other hand, the Temporal Transformer is utilized to further explore long-distance interactions between adjacent clips within a complete video.

3 Methods

In this section, we introduce the proposed overall framework illustrated in Fig. 2, including the generation of Multi-scale Spatial–Temporal PPG map, the two-stage network, and the loss function.

3.1 Multi-scale spatial–temporal representation of PPG

Since the heartbeat signal is sensitive to head movements, light changes, and other disturbances, the untreated face image cannot be directly used to represent the rPPG signal. In order not to be constrained by prior information on ROI selection, inspired by [43], we adopt Multi-scale Spatial–Temporal representation of PPG to fuse multi-region pixel information. As shown in Fig. 3, we first divide a full video into several T - frame video clips with the step size ω. For each video clip, face alignment is performed to obtain facial landmarks. According to the landmarks, set of sub-ROIs $R_t = \{R_{1t}, R_{2t}, \ldots, R_{nt}\}$ is obtained by selecting n informative regions of face such as cheeks, forehead, and jaw. Then, the average pixel values are calculated for all the non-empty subsets of R_t in C color channels. T - frame temporal sequences of averaged pixel values from the same sub-ROI region or combination are arranged into a row. Finally, a max–min normalization is applied to all the temporal sequences in each channel to scale the values into [0,255]. The size of the Multi-scale Spatial–Temporal PPG map is $(2^n - 1) \times T \times C$ for each video clip.

3.2 Overall framework

Our approach is based on the following two assumptions: (1) various video manipulation methods modify different facial regions, and these modifications are also reflected in the PPG map composed of multi-scale facial regions. Highlighting the modified local area may lead the network to better learn the unique rhythmic patterns of each manipulation method and help the network to distinguish between real and fake videos. (2) A single video contains multiple PPG clips, and sufficient interaction of the features from adjacent maps may yield more global information. Therefore, we propose a two-stage network consisting of a Mask-Guided Local Attention module (MLA) to focus on the modified local regions of the PPG map and a Temporal Transformer to exploit long-distance information between adjacent clips.

3.2.1 Mask-guided local attention module

Due to the unique patterns of rPPG signals, we regard Deepfake detection not only as a real-fake discrimination problem, but also as a categorization task of different manipulation methods. To be specific, face swapping methods change the pixels of the entire face area, while expression manipulation methods only modify the pixels of local regions, such as the mouth area. Since the spatial dimension of the PPG maps is arranged by the combination of different facial regions, the spatial–temporal representation of PPG can also reflect the regional discrepancies between each face manipulation method. This assumption is often ignored by previous approaches. Inspired by [25], we proposed a plug-in Mask-Guided Local Attention module (MLA) to highlight the position in the feature map of PPG that corresponds to the modified regions of the face image.

Concretely, the proposed MLA consists of the following steps. As shown in Fig. 3, given a PPG clip $X \in \mathbb{R}^{C\times(2^n-1)\times T}$, where T denotes the clip length, n is the number of face sub-ROIs, and C represents the number of inputted channels. The mid-level feature map F_m derived from the mid-layer of backbone f_{mid} can be formulated as $F_m = f_{\text{mid}}(X) \in \mathbb{R}^{C'\times H\times W}$ where H, W, C' denote the height, width and channel numbers of the feature map, respectively. Then, with F_m as the input, the attention mask $A_{\text{mask}} = \phi(F_m) \in \mathbb{R}^{H\times W}$ can be generated. The weighted feature $F' = A_{\text{mask}} \odot F_m$ is the input of the remaining network layer f_{high}, where $\odot$ denotes pointwise multiplication. Specifically, $\phi(\cdot)$ consists of a convolution operation $\text{Conv}(\cdot)$ for compressing channel dimension and a Sigmoid activation operation $\text{Sigmoid}(\cdot)$ to decide attention weights, which can be formulated as follows:

$$\phi(F_m) = \text{Sigmoid}(\text{Conv}(F_m)) \tag{1}$$

In order to approximate the attention mask A_{mask} with the ground truth manipulation mask A_{gt}, we train the MLA in a supervised manner and add an extra $L1$ loss function L_{mask}:

$$L_{\text{mask}} = \left|A_{\text{mask}} - A_{\text{gt}}\right|_1 \tag{2}$$

Given a pseudo PPG map which is generated from fake videos, its ground truth manipulation mask A_{gt} is calculated from its corresponding original map as a pair. To be elaborate, we first calculate the absolute pixel-wise difference of the PPG map pair in RGB channels to obtain a residual map. Then, the residual map is converted into grayscale, normalized to [0,1] and resized to the same scale of A_{mask}. Finally, a threshold of 0.1 is selected to determine the map as a binary

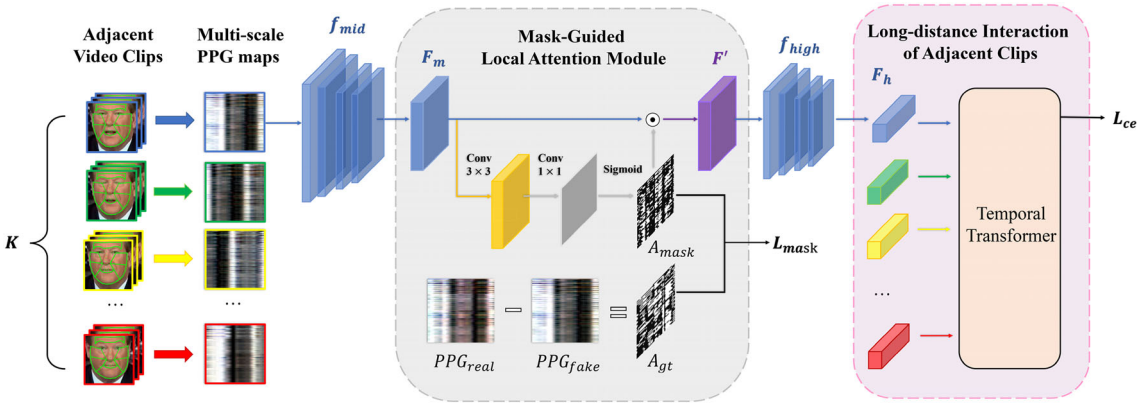

Fig. 2 Proposed pipeline of the two-stage network. ⊙ denotes pointwise multiplication

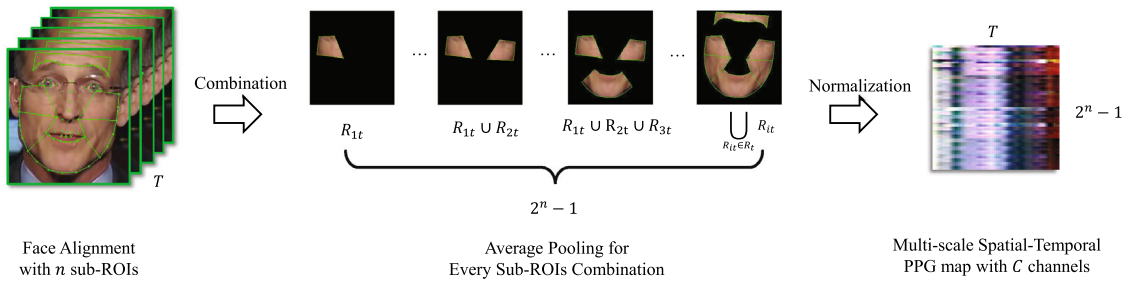

Fig. 3 An illustration of the generation process of the Multi-scale PPG map from an input face video clip of T frames. The procedure includes face alignment, sub-ROI combination, pixel average, and normalization. The final size of the Multi-scale PPG map is $(2^n - 1) \times T \times C$

mask A_{gt}. As for an original PPG map which is generated from real videos, its A_{gt} is set to all zeros because there is no manipulation occurred.

3.2.2 Temporal transformer

Although a single PPG map contains the temporal dimension, we believe that there is still potentially mutually reinforcing information in several adjacent PPG maps of the same video. In order to further mine temporal information, ViT [8] is utilized to interact adjacent clip features with each other in long distance.

As shown in Fig. 4, K adjacent PPG maps are inputted into the backbone network which has been well-trained in stage one, and the high-level features derived from the last convolutional layer of the backbone are denoted as F_h. Then, average pooling and linear operation are performed on F_h to obtain K vectors as D-dimension embedded features $x_i \in \mathbb{R}^D$, $i = 1, 2, \ldots, K$. Similar to the settings of ViT [8], an extra learnable class-token ($Z_0^0 = x_{\text{class}}$) is added to the embedding sequence, whose output is responsible for the final prediction. Meanwhile, the standard 1D learnable position embedding ($E_{\text{pos}} \in \mathbb{R}^{(K+1)\times D}$) is used to record the temporal order of K adjacent feature vectors. The input sequence of the Temporal Transformer can be formulated as follows:

$$Z_0 = [x_{\text{class}}, x_1, x_2, \ldots, x_i]^T + E_{\text{pos}}, \quad i = 1, 2, \ldots, K \tag{3}$$

The Temporal Transformer consists of L Transformer encoder blocks [31], and each encoder block includes a Multi-head Self-Attention operation (MSA) [31] and a Feed-Forward network (FF). The commonly used LayerNorm (LN) is applied before each block. And the structure of the residual connections [44] is utilized after every block. Activation function GELU is also used to ensure nonlinearity. The forward process of the l - th layer can be formulated as follows:

$$Z_l' = \text{MSA}(\text{LN}(Z_{l-1})) + Z_{l-1}, \quad l = 1, 2, \ldots, L \tag{4}$$

$$Z_l = \text{FF}\left(\text{LN}\left(Z_l'\right)\right) + Z_l' \tag{5}$$

To obtain the final prediction score y, MLP head is applied on the class-token output of the last layer ($LN\left(Z_l^0\right)$), which can be formulated as follows:

$$y = MLP\left(LN\left(Z_l^0\right)\right) \tag{6}$$

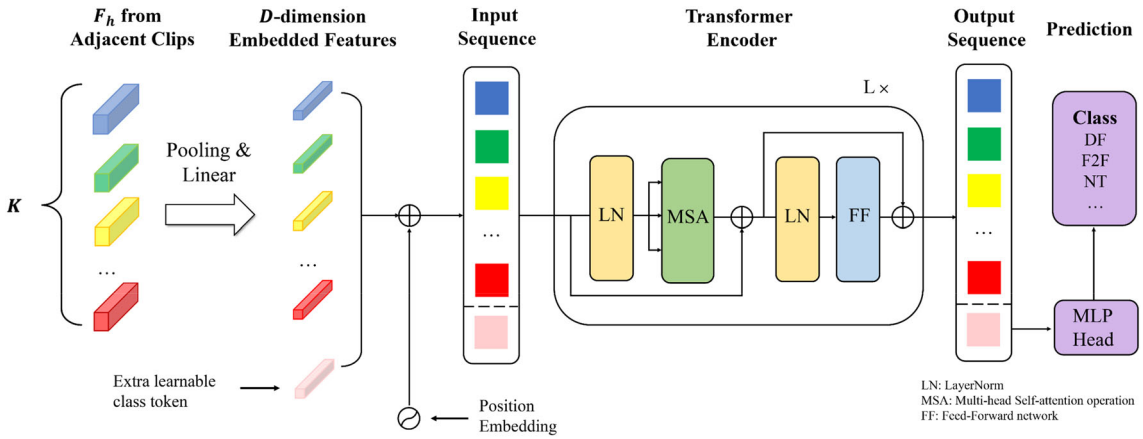

Fig. 4 Structure of the Temporal Transformer. ⊕ denotes sum operation

Table 1 Comparison with other methods on the sub-datasets of FF + +

Method	Binary Face Forgery Detection (real-fake)					Multi-category Source Detection (5 categories)					
	DF	F2F	FS	NT	FSH	Real	DF	F2F	FS	NT	Avg
Xception [23]	99.75	98.53	96.14	91.46	99.97	99.84	99.59	98.55	92.76	89.62	97.11
Ciftci el al. [6]	–	–	–	–	–	97.29	94.66	91.66	92.33	81.93	93.39
FakeCatcher [7]	94.87	96.37	95.75	89.12	–	–	–	–	–	–	–
Boccignone et al. [40]	90.68	94.46	95.39	87.57	98.88	–	–	–	–	–	–
DeepRhythm [41]	100.00	99.50	100.00	–	–	–	–	–	–	–	–
Liang et al. [42]	**100.00**	99.50	**100.00**	97.10	**100.00**	97.59	**99.66**	97.59	98.62	96.55	98.33
Our method	**100.00**	**100.00**	**100.00**	**98.00**	99.28	**100.00**	99.55	**99.33**	**99.67**	**98.33**	**99.38**

Bold values indicate the best results
The left part denotes the results of binary classification task, and the right half is the source detection experiment of 5 categories and the averaged outcome. The metric is accuracy (%), and the best results are highlighted

3.3 Loss function

In the first stage training of the backbone and MLA without the inclusion of ViT, we formulated joint loss function L_{total} including the softmax cross-entropy loss L_{ce} and the attention mask loss L_{mask} as follows:

$$L_{\text{total}} = L_{\text{ce}} + \lambda L_{\text{mask}} \tag{7}$$

where λ is the hyperparameter for balancing classification task and mask regression task. In the second stage, we freeze parameters of the backbone layers and MLA, only use the cross-entropy loss to train the Temporal Transformer.

4 Experiments

In this section, elaborate evaluations are provided to test the effectiveness of the proposed method. First, our method is compared with six benchmark methods on binary face forgery detection and multi-category source detection tasks. Then, detailed ablation studies are performed to show the impact of each component. Moreover, extension experiments are conducted to show the expandability of the proposed method against new manipulation sources. Finally, experiments of video clip length and video compression are also performed for supplement.

4.1 Settings

4.1.1 Dataset

To illustrate the effectiveness of the proposed method on face forgery detection and source detection tasks, we select the most widely used FaceForensics + + (FF + +) dataset [45]. FF + + is a relatively large dataset containing 1000 real videos and 4000 fake videos generated by five different face manipulation methods, i.e., Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and

FaceShifter (FSH). In terms of the modified regions, DF, FS, and FSH swap the whole face, F2F focuses on smaller areas, transfers expressions while keeping the identity of the target face consistent, and NT operates only around the mouth area of the target face. There are three video quality versions in FF++, correspond to different compression rate, i.e., RAW (c0), HQ (c23), and LQ (c40). In addition, to demonstrate the extension capability of the proposed method, experiments on Celeb-DF(v2) dataset [46] are also conducted. Celeb-DF is a more challenging dataset which contains 590 real celebrity videos. 59 subjects swapping faces in pairs to generate 5,639 high-quality fake videos. The forged faces in Celeb-DF are more detailed and convincing because of the usage of a more advanced synthetic process.

4.1.2 Implementation

For the real-fake binary classification task, the training set, test set and validation set of each sub-dataset are divided in the ratio of 8:1:1. As for the source detection task, the dataset is split in the ratio of 7:3, consistent with previous works [6, 7, 42]. We adopt an open-source face detector OpenFace [47] to detect 68 facial landmarks. Following the setting of [43], the number of ROI sub-regions n is 6, and both RGB and YUV color space are used to generate PPG maps; therefore, C is 6. Unless otherwise noted, the video clip length T is 64, the step size ω is 16, and the loss balancing hyperparameter $\lambda = 10$. EfficientNetV2-M [48] pre-trained on ImageNet is adopted as the backbone. The MLA is inserted after the third stage of the backbone. Adjacent PPG clip number K, self-attention heads number, and the embedded features dimension D are set to 5, 8, and 256, respectively. The batch size is set to 32. SGD is used as our optimizer with the initial learning rate of 0.01. The total epoch number is 30. All models are implemented based on PyTorch framework and trained on GTX-1080Ti.

4.1.3 Prediction aggregation

Since a full video contains several video clips, we predict each clip of the video and count the number of real or fake clips. If the number of real clips is greater than fake ones, we identified the video as real and vice versa. As for source detection, majority vetoing is adopted to determine the predict source of each video. All results are based on video classification accuracy.

4.2 Comparison

In order to make a fair and comprehensive comparison, we consider both face-based methods and rPPG-based methods for the selection of baseline. Among face-based approaches, we choose the popular Xception [23]. And all methods utilizing rPPG are included for comparison, i.e., Ciftci et al. [6], FakeCatcher [7], Boccignone et al. [40], DeepRhythm [41], and Liang et al. [42]. Meanwhile, to demonstrate the ability of the proposed method to detect different manipulation sources, other than conventional real-fake binary classification, we also examined the source detection performance with five categories (1 real—4 fakes). All the comparison experiments are conducted on the FF++ dataset. As shown in Table 1, the proposed method achieves the best results in DF, F2F, FS, and NT sub-datasets of binary face forgery task and achieves the state-of-art performance among all rPPG-based methods on the source detection, which is sufficient to demonstrate the effectiveness of the proposed method. Compared with the baseline method using cropped faces as input, our method has more obvious advantages on categorization tasks. In terms of the source detection results on FS and NT categories, our method achieves 99.67 and 98.33% while Xception [23] with face inputs only reaches 92.76 and 89.62%, respectively. This result once again proves the strong ability of PPG maps to preserve the unique rhythmic patterns of different manipulation methods. Moreover, compared with other methods using rPPG, the proposed method has better performance on four categories, indicating the superior capacity of our method for exploiting information of the rPPG signal.

Table 2 Ablation experiments of our method by progressively adding the Multi-scale PPG Spatial–Temporal map (Multi-scale), the Mask-Guided Local Attention module (MLA), and the Temporal Transformer

Method	Avg Acc
POS [49]	85.77
MMSTR [41]	85.65
Multi-scale	98.29
Multi-scale + MLA	99.00
Multi-scale + MLA + LSTM	99.28
Multi-scale + MLA + Self-attention	99.28
Multi-scale + MLA + Transformer	**99.38**

Bold values indicate the best results
POS and MMSTR are two previous representations of the rPPG signal. The metric is average categorization accuracy (%)

4.3 Ablation experiments

To demonstrate the effectiveness of each component of our method, i.e., Multi-scale Spatial–Temporal PPG map, Mask-Guided Local Attention module (MLA), and Temporal Transformer, we conducted detailed ablation experiments of source detection. And the results are shown in Table 2.

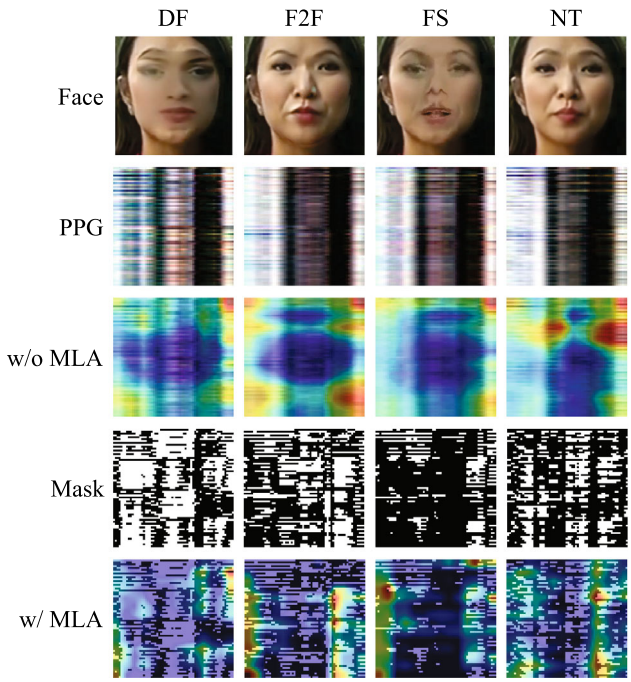

Fig. 5 Visualization of Mask-Guided Local Attention (MLA) on various face manipulation methods. The second and the fourth row show the PPG maps and their corresponding masks of modified regions. The third row is the heat maps of high-level features without MLA in the network, and the last row shows the heat maps when MLA is inserted into the network

4.3.1 Effectiveness of multi-scale PPG map

For a better comparison of the Multi-scale PPG map, two other forms of PPG map from previous works are implemented, i.e., POS [49] and Motion-Magnified Spatial–Temporal Representation (MMSTR) [41]. The first is based on assumptions of skin optic model, and the second utilizes motion magnification algorithm [50]. The amount of data and backbone training settings is the same for three forms of PPG maps. As shown in Table 2, the averaged categorization accuracy by adopting the Multi-scale PPG map is 13% higher than the other two, which proves that the multi-scale combination of facial regions contains more sufficient rhythmic information. Without prior calculation, the Multi-scale Spatial–Temporal representation of PPG can better cope with the deep learning framework in a fully data-driven manner.

4.3.2 Effectiveness of MLA

We then conducted experiments to demonstrate the effectiveness of the MLA which utilizes local attention mechanism. According to the mask examples shown in Fig. 5, the distribution of white pixels which represents the modified regions varies with different face manipulation methods. Mask calculated form DF shows a wider modified area than NT, which is consistent with our hypothesis that PPG maps can reflect spatial and regional differences between each face forgery method.

To present an intuitive interpretation of how MLA works, we also visualize heatmaps of high-level features utilizing Grad-CAM [51]. As shown in Fig. 5, without the mask and local attention operation, the network would focus on the large area near the upper right corner of the PPG map while ignoring the regional differences between specific face forgery methods. After adding MLA into the backbone, the strong-response area successfully converges at specific locations guided by the mask; thus, the rhythmic patterns of each face manipulation method are further distinguished. The average categorization accuracy also improves 0.71% by utilizing MLA. And it is worth noting that the accuracy of NT category, which is usually difficult to classify, improved by 1.33%. This result proves the effectiveness of MLA for detecting local discrepancies.

4.3.3 Effectiveness of temporal transformer

In order to explore the proper temporal model for interacting adjacent feature vectors, we select one layer of the widely used Bidirectional Long Short Term Memory (Bi-LSTM) networks and single head self-attention [31] (denoted as Self-attention) to compare with the standard Transformer encoder [31]. The result in Table 2 shows that the average detection accuracy of utilizing Bi-LSTM is 0.28% higher than not, proving our second assumption that adjacent PPG clips contain temporal-correlated information. Intriguingly, one layer of single head self-attention can be on par with the conventional RNN structure in performance. This result indicates that the recurrent structure is constrained by the contextual order of the sequence, while self-attention mechanism completely relying on the long-distance dependencies of input tokens. With the full structure of MSA and FF block, the Temporal Transformer fully exploits the global information between adjacent features and improves the accuracy by 0.38%. To better demonstrate the effectiveness of the Temporal Transformer over LSTM, we conduct real-fake binary face forgery detection experiments on five sub-datasets of the c23 (HQ) version of FF + + , respectively. As shown in Fig. 6, by adopting Transformer, the accuracy of all five sub-datasets increased by an average of 0.84% in comparison of utilizing LSTM. It should be noted that the superiority of ViT over LSTM is more evident on F2F, FS, and NT sub-datasets, with accuracy improvements of 1.15%, 1.7%, and 0.65%, respectively. These results provide stronger evidence of the superiority of the global attention and long-distance dependencies mechanism of ViT over the local contextual-constrained LSTM.

In addition, the impact of using different numbers of Transformer encoders blocks is also investigated. As shown

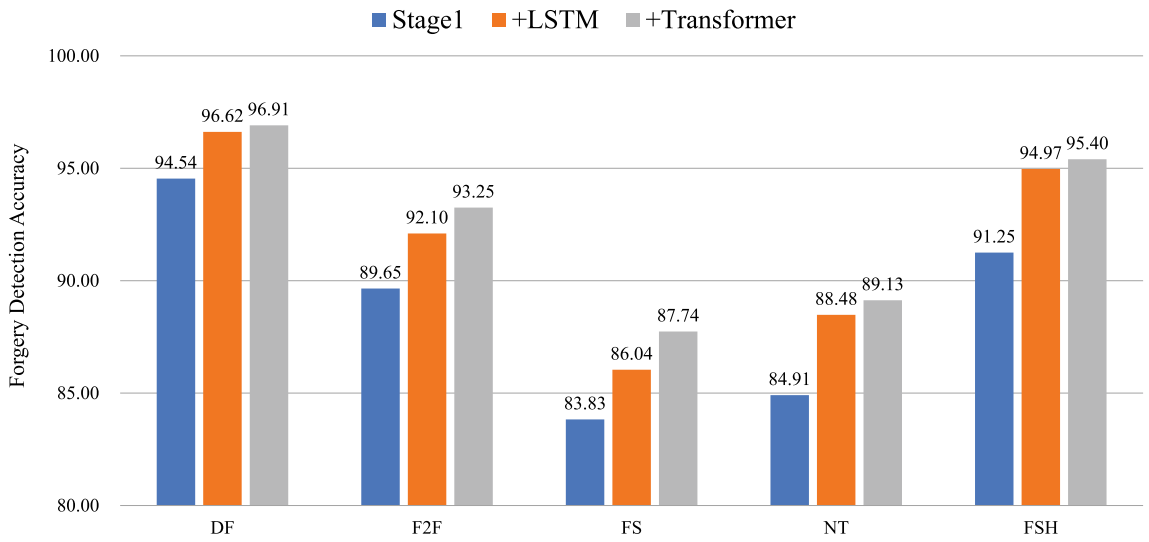

Fig. 6 Binary Face forgery detection (real-fake) accuracy (%) on five sub-datasets of FF + + (c23), including DF, F2F, FS, NT, and FSH, of using different temporal modules

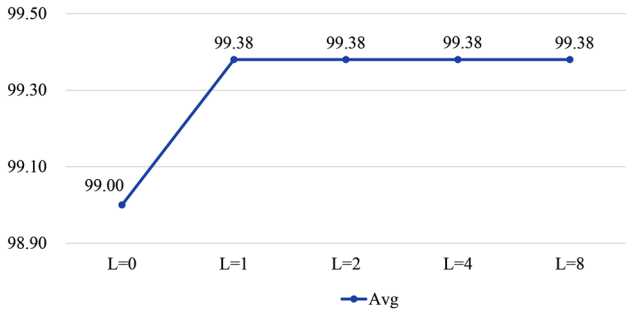

Fig. 7 Average categorization accuracy (%) of using different layers of Transformer encoders

in Fig. 7, without the pre-trained parameters to initialize the second-stage model, adding more layers of Transformer encoder causes additional training parameters, but does not improve the outcome. In the other hand, compared with other scenarios of adopting ViT, the number of embedded features K is much more limited in our work; thus, single layer of the Transformer structure is sufficient.

4.4 Extension experiments

Deepfake generation methods upgrade rapidly, which demands that our method have extension ability against new face forgery methods. Thus, we conducted extension experiments by adding a new category of Celeb-DF (CD) to illustrate generalization performance of the proposed method. 1000 fake videos are selected from Celeb-DF(v2) [46], and PPG maps of each video clip is computed along with their corresponding masks to generate the sixth class for source detection. As shown in Table 3, our method is capable of tracking new sources and the average categorization accuracy of six classes exceeds other two previous methods [6, 42], which confirms the excellent extension ability of the proposed method.

4.5 Video clip length

We also conducted experiments to explore the balance between the length of a single clip and the amount of training data. We assume that the number of PPG maps that can be obtained from a complete video will be large when the clip length is relatively short, but the information contained in individual clips may also be limited. On the other hand, an excessively long clip would greatly reduce the amount of data available for training. Thus, we test the proposed method with different clip length $T = \{32, 64, 128, 256\}$. For a fair comparison, the step size ω is set as 1/4 T to obtain all PPG maps from a full video. As results shown in Table 4, the accuracy reaches the highest score at $T = 64$, but drops sharply at $T = 256$. Not surprisingly, the length of video clips too long results in the limited data size and negative impact on our data-driven approach.

4.6 Video compression

To test the performance of the proposed method against video compression, experiments are conducted on different video quality of FF + + , i.e., HQ (c23) and LQ (c40). As shown in Table 5, the proposed method still reaches the average source detection accuracy of 90.52% on video quality HQ, which demonstrates the robustness against light compression. However, due to the loss of subtle facial color changes

Table 3 Results of extension experiment

Method	Real	DF	F2F	FS	NT	CD	Avg
Cifci et al. [6]	96.89	94.66	91.66	92.66	92.66	92.17	93.69
Liang et al. [42]	95.45	100.00	98.86	96.59	97.73	98.89	97.57
Our method	97.07	99.58	98.04	99.18	96.96	**100.00**	**98.65**

Bold values indicate the best results
1000 videos from Celeb-DF(v2) are added as the sixth category. The metric is forgery categorization accuracy (%)

Table 4 Accuracy of categorization in different video clip length

Clip Length	Real	DF	F2F	FS	NT	Avg
32	99.06	100.00	97.90	99.00	96.22	98.43
64	**100.00**	99.55	**99.33**	**99.67**	**98.33**	**99.38**
128	96.83	100.00	99.33	99.64	96.51	98.64
256	97.47	100.00	100.00	94.44	94.12	97.21

Bold values indicate the best results

Table 5 Accuracy of categorization in different compression rate

Compression	Real	DF	F2F	FS	NT	Avg
HQ (c23)	87.67	98.08	90.48	88.05	88.32	90.52
LQ (c40)	49.55	80.10	50.50	54.18	50.50	56.96

results from the severe compression, the rPPG signal is disrupted under LQ version, which leads to the accuracy of merely 56.96%.

5 Conclusion

In this paper, the Multi-scale Spatial–Temporal PPG map is adopted to further exploit heartbeat signal from multiple facial regions. Motivated by the key observation that rPPG signals produce unique rhythmic patterns in terms of different manipulation methods, a two-stage network is proposed for both face forgery detection and categorization. Concretely, the Mask-Guided Local Attention module (MLA) is designed to locate spatial inconsistencies of modified facial regions reflected on PPG maps. The Temporal Transformer is also adopted to exploit long-distance information between adjacent video clips. Abundant experiments on FaceForensics + + and Celeb-DF(v2) datasets demonstrate the superiority of the proposed method which outperforms all other rPPG-based approaches. Moreover, extension experiment confirms the excellent generalization capability of the method against newly added manipulation model. Furthermore, detailed ablation study and visualization illustrate the effectiveness of each component and different settings.

Acknowledgements The authors greatly appreciate the financial supports of Natural Science Foundation of Shanghai under Grant 22ZR1444700, National Natural Science Foundation of China under Grant 82170110, Science and Technology Commission of Shanghai Municipality under Grant 20DZ2261200.

Data availability statements The data that support the findings of this study are available from the corresponding author, YuZhu, upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.J.A.i.n.i.p.s.: Generative adversarial nets. 27, (2014)
2. Radford, A., Metz, L., Chintala, S.J.a.p.a.: Unsupervised representation learning with deep convolutional generative adversarial networks. (2015)
3. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232. (2017)
4. Hertzman, A.B.: Photoelectric plethysmography of the fingers and toes in man. Proc. Soc. Exp. Biol. Med. **37**(3), 529–534 (1937)
5. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Opt. Express **16**(26), 21434–21445 (2008)
6. Ciftci, U.A., Demir, I., Yin, L.: How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with

biological signals. In: 2020 IEEE international joint conference on biometrics (IJCB), pp. 1–10. IEEE, (2020)
7. Ciftci, U.A., Demir, I., Yin, L., Intelligence, M.: Fakecatcher: Detection of synthetic portrait videos using biological signals. (2020)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 Words: Transformers for Image Recognition at Scale. pp. arXiv:2010.11929 (2020)
9. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational autoencoder for deep learning of images, labels and captions. 29, (2016)
10. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM SIGGRAPH 2003 Papers, pp. 313–318 (2003)
11. Pitié, F., Kokaram, A.C., Dahyot, R.J.C.V., Understanding, I.: Automated colour grading using colour distribution transfer. **107**, 123–137 (2007)
12. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Comput. Graph. Appl. **21**(5), 34–41 (2001)
13. Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C.S., RP, L., Jiang, J.: DeepFaceLab: Integrated, flexible and extensible face-swapping framework. (2020)
14. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.J.a.p.a.: Faceshifter: Towards high fidelity and occlusion aware face swapping. (2019)
15. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395. (2016)
16. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. Acm Trans. Graph. (TOG) **38**(4), 1–12 (2019)
17. Buchana, P., Cazan, I., Diaz-Granados, M., Juefei-Xu, F., Savvides, M.: Simultaneous forgery identification and localization in paintings using advanced correlation filters. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 146–150. IEEE, (2016)
18. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Trans. Inf. Foren. Secur **7**(3), 868–882 (2012)
19. Goljan, M., Fridrich, J.: CFA-aware features for steganalysis of color images. In: Media Watermarking, Security, and Forensics 2015, pp. 279–291. SPIE, (2015)
20. Pan, X., Zhang, X., Lyu, S.: Exposing image splicing with inconsistent local noise variances. In: 2012 IEEE International Conference on Computational Photography (ICCP), pp. 1–10. IEEE, (2012)
21. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE, (2018)
22. Nguyen, H., Yamagishi, J., Echizen, I.: Use of a capsule network to detect fake images and videos. arXiv 2019.
23. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258. (2017)
24. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001–5010. (2020)
25. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, pp. 5781–5790. (2020)
26. Shang, Z., Xie, H., Zha, Z., Yu, L., Li, Y., Zhang, Y.: PRRNet: Pixel-region relation network for face forgery detection. Patt. Recognit **116**, 107950 (2021)
27. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1081–1088. (2021)
28. de Lima, O., Franklin, S., Basu, S., Karwoski, B., George, A.J.a.e.-p.: Deepfake Detection using Spatiotemporal Convolutional Networks. pp. arXiv:2006.14749 (2020)
29. Montserrat, D.M., Hao, H., Yarlagadda, S.K., Baireddy, S., Shao, R., Horváth, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F.: Deepfakes detection with automatic face weighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 668–669. (2020)
30. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. (2014)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. 30, (2017)
32. Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15044–15054. (2021)
33. Xu, Y., Jia, G., Huang, H., Duan, J., He, R.: Visual-Semantic Transformer for Face Forgery Detection. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–7. IEEE, (2021)
34. Khan, S.A., Dai, H.: Video Transformer for Deepfake Detection with Incremental Learning. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1821–1828. (2021)
35. Malolan, B., Parekh, A., Kazi, F.: Explainable deep-fake detection using visual interpretability methods. In: 2020 3rd International Conference on Information and Computer Technologies (ICICT), pp. 289–293. IEEE, (2020)
36. Jayakumar, K., Skandhakumar, N.: A Visually Interpretable Forensic Deepfake Detection Tool Using Anchors. In: 2022 7th International Conference on Information Technology Research (ICITR), pp. 1–6. IEEE, (2022)
37. LIY, C.M., InIctuOculi, L.: ExposingAICreated FakeVideosbyDetectingEyeBlinking. In: 2018IEEEInterG national Workshop on Information Forensics and Security (WIFS). IEEE. (2018)
38. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265. IEEE, (2019)
39. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5039–5049. (2021)
40. Boccignone, G., Bursic, S., Cuculo, V., D'Amelio, A., Grossi, G., Lanzarotti, R., Patania, S.: DeepFakes Have No Heart: A Simple rPPG-Based Method to Reveal Fake Videos. In: International Conference on Image Analysis and Processing, pp. 186–195. Springer, (2022)
41. Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., Zhao, J.: Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4318–4327. (2020)
42. Liang, J., Deng, W.: Identifying Rhythmic Patterns for Face Forgery Detection and Categorization. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE, (2021)
43. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: European Conference on Computer Vision, pp. 295–310. Springer, (2020)
44. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. (2016)

45. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11. (2019)
46. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216. (2020)
47. Baltrušaitis, T., Robinson, P., Morency, L.-P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE, (2016)
48. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106. PMLR, (2021)
49. Wang, W., Den Brinker, A.C., Stuijk, S., De Haan, G.: Algorithmic principles of remote PPG. IEEE Trans. Biomed. Eng. **64**(7), 1479–1491 (2016)
50. Wu, H.Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., Freeman, W.: Eulerian video magnification for revealing subtle changes in the world. ACM Trans. Graph. (TOG) **31**(4), 1–8 (2012)
51. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626. (2017)

Jiahui Wu was born in Zhejiang Province, China and obtained the B.S. degree in School of Information Science and Engineering from East China University of Science and Technology in 2021. He is currently pursuing the M.S. degree in East China University of Science and Technology. His main research interests include Deepfake detection and computer vision.

Yu Zhu *Member IEEE* received the Ph.D degree from Nanjing University of Science and Technology, China, in 1999. She is currently a professor in the department of electronics and communication engineering of East China University of Science and Technology. Her research interests include image processing, computer vision, multimedia communication, and deep learning, especially, for the medical auxiliary diagnosis by artificial intelligence technology. She has published more than 90 papers in journals and conferences.

Xiaoben Jiang is pursuing the Ph.D. degree in East China University of Science and Technology. His current research interests include digital image processing and computer vision. His experience includes the denoising method on chest X-ray images and CT images and detection of COVID-19 cases from denoised CXR images. He has published in journals in the crossing field of medical science and computer vision and has been involved in publicly and privately funded projects.

Yatong Liu received her B.S. degree from East China University of Science and Technology in 2021. She is currently a postgraduate at the school of information science and engineering, East China University of Science and Technology. Her research interests include medical image processing, deep learning, the classification and segmentation of prostate on MRI, and pattern recognition.

Jiajun Lin obtained his Ph.D. degree from Tsinghua University, Beijing. He is a professor at School of Information Science and Engineering, East China University of Science and Technology. His research interests include intelligent information processing and security of industry control systems.