PCA Lab's solution for RePSS 2024 by vision-based self-supervised remote heart rate sensing

Hang Shao¹, Lei Luo^{1,*}, Jianjun Qian¹, Wei Zhuo¹, Chuanfei Hu^{2,3} and Jian Yang^{1,*}

¹PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China

²School of Automation, Southeast University, Nanjing, 211189, China
 ³Nanjing Center for Applied Mathematics, Nanjing, 211135, China

Abstract

Vision-based remote non-contact physiological measurements are crucial for detecting indicators (such as heart rate and blood pressure) that reflect important vital signs. This paper introduces the approach proposed by PCA Lab in the third challenge on Vision-based Remote Physiological Signal Sensing (RePSS) organized within IJCAI 2024. Specifically, we design an end-to-end self-supervised contrastive learning network for remote heart rate detection, which can generalize light reflection changes caused by cardiac activity in subcutaneous capillaries based on unlabeled facial videos. At the same time, we optimize the facial skin region of interest extraction method and remove most of the irrelevant content and redundant information. In addition, we construct the hybrid pipeline temporal attention module and spatiotemporal reconstruction pretraining paradigm to improve the network's ability to model long-distance sequence features. Our network is tested on 1000 samples from 200 participants provided by the Track 1 of this challenge, and it achieve the root mean squared error of 8.96941. The codes and model are publicly available at https://github.com/Sachiel0916/repss-track1-top3/.

Keywords

RePSS, remote heart rate sensing, self-supervision framework, spatiotemporal reconstruction, contrastive learning

1. Introduction

Heart rate (HR), blood pressure (BP), and respiratory frequency are important human vital signs [1, 2]. The accurate detection and analysis of these physiological signals is crucial for the assessment of human health, the prevention of cardiovascular diseases, and the identification of emotions [3, 4, 5, 6]. At present, the widely utilized detection technology is skin-invasive measurement, which requires professional equipment and the sensor probe is in direct contact with the measured personal object. Therefore, these manners not only have limitations in terms of convenience and portability of operation, but long-term wearing will also cause the individual subject to feel uncomfortable. They are not suitable for subjects with low cooperation, such as

© 0000-0002-2452-6985 (H. Shao); 0000-0002-9976-0442 (L. Luo); 0000-0002-0968-8556 (J. Qian); 0009-0007-3109-1290 (W. Zhuo); 0000-0003-1669-9429 (C. Hu); 0000-0003-4800-832X (J. Yang)



IJCAI 2024: International Joint Conference on Artificial Intelligence, August 3–9, 2024, Jeju, South Korea *Corresponding author.

 [☆] shaohang@njust.edu.cn (H. Shao); cslluo@njust.edu.cn (L. Luo); csjqian@njust.edu.cn (J. Qian); weizhuo@njust.edu.cn (W. Zhuo); cfhu@seu.edu.cn (C. Hu); csjyang@njust.edu.cn (J. Yang)
 ♦ https://github.com/Sachiel0916/ (H. Shao)

^{© 2024} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



Figure 1: The whole framework of our self-supervised remote physiological signal sensing network.

the elderly, children, and people with limited mobility and movement disorders, as well as those with skin sensitivities such as burns and rashes. For scenarios where practical deployment difficulties, it is not enough to realize the needs for daily physical sign detection [7, 8]. Through computer vision systems and machine learning algorithms, based on facial videos, remote non-contact detection and analysis of HR, BP, and other physical signs are realized, which has very important practical significance and research value [9, 10].

In order to solve the issues of non-contact biosignal detection, researchers have made a lot of efforts [11, 12, 13, 14, 15]. The workshop and challenge on the Vision-based Remote Physiological Signal Sensing (RePSS)¹ held in conjunction with the International Joint Conference on Artificial Intelligence (IJCAI) 2024, aiming to target this emerging topic. This is the 3rd time RePSS has been organized, after the 1st one in CVPR 2020 [16] and the 2nd one in ICCV 2021 [17], which were oriented towards to learn the cardiac activity and respiratory information contained in videos and images. This workshop will discuss the subtle color and movement changes in the face caused by heartbeat, revealing physiological signals such as remote photoplethysmography (rPPG) [18, 19]. In addition, this challenge is divided into two Tracks: self-supervised HR measurement using unlabeled facial videos and remote BP detection to design more powerful computer vision algorithms and biomedical signal processing methods.

The reason why remote HR monitoring has been able to achieve technological leaps in recent years is due to the rapid development of deep learning, convolutional neural network (CNN) [20, 21], and vision transformer (ViT) [22]. Compared with traditional methods based on facial colorimetric analysis and blind source separation [23, 24], learning-based approaches are driven by large-scale data and over complex scenes and dynamic illuminance to extract and condense color rhythms from facial videos [25]. However, these approaches usually require appropriate labeling. Once the labels are incomplete or incorrect, the learned features will be seriously dissolved. In addition, the above process not only requires a lot of manpower and computing power costs, but also because video and HR tags are obtained from different devices (cameras and electronic sensors), the temporal domain registration issue between multiple devices and the various types of inevitable interference and noise during the tag acquisition process have

¹https://repss-w.github.io/

R Node	R Node	R Node	R Node	R Node	R Node	R Node	R	Node
0 M.(L0:27)	10 M.(L2:48)	20 M.(L4:43)	30 M.(L9:61)	40 M.(L12:37)	50 M.(L14:41)	60 M.(L20:23)	70	M.(M.(L1:40):40)
1 M.(L0:43)	11 M.(L3:40)	21 M.(L4:48)	31 M.(L10:38)	41 M.(L12:38)	51 M.(L14:47)	61 M.(L21:22)	71	M.(M.(L2:41):2)
2 M.(L0:48)	12 M.(L3:41)	22 M.(L5:31)	32 M.(L10:54)	42 M.(L12:47)	52 M.(L14:54)	62 M.(L40:41)	72	M.(M.(L2:41):41)
3 M.(L1:40)	13 M.(L3:43)	23 M.(L5:32)	33 M.(L10:60)	43 M.(L12:54)	53 M.(L15:40)	63 M.(L42:41)	73	M.(M.(L3:43):3)
4 M.(L1:41)	14 M.(L3:48)	24 M.(L5:48)	34 M.(L11:37)	44 M.(L13:40)	54 M.(L15:41)	64 M.(L43:49)	74	M.(M.(L13:47):13)
5 M.(L1:43)	15 M.(L4:17)	25 M.(L6:31)	35 M.(L11:38)	45 M.(L13:41)	55 M.(L15:47)	65 M.(L44:50)	75	M.(M.(L14:41):14)
6 M.(L2:17)	16 M.(L4:27)	26 M.(L6:48)	36 M.(L11:54)	46 M.(L13:47)	56 M.(L16:36)	66 M.(L45:51)	76	M.(M.(L14:41):41)
7 M.(L2:40)	17 M.(L4:30)	27 M.(L6:64)	37 M.(L12:26)	47 M.(L13:54)	57 M.(L16:47)	67 M.(L46:52)	77	M.(M.(L15:40):15)
8 M.(L2:41)	18 M.(L4:31)	28 M.(L7:63)	38 M.(L12:33)	48 M.(L14:26)	58 M.(L16:54)	68 M.(L47:53)	78	M.(M.(L15:40):40)
9 M.(L2:43)	19 M.(L4:32)	29 M.(L8:62)	39 M.(L12:36)	49 M.(L14:40)	59 M.(L19:24)	69 M.(M.(L1:40):1)	79-12	7 Global face

 Table 1

 Landmark configuration information for facial spatiotemporal mapping construction.

also been plaguing researchers to further promote and deepen applications.

To solve the above problems, our team at the PCA Lab proposes a novel remote HR measurement network based on self-supervised contrastive learning in this paper to realize the mining of faint color changes with the blood volume pulse feedback in unlabeled facial videos. Specifically, first, to overcome the redundant skin information, we design a new facial region of interest extraction method that focuses on areas rich in facial muscles and capillaries, while ignoring the interference of explicit edges, corners, and texture changes. Second, we convert the input video segments into spatiotemporal mappings to guide contrastive learning between and within instances through the vast enrichment of positive and negative sample pairs in the pretraining stage. Third, we improve the traditional rPPG waveform regression into spatiotemporal reconstruction to further improve the robustness of our model by focusing on the interaction of temporal features between different sub-regions of the face during the fine-tuning stage.

We pretrain the proposed model on the VFHQ [26], CelebV-HQ [27], and MAHNOB-HCI [28] datasets using facial videos without any biosignal annotation, and fine-tune it on the VIPL-HR-V2 dataset [16]. We test our model on 1000 video clips containing OBF [29] and VIPL-HR-V2 datasets provided by the organizer, with an average root mean squared error (RMSE) of 8.96941, and achieve the top-3 result on the Track 1 of the 3rd RePSS challenge.

2. Methodology

The overall architecture of our self-supervised remote HR sensing network is shown in Figure 1. In this section, we will introduce in detail three aspects: facial spatiotemporal mapping construction (Sec. 2.1), pretraining stage based on contrastive learning (Sec. 2.2), and fine-tuning stage based on spatiotemporal reconstruction (Sec. 2.3).

2.1. Facial spatiotemporal mapping construction

Early deep learning methods [30, 31, 32] for remote HR detection often directly utilized 3D CNN to calculate facial video clips. However, such operations ignore the interactions between long-range rhythms [33]. Inspired by RhythmNet [34] and NEST [35], we improve our input pattern into the spatiotemporal mapping [36]. Furthermore, we find that simply calculating facial frames or thresholded skin areas not only leads to redundant information, but also interference from corner semantics invades the learning path. Inspired by THR-Net (HR-GCN) [37] and RADIANT



Figure 2: The illustration of our facial spatiotemporal mapping, which can feedback more rhythmic information at the beginning of preprocessing than traditional augmentation manners.

[38], we divide the face into multiple sub-regions. The difference is that THR-Net only employs four natural light regions and RADIANT algorithm adopts non-overlapping patches, while we design 79 mutually entangled skin sub-region blocks to exploit the capillary-rich parts [39] of the face as much as possible.

The specific selection of 79 skin sub-regions (R) is based on the 68 landmarks at positions (0-59, 61-65, and 68-70, respectively) in the Python face recognition framework². To impress, we list the operable configurations in Table 1, where "M." represents the midpoint of two landmarks (L). Each sub-region is a rectangular block surrounded by 1/10 of the long side of the entire facial corresponding to the point (Node). Moreover, in order not to lose too much spatially discriminative pixels, we construct additional 49 global face downsampling blocks (that is, the resolution is 7×7). We average each block within the frame, which can be converted into 128 RGB three-channel feature values for a single facial frame.

Afterwards, we convert a video segment into a spatiotemporal mapping based on the length of the required input frames. We perform the temporal normalization independently on each sub-patch feature dimension, and then perform YUV color space conversion which has been proven to have the ability to enhance rPPG imaging beyond mappings such as YCrCb, HSV, and raw RGB. Figure 2 is a visual comparison of our facial spatiotemporal construction approach and the traditional method. It can be seen that our algorithm can harvest a certain degree of rhythm and color change characteristics without relying on the learning model in this preprocessing and augmentation stage.

2.2. Contrastive learning-based pretraining

The design idea based on self-supervision is an important means to solve the lack of data and unstable labeling in learning-based remote physiological estimation tasks [40, 41]. Currently, a widely used self-supervised strategy is contrastive learning [42], which builds different pairs

²https://github.com/ageitgey/face_recognition/

of positive and negative samples to pull in the representations between samples with the same attributes and separate them from those that are different. Based on the remarkable achievements of contrastive learning in the rPPG task [43, 44], we design a novel self-supervised HR sensing network and training paradigm.

Following Contrast-Phys+ [45], instead of constructing pairs of positive and negative samples within the same object or between shifting sequences [46, 47], we build them between different individual instances [48]. This is because pulse waveforms vary from subject to subject. Specifically, we randomly shuffle the arrangement of feature points in the spatiotemporal mapping to increase topological diversity. After that, we randomly select one from the original mapping and the *m* newly generated mappings as a positive sample \mathbf{x}^{pos} , and perform partial block extraction and size adjustment again for the other mappings as supplementary samples $\mathbf{x}_m^{\text{sup}}$. Since the supplementary set and the positive sample differ only in space and not in time and frequency domains, we combine the above as a positive set $\mathbf{X}_{\text{pos}} = {\mathbf{x}^{\text{pos}}, \mathbf{x}_m^{\text{sup}}}$. Meanwhile, we randomly select the *n* instances in the training set, perform random feature space shuffling and reselection as above, and let them enter a negative set $\mathbf{X}_{\text{neg}} = {\mathbf{x}_n^{\text{neg}}}$.

Our pretrained module employs PhysNet [49] as the backbone, and embeds attention operations for long-distance temporal perception and interaction (as shown in Figure 1). The temporal attention module is not limited to the hybrid scale of the input, but squeezes its features into a one-dimensional vector related only to time, calculates the attention score and then expands it to regress the backbone path. For the input mapping from the positive and negative sample sets, the corresponding output is a one-dimensional vector related to time respectively. At this stage, to guide the training, we calculate the power spectral densities and HR values between different outputs respectively, where the frequency contrast loss \mathcal{L}_{cf} is expressed as:

$$\mathcal{L}_{\rm cf} = \log\left(\frac{\sum_{i_1=1}^{m} \sum_{\substack{i_2=1\\i_1 \neq i_2}}^{m} \exp\left(M(\mathbf{y}_{i_1}^{\rm pos}, \mathbf{y}_{i_2}^{\rm pos})/\tau\right)}{\sum_{i=1}^{m} \sum_{j=1}^{n} \exp\left(M(\mathbf{y}_{i}^{\rm pos}, \mathbf{y}_{j}^{\rm neg})/\tau\right)} + 1\right)$$
(1)

where \mathbf{y}^{pos} and \mathbf{y}^{neg} correspond to the output of samples within positive and negative sets \mathbf{X}_{pos} and \mathbf{X}_{neg} respectively, τ is the temperature hyperparameter (we set it to 0.08 in this paper), and M is the frequency difference between the two output vectors. Correspondingly, the waveform contrast loss function \mathcal{L}_{cp} is:

$$\mathcal{L}_{\rm cp} = \log\left(\frac{\sum_{i_1=1}^{m} \sum_{\substack{i_2=1\\i_1 \neq i_2}}^{m} \exp\left(D(\mathbf{y}_{i_1}^{\rm pos}, \mathbf{y}_{i_2}^{\rm pos})/\tau\right)}{\sum_{i=1}^{m} \sum_{j=1}^{n} \exp\left(D(\mathbf{y}_{i}^{\rm pos}, \mathbf{y}_{j}^{\rm neg})/\tau\right)} + 1\right)$$
(2)

where D represents the correlation of the two pulse waveforms (taking $\mathbf{y}_1^{\text{pos}}$ and $\mathbf{y}_2^{\text{pos}}$ as the example):

$$D(\mathbf{y}_{1}^{\text{pos}}, \mathbf{y}_{2}^{\text{pos}}) = 1 - \frac{T \sum_{i=1}^{T} \mathbf{y}_{1_{i}}^{\text{pos}} \mathbf{y}_{2_{i}}^{\text{pos}} - \sum_{i=1}^{T} \mathbf{y}_{1_{i}}^{\text{pos}} \sum_{i=1}^{T} \mathbf{y}_{2_{i}}^{\text{pos}}}{\sqrt{\left(T \sum_{i=1}^{T} (\mathbf{y}_{1_{i}}^{\text{pos}})^{2} - (\sum_{i=1}^{T} \mathbf{y}_{1_{i}}^{\text{pos}})^{2}\right) \left(T \sum_{i=1}^{T} (\mathbf{y}_{2_{i}}^{\text{pos}})^{2} - (\sum_{i=1}^{T} \mathbf{y}_{2_{i}}^{\text{pos}})^{2}\right)}}$$
(3)

where T represents the temporal dimension of our facial spatiotemporal mapping. Meanwhile, the pretraining loss \mathcal{L}_{P} is expressed as:

$$\mathcal{L}_{\mathrm{P}} = \mathcal{L}_{\mathrm{cf}} + \mathcal{L}_{\mathrm{cp}}$$
 (4)

2.3. Spatiotemporal reconstruction-based fine-tuning

As a complement to large-scale self-supervised pretraining, we design a supervised pipeline for model fine-tuning in addition to the contrastive path mentioned above. Since the amount of data in fine-tuning process is relatively small, to fully explore the spatiotemporal interaction between different feature points, we improve the traditional signal regression into spatiotemporal reconstruction supervision. Specifically, we construct a U-shaped structure [50] consisting of encoder, latent module, and decoder. The encoder takes the downsampling part of PhysNet as the backbone. The difference from the pretrained module is that the fine-tuned encoder converts the input mapping of scale ($3 \times T \times 128$) into a tensor of ($512 \times T/4 \times 8$), where the three dimensions correspond to channel, time, and intra-frame feature scale respectively.

After that, there is a latent module that outputs pulse waveforms, HR values, and features that will be fed to the decoder. For the decoder, it is the opposite of the encoder counterpart. Meanwhile, we construct the grayscale ground truth mapping at the same scale as the input facial spatiotemporal mapping of the ground truth pulse label. We use L1 loss as the constraints for spatiotemporal mapping and HR ($\mathcal{L}_{\rm fm}$ and $\mathcal{L}_{\rm fh}$). Since we randomly disrupt the spatial relationship of the original input feature points and reselect the regions, each dimension is considered an independent constraint process, and the same constraints can also be shared between different dimensions. This greatly improves the robustness of our model. Moreover, we embed a serial of temporal attention modules at the skip layer connections of the U structure. They use a multi-head attention method (we set it to 8), which calculates the global self-attention score within the features of the encoder and concatenates it with the main path features to the decoder. Finally, the loss on predicted waveforms is the negative Pearson's correlation $\mathcal{L}_{\rm fp}$:

$$\mathcal{L}_{\rm fp} = 1 - \frac{T \sum_{t=1}^{T} p(t)g(t) - \sum_{t=1}^{T} p(t) \sum_{t=1}^{T} g(t)}{\sqrt{\left(T \sum_{t=1}^{T} \left(p(t)\right)^2 - \left(\sum_{t=1}^{T} p(t)\right)^2\right) \left(T \sum_{t=1}^{T} \left(g(t)\right)^2 - \left(\sum_{t=1}^{T} g(t)\right)^2\right)}}$$
(5)

where p(t) is the predicted pulse on the learning pipeline, and g(t) is the ground truth label for the fine-tuning phase. The fine-tuning loss $\mathcal{L}_{\rm F}$ is expressed as:

$$\mathcal{L}_{\rm F} = 0.5 \times (\mathcal{L}_{\rm cf} + \mathcal{L}_{\rm cp}) + \mathcal{L}_{\rm fm} + \mathcal{L}_{\rm fh} + \mathcal{L}_{\rm fp}$$
(6)

3. Experiments

3.1. Datasets

We pretrain our network on three publicly available datasets: VFHQ [26], CelebV-HQ [27], and MAHNOB-HCI [28]. Regarding them, VFHQ dataset contains over 16000 high-fidelity

	-		
Rank	Team	Affiliation	RMSE (bpm)
1	Face Al	Agency for Science, Technology and Research of Singapore	8.50693
2	HFUT-VUT	Hefei University of Technology	8.85277
3	PCA_Vital (Ours)	Nanjing University of Science and Technology	8.96941
4	Hash Brown	Beijing University of Posts and Telecommunications	9.26198
5	AIIA	Harbin Institute of Technology	9.28902
6	SHDMIC	Ruijin Hospital of Shanghai Jiao Tong University	10.74201
7	HFUT-BCDH	Hefei University of Technology	11.77657
8	NeuroAI_KW	Kwangwoon University	14.47930
9	NUIST	Nanjing University of Information Science & Technology	15.79680
10	SCUT_rPPG	South China University of Technology	15.88228

Table 2The public leaderboard for this challenge.

facial clips of different interview scenes, CelebV-HQ has 35666 videos, and MAHNOB-HCI dataset consists of 527 videos of 30 objects. None of the three facial continuous frame datasets involve rPPG labels for fine-tuning and supervised HR learning. We fine-tune our model on the VIPL-HR-V2 [16] dataset, which contains 2000 videos of 400 objects.

3.2. Implement and evaluation metric

Our model is deployed based on the PyTorch framework and runs on a device equipped with four GeForce RTX 4090 GPUs. The number of frames of the input video segment is 300 (that is, the scale of the spatiotemporal mapping is (300×128)), the batch size is 20, the pretraining round is 200, and the fine-tuning epoch is 50. Our model is optimized using AdaMax, the initially learning rate is 1×10^{-5} and will drop to 0.5×10^{-5} at the 50th epoch. We use RMSE as the evaluation metric to calculate the gap of beats per minute (bpm) between the ground truth HR_{gt} and the average predicted HR_{pred} for each video segment (total number is N), which is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (HR_{gt_i} - HR_{pred_i})^2}{N}}$$
(7)

3.3. Results

Our team (the name is PCA_Vital) won the third place on the 3rd Vision-based RePSS challenge Track 1. We list the RMSE results published by the organizer³ of the top ten teams as shown in Table 2, it can be seen that our score is 8.96941. Furthermore, we are just behind second place 0.11664 and significantly ahead of fourth place 0.29257.

In addition, we embed the contrastive loss in the fine-tuning stage. Regarding the contrastive loss coefficient in the overall loss function (Equ. 6), we set it as ξ and conduct an ablation study with different hyperparameters to verify our settings. The RMSE results are shown in Table 3. It can be seen that introducing contrastive loss in the overall loss can improve the model performance to a certain extent. We analyze that this is due to the limited data scale of

³https://repss-w.github.io/Challenge.html

Table 3Ablation study on loss function hyperparameters.

Param.	RMSE (bpm)		
ξ=2.0	9.99876		
<i>ξ</i> =1.5	9.79554		
<i>ξ</i> =1.0	9.10092		
ξ=0.5	8.96941		
<i>ξ</i> =0.2	8.99006		

the fine-tuning dataset, adding contrastive discrimination on its basis can effectively prevent overfitting and increase the network's prediction ability for unseen samples.

4. Conclusion

Facial vision-based remote HR measurement has been booming over the past decade, but its deployment is hampered by the lack of labels and incomplete learning paradigms. This paper proposes a novel self-supervised approach to solve them. It uses facial spatiotemporal reconstruction and contrast learning to mine commonalities of subtle changes in facial skin color with heartbeats between different frames, thereby improving robustness and achieving excellent results in the 3rd RePSS challenge. Although this challenge has ended, our research will continue on the optimization of related rPPG learning processes and strategies.

Acknowledgments

Thanks to Zhejiang University, University of Oulu, Institute of Computing Technology of CAS, and INRIA of France for organizing the 3rd RePSS.

This work was supported by the National Natural Science Foundation of China under Grant 62176124, Grant 62276135, and Grant 62361166670.

References

- [1] S. Guler, A. Golparvar, O. Ozturk, H. Dogan, M. K. Yapici, Optimal digital filter selection for remote photoplethysmography (rppg) signal conditioning, Biomedical Physics and Engineering Express 9 (2023) 027001.
- [2] W. Othman, A. Kashevnik, A. Ali, N. Shilov, D. Ryumin, Remote heart rate estimation based on transformer with multi-skip connection decoder: Method and evaluation in the wild, Sensors 24 (2024) 775.
- [3] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, ... J. Zhao, Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms, in: Proceedings of the ACM International Conference on Multimedia (ACM MM), Seattle, USA, 2020, pp. 4318–4327.
- [4] Y. C. Wu, L. W. Chiu, C. C. Lai, B. F. Wu, S. S. Lin, Recognizing, fast and slow: Complex emotion recognition with facial expression detection and remote physiological measurement, IEEE Transactions on Affective Computing 14 (2023) 3177–3190.

- [5] C. Á. Casado, M. L. Cañellas, M. B. López, Depression recognition using remote photoplethysmography from facial videos, IEEE Transactions on Affective Computing 14 (2023) 3305–3316.
- [6] Y. Ru, P. Li, M. Sun, Y. Wang, K. Zhang, Q. Li, ... Z. Sun, Sensing micro-motion human patterns using multimodal mmradar and video signal for affective and psychological intelligence, in: Proceedings of the ACM International Conference on Multimedia (ACM MM), Ottawa, Canada, 2023, pp. 5935–5946.
- [7] M. Besirli, K. Ture, M. Beghetti, F. Maloberti, C. Dehollain, M. Mattavelli, D. Barrettino, An implantable wireless system for remote hemodynamic monitoring of heart failure patients, IEEE Transactions on Biomedical Circuits and Systems 17 (2023) 688–700.
- [8] C. Pham, K. Poorzargar, D. Panesar, K. Lee, J. Wong, M. Parotto, F. Chung, Video plethysmography for contactless blood pressure and heart rate measurement in perioperative care, Journal of Clinical Monitoring and Computing 38 (2024) 121–130.
- [9] Z. Sun, A. Vedernikov, V. L. Kykyri, M. Pohjola, M. Nokia, X. Li, Estimating stress in online meetings by remote physiological signal and behavioral features, in: Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and the ACM International Symposium on Wearable Computers (UbiComp/ISWC), New York, USA, 2022, pp. 216–220.
- [10] A. Helwan, D. Azar, M. K. S. Ma'aitah, Conventional and deep learning methods in heart rate estimation from rgb face videos, Physiological Measurement 45 (2024) 02TR01.
- [11] M. Lewandowska, J. Rumiński, T. Kocejko, J. Nowak, Measuring pulse rate with a webcam - A non-contact method for evaluating cardiac activity, in: Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), Szczecin, Poland, 2011, pp. 405–410.
- [12] M. Artemyev, M. Churikova, M. Grinenko, O. Perepelkina, Neurodata lab's approach to the challenge on computer vision for physiological measurement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, USA, 2020, pp. 316–317.
- [13] Y. Dong, G. Yang, Y. Yin, Time lab's approach to the challenge on computer vision for remote physiological measurement, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, Canada, 2021, pp. 2398–2403.
- [14] X. Liu, X. Yang, Z. Meng, Y. Wang, J. Zhang, A. Wong, Manet: A motion-driven attention network for detecting the pulse from a facial video with drastic motions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, Canada, 2021, pp. 2385–2390.
- [15] C. Hu, K. Y. Zhang, T. Yao, S. Ding, J. Li, F. Huang, L. Ma, An end-to-end efficient framework for remote physiological signal sensing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, Canada, 2021, pp. 2378– 2384.
- [16] X. Li, H. Han, H. Lu, X. Niu, Z. Yu, A. Dantcheva, ... S. Shan, The 1st challenge on remote physiological signal sensing (repss), in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, USA, 2020, pp. 314–315.
- [17] X. Li, H. Sun, Z. Sun, H. Han, A. Dantcheva, S. Shan, G. Zhao, The 2nd challenge on

remote physiological signal sensing (repss), in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, Canada, 2021, pp. 2404–2413.

- [18] S. Liu, X. Lan, P. Yuen, Temporal similarity analysis of remote photoplethysmography for fast 3d mask face presentation attack detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass, USA, 2020, pp. 2608–2616.
- [19] L. Kong, K. Xie, K. Niu, J. He, W. Zhang, Remote photoplethysmography and motion tracking convolutional neural network with bidirectional long short-term memory: Noninvasive fatigue detection method based on multi-modal fusion, Sensors 24 (2024) 455.
- [20] X. Liu, Z. Sun, X. Li, R. Song, X. Yang, Vidbp: Detecting blood pressure from facial videos with personalized calibration, in: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Sydney, Australia, 2023, pp. 1–5.
- [21] L. Jiao, M. Wang, X. Liu, L. Li, F. Liu, Z. Feng, ... B. Hou, Multiscale deep learning for detection and recognition: A comprehensive survey, IEEE Transactions on Neural Networks and Learning Systems (2024).
- [22] H. Shao, L. Luo, J. Qian, S. Chen, C. Hu, J. Yang, Tranphys: Spatiotemporal masked transformer steered remote photoplethysmography estimation, IEEE Transactions on Circuits and Systems for Video Technology 34 (2024) 3030–3042.
- [23] Z. Zhang, C. H. Fu, L. Zhang, H. Hong, Near infrared video heart rate detection based on multi-region selection and robust principal component analysis, in: Proceedings of the International Conference on Image and Graphics (ICIG), Nanjing, China, 2023, pp. 37–47.
- [24] K. Balaraman, A. Claret, Cardiac signal monitoring system using noncontact measurement for physiological features, in: Proceedings of the International Conference on Information Management and Machine Intelligence (ICIMMI), Jaipur, India, 2023, pp. 1–6.
- [25] H. Shao, L. Luo, S. Chen, C. Hu, J. Yang, Hyperbolic embedding steered spatiotemporal graph convolutional network for video-based remote heart rate estimation, Engineering Applications of Artificial Intelligence 124 (2023) 106642.
- [26] L. Xie, X. Wang, H. Zhang, C. Dong, Y. Shan, Vfhq: A high-quality dataset and benchmark for video face super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, USA, 2022, pp. 657–666.
- [27] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, ... C. C. Loy, Celebv-hq: A large-scale video facial attributes dataset, in: Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 2022, pp. 650–667.
- [28] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Transactions on Affective Computing 3 (2012) 42–55.
- [29] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, ... G. Zhao, The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG), Xi'an, China, 2018, pp. 242–249.
- [30] F. Bousefsaf, A. Pruski, C. Maaoui, 3D convolutional neural networks for remote pulse rate measurement and mapping from facial video, Applied Sciences 9 (2019) 4364.
- [31] J. Speth, N. Vance, P. Flynn, K. Bowyer, A. Czajka, Remote pulse estimation in the presence of face masks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR) Workshops, New Orleans, USA, 2022, pp. 2086–2095.

- [32] R. Karthick, M. S. Dawood, P. Meenalochini, Analysis of vital signs using remote photoplethysmography (rppg), Journal of Ambient Intelligence and Humanized Computing 14 (2023) 16729–16736.
- [33] D. Q. Le, J. C. Chiang, W. N. Lie, Remote ppg estimation from rgb-nir facial image sequence for heart rate estimation, in: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Austin, USA, 2022, pp. 2077–2081.
- [34] X. Niu, S. Shan, H. Han, X. Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, IEEE Transactions on Image Processing 29 (2020) 2409–2423.
- [35] H. Lu, Z. Yu, X. Niu, Y. C. Chen, Neuron structure modeling for generalizable remote physiological measurement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 2023, pp. 18589–18599.
- [36] J. Du, S. Q. Liu, B. Zhang, P. C. Yuen, Dual-bridging with adversarial noise generation for domain adaptive rppg estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 2023, pp. 10355– 10364.
- [37] Z. Yue, S. Ding, S. Yang, L. Wang, Y. Li, Multimodal information fusion approach for noncontact heart rate estimation using facial videos and graph convolutional network, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–13.
- [38] A. K. Gupta, R. Kumar, L. Birla, P. Gupta, Radiant: Better rppg estimation using signal embeddings and transformer, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, USA, 2023, pp. 4976–4986.
- [39] A. Ni, A. Azarang, N. Kehtarnavaz, A review of deep learning-based contactless heart rate measurement methods, Sensors 21 (2021) 3719.
- [40] N. Zhang, H. M. Sun, J. R. Ma, R. S. Jia, A self-supervised learning network for remote heart rate measurement, Measurement 228 (2024) 114379.
- [41] D. Gupta, A. Etemad, Remote heart rate monitoring in smart environments from videos with self-supervised pretraining, IEEE Internet of Things Journal 11 (2024) 10279–10294.
- [42] T. Qiao, S. Xie, Y. Chen, F. Retraint, X. Luo, Fully unsupervised deepfake video detection via enhanced contrastive learning, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
- [43] M. Cao, X. Cheng, X. Liu, Y. Jiang, H. Yu, J. Shi, St-phys: Unsupervised spatio-temporal contrastive remote physiological measurement, IEEE Journal of Biomedical and Health Informatics (2024).
- [44] J. Peng, W. Su, H. Chen, J. Sun, Z. Tian, Cl-spo2net: Contrastive learning spatiotemporal attention network for non-contact video-based spo2 estimation, Bioengineering 11 (2024) 113.
- [45] Z. Sun, X. Li, Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
- [46] L. Birla, S. Shukla, A. K. Gupta, P. Gupta, Alpine: Improving remote heart rate estimation using contrastive learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, USA, 2023, pp. 5029–5038.

- [47] Z. Yue, M. Shi, S. Ding, Facial video-based remote physiological measurement via selfsupervised learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 13844–13859.
- [48] H. Wang, E. Ahn, J. Kim, Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Palo Alto, USA, 2022, pp. 2431–2439.
- [49] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, in: Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 2019, pp. 1–12.
- [50] A. M. Shaker, M. Maaz, H. Rasheed, S. Khan, M. H. Yang, F. S. Khan, Unetr++: Delving into efficient and accurate 3d medical image segmentation, IEEE Transactions on Medical Imaging (2024).