A two-stage remote blood pressure estimation method based on selective state space model

Zizheng Guo, Bochao Zou* and Huimin Ma

University of Science and Technology Beijing, Beijing, China

Abstract

Blood pressure (BP) estimation based on remote photoplethysmography is an innovative and challenging task that enables non-contact BP monitoring through facial videos using ordinary cameras, facilitating long-term health monitoring. Most existing methods require first extracting blood volume pulse waves from facial videos, followed by estimating BP from the pulse wave characteristics. This process means that the quality of pulse wave extraction significantly constrains BP estimation. However, these methods generally overlook long-range context modeling, resulting in suboptimal precision and granularity in pulse wave extraction. This paper introduces the solution we proposed for the third Remote Physiological Signal Sensing (RePSS) challenge in IJCAI'24. Our approach is an end-to-end method comprising two stages: blood volume pulse estimation and blood pressure estimation. Using the selective state space model allows for capturing long-range dependencies while preserving linear complexity. We conducted intra-dataset experiments, cross-dataset experiments, and ablation studies for evaluation. The proposed method achieved 13.59 mmHg RMSE and ranked in third place in the challenge.

Keywords

Blood Pressure Estimation, Remote Photoplethysmography, Video Understanding, State Space Model

1. Introduction

Blood Volume Pulse (BVP) is a crucial physiological signal from which vital signs such as blood pressure (BP), heart rate (HR), and heart rate variability (HRV) can be derived. The principle behind extracting BVP using photoplethysmography (PPG) is based on the changes in light absorption and scattering caused by variations in blood volume in the subcutaneous vessels during cardiac cycles. This results in minute periodic changes in color signals, invisible to the naked eye, which can be captured by imaging sensors. Traditionally, BVP is obtained using contact-based sensors, which can lead to various inconveniences and limitations. Recently, non-contact methods for extracting BVP, such as remote photoplethysmography (rPPG), have gained increasing attention [1, 2, 3].

Among these physiological signals, estimating blood pressure is particularly challenging, with no current theoretical framework adequately explaining the underlying mechanism for BP estimation from BVP wave characteristics. Most existing rPPG-based BP estimation methods have evolved from PPG-based techniques and primarily include two types of approaches: the first measures BP using pulse waves and their features obtained from facial videos [4, 5, 6, 7]; the

The 3rd Vision-based Remote Physiological Signal Sensing (RePSS) Challenge & Workshop, August 3–9, 2024, Jeju, South Korea (IJCAI'24)

*Corresponding author.

guozizheng@xs.ustb.edu.cn (Z. Guo); zoubochao@ustb.edu.cn (B. Zou); mhmpub@ustb.edu.cn (H. Ma)

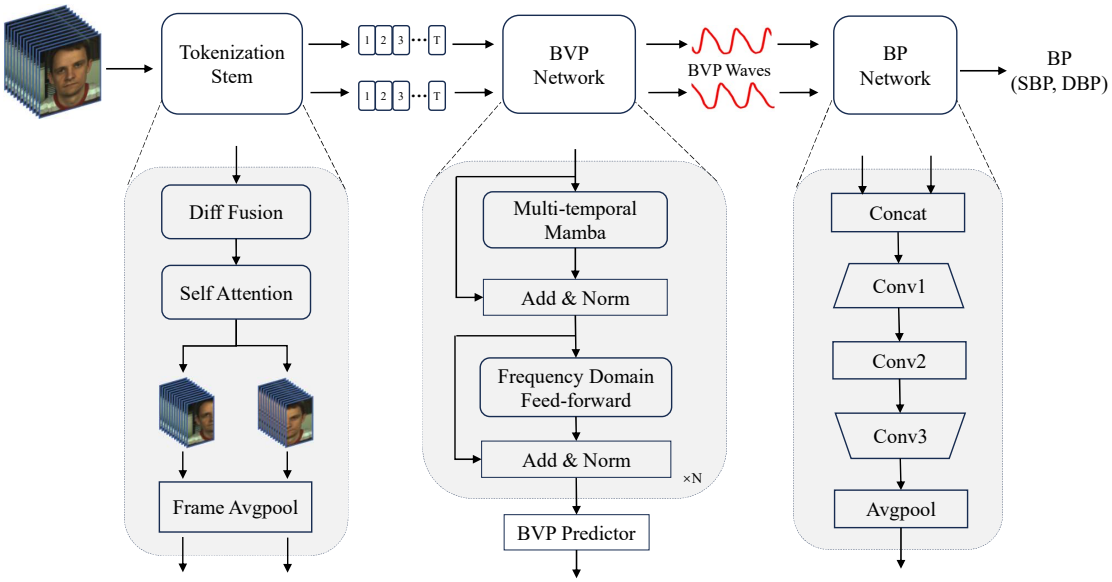

Figure 1: The overall framework of the proposed method includes Tokenization Stem, BVP Network, and BP Network. The process begins with video input, from which the Tokenization Stem extracts temporal token sequences from both the left and right facial regions. Subsequently, the BVP Network reconstructs BVP waveforms separately from the two temporal token sequences. Finally, the BP Network estimates the blood pressure values based on the two reconstructed BVP waveforms.

second calculates BP using pulse transit time (PTT) derived from BVP waves extracted between two body parts [8, 9, 10]. These methods require the initial estimation of BVP waves from videos, and the quality of pulse wave estimation significantly affects the accuracy of BP estimation. However, the existing methods are predominantly based on traditional signal processing or convolutional neural networks. These approaches overlook long-range context modeling, resulting in suboptimal precision and granularity in BVP wave extraction [11]. Recently, Mamba has emerged with the selective state space model, striking a balance between maintaining linear complexity and facilitating long-term dynamic modeling [12, 13, 14]. This provides a novel option for rPPG, which is suitable for practical deployment.

In this paper, we further explore remote BP estimation methods and propose a two-stage BP estimation approach based on the selective state space model, comprising BVP estimation and BP estimation stages. By imposing constraints on BVP estimation across multiple temporal scales in both temporal and frequency domains, we aim to refine the granularity of BVP estimation and improve the accuracy of BP estimation. We evaluate the precision of BVP recovery by assessing the accuracy of heart rate calculated from the BVP waves and conduct intra-dataset experiments, cross-dataset experiments, and ablation studies to validate the proposed method.

2. Methodology

2.1. The overall framework

Our proposed method is an end-to-end framework that takes video as input to predict blood pressure values as output. Directly predicting blood pressure from facial video may not yield optimal results [see Section 3.5]. Therefore, we divide the blood pressure estimation process into two stages within the model: 1) estimating the corresponding BVP waves from the left and right halves of the face, and 2) estimating the BP values from these two BVP waves.

As depicted in Figure 1, the overall framework of the proposed method mainly consists of three parts: Tokenization Stem, BVP Network, and BP Network. For the input, spatial information from the left and right faces is initially extracted into token channels separately via the Tokenization stem, forming two temporal token sequences. Specifically, given an RGB video input $X \in \mathbb{R}^{3\times T\times H\times W}$, the Tokenization Stem output $X_{token_1}, X_{token_2} \in \mathbb{R}^{C\times T/2}$, and C, T, W, H indicate channel, sequence length, width, and height, respectively.

The two token sequences obtained from the Tokenization Stem are each processed by the BVP Network. The BVP Network is based on RhythmMamba [15], employing multi-temporal Mamba and frequency domain feed-forward. It constrains a state space model at multiple temporal scales in both temporal and frequency domains. The input token sequences will be partitioned into sequences of different temporal lengths. These sequences will undergo extraction of implicit information using selective state-space models, and then they will be added according to temporal correspondence. Subsequently, multi-scale information within multi-channel frequency domains will be interacted in the frequency domain. The dimensions of the feed-forward output and the Mamba output are identical to the output of the stem $X_{token} \in \mathbb{R}^{C\times T/2}$. Finally, the rPPG features are upsampled and projected into BVP waves through the BVP predictor.

The BP Network estimates blood pressure based on the two BVP waves obtained from the left and right faces by the BVP Network. The BVP waves predicted from the left and right faces are first concatenated along the channel dimension and then processed through three convolution layers. The first two convolution layers are followed by batch normalization and ELU activation layers. Finally, the BP values are obtained through average pooling.

Tokenization Stem. Tokenization Stem comprises three components: diff-fusion, self-attention, and frame average pooling. The diff-fusion module is used to incorporate inter-frame differences into the raw frames, allowing the raw frames to be aware of BVP wave variations, thereby enhancing the rPPG representation. Specifically, given the input video X_t, temporal shift is initially applied to obtain X_{t-2}, X_{t-1}, X_t, X_{t+1} and X_{t+2}. Subsequently, frame differences between consecutive frames are computed in reverse chronological order, yielding D_{-2}, D_{-1}, D_1 and D_2. The frame differences and raw frames are processed through $Stem_1$ to reduce resolution and perform primary feature extraction.

$$\begin{aligned} X_{diff} &= Stem_1(Concat\,(D_{-2}, D_{-1}, D_1, D_2)), \\ X_{origin} &= Stem_1(X_t). \end{aligned} \tag{1}$$

Next, the raw frames and frame differences are combined and processed through $Stem_2$ to

enhance feature representation.

$$X_{fusion} = \alpha \cdot Stem_2(X_{origin}) + \beta \cdot Stem_2(\alpha \cdot X_{origin} + \beta \cdot X_{diff}). \tag{2}$$

Subsequently, the self-attention module focuses on regions rich in rPPG information, aiming to enhance the weight of rPPG information in the subsequent frame-level average pooling. The self-attention mask can be represented as:

$$Mask = \frac{(H/8)(W/8) \cdot \sigma(Stem_3(X_{fusion}))}{2||\sigma(Stem_3(X_{fusion}))||_1}. \tag{3}$$

After self-attention, the left and right halves of the face are split and the rPPG information from each half is extracted through subsequent networks. Frame-level average pooling is applied to tokenize the single-frame images of the left and right halves of the face.

Multi-temporal Mamba. Multi-temporal Mamba leverages multiple temporal scales, enabling a selective state space model to simultaneously learn the short-term trends and periodic patterns of BVP waves. Specifically, the input token sequence X_{in} is projected and then processed through four separate paths. For the i_{th} path, the sequence is divided into 2^{i-1} sub-sequences, each of which undergoes sequential processing through a convolution layer, activation layer, and selective state space model. Subsequently, they are recombined into a sequence of the original length, forming the output of the i_{th} path, denoted as X_{path_i}. The output before projection can be represented as follows:

$$X_{out} = \sum_{i=1}^{4} X_{path_i} \times \sigma(Proj(X_{in})). \tag{4}$$

Frequency domain feed-forward. The frequency domain feed-forward facilitates channel interaction in the frequency domain. It consists of three components: domain conversion, channel interaction, and domain inversion. There is a linear layer before and after each component for projection. Frequency domain conversion and inversion are carried out using FFT and iFFT, respectively. After frequency domain conversion, the signal becomes complex, so channel interaction involves complex arithmetic operations. Specifically, for complex input $H \in \mathbb{R}^{T/2 \times C}$, given complex weight matrix $W \in \mathbb{R}^{C \times C}$ and complex bias $B \in \mathbb{R}^{C}$, according to the rules of complex multiplication, it is expressed as follows:

$$\begin{aligned} H'_{re} &= H_{re}W_{re} - H_{im}W_{im} + B_{re}, \\ H'_{im} &= H_{re}W_{im} + H_{im}W_{re} + B_{im}, \\ H' &= H'_{re} + j \cdot H'_{im}. \end{aligned} \tag{5}$$

The subscripts re and im denote the real and imaginary components of the corresponding complex data, respectively.

2.2. Training Procedure

The training procedure is divided into two stages.

BVP estimation stage. During the BVP estimation stage, the parameters of the BP Network are not updated. The loss is calculated based on the distance between the output of the BVP predictor and the ground truth BVP waveforms. This stage's loss is a hybrid loss that constrains the BVP waveforms in both temporal and frequency domains. The temporal-domain loss is computed using the negative Pearson correlation coefficient between the predicted and ground truth BVP waveforms:

$$\mathcal{L}_{Time} = 1 - \frac{\sum_{i=1}^{n} \left(BVP_{gt_i} - \overline{BVP_{gt}}\right)\left(BVP_{pred_i} - \overline{BVP_{pred}}\right)}{\sqrt{\sum_{i=1}^{n} \left(BVP_{gt_i} - \overline{BVP_{gt}}\right)^2}\sqrt{\sum_{i=1}^{n} \left(BVP_{pred_i} - \overline{BVP_{pred}}\right)^2}} \tag{6}$$

The frequency domain loss is calculated using the cross-entropy between the heart rates derived from the predicted BVP waves and the ground truth BVP waves' frequency domain. The frequency loss $\mathcal{L}_{Freq}$ is then expressed by:

$$\mathcal{L}_{Freq} = CE(maxIndex(PSD(BVP_{gt})), PSD(BVP_{pred})) \tag{7}$$

Where CE denotes cross-entropy, maxIndex represents the index of the maximum value, and PSD stands for power spectral density analysis. The first stage loss is expressed by:

$$\mathcal{L}_{BVP} = a \cdot \mathcal{L}_{Time} + b \cdot \ \mathcal{L}_{Freq} \tag{8}$$

The predicted BVP waveform requires post-processing. In the post-processing phase, a second-order Butterworth filter (with cutoff frequencies of 0.75 and 2.5 Hz) was applied to the predicted BVP waveform. Subsequently, the Welch algorithm was used to compute the power spectral density for further heart rate estimation.

BP estimation stage. During the BP estimation stage, the loss is calculated based on the distance between the predicted diastolic and systolic blood pressure values and the ground truth values. The specific loss function is defined as follows:

$$\mathcal{L}_{BP} = 0.5 \cdot (SBP_{gt} - SBP_{pred})^2 + 0.5 \cdot (DBP_{gt} - DBP_{pred})^2 \tag{9}$$

3. Experiment

3.1. Dataset and Performance Metric

The Vital Videos dataset [16] was used for training, and the challenge provided 200 videos from 100 subjects in the OBF dataset [17] for testing. The Vital Videos dataset was recorded at a resolution of 1920 x 1200 pixels and 30 fps. Each subject was recorded in two 30-second video clips. Participants were recorded under various indoor lighting conditions, including artificial and natural light sources, in multiple indoor environments such as libraries and shopping malls across ten locations in Western Europe, covering six skin tones and age groups from 5 to 95 years. The Vital Videos dataset includes three versions: VV-small, VV-Medium, and

VV-Large, with only VV-small (a subset of 100 subjects exhibiting the greatest demographic diversity) can be used. The OBF dataset contains 200 five-minute-long videos recorded at a frame rate of 60 fps, documenting 100 healthy adults under consistent environmental and setting conditions. Five commonly used metrics were employed for evaluation: MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), Pearson correlation coefficient ρ, and SNR (Signal-to-Noise Ratio).

3.2. Implementation Details

We utilized the rPPG toolbox [18], a PyTorch-based open-source toolbox, to implement our proposed method. For video segments, facial regions were cropped and resized to 128×128 pixels in the initial frame, remaining fixed in subsequent frames. During the first stage, we adhered to the outlined protocol [19], incorporating random upsampling, downsampling, and horizontal flipping for data augmentation. Additionally, we continued training based on a pre-trained model from the PURE dataset. Parameters for both phases remained consistent: a learning rate of 3e-4, batch size of 16, and 30 epochs for training. Network training was conducted on a single NVIDIA RTX 3090.

3.3. HR Evaluation

We used HR estimation results as an indicator for BVP estimation. The VV-small dataset was sequentially partitioned into training, validation, and test sets in a 7:1:2 ratio. As shown in the first two rows of Table 1, RhythmMamba significantly outperformed the commonly employed POS in the past. Moreover, the BVP estimation demonstrated relatively favorable outcomes, with 2.90 bpm MAE for heart rate. We conducted a simple cross-validation by exchanging the validation and test sets, as indicated in the last two rows of Table 1, revealing a nearly tenfold difference in results. This discrepancy may stem from the considerable distribution variation within the VV-small dataset, which comprises only 200 videos.

Table 1
HR estimation results.

	MAE↓	RMSE↓	MAPE↓	ρ ↑	SNR↑
POS[20]	5.66	10.10	7.15	0.44	-5.79
RhythmMamba[15]	2.90	7.33	3.63	0.74	8.56
Cross-validation	0.32	0.42	0.46	0.99	18.71

3.4. BP Evaluation

Intra-dataset experiments on VV-small. The BP estimation still adheres to the dataset partitioning criteria established during HR estimation. As indicated in the first row of Table 2, the results are less than satisfactory. This may be attributed to severe limitations imposed by the quantity and distribution of the dataset. While the performance is acceptable for HR tasks with lower precision requirements, it demonstrates unsatisfactory outcomes for BP tasks with higher precision demands.

Blood pressure consists of systolic blood pressure (SBP) and diastolic blood pressure (DBP). PTT is linearly correlated with SBP [21], while its correlation with DBP is relatively low. We separately present the test results for SBP and DBP. As shown in the last two rows of Table 2, DBP demonstrates higher accuracy compared to SBP. However, this may also be due to the larger fluctuation range of SBP compared to DBP. Additionally, it can be observed that the ρ of BP estimation is significantly higher compared to SBP and DBP. This is because the metrics for BP are calculated by concatenating the SBP and DBP sequences.

Cross-dataset experiments on OBF. We conducted testing on the videos of 100 subjects from the OBF dataset provided in the challenge, which can be regarded as a form of cross-dataset testing. The training dataset was the VV-small dataset, partitioned sequentially into training and validation sets in an 8:2 ratio. The RMSE of the test results was 13.59. The cross-dataset results were nearly identical to those from testing within the VV-small dataset, yet neither set of results was particularly satisfactory. This aligns with the analysis from the HR evaluation, where the small volume and significant distribution differences in the dataset were identified as key issues. The distribution gap between the training and test sets within VV-small is comparable to the cross-dataset gap.

Table 2
BP estimation results of intra-dataset.

	MAE↓	RMSE↓	MAPE↓	ρ ↑
BP	10.18	14.77	9.47	0.86
SBP	14.45	19.47	11.67	0.28
DBP	5.92	7.57	7.27	0.1

3.5. Ablation Study

We conducted ablation studies using the same partitioning strategy as in the intra-dataset evaluation.

Impact of BVP stage. We attempted to estimate BP directly from facial videos by removing the BVP stage and the BVP predictor in the network. As shown in the first and last rows of Table 3, directly estimating BP from facial videos does not yield satisfactory results.

Impact of BVP extraction algorithm. The BVP extraction algorithm was replaced from POS to RhythmMamba. The results are shown in the second and last rows of Table 3. Comparing these results with the HR estimation results in Table 1, it appears that the quality of PPG extraction only slightly affects BP outcomes.

Impact of face separate estimation. Theoretically, extracting BVP separately from the left and right halves of the face and merging them into two channels for input into the neural network should provide not only BVP feature information but also PTT information. This approach is expected to outperform extracting a single-channel BVP from the whole face. To investigate this, we conducted an ablation study. As shown in the third and last rows of Table 3, the performance improvement from separate extraction of the left and right facial BVPs was not significant. Further visualization of the BVP waveforms extracted from the left and right sides

of the face revealed that the waveforms were extremely similar, providing almost no additional PTT information.

Impact of BP separate estimation. Since SBP is highly correlated with PTT and DBP is less correlated with PTT [21], we attempted to estimate SBP and DBP separately using two BP networks. The results, as shown in the fourth and last rows of Table 3, indicate that separate estimation has little impact.

Impact of face detection. The default face detection algorithm is the Haar Cascade. The videos are split into 160-frame segments, performing detection on the first frame of each segment and maintaining that position for subsequent frames. To investigate the impact of the face detection algorithm, we used the more recent RetinaFace [22], splitting the video into 300-frame segments and performing face detection every 4 frames. As shown in the last two rows of Table 3, the performance actually declined, which may be attributed to the reduced data volume resulting from the longer video segments.

Table 3
Results of ablation studies.

BVP Stage	BVP Algo.	Face Sep.	BP Sep.	Face Detection	MAE	RMSE	MAPE	ρ
w/o	-	w.	w/o	Default	15.26	19.01	13.93	0.79
w.	[20]	w.	w/o	Default	10.72	15.21	9.90	0.85
w.	[15]	w/o	w/o	Default	**10.13**	15.17	**9.23**	0.85
w.	[15]	w.	w.	Default	10.30	**14.66**	9.48	0.86
w.	[15]	w.	w/o	Customized	13.45	19.15	11.84	0.80
w.	[15]	w.	w/o	Default	10.18	14.77	9.47	**0.86**

4. Conclusion

In the challenge, we proposed a two-stage method for BP estimation based on the selective state space model which achieves a RMSE of 13.59 mmHg. However, the model's performance is still not adequate for practical application. In future work, we hope to integrate BP priors into the model design and explore the feasibility of pre-training with PPG datasets for better performance on the rPPG dataset.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62206015, 62227801, U20B2062), the Fundamental Research Funds for the Central Universities (FRF-TP-22-043A1), and the Young Scientist Program of The National New Energy Vehicle Technology Innovation Center (Xiamen Branch).

References

[1] D. McDuff, Camera measurement of physiological vital signs, ACM Computing Surveys 55 (2023) 1–40.

[2] B. Huang, S. Hu, Z. Liu, C.-L. Lin, J. Su, C. Zhao, L. Wang, W. Wang, Challenges and prospects of visual contactless physiological monitoring in clinical study, NPJ Digital Medicine 6 (2023) 231.

[3] 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).

[4] H. Luo, D. Yang, A. Barszczyk, N. Vempala, J. Wei, S. J. Wu, P. P. Zheng, G. Fu, K. Lee, Z.-P. Feng, Smartphone-based blood pressure measurement using transdermal optical imaging technology, Circulation: Cardiovascular Imaging 12 (2019) e008857.

[5] M. Rong, K. Li, A blood pressure prediction method based on imaging photoplethysmography in combination with machine learning, Biomedical Signal Processing and Control 64 (2021) 102328.

[6] F. Bousefsaf, T. Desquins, D. Djeldjli, Y. Ouzar, C. Maaoui, A. Pruski, Estimation of blood pressure waveform from facial video using a deep u-shaped network and the wavelet representation of imaging photoplethysmographic signals, Biomedical Signal Processing and Control 78 (2022) 103895.

[7] F. Schrumpf, P. Frenzel, C. Aust, G. Osterhoff, M. Fuchs, Assessment of non-invasive blood pressure prediction from ppg and rppg signals using deep learning, Sensors 21 (2021) 6022.

[8] I. C. Jeong, J. Finkelstein, Introducing contactless blood pressure assessment using a high speed video camera, Journal of medical systems 40 (2016) 1–10.

[9] X. Fan, T. Tjahjadi, Robust contactless pulse transit time estimation based on signal quality metric, Pattern Recognition Letters 137 (2020) 12–16.

[10] J.-s. Park, K.-S. Hong, Robust blood pressure measurement from facial videos in diverse environments, Heliyon (2024).

[11] B. Zou, Z. Guo, J. Chen, H. Ma, Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer, arXiv preprint arXiv:2402.12788 (2024).

[12] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752 (2023).

[13] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, Vmamba: Visual state space model, arXiv preprint arXiv:2401.10166 (2024).

[14] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, Y. Qiao, Videomamba: State space model for efficient video understanding, arXiv preprint arXiv:2403.06977 (2024).

[15] B. Zou, Z. Guo, X. Hu, H. Ma, Rhythmmamba: Fast remote physiological measurement with arbitrary length videos, arXiv preprint arXiv:2404.06483 (2024).

[16] P.-J. Toye, Vital videos: A dataset of videos with ppg and blood pressure ground truths, arXiv preprint arXiv:2306.11891 (2023).

[17] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, G. Zhao, The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection, in: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 242–249.

[18] X. Liu, G. Narayanswamy, A. Paruchuri, X. Zhang, J. Tang, Y. Zhang, S. Sengupta, S. Patel, Y. Wang, D. McDuff, rPPG-toolbox: Deep remote PPG toolbox, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL: https://openreview.net/forum?id=q4XNX15kSe.

[19] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching, IEEE Signal Processing Letters 27 (2020) 1245–1249.

[20] W. Wang, A. C. Den Brinker, S. Stuijk, G. De Haan, Algorithmic principles of remote ppg, IEEE Transactions on Biomedical Engineering 64 (2016) 1479–1491.

[21] Y. Zhang, M. Berthelot, B. Lo, Wireless wearable photoplethysmography sensors for continuous blood pressure monitoring, in: 2016 IEEE Wireless Health (WH), IEEE, 2016, pp. 1–8.

[22] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5203–5212.