Model-agnostic Method: Exposing Deepfake using Pixel-wise Spatial and Temporal Fingerprints

Jun Yang, Yaoru Sun, Maoyu Mao, Lizhi Bai, Siyu Zhang and Fang Wang

***Abstract*—Deepfake poses a serious threat to the reliability of judicial evidence and intellectual property protection. Existing detection methods either blindly utilize deep learning or use biosignal features, but neither considers spatial and temporal relevance of face features. These methods are increasingly unable to resist the growing realism of fake videos and lack generalization. In this paper, we identify a reliable fingerprint through the consistency of AR coefficients and extend the original PPG signal to 3-dimensional fingerprints to effectively detect fake content. Using these reliable fingerprints, we propose a novel model-agnostic method to expose Deepfake by analyzing temporal and spatial faint synthetic signals hidden in portrait videos. Specifically, our method extracts two types of faint information, i.e., PPG features and AR features, which are used as the basis for forensics in temporal and spatial domains, respectively. PPG allows remote estimation of the heart rate in face videos, and irregular heart rate fluctuations expose traces of tampering. AR coefficients reflect pixel-wise correlation and spatial traces of smoothing caused by up-sampling in the process of generating fake faces. Furthermore, we employ two ACBlock-based DenseNets as classifiers. Our method provides state-of-the-art performance on multiple deep forgery datasets and demonstrates better generalization.**

***Index Terms*—Deepfake detection, photoplethysmography (PPG), auto-regressive (AR), temporal and spatial, fingerprint, deep learning.**

I. Introduction

FAKE face videos generated by deep learning technology, also known as Deepfake videos [51], are characterized by rapid technological innovation and an unprecedented level of authenticity. These videos, which often contain dangerous yet compelling content, such as depicted in Fig. 1, pose a serious threat to national political security and social stability [28]. Therefore, it is imperative to develop a universal Deepfake detection algorithm. Deepfake detection methods can be categorized into two types: fake image detection and fake video detection [9], [23]. While video compression may result in degraded frame data and changes in timing characteristics between frames, image detection algorithms cannot be directly applied to video detection [1]. Considering the temporal structure of fake videos, strategies for fake video detection can be further classified into those that rely on visual artifacts within individual frames [1], [30], [33], [39], [57] and those that analyze temporal features across frames [2], [21], [31], [47]. To achieve detection, methods based on artifacts within frames usually combine deep or shallow classifiers with image frame features. The underlying principle of methods that analyze temporal features across frames is that Deepfake videos mostly rely on frame-by-frame manipulation, which means that the temporal features between video frames are not preserved seamlessly [49]. Intuitively, these methods can adopt deep recurrent neural networks (RNNs) [6], [14], [21], [47], [60] or attention mechanism [10], [63] for end-to-end detection. However, few previous studies have evaluated the performance of these networks or explored effective features for determining the authenticity of a face video. Furthermore, pixel-level methods are inadequate in coping with increasingly realistic Deepfake videos. Interestingly, methods utilizing temporal changes in biological traits have garnered substantial interest from the research community. For example, the presence of eye blinking as a physiological signal, as presented in [31], can expose fake face videos using deep neural networks. Another approach, as demonstrated in [2], assesses the fidelity of lip movement in face videos by combining sound synchronization analysis to determine the authenticity of the video.

Generative models of Deepfake can learn visual features at the pixel level to achieve deception, but they struggle to replicate faint signals accurately. One example of such a signal is the heart rate (HR), which is present in every living organism. In 2008, Verkruysse et al. [53] first propose remote photoplethysmography (rPPG) technology. rPPG is an optical technique that monitors various vital signs by utilizing photoelectric sensors to detect and record variations in light absorption or reflection intensity on the human skin [17], [19]. As these variations are caused by changes in blood volume during the heartbeat cycle, it allows for the extraction of remote HR signals. Movements can cause changes in distance and angle between region of interest (ROI), camera and light source, creating noise interference. To overcome motion-induced interference, research on rPPG-based HR detection can be classified into three categories: blind source separation (BSS)-based methods [3], [20], [41], [50], [59], model-based methods [11], [22], [55], and other methods [50]. Chrominance-based PPG (CPPG) is a model-based method that achieves a mean square error half that of the BSS-based model [53]. PPG technology has garnered significant attention and is widely applied in the fields of medical imaging research [4], emotion recognition [38] and face anti-spoofing [35]. Recent experiments indicate that temporal consistency in the

This work was supported by the grants from the National Key R&D Program of China (2019YFC1906201), the National Natural Science Foundation of China (91748122), the Xianyang Key R&D Program (S2021ZDYF-SF-0739) and China Scholarship Council (CSC).

Jun Yang, Yaoru Sun, Maoyu Mao, Lizhi Bai and Siyu Zhang are with the Department of Computer Science and Technology, Tongji University, Shanghai 200092, China. (e-mail: junyang@tongji.edu.cn, yaoru@tongji.edu.cn, maomy@tongji.edu.cn, bailzh@foxmail.com, xuehua93@126.com).

Fang Wang is with the Department of Computer Science, Brunel University, Uxbridge UB8 3PH, United Kingdom.

Corresponding author: Maoyu Mao; Yaoru Sun.

heart rate signal of biological signals can be used to achieve Deepfake detection [8], marking the first time HR signals have been incorporated into this field. However, HR signal reflects the temporal characteristics of video, and this method lacks spatial dimension-based judgment.

Another important observation is that even when state-of-the-art models like generative adversarial networks (GANs) are used to generate Deepfake images, the upsampling process still leaves traces in the image pixels [5], [58]. However, there is still a lack of research focused on detecting these upsampling traces when generating fake faces using deep networks. An autoregressive (AR) model is a statistical model that predicts a variable at a specific time point based on its previous values. In the context of image processing, the autoregressive coefficient in an AR model represents the strength of dependence between the current value of a pixel and its past values [37]. The AR model captures the correlation between image pixels, assuming that the value of a given pixel is influenced by the values of its neighboring pixels. Kang et al. [26] propose using an AR model by converting the image into a one-dimensional signal, achieving robustness in image median filtering detection. Building upon this work, Yang et al. [56] extend the autoregressive (AR) model to two-dimensional space, enhancing its applicability in two-dimensional signal processing, such as images. This inspires us to consider that the correlation between pixels can be used to distinguish between fake and authentic data. When the correlation between pixels is low, it is highly likely that the image has been tampered with.

In total, previous studies are mainly focused on capturing biosignal features as the basis for distinguishing between real and fake faces. However, these features change over time, reflecting the temporal features and to a large extent lacking the spatial features, let alone capturing pixel-level details [2], [31], [53]. In fact, pixel-level information is difficult to manipulate, even through upsampling processes, making it a reliable indicator of the authenticity of an image. Therefore, considering the relationships among pixels in facial regions can improve the ability to distinguish between real and fake images based on local features. Moreover, combining pixel-level information from different regions can effectively capture spatial features within an image. By incorporating information from the temporal, spatial, and synthetic domains hidden in portrait videos, we can overcome the limitations of using solely temporal information.

In this paper, our proposed approach involves extracting biosignals in the temporal domain and pixel-level information in the spatial domain as fingerprints to distinguish between real and fake faces. After identifying two dimensions of new reliable feature for Deepfake detection, we investigate the improved PPG signal and modeling of AR coefficient as prior implicit fingerprints. By combining these two types of fingerprints with asymmetric convolution block (ACBlock)-based densely connected networks (DenseNets), we propose a model-agnostic deep forgery forensics method to solve facial manipulation behaviors. Concretely, we first divide the facial video into segments with a fixed number of frames and divide the ROI of each frame into identical sub-regions. To obtain PPG fingerprints, we sequentially extract the PPG signals from each sub-region and arrange them as a temporal-spatial feature map. For obtaining AR fingerprints, we extract the pixel values of the ROI and create a one-dimensional matrix for AR modeling. We then arrange the AR coefficients of each frame column by column to form a spatial-temporal feature map. Furthermore, we utilize two structurally identical improved DenseNets to autonomously detect the authenticity of facial videos based on the PPG and AR fingerprints. Simulation results demonstrate that our proposed fusion of the deep forgery forensics model improves detection accuracy, reduces detection delay, and enhances universality.

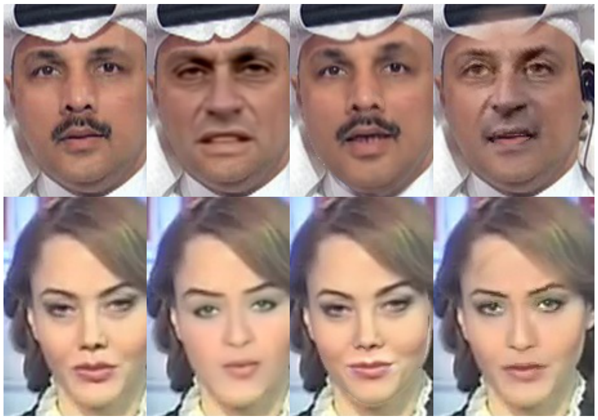


Fig. 1. Portrait video screenshots of real and Deepfake faces. The first column shows the real faces, the rest are fake faces.

Our main contributions of this paper can be summarized as follows:

- We combine two types of faint information hidden within facial pixels – novel biological signals in the form of PPG and AR coefficients – as spatial and temporal fingerprints for Deepfake forensics. AR fingerprinting is the first in this field to be used as a pixel-level spatial domain feature.
- We combine two types of faint information hidden between face pixels-novel biological signal, PPG, and AR coefficients as priori spatial and temporal fingerprints for Deepfake forensics. This is a novel method for Deepfake detection, which fully consider the temporal and spatial domain as well as pixel-wise features.
- We utilize an improved DenseNet with ACBlock to analyze two types of fingerprint information. Its asymmetric convolutional structure enhances the robustness of classifier to upside-down and left-right inversion, and avoids interference of feature stitching order.
- We develop a model-agnostic deep forgery forensic scheme based on faint facial information, utilizing subtle biometric features and using AR coefficient features for the first time in Deepfake detection. The proposed scheme improves detection method, saves computational resources, and enhances generality.

II. Related work

Despite deep learning models being capable of generating fake face, these generated synthetic images are still detectable.

TABLE I
SUMMARY OF MAIN SYMBOLS AND NOTATION

Notation	Definition
$\mathcal{G}$	Face video frame sequences
$\boldsymbol{G(X,Y)}^t$	t-th face video frame
$(\boldsymbol{X},\boldsymbol{Y})$	Image pixel coordinate
$\mathcal{V}\left(\boldsymbol{G}(x,y)_i^t\right)$	Intensity of reflected light from the skin surface with i-th pixel
$\mathcal{R}$	Reflectance of the skin surface
$\mathcal{I}\left(\boldsymbol{G}(x,y)_i^t\right)$	Illumination intensity of the light source
s	Specular reflection
ρ	Diffuse reflection
ξ	Color channel
$\mathcal{V}_\xi\left(\boldsymbol{G}(x,y)_i^t\right)$	Intensity of reflected light for each color channel ξ
$\mathcal{I}_\xi\left(\boldsymbol{G}(x,y)_i^t\right)$	Illumination intensity of the light source for each color channel ξ
$M \times N$	Size of ROI
$\mathcal{C}_\xi^t$	Chrominance signal of each color channel
$\overline{\mathcal{C}}_\xi^t$	Corrected Chrominance signal
$\overline{\chi}^t$	Orthogonal Chrominance signal 1
$\overline{\overline{\varpi}}^t$	Orthogonal Chrominance signal 2
Ξ^t	Ratio of two orthogonal signals
Υ^t	Chrominance-based PPG signal
$\boldsymbol{L}^{(o)}$	Output variable at time o
$\mathcal{P}$	Order of AR model
φ_p	Autocorrelation coefficient of p-th orders
ϵ_p	White noise
$\boldsymbol{\phi}$	Set of AR coefficients
$\boldsymbol{L}_{row}\left(\boldsymbol{G}(x,y)_i^t\right)$	Pixel value of frame t calculated by row
$\boldsymbol{L}_{col}\left(\boldsymbol{G}(x,y)_i^t\right)$	Pixel value of frame t calculated by column
$\mathcal{R}_o$	ROI of cheek region
$\mathcal{R}_n$	Regular rectangle ROI

A ubiquitous ProGAN adopted in [54] generates a large number of forged images and accomplishes detection, showing impressive generalization on a wide range of GAN-generated face images. From the perspective of spatial information, global texture information verified in [36] affords a generality method for synthetic face detection. Jia et al. [25] observe that features of real face are compactly distributed, while features of fake face are crowded within domains and scattered between domains. Therefore, they develop a single-side domain generalization framework for judging feature distribution and detecting Deepfake images. Durall et al. [15] claim that models that rely on convolution-based up-sampling, e.g., popular GAN architectures, cannot reconstruct the same spectral distribution of natural data, and that this effect enables full detection of Deepfakes images. Proximity points raised in [18], [44] arouse widepublic attention and our deep thinking. In terms of temporal domain information, facial physiological properties-based methods become a hot topic in the field of Deepfake detection. Local physiological characteristics of face, such as blink frequency of eyes, consistency of iris color, shape of tooth or hair and shadows on either side of nose are all discussed for Deepfake detection [39]. With regards to global biometrics, face X-ray represented in [30] indicates whether input image can be decomposed into the blending of two images from different sources. Motion-magnified spatial-temporal representation and dual-spatial-temporal attentional network utilized in [43] facilitate a better exhibition of heart-beat rhythms, and expose the fake face videos automatically.

III. System model

The system model of the proposed Deepfake detection scheme is shown in Fig. 2. The process begins by extracting frame sequences from a face video, which are then divided into multiple segments of fixed frame length (Frames that are shorter than the fixed length are discarded). The cheek region is selected as the ROI to generate fingerprints based on PPG signal and AR coefficient. Subsequently, the two types of fingerprints are fed into the ACBlock-based DenseNet for training. Finally, model fusion is conducted to achieve segment-level and video-level detection.

TABLE II
Performance comparison between traditional rPPG algorithms and the contact sensor signal for static subjects.

	GPPG	BSS-based		CPPG
		ICA	PCA	
Std. deviation	3.50	2.60	1.80	0.90
RMSE	1.70	1.10	0.90	0.40
Accuracy	0.8514	0.8770	0.8966	0.9595

A. Biological fingerprint of CPPG signal

In this paper, as presented in Tab. II, we compare the root-mean-square error (RMSE) and accuracy between the contact sensor signals and the *r*PPG signals, such as Green-channel (GPPG) [62], independent component correlation algorithm (ICA) [42], principal components analysis (PCA) [29], and CPPG [53] methods. Based on the comparison results, we adopt the CPPG method to capture the biological fingerprints.

Given frame sequence $\mathcal{G} = \left\{\boldsymbol{G(X,Y)}^t\right\}_{t=1}^{T}$ of a video recorded by the light sensitive sensor in a camera, for the t-th frame $\boldsymbol{G(X,Y)}^t$ with pixel coordinate $(\boldsymbol{X},\boldsymbol{Y}) = [(x,y)_i]_{i=1}^{M\times N} \in \mathbb{R}^{M\times N}$, the intensity of reflected light from the skin surface can be expressed as

$$\mathcal{V}\left(\boldsymbol{G}(x,y)_i^t\right) = \mathcal{I}\left(\boldsymbol{G}(x,y)_i^t\right)\mathcal{R}\,, \tag{1}$$

where $\mathcal{R}$ is the reflectance of the skin surface. $\mathcal{I}\left(\boldsymbol{G}(x,y)_i^t\right)$ is the illumination intensity of the light source at i-th pixel coordinate $(x,y)_i \in \mathbb{R}^{M\times N}$. Light produces specular and diffuse reflections on the facial skin. Therefore, the reflectance $\mathcal{R}$ can be further decomposed as

$$\mathcal{R} = s + \rho, \tag{2}$$

where s and ρ denote specular reflection and diffuse reflection, respectively. Analyzing the interaction of the structural level of the skin with light, the diffuse reflection ρ is divided into a direct current part ρ_{DC} and an alternating current part ρ_{AC}, i.e.,

$$\rho = \rho_{DC} + \rho_{AC}\,. \tag{3}$$

For the selected ROI, the intensity of reflected light $\mathcal{V}\left(\boldsymbol{G}(x,y)_i^t\right)$ for each color channel $\xi \in \{R, G, B\}$ can be modeled as

$$\mathcal{V}_\xi\left(\boldsymbol{G}(x,y)_i^t\right) = \mathcal{I}_\xi\left(\boldsymbol{G}(x,y)_i^t\right)\left(s + \rho_{DC_\xi} + \rho_{AC_\xi}\right)\,, \tag{4}$$

where s is identical for all the color channels. ρ_{DC_ξ} is the stationary reflection coefficient of the skin, while ρ_{AC_ξ} is the time-varying physiological waveform attributed to cardiac synchronous changes in blood volume.

To reduce motion artifacts and other noise effects parallel to the imaging plane, a group of pixels $(\boldsymbol{X},\boldsymbol{Y})$ in ROI of size $M\times$

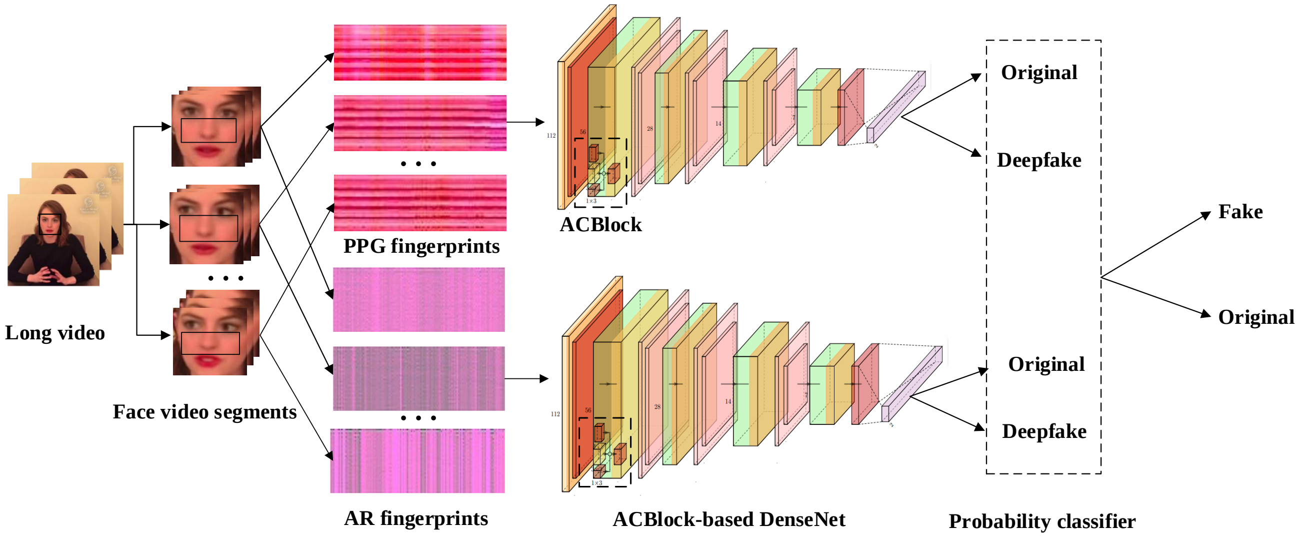


Fig. 2. System model of proposed Deepfake detection scheme.

N is selected for averaging pooling. Then the chrominance signal of each color channel is

$$\mathcal{C}_\xi^t = \frac{\sum_{y=0}^{N\text{-}1}\sum_{x=0}^{M\text{-}1}\mathcal{V}_\xi\left(\boldsymbol{G}(x,y)_i^t\right)}{MN}, \quad \xi\in\{R,G,B\}. \tag{5}$$

In a white-light illumination environment, due to differences in the mapping of skin tones of different individuals, we correct the chrominance signal $\mathcal{C}_\xi^t$ with the help of normalization of the differences as

$$\overline{\mathcal{C}}_\xi^t = \sigma^{\text{-}1}\mathcal{C}_\xi^t, \quad \xi\in\{R,G,B\}, \tag{6}$$

where $\sigma = \sqrt{{\mathcal{C}_R^t}^2 + {\mathcal{C}_G^t}^2 + {\mathcal{C}_B^t}^2}$. Assuming a fixed skin-tone the standard skin color vector, $[\mathcal{C}_R, \mathcal{C}_G, \mathcal{C}_B]$ is noted as $[0.7682, 0.5121, 0.3841]$ [29] under white light illumination. In the selected color space, the distribution of angles between different skin colors and the standard skin color vector remains relatively consistent across the entire range of skin types. Therefore, the corrected chrominance signal can be expressed as

$$\left[\overline{\mathcal{C}}_R^t, \overline{\mathcal{C}}_G^t, \overline{\mathcal{C}}_B^t\right] = \left[0.7682\mathcal{C}_R^t, 0.5121\mathcal{C}_G^t, 0.3841\mathcal{C}_B^t\right]. \tag{7}$$

Two orthogonal chrominance signals are utilized to eliminate specular reflection component in Eq. (4), i.e.,

$$\begin{aligned}\overline{\chi}^t &= \frac{\overline{C}_R^t - \overline{C}_G^t}{0.7682 - 0.5121} \approx 3\overline{C}_R^t - 2\overline{C}_G^t\\ \overline{\varpi}^t &= \frac{\overline{C}_R^t + \overline{C}_G^t - 2\overline{C}_B^t}{0.7682 + 0.5121 - 2\times 0.3841}\\ &\approx 1.5\overline{C}_R^t + \overline{C}_G^t - 1.5\overline{C}_B^t\end{aligned} \tag{8}$$

When the skin surface moves with respect to the light source, the illumination intensity in Eq. (4) will change and affect the chrominance intensity. However, it is important to note that these intensity modulations affect all channels equally. As a result, the impact of this motion can be mitigated by calculating the ratio of two filtered orthogonal signals. This ratio can be represented as

$$\Xi^t = \frac{\overline{\chi}^t}{\overline{\varpi}^t} - 1\,. \tag{9}$$

To deform the signal from a ratio to a linear combination, Eq. (13) is rewritten as

$$\log\left(1+\Xi^t\right) = \log\left(\frac{\overline{\chi}^t}{\overline{\varpi}^t}\right) = \log\left(\overline{\chi}^t\right) - \log\left(\overline{\varpi}^t\right). \tag{10}$$

Since all the arguments of log in Eq. (10) are close to 1, according to Taylor expansion, the approximate ratio $\hat{S}(t)$ is

$$\begin{aligned}\widehat{\Xi}^t &\approx \overline{\chi}^t - \overline{\varpi}^t\\ &= 1.5\overline{C}_R^t - 3\overline{C}_G^t + 1.5\overline{C}_B^t\,.\end{aligned} \tag{11}$$

Standard skin colour is calculated to eliminate the effect of different coloured lights on the skin, which can lead to an effect on the magnitude of changes of $\overline{\chi}^t$ and $\overline{\varpi}^t$. So the Eq.(11) can be rewrited as

$$\widehat{\Xi}^t \approx \overline{\chi}^t - \alpha\overline{\varpi}^t, \tag{12}$$

with

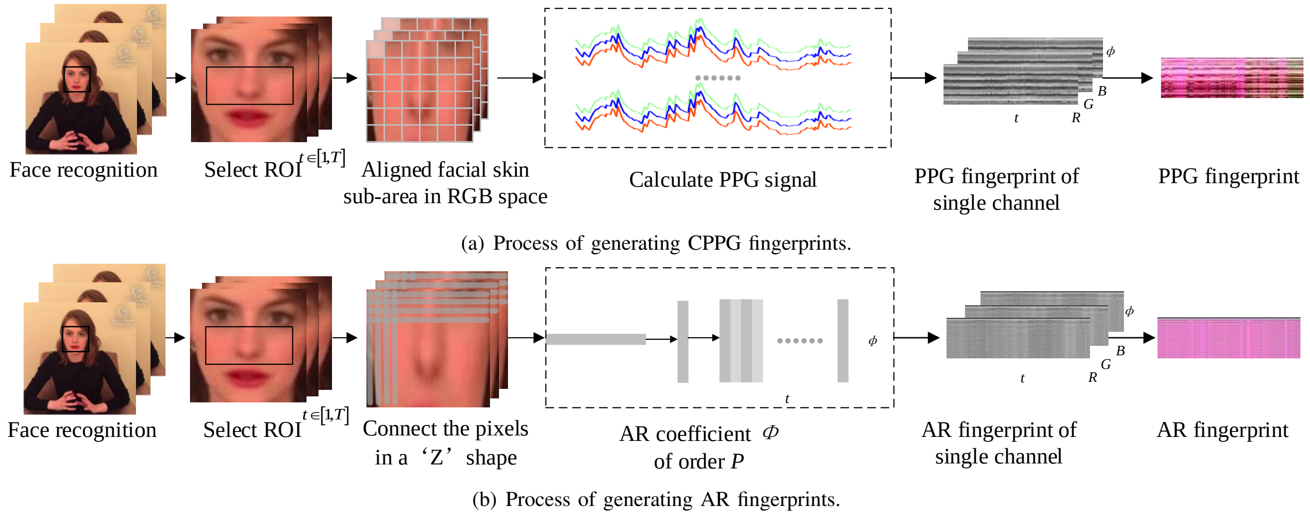
$$\alpha = \frac{\partial(\overline{\chi}^t)}{\partial(\overline{\varpi}^t)}. \tag{13}$$

$\partial(\overline{\chi}^t)$ is the standard deviation of $\overline{\chi}^t$, $\partial(\overline{\varpi}^t)$ is the standard deviation of $\overline{\varpi}^t$. This algorithm allows the effect of standardised differences in skin colour by race to be negligible and ensures the fairness of the underlying assumption.

Finally, difference between two frames of signals is used to reduce the influence of pigment absorption on diffuse reflection, and a chrominance-based PPG signal is obtained as

$$\Upsilon^t = \widehat{\Xi}^t - \widehat{\Xi}^{t-1}\,. \tag{14}$$

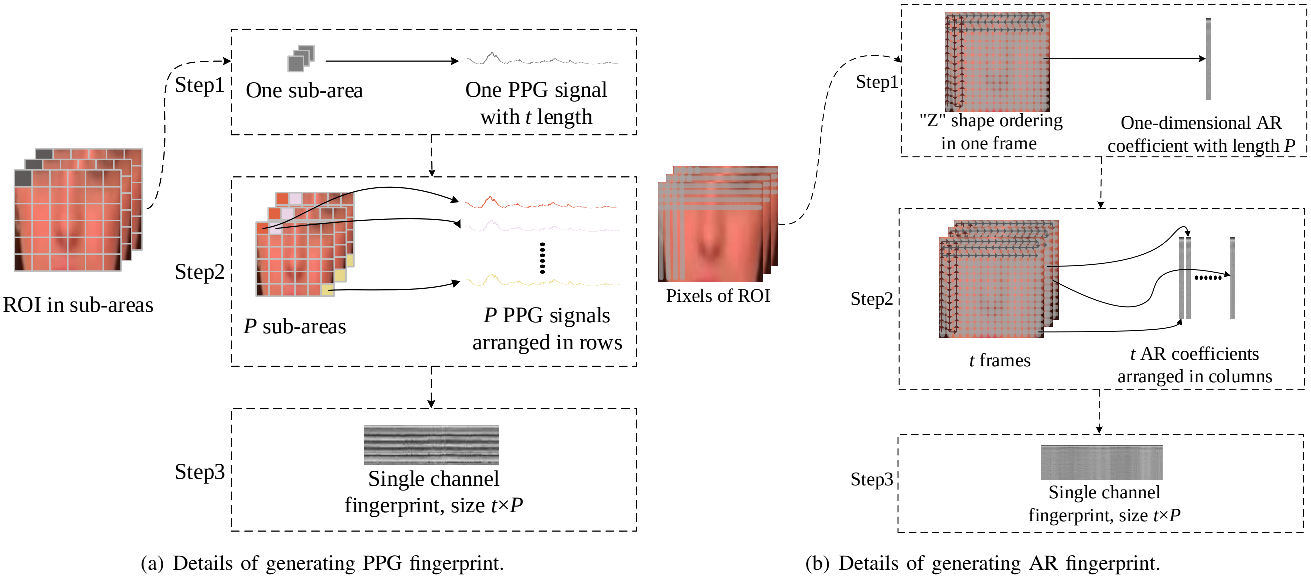
Υ^t is the CPPG signal containing blood volume variation information. If needed, based on the multiple signal classification (MUSIC) algorithm, an estimate value of heart rate is obtained by calculating the pseudo-spectral peak in signal subspace.


Face recognition
Select ROI
$t\in[1,T]$
Aligned facial skin sub-area in RGB space
Calculate PPG signal
ϕ
B
G
R
t
PPG fingerprint of single channel
PPG fingerprint
Face recognition
Select ROI
$t\in[1,T]$
Connect the pixels in a 'Z' shape
AR coefficient Φ of order P
ϕ
t
AR fingerprint of single channel
AR fingerprint

(a) Process of generating CPPG fingerprints.

(b) Process of generating AR fingerprints.

Fig. 3. Process of generating fingerprints. With $\mathcal{P} = 36$ and $t = 128$, fingerprints of size 36×128 are obtained. 'Z' shape ordering is shown in Fig. 5


Step1
One sub-area
One PPG signal with t length
ROI in sub-areas
Step2
P sub-areas
P PPG signals arranged in rows
Step3
Single channel fingerprint, size $t\times P$
Pixels of ROI
Step1
"Z" shape ordering in one frame
One-dimensional AR coefficient with length P
Step2
t frames
t AR coefficients arranged in columns
Step3
Single channel fingerprint, size $t\times P$

(a) Details of generating PPG fingerprint.

(b) Details of generating AR fingerprint.

Fig. 4. Detailed process of generating PPG and AR fingerprints in a single channel. The PPG signal is used to reflect the remote HR, and its fluctuation is regarded as temporal-domain features; the AR model is used to reflect the pixel-wise correlation, which is considered as a spatial-domain features.

B. Pixel-level fingerprint of AR coefficient

By employing AR models, the conditional distribution of each pixel can be modeled based on its previous neighbors. This allows the AR models to capture spatial dependencies in the image, including edges, textures, and other structures. As previously discussed, the prediction of each pixel is determined by a linear combination of its preceding neighbors, with autoregressive coefficients serving as weights. These coefficients represent the strength of the relationship between each pixel and its neighbors. Once the coefficients are learned, the AR model can be utilized for various image processing and detection tasks [48], [61].

For a stationary non-white noise sequence, a linear model is usually established to fit the trend of the sequence in statistics, and the useful information in the sequence is extracted by this. AR model is a representation of a type of random process. At time o, the output variable $\boldsymbol{L}^{(o)}$ depends linearly on its own previous values and a stochastic term, i.e.,

$$\boldsymbol{L}^{(o)} = c + \sum_{p=1}^{\mathcal{P}} \varphi_p \boldsymbol{L}^{(o-p)} + \epsilon_p \,. \tag{15}$$

$\mathcal{P}$ is the order of AR model. φ_p is the autocorrelation coefficient of p-th orders, where $1 \le p \le \mathcal{P}$. c is a constant, and ϵ_p is white noise. In the established model, the calculated correlation coefficient $\boldsymbol{\phi} = [\varphi_p]_{1\le p\le \mathcal{P}}$ can be used to evaluate the correlation between current sample and previous sample.

Therefore, the AR model can be employed as a linear prediction model that predicts the current sample by assessing the correlation between the current sample and the preceding sample. Taking into account the principles of camera imaging, we assume that pixel points are generated sequentially. In real facial images, pixels exhibit strong correlations, which can be modeled separately for rows and columns as stable AR processes.

$$\boldsymbol{L}_{row}\left(\boldsymbol{G}(x,y)_i^t\right) = c_r + \sum_{p_r=1}^{\mathcal{P}} \varphi_{p_r}\boldsymbol{L}_{row}\left(\boldsymbol{G}(x\text{-}p,y)_i^t\right) + \epsilon_{p_r}\,, \tag{16}$$

or

$$\boldsymbol{L}_{col}\left(\boldsymbol{G}(x,y)_i^t\right) = c_c + \sum_{p_c=1}^{\mathcal{P}} \varphi_{p_c}\boldsymbol{L}_{col}\left(\boldsymbol{G}(x,y\text{-}p)_i^t\right) + \epsilon_{p_c}\,. \tag{17}$$

$\boldsymbol{L}_{row}\left(\boldsymbol{G}(x,y)_i^t\right)$ is the pixel value of frame t calculated by row. $\boldsymbol{L}_{col}\left(\boldsymbol{G}(x,y)_i^t\right)$ is the pixel value of frame t calculated by column. p_r and p_c represent the order of AR computed by row and column, respectively. c_r and c_c are constants.

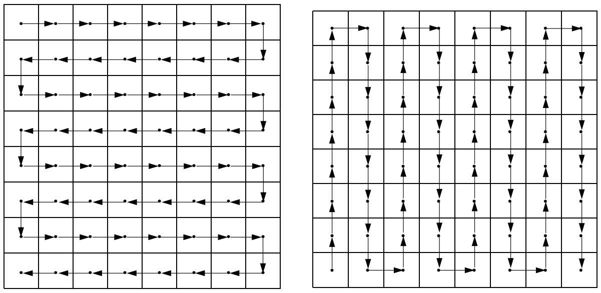
C. Generation of fingerprint image

It is worth reiterating that in this paper, we utilize two distinct fingerprints generated from AR coefficients and the extended PPG signal. Deepfake videos contain volatile temporal and spatial information. PPG is adept at detecting temporal-domain variations, while AR excels at capturing spatial-domain relationships. Therefore, this article presents an innovative approach by combining these two signals to obtain fusion fingerprint information based on the temporal-spatial domain.

The pure PPG signal exhibits stable and periodic fluctuations within the heart rate range. However, fake face videos generated frame by frame disrupt the quasi-periodicity of the PPG signal, leading to abrupt changes in its instantaneous phase. As a result, the anomalous variations in the PPG signal can be regarded as fingerprints of a Deepfake face. The process of generating a PPG fingerprint is shown in Fig. 3 (a), and a more detailed process is shown in Fig. 4 (a). Specifically:

i. Extract a collection of face images $\mathcal{G} = \left\{\boldsymbol{G}(\boldsymbol{X},\boldsymbol{Y})^t\right\}_{t=1}^{T}$ from a video of frame length T using face detector.
ii. Select the pixels enclosed by four pixel coordinates of the cheek region as ROI $\mathcal{R}_o$ for feature extraction.
iii. Construct point-line relationship by Triangulation and stretch $\mathcal{R}_o$ into a regular rectangle $\mathcal{R}_n$ by means of affine transformation. Divide each frame of $\mathcal{R}_n$ uniformly into 36 equal area non-overlapping sub-regions $\mathbf{R}_n = \left[\mathcal{R}_n^k\right]_{k=1}^{36}$.
iv. Align facial skin ROI $\mathcal{R}_n^k$ in color space of RGB. Separate three dimensional chromaticity signals of $\xi \in \{R, G, B\}$. According to Eq.(14), calculate the PPG signal Υ_ξ^k of each sub-region k over a range of frames of fixed length $t \in [1, T]$, and normalize it to the range of $[0 - 255]$. Replace B channel that contains the least obvious HR information with Υ_ξ^k.
v. Calculate PPG signal value of length t from each sub-region $\mathcal{R}_n^k$. Construct a matrix with 36 rows and t columns. The R and G channel values are merged directly with Υ_ξ^k to obtain a color image of size $[3 \times 36 \times t]$ as a PPG fingerprint.

Notably, in step iv., different from 1-D pseudo-color image PPG map generated in [8], our method employs 3-D RGB image. Specifically, CPPG signal is used to replace chrom-signal of B channel. R and G channels are retained. This is done for two reasons. a) Using three channels allows the maximum possible retention of HR as reflected by chrom-information. The reason for replacing B channel is that, among three channels of the camera sensor, B spectrum has the worst relative sensitivity at longer wavelengths (i.e., wavelength values prone to diffuse reflection). b) The 3D PPG fingerprint shares the same dimensions as the AR fingerprint, allowing for the use of a single model structure to perform detection. This also simplifies the comparison of models in ablation experiments.



(a) 'Z' shape ordering in row. (b) 'Z' shape ordering in column.

Fig. 5. 'Z' shape ordering. We stitch the pixels by row and by column, and finally combine them together for AR detection.

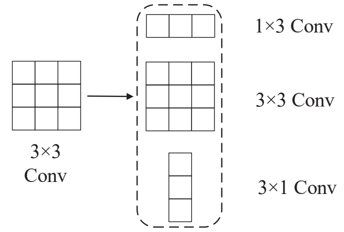


Fig. 6. Structure of ACBlock.

Each row or column of pixels in image has a certain correlation, so it can be captured by an AR model. Coefficient of AR can be used as a one-dimensional AR signal to describe relevant information between pixels. Process of generating AR fingerprint is shown in Fig. 3 (b), and a more detailed process is shown in Fig. 4 (b). Specifically:

i.-iii. The same as the process of generating PPG fingerprint.
iv. Traverse the pixel values $(x,y)_i$ of rows $[1, 3, ..., (2q-1)]_{q \le \frac{M}{2}}$ in the order from left to right, and traverse the pixel values of rows $[2, 4, ..., 2p]_{p \le \frac{N}{2}}$ in the order from right to left when building an AR model. They are connected row by row. According to CMOS camera, videos are captured by using a rolling shutter sampling mechanism [52]. Pixels are stitched in a 'Z' shape shown in Fig. 5, i.e., the pixels at the end of an odd numbered row are more correlated with the pixels at the end of the next even numbered row. Same thing with column traversal. Finally, the grey values in both directions will be joined by means of a head-to-tail connection to give a one-dimensional sequence of pixel values $(\boldsymbol{X},\boldsymbol{Y}) = [(x,y)_i]_{i=1}^{M\times N} \in \mathbb{R}^{M\times N}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TBDATA.2023.3284272, IEEE Transactions on Big Data

v. Set the order of AR to the number of blocks in PPG sub-region, i.e., $\mathcal{P} = 36$. Calculate the one-dimensional model coefficient matrix $\boldsymbol{\phi} = [\varphi_p]_{1 \leq p \leq 36}$ of the AR model, according to Eq. (17).

vi. Calculate the AR coefficient matrix $\boldsymbol{\phi}$ within a fixed frame length $t \in [1, T]$, and construct a $[36 \times t]$ gray-scale image. Therefore, $[3 \times 36 \times t]$ color images based on RGB colour space are obtained as a AR fingerprint.

In Fig. 4 (a), using a single channel as an example, the generation of the PPG fingerprint can be divided into three steps. In step 1, a PPG signal with a length of t is obtained from a single sub-area. In step 2, $\mathcal{P}$ PPG signals from $\mathcal{P}$ sub-areas are arranged in rows. In step 3, the $\mathcal{P}$ PPG signals with a length of t are combined to form a single-channel PPG fingerprint with a size of $t \times \mathcal{P}$. The PPG signals are used to reflect the remote heart rate, and their fluctuations are considered as temporal-domain features. Similarly, in Fig. 4 (b), take a single channel as an example, the generation of AR fingerprint can be divided into three steps. In step1, in each frame, pixels are connected by row and by column through 'Z' shape ordering into a sequence. The sequence is used to generate a $\mathcal{P}$ order AR model. $\mathcal{P}$ AR coefficients ϕ are generated into a one-dimensional feature sequence. In step2, t AR coefficient sequences generated by t frames are arranged in columns. In step3, t AR sequences with $\mathcal{P}$ length are stitched into the AR fingerprint with size of $t \times \mathcal{P}$.

Fig. 7 illustrates sample facial images generated by the original and each deepfake method (top), along with an example of the PPG fingerprint generated from the same window (middle) and the AR fingerprint generated from the same window (bottom). The fake face fingerprints exhibit overall background chromaticities similar to those of the original fingerprint, but with distinct local distributions. Among the PPG fingerprints, the one that bears the closest resemblance to the original is the fingerprint obtained by the NeuralTextures method. This is because NeuralTextures method utilizes a face model to track and render the corresponding UV mask, optimizing the neural texture to regenerate the final face. This optimization process minimizes facial chroma mutation, setting it apart from the other three deepfake methods. On the other hand, the AR fingerprint that closely resembles the original is the fingerprint obtained by the Face2Face method, whereas the others exhibit noticeable local differences. This is because the Face2Face method fits two 3D models of the face in both the source and target domains, performing a pixel-to-pixel conversion of key facial landmark points, which is then further rendered. This forgery method effectively "smooths out" pixel-level image differences and mutations, resulting in a fingerprint that closely resembles the original. The performances of these indistinguishable fingerprints are also consistent with the results presented in Tab. IV.

D. Authenticity discrimination with ACBlock-based DenseNet

An overview of the network architecture for Deepfake detection is depicted in Fig. 2. To identify deep face forgery, a novel convolutional layer structure inspired by asymmetric convolutional networks (ACNets) [12] is incorporated on top of DenseNet. In typical networks, 3×3 convolution kernels are commonly used. However, ACNet introduces asymmetric convolution, which splits the 3×3 convolution into asymmetric convolutions. This asymmetric convolution structure enhances the network's robustness to vertical and horizontal flipping of the input feature images. Consequently, the detection results remain unaffected regardless of the order in which the input fingerprint information is stitched together.

Specifically, in the training phase, we replace each 3×3 convolutional kernel with an ACBlock (shown in Fig. 6) containing 3×3, 1×3 and 3×1 convolutional kernels in our classifier, two *DenseNet-121*. We perform the convolutional operations separately and sum up their final results. Once the convolutional kernels have been fused, the same 3×3 structure is used as the normal convolutional kernels in the inference phase.

We leverage two *DenseNet-121* networks [24] as our classifier, where we replace the 3×3 convolution kernels in the network with ACBlock to enhance the robustness of image inversion up and down and left and right, thereby eliminating the effect of the different stitching order of PPG or AR features. According to [12], the performance of *DenseNet-121* is boosted without introducing any computation and memory increase in the inference phase.

In the model fusion stage, the arithmetic mean of the probability values from the two networks is calculated to determine the video segment classification accuracy. Subsequently, the segment labels are aggregated into video labels through majority voting, which helps compensate for the impact of incorrect frames.

It is worth noting that our method differs from classical approaches that heavily rely on complex deep learning structures. Instead, our method is model-agnostic and only requires an improved version of *DenseNet-121*. In other words, our approach focuses on extracting discriminative features from real and fake faces, which can be effectively utilized in combination with model-based methods for enhanced performance.

IV. Simulation results

The system is implemented in *Python*, using the *dlib* library for face detection, *OpenCV* for image processing. The networks are trained and tested on 4 *Tesla V100* GPUs, resulting in short training times. The extraction of PPG and AR fingerprints from large datasets is the most computationally intensive part of the system. However, since our method is independent of the deep learning model, the generation of PPG and AR fingerprints only needs to be done once. Consequently, our approach has significantly smaller computational complexity compared to the model-based state-of-the-art approach.

A. Datasets and Implement details

During our experiments, we use five types of data sets. *FaceForensics++* [46] is a facial forgery dataset that enables researchers to train deep learning-based methods in a supervised manner. The main experiments are performed on the set *c40* with high compression rate, which includes 4000 fake face videos from 1000 authentic vdieos, separately

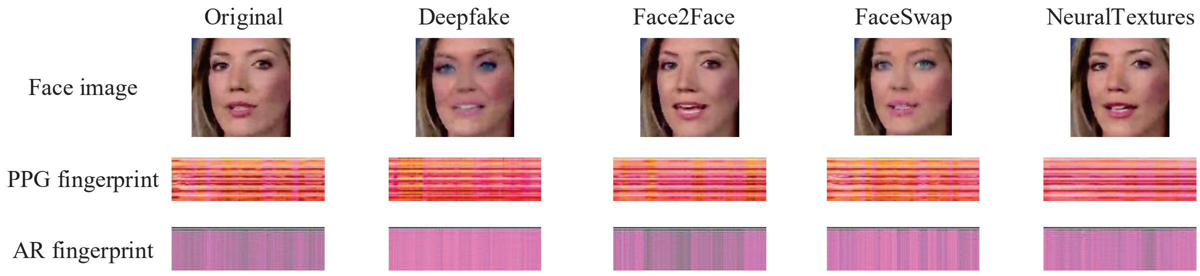


Fig. 7. Examples of face images and their PPG fingerprints, AR fingerprints. Real faces and faces generated by Deepfake, Face2Face, FaceSwap, NeuralTextures models (top) are extracted to PPG fingerprints (middle) and AR fingerprints (bottom) with our scheme.

TABLE III
QUALITATIVE COMPARISON: AUTHENTICITY DETECTION ACCURACY OF DIFFERENT FORGERY DETECTION SCHEMES.

Methods		FaceForensic++					FaceForensic	CelebDF	DFDC	FakeAVCeleb
		Deepfakes	Face2Face	FaceSwap	NeuralTextures	All				
Deep network-based	Xception [7]	0.9428	0.9156	0.9370	0.8211	0.9041	0.8781	0.4820	0.6120	0.6790
	MesoNet [1]	0.8952	0.8444	0.8356	0.7574	0.9331	0.8213	0.5480	0.7540	0.5730
	Slowfast [16]	0.9242	0.9493	0.9501	0.8255	0.9053	-	0.6624	0.7100	-
	Multi-attention [63]	0.9510	0.9336	0.9354	0.8137	0.8869	-	0.6744	-	**0.7760**
	3-D CNN [40]	0.9537	0.8972	0.9220	0.8105	0.9450	-	-	-	-
Similar feature-based	DeepRhythm [43]	0.9372	0.9314	0.9262	-	0.9493	0.9105	-	0.6410	-
	SPSL [34]	0.9348	0.8602	0.9226	0.7678	0.8157	-	0.6940	0.6516	-
	Fakecatcher [8]	0.9487	**0.9600**	0.9575	-	0.9465	0.9066	0.9150	-	0.6947
Our method combining other network structures	Ours(VGG-16)	0.9497	0.9424	0.9415	0.8561	0.9488	0.9208	0.9216	0.7102	0.7327
	Ours(Xception)	0.9511	0.9572	0.9565	0.8677	0.9602	0.9261	0.9244	0.7543	0.7344
	Ours(DenseNet121)	0.9541	0.9553	0.9577	0.8701	0.9551	0.9310	0.9298	0.7618	0.7305
	Ours(ACBlock-based DenseNet121)	**0.9644**	0.9582	**0.9595**	**0.8797**	**0.9613**	**0.9401**	**0.9341**	**0.7740**	0.7531

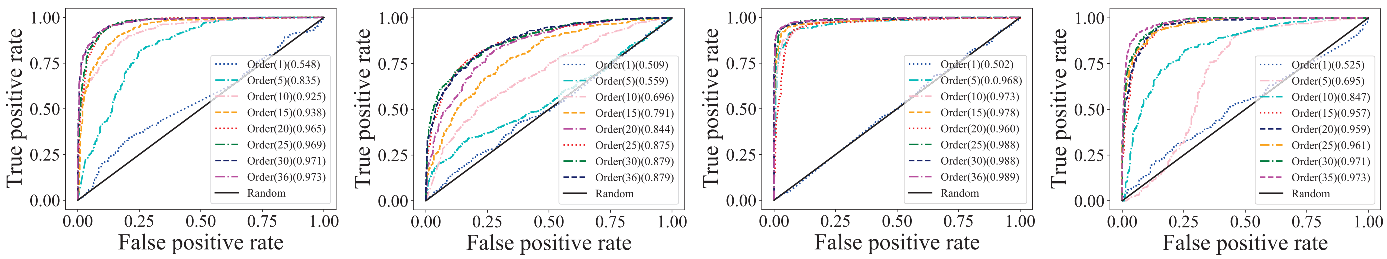
TABLE IV
ABLATION EXPERIMENTS: DETECTION ACCURACY (%) OF ORIGINAL RGB SIGNAL WITH ACNET, REMOVING AR FINGERPRINT (I.E., PPG FINGERPRINT WITH ACBLCOK), REMOVING PPG FINGERPRINT, REMOVING ACBLCOK (I.E., AR FINGERPRINT AND PPG FINGERPRINT WITH CNN), AND OUR PROPOSED SCHEME.

Manipulation	RGB	-AR	-PPG	-ACBlock	Ours
Deepfakes	85.32	96.01	93.20	95.50	96.44
Face2Face	76.61	88.93	82.17	89.40	95.82
FaceSwap	80.01	92.29	94.58	95.70	95.95
NeuralTextures	70.57	82.04	87.26	86.59	87.97
All	78.75	93.38	93.65	95.15	96.13

generated by four submethods: *DeepFake (DF), Face2Face (F2F), FaceSwap (FS)* and *NeuralTextures (NT)*. *Celeb-DF* [32] dataset contains 590 real and 5639 DeepFake synthesized videos with subjects of different ages, ethic groups and genders. *FaceForensics* [45] dataset is the predecessor of FaceForensics++. It contains over 1000 original videos, and their faked videos processed with the Face2Face method. The Deepfake Detection Challenge (DFDC) [13] dataset has 19,197 videos of real footage shot by approximately 430 actors and the remaining 100,000 videos are fake face videos generated from real videos. Fake faces are generated using DeepFakes, GAN-based and partially non-learned methods. FakeAVCeleb [27] generates 19,500 deepfake videos using a base set of 490 authentic videos featuring individuals from various ethnic groups - Caucasian, Black, South Asian, and East Asian. Each ethnic group has 100 genuine videos of 100 celebrities, with an equal 1:1 male and female ratio. The creators utilize Faceswap and FSGAN to produce the swapped deepfake videos.

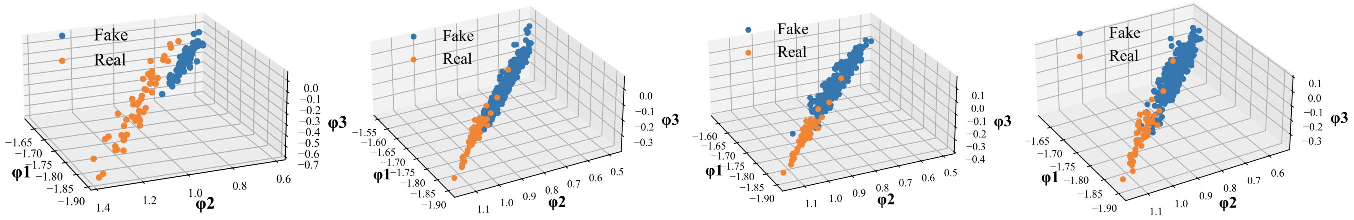
Simulations are performed to evaluate the performance of our methods with fixed frame length $t = 128$ in each video segment. Referring to the experimental evaluation in FakeCatcher [8], $\mathcal{R}_n^k$ is divided into 36 subareas, thus the size of each fingerprint image is $[36 \times 128]$. For a more objective evaluation, in addition to cross-domain and cross-method evaluations, we utilize a 10-fold cross-validation strategy to divide the entire dataset into 10 groups. The training and testing sets are divided in a ratio of 7:3, and the models are trained and tested accordingly. This entire process is repeated ten times, with different groups assigned as training and testing sets in each iteration. Ultimately, the accuracy of each fold is recorded and the average accuracy is calculated as the final accuracy score. In the case of cross-domain and cross-method evaluations, we train 10 models on the corresponding overall dataset for detection, and the average accuracy of the 10 models is considered as the final accuracy score. Additionally, in order to calculate the video detection accuracy, when randomly selecting, it is necessary to ensure that the PPG and AR fingerprint images extracted from the same video are either all in training set or all in test set.

We train our networks using Tensorflow for 80 epochs with batches of size 32. For optimization, we use Adam with learning rates of 0.0005 and a small weight decay of 0.1. To reflect the generalisation of the method attributed to the spatio-temporal domain features themselves rather than to



(a) AR orders of Deepfakes. (b) AR orders of Face2Face. (c) AR orders of FaceSwap. (d) AR orders of NeuralTextures.

Fig. 8. ROC curves and AUC values with different AR orders.



(a) Deepfakes. (b) Face2Face. (c) FaceSwap. (d) NeuralTextures.

Fig. 9. Visualization of the first three coefficients of 36-order AR model obtained from different database.

the complex model, we do not use any image enhancement methods such as random scaling, cropping, and flipping.

B. *Order analysis of AR*

Generative models for fake face images such as GANs, still leaves forgery traces in the image in the form of padding between adjacent pixel points during upsampling process. The upsampling process allows AR to perform tamper detection with the help of correlation between pixel points determined. Receiver operating characteristic (ROC) curves remain credibly evaluated on the unbalanced data set because they are not affected by the proportion of the category distribution. Calculated by scratch CNN model on *FaceForensics++*, ROC curves with $\mathcal{AR}_{orders} = [1, 5, 10, 15, 20, 25, 30, 36]$ are shown in Fig. 8. We can observe that the area under curve (AUC) rises until the 25-th order, beyond which further increases in order do not significantly impact the AUC of the model. Even when the order goes to $\mathcal{AR}_{orders} = 36$, the AUC is stable and saturated. For example, when AR order of Deepfakes is 5, the AUC is only 0.835; when the order is up to 20, the AUC reaches 0.965, and when the order is 36, the AUC remains stable at 0.973. It is important to note that the AUC has reached a plateau, and increasing the order further will not enhance the performance of our model. In terms of AR order selection, on the one hand, the difference of fingerprints between real and fake faces is pronounced when PPG order is 36. On the other hand, after the AR order is greater than 20, the increase in order does not improve the validity of AR fingerprint. Therefore, the AR order of 36 is selected to ensure the validity and to be the same as the PPG order. To visualize the validity of the AR model, we plot the first three of the 36 AR coefficients obtained from real and fake faces. The fake faces are generated from the corresponding real faces. As depicted in Fig. 9, it is evident that the AR coefficients of real and fake faces can be distinctly differentiated.

C. *Quantitative comparison*

The quantitative analysis of the proposed model-agnostic method in comparison to the deep network-based approach, the similar feature-based approach, and our method combining other network structures is presented in Tab. III.

In Tab. III, it is evident that our method achieves the highest overall performance across most types of Deepfake datasets. For example, on all manipulated videos from FaceForensics++, the video accuracy of our method increases by 2.82% over the MesoNet [1], which is the best performer among the depth-based models. On the FaceForensics++ dataset, our proposed method outperforms the deep network-based approach that also considers spatio-temporal information, 3-D CNN, by 1.63%. Despite 3-D CNN considering both temporal and spatial information, its utilization of 3D convolution kernels entails significant computational overhead in the convolutional layer. In contrast, our method employs a lightweight network with pre-obtained spatio-temporal fingerprint inputs, resulting in significantly reduced computation costs and increased flexibility with measurable datasets. Our method performs best among similar biosignal-based methods. For instance, on the FaceForensics++ dataset with four manipulations, our method surpasses the SPSL [34] method by 14.56% in video accuracy and outperforms the Fakecatcher [8] method by 1.48%. Similarly, on the FakeAVCeleb dataset, our method exhibits a 5.84% improvement over the Fakecatcher method. This improvement can be attributed to our enhancement in generating PPG fingerprints and our utilization of the AR model to consider pixel correlations within frames. Additionally, our proposed method demonstrates the highest accuracy among other state-of-the-art methods on the challenging DFDC dataset, indicating its efficiency in handling large datasets and its high degree of generalization. To further demonstrate the effectiveness of PPG and AR and to showcase the model-agnostic nature of our proposed method, we also evaluate the

TABLE V
VIDEO ACCURACY OF CROSS DEEPFAKE GENERATION MODEL AND CROSS FAKE FACE VIDEO DATA DOMAIN.

	Train set	Test set	Accuracy
Cross-model	FF++ - DF	DF	0.9575
	FF++ - F2F	F2F	0.9247
	FF++ - FS	FS	0.9615
	FF++ - NT	NT	0.8525
Cross-domain	FF++	FF	0.9066
	FF++	Celeb-DF	0.8657
	FF++	DFDC	0.7462
	FF++	FakeAVCeleb	0.7254
	FF	FF++	0.8793
	Celeb-DF	FF++	0.9019
	DFDC	FF++	0.9223
	FakeAVCeleb	FF++	0.8774

performance of our method in combination with other network structures. Whether combining VGG, Xception or DenseNets, the proposed method demonstrates better detection accuracy. For instance, when combined with Xception, our method improves accuracy by 44.24% in the CelebDF dataset and by 14.23% in the DFDC dataset compared to using the Xception model directly for end-to-end detection. This highlights the superior performance of our method across different network structures and underscores its model-agnostic nature. Moreover, in the experiments involving various network structures for detection, the combination of ACBlock-based DenseNet yields the best performance. This indicates that the classifier exhibits greater robustness in terms of feature alignment when utilizing this specific network structure.

In a more detailed analysis, we evaluate the performance of our proposed method against four different generative models in the FaceForensic++ dataset. Our solutions outperform the other methods, with the exception of the Face2Face-based generative model, which exhibits lower video detection accuracy compared to Fakecatcher. In addition, for the FakeAVCeleb dataset, our method performed worse than the Deep network-based approach - Multiple-attention. This is because feeding the original facial images into more complex models indeed allows for better capturing and analyzing facial image information, but this requires massive amounts of data for extensive training, which is time-consuming and computationally expensive. However, in our approach, AR fingerprints compensate for the shortcomings of PPG fingerprints in terms of sensitivity to chromatic changes. Thus, acceptable detection results can be achieved with minimal computation.

D. Ablation experiments

In order to showcase the effectiveness of the various components in our proposed scheme, we conducted ablation experiments with different combinations of these components, as shown in Tab. IV. In summary, when compared to the ablation experiments that remove specific components, our proposed method demonstrates higher video detection accuracy across different fake face generation models and overall authenticity detection. Specifically, on Face2Face, the accuracy of our method improves by 6.42% over removing the ACBlock part (replacing it with a simple architecture CNN network). Even without ACBlock structure, the performance of the proposed method is still better than that of FakeCatcher [8] method. This reflects the improvement of the robustness of the ACNet asymmetric convolution structure to the selected feature fingerprints upside-down and left-right upside down, eliminates the influence of the feature arrangement order. Taking the last column as a reference, the overall accuracy of the second and third columns has increased by 2.75% and 2.48%, respectively. This respectively reflects the gain effect of fingerprint features based on AR coefficients in the space domain, and fingerprint features based on PPG signals in the time domain. Neither of the first two columns in the table can accurately determine the authenticity of the face video. For instance, in an experiment based solely on $\mathcal{RGB}$ channels without any feature preprocessing, the video detection accuracy is only 78.75%. In experiments where the entire frame pixels are directly used for detection, the authenticity detection method is almost ineffective. These experimental results reaffirm two perspectives:

i. Feature preprocessing based on the sparsity of AR coefficients is a crucial factor in authenticity detection.
ii. In fake face detection experiments, it is necessary to extract the face in advance rather than using the entire video frame to improve detection performance.

E. Cross-model and cross-domain evaluation

In this section, we perform cross-model and cross-domain evaluations to assess the generalization ability of our proposed model. The results are presented in Tab. V. In the cross-model evaluation, except for NeuralTextures, the remaining models achieve high detection performance. For instance, when the model is trained using the other categories except for FaceSwap, and then the fingerprint data of the FaceSwap-generated model is used for detection, the video detection accuracy reaches 96.15%. The cross-model accuracy of NeuralTextures is only 85.25% due to its fundamentally different nature compared to other generative models.

In addition, the second part of Tab. V shows the results of the cross-domain evaluation of *FaceForensic++* with *FaceForensic*, *Celeb-DF*, *DFDC* and *FakeAVCeleb*. When tested on a more realistic database such as *Celeb-DF*, the performance of the proposed method slightly decreases due to the presence of highly realistic fake faces in the dataset. However, the performance remains acceptable, indicating the generalization ability and usefulness of our method in various real-world deepfake detection scenarios. However, the model trained using *FaceForensic++* performs poorly on *DFDC* and *FakeAVCeleb* datasets. Because these datasets contain not only deepfake faces, but also faces generated using GAN-based and other non-learning methods. In addition, *FakeAVCeleb* is a large-scale multi-ethnicity dataset, where skin color information has not been fully pre-trained and learned in the *FaceForensic++* dataset. Although AR fingerprints partially compensate for the shortcomings, the judgment ability of the model is still affected. Conversely, training on diverse datasets and performing detection on *FaceForensic++* dataset proves

TABLE VI
EVALUATION ON *FaceForensic*++ WITH DIFFERENT COMPRESSION RATES. THE $c0$, $c23$ AND $c40$ REPRESENT ORIGINAL VIDEOS, HIGH QUALITY (LIGHT COMPRESSION) AND LOW QUALITY (HEAVY COMPRESSION), RESPECTIVELY.

	Deepfake			Face2Face			FaceSwap			NeuralTextures			All		
	PPG	AR	All	PPG	AR	All	PPG	AR	All	PPG	AR	All	PPG	AR	All
c0	0.9947	0.9922	0.9959	0.9647	0.9725	0.9774	0.9964	0.9961	0.9911	0.9290	0.9331	0.9303	0.9865	0.9792	0.9902
c23	0.9791	0.9819	0.9846	0.9599	0.9578	0.9687	0.9713	0.9662	0.9732	0.8962	0.9007	0.9111	0.9559	0.9527	0.9757
c40	0.9489	0.9613	0.9644	0.9308	0.9490	0.9582	0.9577	0.9498	0.9595	0.8485	0.8719	0.8797	0.9338	0.9365	0.9613

TABLE VII
CROSS COMPRESSION EVALUATION ON *FaceForensic*++.

Cross compression rate		FaceForensics++				
Train	Test	Deepfakes	Face2Face	FaceSwap	NeuralTextures	All
C0	C40	0.9973	0.9782	0.9883	0.9361	0.9823
C40	C0	0.9961	0.9330	0.9782	0.9217	0.9452

TABLE VIII
VIDEO ACCURACY ON THE OVERALL *FaceForensic*++ DATASET WITH DIFFERENT GAUSSIAN FILTERS. THE KERNEL SIZES OF GAUSSIAN FILTERS ARE $(3, 5, 7, 9, 11)$.

Kernel	PPG	AR	All
N/A	0.9865	0.9792	0.9902
3×3	0.9804	0.9543	0.9883
5×5	0.9537	0.8823	0.9604
7×7	0.8729	0.7867	0.8897
9×9	0.7719	0.6776	0.7744
11×11	0.6142	0.5902	0.6237

TABLE IX
VIDEO ACCURACY ON THE FACEFORENSICS++ DATASET WITH DIFFERENT SIZES OF NON-OVERLAPPING SUBREGIONS $\mathcal{R}_n^k$ SETTINGS.

Sub-region	$\mathcal{R}_n^k$=25	$\mathcal{R}_n^k$=36	$\mathcal{R}_n^k$=49	$\mathcal{R}_n^k$=64
Deepfakes	0.9637	**0.9644**	0.9412	0.8897
Face2Face	0.9551	**0.9582**	0.9233	0.8523
FaceSwap	0.9512	**0.9595**	0.9326	0.8747
NeuralTextures	0.8546	**0.8797**	0.8327	0.6949
All	0.9601	**0.9613**	0.9096	0.8288

advantageous, as it is a widely used and generic deepfake dataset.

It is worth noting that the effectiveness of the proposed method is confirmed by combining the cross-model and cross-domain evaluations. The information captured by the AR model and PPG fingerprints reflects the fundamental differences between real and fake faces, making it applicable to deepfake detection in a generalized manner, independent of the dataset.

F. Compression rate experiment

To make a fair comparison with other SOTA models, the dataset used in our paper is the c40 compressed version (low quality videos). Additionally, we also test the efficacy of our method on the original videos (c0) and high quality videos (c23). The results, as shown in Tab. VI, demonstrate that our proposed AR fingerprint significantly improves performance. Regardless of whether the dataset is compressed or not, our AR fingerprint substantially enhances the accuracy of the model. This improvement is attributed to the fact that the distinction between real and fake faces is clearly reflected in the pixel-wise correlation, which is independent of the dataset.

We also conducted cross-compression experiments using the uncompressed (c0) and highest compressed (c40) versions of the *FaceForensic*++ dataset, and the resulting accuracies are presented in Tab. VII. The table demonstrates that the proposed method achieves successful detection in both experiments: using c0 as the training set and c40 as the testing set, as well as using c40 as the training set and c0 as the testing set. This success can be attributed to the compression method used for c40, which is intra-frame compression H264. This method preserves the temporal information between frames with minimal disturbance, resulting in largely undisturbed time-domain features. Additionally, although the complete image is re-encoded in the spatial domain, the overall distribution of pixels and chrominance remains similar, leading to similar feature fingerprints. Consequently, the neural network smoothly fits the approximate distribution results and successfully detects deepfake videos. Furthermore, when c0 is used as the testing set, the detection performance of c40 is generally higher, as the training set c0 contains more information.

G. Blurring experiments

To verify the robustness of the proposed method, we conduct an analysis of its detection performance under Gaussian blurring operations, as demonstrate in Tab. VIII. We convolve uncompressed face images (c0) from the overall *FaceForensic*++ dataset with Gaussian kernels of varying sizes to simulate blur effects. The results of the forgery detection using AR fingerprint, PPG fingerprint, and spatial-temporal fusion fingerprint are shown in Tab. VIII. The PPG fingerprint demonstrate a generally high resistance to the blurring of 7×7 Gaussian kernel, while the AR fingerprint shows a higher resistance to 5×5 size. Meanwhile, the spatial-temporal fusion fingerprint can resist blurring of 7×7 Gaussian kernel and demonstrate the highest accuracy at each level of blur. As the blurring increases, the accuracy decreases gradually. In fact, when the image is processed with a very large Gaussian kernel, the blur visible to the naked eye made it difficult to distinguish between real and fake faces, rendering the detection meaningless.

H. Sub-regional division experiment

The division of face sub-regions plays a crucial role in capturing the representativeness and subspace correlation of the PPG signal. Choosing sub-regions that are too small may

TABLE X
VIDEO ACCURACY ON THE FACEFORENSICS++ DATASET WITH DIFFERENT SIZES OF FRAGMENT LENGTH EXPERIMENT t SETTINGS.

Segment length	t=64		t=128		t=256		t=512	
	s. acc.	v. acc.	s. acc.	v. acc.	s. acc.	v. acc.	s. acc.	v. acc.
Deepfakes	0.9015	0.9276	**0.9371**	**0.9644**	0.9299	0.9607	0.9106	0.9174
Face2Face	0.8506	0.8523	0.9296	**0.9582**	**0.9304**	0.9417	0.8583	0.8600
FaceSwap	0.8644	0.8785	**0.9545**	**0.9595**	0.9001	0.9333	0.8633	0.8942
NeuralTextures	0.7312	0.7749	**0.8566**	**0.8797**	0.8371	0.8687	0.7871	0.7921
All	0.8687	0.8842	**0.9293**	**0.9613**	0.9131	0.9556	0.8264	0.8288

result in capturing mostly noise, while selecting sub-regions that are too large may lead to less representative features. To strike a balance, we experimentally divided the face region into non-overlapping sub-regions using $\mathcal{R}_n^k = \{16, 25, 36, 49, 64\}$. As shown in Tab. IX, the results are approximate for $\mathcal{R}_n^k = 25$ versus $\mathcal{R}_n^k = 36$. But the best value is at $\mathcal{R}_n^k = 36$ where the computational complexity is low. Since then, as $\mathcal{R}_n^k$ increases, the accuracy decreases.

I. Fragment length experiment

Our method calculates video-level accuracy while segment accuracy is calculated using a fixed number of segment frames, specifically $t = 128$. This choice aligns with the original PPG-based method proposed by [8].

The duration of the segment is crucial for extracting a stable PPG signal. If the segment is too short, important PPG frequencies may be missed, making extraction impossible. On the other hand, if the segment is too long, it may contain excessive noise, obscuring the actual signal. We evaluate segments with frame lengths of $t = \{64, 128, 256, 512\}$, and the results in Tab. IX shows the segment accuracy (s. acc.) and video accuracy (v.acc.). The best video accuracy is obtained when the segment length is set to 128.

V. CONCLUSION

The existing Deepfake detection methods often fail to investigate the fundamental differences between real and fake faces. That is, these methods either blindly utilize deep learning or use biosignal features, but none of them consider the spatial and temporal relevance of face features. Consequently, these models lack interpretability and struggle to generalize well, often resulting in poor performance on cross-domain tests.

In this paper, our first contribution is to identify two new reliable fingerprints. The first one is obtained from pixel-wise correlation based on AR model, which can be utilized as an latent descriptor of authenticity for face video. The second one is the improved PPG signal, which is the extended 3-dimensional signal for effective catching fake content. Our second contribution is to propose a novel temporal-spatial domain Deepfake detection method, which is based on fingerprints from PPG signals and AR coefficients. Notably, the PPG signal is used to reflect the remote HR, and its fluctuation is regarded as temporal-domain features; the AR model is used to reflect the inter-pixel correlation, which is considered as a spatial-domain features. To evaluate these temporal-spatial fingerprints, we employ an ACBlock-based DenseNet, enabling automatic authenticity detection.

Our method is model-agnostic and emphasizes the extraction of discriminative pixel-wise features for real and fake faces through fingerprints. By generating the PPG and AR fingerprints once, we can achieve state-of-the-art performance even with a simple CNN architecture. As a result, our model trains faster than traditional complex model-dependent methods.

Based on the simulation results, our proposed method effectively enhances the generalization capability of authenticity detection, making it reliable for judicial forensics and intellectual property protection purposes.

VI. DISCUSSION

The Deepfake generation poses a serious threat to society, law, and privacy. Our research is devoted to reducing reputation damage, disclosure of state secrets and the issue of social trust crisis caused by such tampering. To the best of our knowledge, recent studies have paid limited attention to identifying valid fingerprints to distinguish real faces from fake ones. Furthermore, the exploration of temporal and spatial domain information in Deepfake detection remains insufficient. Therefore, our findings contribute to mitigating these problems by introducing two reliable fingerprints that serve as effective guides in the field of authenticity detection. Our proposed model-agnostic fake face forensics method effectively improves authenticity detection performance, ensuring the reliability of judicial forensics and intellectual property protection. Future research efforts will be dedicated to exploring more complex information for face authenticity detection. Additionally, we are highly interested in exploring whether skin color and ethnicity have an impact on Deepfake detection.

REFERENCES

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPRW*, June 2019.

[3] Ali Alnaji, Asanka Perera, and Javaan Chahl. Remote monitoring of cardiorespiratory signals from a hovering unmanned aerial vehicle. *Biomedical engineering online*, 16(1):1–20, 2017.

[4] Christoph Brüser, Christoph Hoog Antink, Tobias Wartzek, Marian Walter, and Steffen Leonhardt. Ambient and unobtrusive cardiorespiratory monitoring techniques. *IEEE Reviews in Biomedical Engineering*, 8:30–43, 2015.

[5] Beijing Chen, Xin Liu, Yuhui Zheng, Guoying Zhao, and Yun-Qing Shi. A robust gan-generated face detection method based on dual-color spaces and an improved xception. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[6] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE J. of Selected Topics in Signal Process.*, 14(5):1024–1037, 2020.

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[8] Umur Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of synthetic portrait videos using biological signals. *TPAMI*, pages 1–1, 2020.

[9] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021.

[10] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, pages 5781–5790, 2020.

[11] Gerard De and Arno Van Leest. Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014.

[12] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *ICCV*, pages 1911–1920, 2019.

[13] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[15] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, pages 7890–7899, 2020.

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[17] Litong Feng, Lai-Man Po, Xuyuan Xu, Yuming Li, and Ruiyi Ma. Motion-resistant remote imaging photoplethysmography based on the optical properties of skin. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):879–891, 2014.

[18] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, pages 3247–3258, 2020.

[19] Yu Gu, Yifan Zhang, Jie Li, Yusheng Ji, Xin An, and Fuji Ren. Sleepy: Wireless channel data driven sleep monitoring via commodity wifi devices. *IEEE Transactions on Big Data*, 6(2):258–268, 2018.

[20] Zhenyu Guo, Z Jane Wang, and Zhiqi Shen. Physiological parameter monitoring of drivers based on video data and independent vector analysis. In *ICASSP*, pages 4374–4378, 2014.

[21] David Güera and Edward J. Delp. Deepfake video detection using recurrent neural networks. In *AVSS*, pages 1–6, 2018.

[22] Gerard Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Trans. on Biomedical Engineering*, 60(10):2878–2886, 2013.

[23] Juan Hu, Xin Liao, Wei Wang, and Zheng Qin. Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[24] Gao Huang, Zhuang Liu, Laurens Maaten, and Kilian Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[25] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *CVPR*, pages 8484–8493, 2020.

[26] Xiangui Kang, Matthew C Stamm, Anjie Peng, and Ray Liu. Robust median filtering forensics using an autoregressive model. *IEEE Trans. on Info. Forensics and Security*, 8(9):1456–1468, 2013.

[27] Hasam Khalid, Shahroz Tariq, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2021.

[28] Chenqi Kong, Baoliang Chen, Wenhan Yang, Haoliang Li, Peilin Chen, and Shiqi Wang. Appearance matters, so does audio: Revealing the hidden face via cross-modality transfer. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):423–436, 2021.

[29] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410, 2011.

[30] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-Ray for more general face forgery detection. In *CVPR*, pages 5000–5009, 2020.

[31] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In *WIFS*, pages 1–7, 2018.

[32] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *CVPR*, pages 3204–3213, Seattle, WA, United States, 2020.

[33] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[34] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[35] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398, 2018.

[36] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *CVPR*, pages 8060–8069, 2020.

[37] Jianchang Mao and Anil K Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern recognition*, 25(2):173–188, 1992.

[38] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. CLAS: A database for cognitive load, affect and stress recognition. In *BIA*, pages 1–4, 2019.

[39] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose Deepfakes and face manipulations. In *WACVW*, pages 83–92, 2019.

[40] Xuan Hau Nguyen, Thai Son Tran, Kim Duy Nguyen, Dinh-Tu Truong, et al. Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques. *Forensic Science International: Digital Investigation*, 36:301108, 2021.

[41] Ming Poh, Daniel McDuff, and Rosalind Picard. Advancements in non-contact, multiparameter physiological measurements using a webcam. *IEEE trans. on biomedical engineering*, 58(1):7–11, 2010.

[42] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010.

[43] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4318–4327, 2020.

[44] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103, 2020.

[45] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

[46] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019.

[47] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1):80–87, 2019.

[48] Anne H Schistad Solberg and Anil K Jain. Texture analysis of sar images: a comparative study. *IEEE Transactions on Geoscience and Remote Sensing*, 35:25–32, 1997.

[49] Zhonglin Sun, Li Sun, and Qingli Li. Investigation in spatial-temporal domain for face spoof detection. In *ICASSP*, pages 1538–1542, 2018.

[50] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, João Jorge, DA Clifton, and Chris Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5):807, 2014.

[51] Ruben Tolosana, Ruben Vera, Julian Fierrez, Aythami Morales, and Javier Ortega. Deepfakes and beyond: A survey of face manipulation and fake detection. *Info. Fusion*, 64:131–148, 2020.

[52] Saffet Vatansever, Ahmet Emir Dirik, and Nasir Memon. Analysis of rolling shutter effect on ENF-based video forensics. *IEEE Trans. on Inform. Forensics and Security*, 14(9):2262–2275, 2019.

[53] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[54] Shengyu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, pages 8695–8704, 2020.

[55] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De. Algorithmic principles of remote ppg. *IEEE Trans. on biomedical engineering*, 64(7):1479–1491, 2016.

[56] Jianquan Yang, Honglei Ren, Guopu Zhu, Jiwu Huang, and Yunqing Shi. Detecting median filtering via two-dimensional AR models of multiple filtered residuals. *Multimedia Tools and Applications*, 77(7):7931–7953, 2018.

[57] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, pages 8261–8265, 2019.

[58] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *CVPR*, pages 7556–7566, 2019.

[59] Sun Yu, Sijung Hu, Vicente Azorin, Jonathon Chambers, Yisheng Zhu, and Stephen Greenwald. Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise. *J. of biomedical optics*, 16(7):077010, 2011.

[60] Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[61] Xiangjun Zhang and Xiaolin Wu. Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation. *IEEE Transactions on Image Processing*, 17(6):887–896, 2008.

[62] Changchen Zhao, Chun-Liang Lin, Weihai Chen, and Zhengguo Li. A novel framework for remote photoplethysmography pulse extraction on compressed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1299–1308, 2018.

[63] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, pages 2185–2194, 2021.

JUN YANG is currently pursuing the Ph.D. degree in computer science, Tongji University, Shanghai, China. He was a Visiting Scholar with the Department of Mathematics at Purdue University, West Lafayette, IN, USA. His research interests include deep learning, data analysis, and computer vision.

YAORU SUN received the Ph.D. degree in artificial intelligence from the University of Edinburgh. He is currently a Full Professor with the Department of Computer Science and Technology, Tongji University, China. His research interests include brain-like computation, machine intelligence, and cognitive neuroscience.

MAOYU Mao is currently pursuing the Ph.D. degree in computer science, Tongji University. Her research interests include Deep learning, and pattern recognition.

Lizhi Bai received the Master degree in Mechanical Engineering from Nanjing University of Aeronautics and Astronautics. He is currently a Ph.D. student in Department of Computer Science and Technology, Tongji University, China. His research interests include semantic segmentation and visual SLAM.

Siyu Zhang is currently pursuing the Ph.D degree in Computer science and technology from Tongji University, Shanghai, China. Her research interests include signal processing and computer vision.

Fang Wang is a Senior Lecturer with the Department of Computer Science, Brunel University, Uxbridge UB8 3PH, United Kingdom. Her research interests cover software agents, cognitive neuroscience and distributed computing.