Detection of Photoplethysmography Manipulation in Video Forgery

Yuzhen Lin
Guangdong Key Lab of Intelligent Information Processing,
Shenzhen Key Lab of Media Security, Shenzhen University
Shenzhen, China
linyuzhen2020@email.szu.edu.cn

Emanuele Maiorana
Dept. of Industrial, Electronic and Mechanical Engineering
Roma Tre University
Rome, Italy
emanuele.maiorana@uniroma3.it

Bin Li
State Key Laboratory of Radio Frequency
Heterogeneous Integration, Shenzhen University
Shenzhen, China
libin@szu.edu.cn

Patrizio Campisi
Dept. of Industrial, Electronic and Mechanical Engineering
Roma Tre University
Rome, Italy
patrizio.campisi@uniroma3.it

***Abstract*—Recent studies have shown that physiological signals related to blood pressure and heart rate can be estimated in a contactless modality from facial videos using remote photoplethysmography (rPPG). This has paved the way to the development of techniques that can acquire and manipulate the rPPG signals recoverable from facial videos without affecting their visual appearance. The goal of this paper is to analyze the detectability of this new kind of forgery, here referred to as rPPG deepfake videos. Specifically, we propose a two-stream method based on the analysis of rPPG deepfake videos at both spatial and temporal levels. The experimental results obtained from tests performed on samples taken from two distinct databases demonstrate that our method performs better than popular deep learning methods in the rPPG deepfake detection task.**

***Index Terms*—Remote photoplethysmography (rPPG), Face forgery detection, Video forensics.**

I. Introduction

Physiological signals related to blood pressure and heart rate (HR) are important vital signs that can be measured in many circumstances, especially for healthcare or medical purposes. Traditionally, electrocardiography (ECG) and photoplethysmography (PPG) [1] are the two most common ways for evaluating heart activities and the corresponding physiological signals. However, both ECG and PPG sensors need to be attached to the body, thus causing discomfort and inconvenience for users in everyday life. In order to mitigate this issue, remote photoplethysmography (rPPG) has been developed in the recent years, to estimate, in a contactless modality, relevant heart-related parameters from facial videos. Therefore, rPPG can be used in many applications, ranging from remote healthcare to affective computing [2].

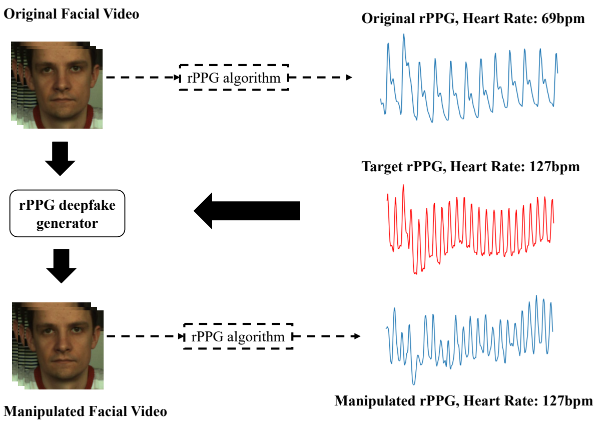
Original Facial Video
Original rPPG, Heart Rate: 69bpm
rPPG algorithm
Target rPPG, Heart Rate: 127bpm
rPPG deepfake generator
rPPG algorithm
Manipulated Facial Video
Manipulated rPPG, Heart Rate: 127bpm

Fig. 1. The pipeline of rPPG deepfake generation. The original facial video is processed so that the manipulated facial video will be visually identical to the original video, while the associated rPPG signal is modified.

Given the relevance that rPPG techniques are recently gaining, methods have also been proposed to manipulate the rPPG signals that can be acquired from facial videos. For instance, methods to conceal rPPG in facial videos have been explored in [3], [4]. In [5]–[7], it has been also proposed to tamper the rPPG signals form facial videos in order to obtain a specific estimated heart rate(see Figure1). Although the motivations behind such methods may be based on legitimate needs, such as privacy protection [6] or the demand for expanding the diversity of rPPG datasets [7], it is worth remarking that rPPG deepfake videos, where the physiological signals estimated from a facial video are intentionally edited, may also represent a notable potential threat, since they could be used to impersonate or deceive medical professionals, or even for criminal activities.

To address this issue, we here propose a two-stream method for the detection of rPPG deepfake videos, by learning spatial- and temporal-level features. In more details, given the very limited perturbations in the skin color space brought by rPPG deepfake generation, our approach relies on the suppression of the image content, in order to capture possible subtle artifacts. To this aim, we employ a steganalytic network [8], which first extracts the image residuals using high-pass filters and then feeds them into a convolutional neural network (CNN) model, thus extracting spatial-level stream features. On the other hand, we note that the heart rate estimated from fake videos may be inconsistent in different subregions of the face. Thus, we first transform the video clips into spatial-temporal maps, which encode the HR signals from multiple region of interest (ROI) volumes on faces and then feed them into a CNN-gated recurrent unit (GRU) model [9], to extract temporal-level stream features. The two models are trained independently, and their outputs are fused to take the final decisions. Experimental tests carried out on two public databases prove the effectiveness of our proposed method in the rPPG deepfake detection task. The main contributions of this paper are summarized as follows:

- to the best of our knowledge, this is the first work that attempts to detect rPPG deepfakes;
- we analyze the properties of rPPG deepfake videos created with state-of-the-art algorithms and propose a two-stream learning framework for their detection;
- the obtained experimental results demonstrate that our method outperforms several popular competitors proposed to detect standard deepfake videos.

II. Related Works

A. Remote Photoplethysmography

Remote photoplethysmography is a non-invasive technique for estimating heart-related physiological signals by analyzing changes in skin color caused by blood flow from facial videos. Traditional approaches [10]–[12] have evaluated the feasibility of extracting the information of interest under different prior assumptions for the considered facial videos. With the development of deep learning technology, many learning-based methods [9], [13]–[16] have also been developed for rPPG estimation, currently achieving reliable performance.

B. Deepfake Video Detection

Deepfake techniques refer to a series of deep-learning-based forgery techniques that can swap or reenact the face of one person in a video in another [17]. The past five years have witnessed a wide variety of methods proposed to counteract the malicious usage of deepfakes. Early works focus on handcrafted features such as eyes-blinking and visual artifacts. Due to the tremendous success of deep learning, CNNs have been widely used to detect deepfakes [18]–[20], commonly achieving better performance than traditional methods. In addition, approaches relying on rPPG signals have been proposed for the detection of deepfakes [21]. In this work, we take a different perspective and attempt to detect for the first time a specific kind of deepfake, consisting in the modification of the rPPG signals.

C. Biometric Recognition with rPPGs

The analysis of rPPG signals has been used within the field of biometrics for liveness detection purposes [22], [23], with the aim of evaluating whether a subject performing a face-based recognition process is wearing a mask or not [24]. The rPPG signal recovered from a facial video has been also exploited as a biometric identifier [25]. The possibility of generating rPPG deepfakes, as proposed in the works mentioned in the following section, is therefore a relevant threat in the biometric framework.

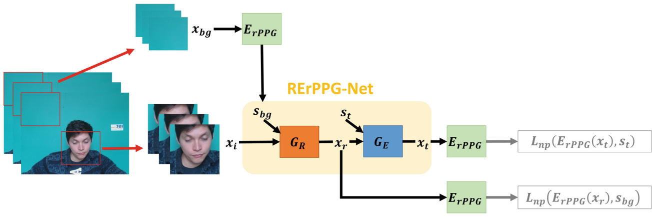

Fig. 2. Architecture of RErPPG-Net [7].

D. rPPG Deepfake Video Generation

Given the several applications of rPPG that are rapidly emerging, novel deepfake algorithms, able to edit the physiological signals detectable from facial videos without affecting their visual appearance, have been also recently proposed [3]–[7], as mentioned in Section I.

The state-of-the-art rPPG manipulation approach, considered for the generation of rPPG deepfake videos in the tests we have performed, consists in RErPPG-Net [7], an architecture operating under the cycle generative adversarial learning framework. The goal of RErPPG-Net is to augment existing rPPG datasets by embedding ground-truth rPPG signals into any existing facial videos. As shown in Figure 2, the proposed RErPPG-Net consists of a Removal-Net G_R and an Embedding-Net G_E. It aims to remove any inherent rPPG signals existing in the input videos and then to embed the target rPPG signals into the rPPG-removed videos. To train the model from unpaired videos, RErPPG-Net proposes a novel double-cycle consistent learning approach to enforce the Embedding-Net G_E and the Removal-Net G_R to learn how to robustly and accurately embed and remove the delicate rPPG signals. As detailed in Section IV, we employ RErPPG-Net on the PURE [26] and UBFC-rPPG [27] datasets to generate the rPPG deepfake videos used in our experiments.

III. Proposed rPPG Deepfake Detector

The proposed rPPG detection approach relies on the analysis of specific characteristics of rPPG deepfake videos. In more details, the goodness of rPPG deepfake videos, generated as in [7], can be evaluated through the perceptual quality of the manipulated videos, measured in terms of peak signal to noise ratio (PSNR) and structural similarity metric (SSIM) between the original and edited videos. Moreover, in order to determine if a target rPPG signal has been properly embedded into the modified video, the mean absolute error (MAE) between the heart rate detected in the manipulated videos, measured by PhysNet [13], and the target one, can be computed.

As shown in Table I, high PSNR and SSIM indexes suggest that RErPPG-Net hardly introduces perceptual distortion on the subject's appearance. This fact demonstrates that RErPPG-Net perturbs the skin pixels in the video frames by a small amount when editing the rPPG signals, thus introducing a kind of non-stationary noise, similarly to how several steganography methods manipulates images [28], [29].

TABLE I
PERFORMANCES OF THE GENERATED rPPG DEEPFAKE VIDEOS.

Datasets	PSNR	SSIM	MAE
PURE [26]	50.73	0.9921	4.39
UBFC-rPPG [27]	52.71	0.9947	2.66

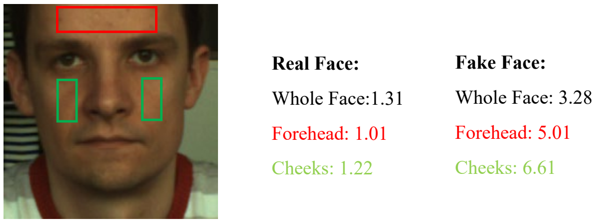

Fig. 3. MAE performance of HR estimation from the two facial subregions (forehead and cheeks).

It has also to be observed that, for original face videos, the results of rPPG estimation are typically consistent over different face subregions. To examine this property, we analyze the MAE of rPPG signals in two facial subregions, i.e., forehead and cheeks, and estimate the heart rate using the plane-orthogonal-to-skin (POS) algorithm [12] included in the rPPG-Toolbox [30]. As shown in Figure 3, the MAE measured by different face subregions in the rPPG deepfake video varies greatly, demonstrating that rPPG deepfakes cannot guarantee a global consistency of rPPG information on the face.

In this work, we propose a two-stream framework, depicted in Fig. 4, to detect rPPG deepfake videos. Details on the proposed spatial- and temporal-level streams, and on their combined use, are given in the following.

A. Spatial-level Stream

As previously mentioned, the manipulation on rPPG signals brings very subtle perturbations to the video. Therefore, we employ UCNet [8], that is, a universal steganalysis network for color images, to extract spatial-level features. Specifically, UCNet consists of preprocessing, convolutional, and classification modules. The preprocessing module first extracts the image residuals from each color channel with 62 fixed high-pass filters, and then concatenate them for the subsequent modules. The convolutional module contains three carefully-designed types of layers with different shortcut connections and group convolution structures, to further learn the high-level steganalytic features. The classification module consists of a global average pooling and a fully connected layer for classification. For each video, the average of all the prediction scores over the processed frames are computed as a final prediction score r_1. The binary cross-entropy loss $\mathcal{L}_{BCE}$ is applied to train the proposed spatial-level stream.

B. Temporal-level Stream

In order to detect potential inconsistencies in rPPG estimation created by possible manipulations, we employ a spatial-temporal map (STmap) [9], designed to learn robust rPPG features from the facial ROI-based spatio-temporal signal map. Such approach can enforce the input to the succeeding network to be as specific as possible to the heart rate signal, and leverage CNN to learn informative representation for it.

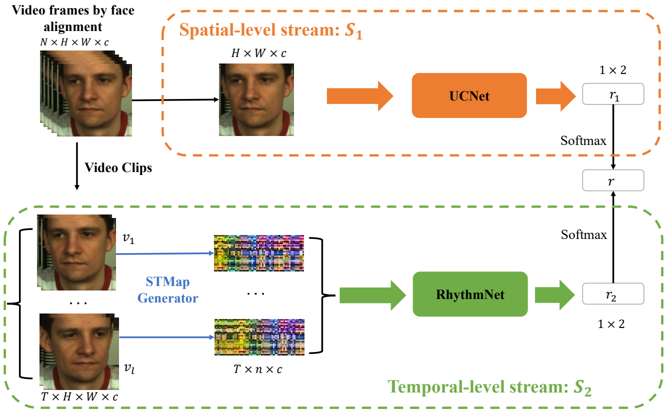

Fig. 4. Overall framework of our proposed method. We first perform face and landmark detection on each frame for face alignment to obtain an input facial video sequence with the size of $N \times H \times W \times c$.

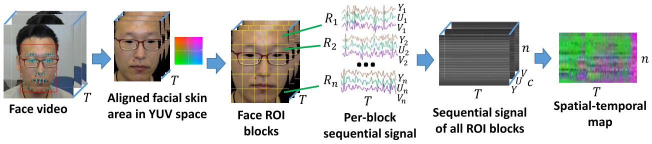

Fig. 5. STmap generator proposed in [9]. Then the facial area is divided into n ROI blocks $R_1, \cdots, R_n$, and the average color value is computed for each color channel within each are concatenated into a sequences, i.e., $Y_1, U_1, V_1, \cdots Y_n, U_n, V_n$. These $n \times c$ sequences are placed into rows to form a spatial-temporal map with the size of $T \times n \times c$.

Specifically, we first divide the input facial video sequence into multiple short video clips $\{v_1, \cdots, v_l\}$. For each video clip with T frames and c color space dimensions, we generate the STmap using the pipeline shown in Figure 5. Eventually, we get a spatial-temporal representation from the video clip with size of $T \times n \times c$ as input to the subsequent network.

We employ RhythmNet [9], that has been proposed to learn the rPPG information in STmaps, as backbone network. RhythmNet is a spatial-temporal framework consisting of a cascade of a ResNet-18 [31] and a one-layer GRU [32]. This latter is a recurrent cell consisting of a reset gate and an update gate, employed to model the temporal relationship between succeeding predictions. The output of the GRU is fed into a fully-connected layer to compute the prediction scores for each individual video clip. For a facial video, the average of all the prediction scores over individual video clips is computed as the final prediction score r_2. The binary cross-entropy loss $\mathcal{L}_{BCE}$ is applied to train the temporal-level stream.

C. Two-Stream Score Fusion

The classification results of the processed videos are acquired from the two streams. The final score r for a video is obtained by combining the output scores of the two streams, $r = \text{softmax}(r_1) + \text{softmax}(r_2)$. Both r_1 and r_2 are vectors containing two coefficients, representing the probability of having an original or a fake facial video. The same information, yet based on both a spatial- and temporal-level analysis, can be retrieved from r, whose values are used to take a final decision on the considered video.

TABLE II
VIDEO-LEVEL SPLIT SETTING USED IN OUR EXPERIMENTS (FAKE:REAL).

Datasets	Training	Testing	Description
PURE	105:105	72:72	First 6 subjects for training, remaining 4 subjects for testing
UBFC-rPPG	90:90	36:36	First 30 subjects for training, remaining subjects for testing
Total	195:195	108:108	Video size: 600× 200× 200

IV. EXPERIMENTS

A. Experimental Settings

1) Datasets: We have employed two popular public rPPG datasets, i.e., PURE [26] and UBFC-rPPG [27], to generate the rPPG deepfake videos in our experiments.

The PURE dataset [26] contains 6 videos from each of 10 subjects. Videos are recorded with subjects in different conditions, namely (1) sitting still, (2) talking, (3) slowly moving the head, (4) quickly moving the head, (5) rotating the head with 20 degree angles, and (6) rotating the head with 35 degree angles. All the videos are captured by an evo274CVGE camera with resolution of 640 × 480 pixels and 30 fps. We split the dataset into two disjoint training and testing sets, respectively containing videos from the first 6 and the remaining 4 subjects.

The UBFC-rPPG dataset [27] contains a single video taken from each of 42 individuals. All the videos are recorded by a Logitech C920 HD Pro with resolution of 640 × 480 pixels in uncompressed 8-bit format, and 30 fps. We split the available data into a training set with data from the first 30 subjects, and a testing set with videos from the remaining 12 ones.

As in [7], the considered videos are cropped to a 200×200 ROI containing the subjects' faces. Furthermore, to increase the data diversity of both databases, from each original video we select the first, middle and last 600 frames, and then we get three video clips, each lasting 20 seconds. We shuffle the corresponding rPPG signals of each of the video clips in the dataset to set the target rPPG signal. Finally, we feed the video clips and the target rPPG into the RErPPG-Net to generate the fake videos. We thus obtain, for our experiments, a total of 606 real and fake videos, as summarized in Table II.

2) Evaluation metrics: In this work, all experimental results are mainly reported in terms of accuracy (ACC), area under curve (AUC) of a receiver-operating-characteristic (ROC) curve, and EER (Equal Error Rate) at video level, for the considered binary classification task on the testing set.

3) Implementation details: We performed all experiments with PyTorch on a workstation equipped with one NVIDIA Tesla V100 GPU (32GB memory). We implemented the RErPPG-Net by their official code[1]. For training our proposed method, we used the Adam optimizer with an initial learning rate of 2×10^{-4} and the weight decay of 5×10^{-4}. The batch size was set to 8.

B. Experimental Results

TABLE III
COMPARISON WITH THE PREVIOUS METHODS

Methods	ACC	AUC	EER
MesoNet [33]	0.5185	0.5571	0.4314
ResNet18 [31]	0.5092	0.5321	0.4713
Xception [34]	0.5741	0.6152	0.4149
3D-ResNet18 [35]	0.6019	0.6721	0.3541
PhysNet-3DCNN [13]	0.7037	0.7846	0.2642
Our method	**0.8892**	**0.9484**	**0.1255**

TABLE IV
PERFORMANCES OF THE ABLATION STUDIES

Methods	ACC	AUC	EER
Our method	**0.8892**	**0.9484**	**0.1255**
Spatial only (UCNet [8])	0.8611	0.9170	0.1503
Temporal only (RhythmNet [9])	0.8333	0.9086	0.1786

The effectiveness of the proposed rPPG deepfake detector is evaluated by comparing our approach against several popular methods designed for fake video detection. Specifically, MesoNet [33], ResNet18 [31], and Xception [34] are 2DCNN-based approaches performing well in traditional deepfake detection task. For them, we average the predictions on all frames as final video prediction. The 3D-ResNet18 [35] is a network commonly used for video analysis, while the PhysNet-3DCNN [13] is frequently used for rPPG signal estimation and recovery. We adapt the PhysNet-3DCNN network for identifying rPPG deepfake videos directly. As shown in Table III, the proposed method consistently achieves the best performance among the considered methods over the employed data.

An ablation study has been also performed in order to evaluate the effectiveness of each proposed stream. As shown in Table IV, spatial-level features are more informative than the temporal-level ones, yet the proposed combined use of both schemes notably improves the achievable performance, indicating that both spatial and temporal information should be taken into account to perform rPPG deepfake detection.

To understand what visual clues our method rely on to detect the rPPG deepfake videos, we visualize in Figure 6 the class activation map (CAM) obtained using Smooth Grad-CAM++ [36]. We can observe that the high activation probability regions are more evenly distributed over the whole face for real images, while for fake images specific regions tend to be highlighted. This is understandable since, as discussed in Section III, rPPG deepefakes cannot guarantee a global consistency of rPPG information on the face.

V. CONCLUSION

In this paper, we have presented a method to detect face forgery videos that manipulate rPPG signals. The proposed approach relies on the analysis of both spatial- and temporal-level characteristics of the considered facial videos when performing the required decisions. Experimental results obtained on samples from two public databases demonstrate that our proposed method can achieve promising performances for the considered detection task.

ACKNOWLEDGMENTS

This work was supported in part by NSFC (Grant U22B2047), Guangdong Basic and Applied Basic Research

[1] https://github.com/nthumplab/RErPPGNet

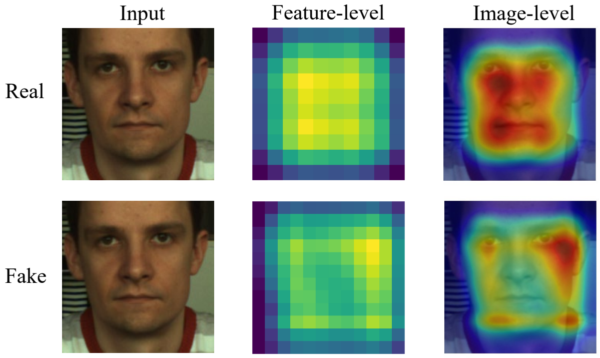

Fig. 6. Visualization of feature and image level class activation maps.

Foundation (Grant 2019B151502001), Shenzhen R&D Program (Grant JCYJ20200109105008228), the China Scholarship Council (file No. 202108440342).

References

[1] E. Maiorana, C. Romano, E. Schena, and C. Massaroni, "Biowish: Biometric recognition using wearable inertial sensors detecting heart activity," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[2] Z. Yu, X. Li, and G. Zhao, "Facial-Video-Based Physiological Signal Measurement: Recent advances and affective applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 50–58, Nov. 2021.

[3] W. Chen and R. W. Picard, "Eliminating Physiological Information from Facial Videos," in *12th IEEE International Conference on Automatic Face & Gesture Recognition*, May 2017, pp. 48–55.

[4] L. Li, C. Chen, L. Pan, Y. Tai, J. Zhang, and Y. Xiang, "Hiding Your Signals: A Security Analysis of PPG-based Biometric Authentication," *arXiv preprint*, Jul. 2022.

[5] M. Chen, X. Liao, and M. Wu, "PulseEdit: Editing Physiological Signals in Facial Videos for Privacy Protection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 457–471, 2022.

[6] Z. Sun and X. Li, "Privacy-Phys: Facial Video-Based Physiological Modification for Privacy Protection," *IEEE Signal Processing Letters*, vol. 29, pp. 1507–1511, 2022.

[7] C.-J. Hsieh, W.-H. Chung, and C.-T. Hsu, "Augmentation of rPPG Benchmark Datasets: Learning to Remove and Embed rPPG Signals via Double Cycle Consistent Learning from Unpaired Facial Videos," in *ECCV 2022*, 2022, pp. 372–387.

[8] K. Wei, W. Luo, S. Tan, and J. Huang, "Universal Deep Network for Steganalysis of Color Image Based on Channel Representation," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3022–3036, 2022.

[9] X. Niu, S. Shan, H. Han, and X. Chen, "RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2020.

[10] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[11] G. de Haan and V. Jeanne, "Robust Pulse Rate From Chrominance-Based rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.

[12] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic Principles of Remote PPG," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2017.

[13] Z. Yu, X. Li, and G. Zhao, "Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks," in *British Machine Vision Conference (BMVC)*, 2019.

[14] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. S. Torr, and G. Zhao, "PhysFormer: Facial Video-Based Physiological Measurement With Temporal Difference Transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4186–4196.

[16] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Cardiac Measurement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5008–5017.

[17] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[18] Y. Lin, H. Chen, B. Li, and J. Wu, "Towards Generalizable DEEPFAKE Face Forgery Detection with Semi-Supervised Learning and Knowledge Distillation," in *2022 IEEE International Conference on Image Processing (ICIP)*, Oct. 2022, pp. 576–580.

[19] H. Chen, Y. Lin, and B. Li, "Exposing Face Forgery Clues via Retinex-based Image Enhancement," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 602–617.

[20] H. Chen, Y. Li, D. Lin, B. Li, and J. Wu, "Watching the BiG artifacts: Exposing DeepFake videos via Bi-granularity artifacts," *Pattern Recognition*, vol. 135, p. 109179, Mar. 2023.

[21] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, *DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-frame*. Cham: Springer International Publishing, 2022, pp. 255–273.

[22] G. Heusch and S. Marcel, "Pulse-based features for face presentation attack detection," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018.

[23] R. Padnevych, D. Semedo, D. Carmo, and J. Magalhães, "Improving face liveness detection robustness with deep convolutional generative adversarial networks," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022.

[24] S.-Q. Liu, X. Lan, and P. C. Yuen, "Learning temporal similarity of remote photoplethysmography for fast 3d mask face presentation attack detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3195–3210, 2022.

[25] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Heartbeat signal from facial video for biometric recognition," in *Image Analysis*, R. R. Paulsen and K. S. Pedersen, Eds., 2015.

[26] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, Aug. 2014, pp. 1056–1062.

[27] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, Jun. 2019.

[28] X. Zhang, R. Wang, D. Yan, L. Dong, and Y. Lin, "Selecting optimal submatrix for syndrome-trellis codes (stcs)-based steganography with segmentation," *IEEE Access*, vol. 8, pp. 61 754–61 766, 2020.

[29] X. Mo, S. Tan, W. Tang, B. Li, and J. Huang, "ReLOAD: Using Reinforcement Learning to Optimize Asymmetric Distortion for Additive Steganography," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1524–1538, 2023.

[30] X. Liu, X. Zhang, G. Narayanswamy, Y. Zhang, Y. Wang, S. Patel, and D. McDuff, "Deep physiological sensing toolbox," *arXiv preprint arXiv:2210.00716*, 2022.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[32] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[33] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2018, pp. 1–7.

[34] F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[35] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[36] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models," *arXiv preprint*, 2019.