An end-to-end contrastive deep-learning framework for remote physiological signal measurement

Bingjie Wu[1], Menghan Zhou[1], Wei Liu[1], Xingyao Wang[1], Xingjian Zheng[1], Yiping Xie[2], Chaoqi Luo[3] and Liangli Zhen[1,*]

[1]*Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore*
[2]*College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.*
[3]*School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China.*

Abstract

Heart rate measurements based on remote physiological signals could significantly facilitate health monitoring in daily life. However, the ground-truth labels of the physiological signals are expensive and hard to collect. In this paper, we present a contrastive self-supervised learning framework to extract discriminative remote physiological features by leveraging periodic signal priors without ground-truth labels in the pre-training stage. Specifically, a ranking loss and a contrastive learning loss are constructed to extract knowledge with resampling of the video clips. In addition, data augmentation and ensemble learning strategies are designed to fine-tune the pre-trained model and fuse the results to improve the heart rate measurement. Our final solution achieves the 1st place in track 1 of the 3rd Vision-based Remote Physiological Signal Sensing (RePSS) Challenge.

1. Introduction

Remote photoplethysmography (rPPG) has emerged as a promising technology in the pursuit of non-invasive methods of monitoring vital signs. By leveraging the principles of light absorption and reflection, rPPG enables the extraction of vital sign information, such as heart rate [1, 2], respiratory rate [3, 4], and blood pressure [5, 6], remotely and without physical contact with the subject. Traditional vital sign monitoring is confined to clinical settings with cumbersome wired sensors, the rPPG heralds a paradigm shift by allowing remote physiological signal measurement with common cameras. This freedom from physical attachments opens avenues for continuous monitoring in scenarios where traditional sensors are impractical or uncomfortable. This revolutionary technology holds immense potential for applications of telemedicine, health monitoring, fitness tracking, and driver monitoring.

Pioneer studies propose developing skin models to predict the rPPG signals [7, 8] or through mathematical analysis to decompose the rPPG signals [9]. Recently, some studies have shown that deep learning algorithms are promising in rPPG prediction [10, 11]. Compared with wrist and leg, the facial rPPG signals were typically stronger, especially on the forehead [12]. However,

---

*Corresponding author: Liangli Zhen (email: llzhen@outlook.com).

the collection of facial videos with ground-truth labels for these methods is time-consuming and expensive.

In this study, we propose a contrastive self-supervised method to extract physiological signals from unlabeled facial videos in the pre-training stage. As rPPG signals are periodic signals, the heart rate will vary if the frame rate is changed. Based on this consideration, a ranking loss for heart rate and a contrastive learning loss are designed to extract physiological signals based on the resampling of the video clips without ground-truth labels. In the fine-tuning stage, we formulate the heart rate prediction into a classification problem. Moreover, data augmentation and ensemble learning strategies are adopted to reduce the prediction error. The experimental results demonstrate that our proposed method can achieve the average root mean squared error (RMSE) of 19.84 on VIPL-HR-V2 and our final solution achieves the average RMSE of 8.50693, indicating the effectiveness of our proposed solution.

2. Related Work

Recent self-supervised learning methods have achieved remarkable results in estimating heart rate on the rPPG public datasets. Among self-supervised methods, contrastive learning methods have been increasingly popular. Some contrastive methods leverage spatial and temporal characteristics to extract the invariant features. For instance, Wang et al. [13] proposed a self-supervised representation method using spatiotemporal augmentation. The spatial augmentation is based on the seven regions of interest (ROIs) on faces that represent similar heart rates. Several frame strides are used to extract video frames to augment each video along the time axis. A contrastive learning loss is constructed to attract the spatiotemporal augmented clips from the same sample and resist the clips from different samples. Additional spatial and temporal classifiers are designed to constrain the learning process.

Another notable contrastive-based unsupervised remote physiological measurement framework is Contrast-Phys [14], which extracts physiological signals from facial videos. The model generates positive pairs from different spatiotemporal locations of the same video and negative pairs from two different videos. Based on contrastive learning of positive and negative pairs, Contrast-Phys achieves comparable results to supervised methods.

Moreover, some contrastive methods are based on periodic signal characteristics. For example, Gideon and Stent [15] introduced a multi-view triplet loss for contrastive training. The positive samples are taken from subset views of anchor samples. The negative samples are generated through a trilinear resampler. This method achieves comparable results on four rPPG datasets when compared to several supervised deep-learning methods.

In addition to contrastive learning methods, Speth *et al.* proposed a non-contrastive unsupervised deep learning method [16]. It learns the rPPG features by shaping the frequency spectrum through three loss functions: 1) The bandwidth loss forces the model to focus on the bandlimits between 0.66Hz-3Hz, which corresponds to a heart rate of 40 bpm to 180 bpm; 2) The sparsity loss penalizes the model if power spectral density is not near the spectral peak; and 3) The variance loss is utilized to enforce diverse outputs to prevent model collapse.

3. Our Proposed Method

Our proposed solution includes two stages, as shown in Fig. 1. During the pre-training stage, we propose a contrastive deep learning method called RankContrast to extract the rPPG-related features. In the fine-tuning stage, a supervised method with data augmentation and ensemble learning is utilized to train the model based on a limited number of labeled facial videos.
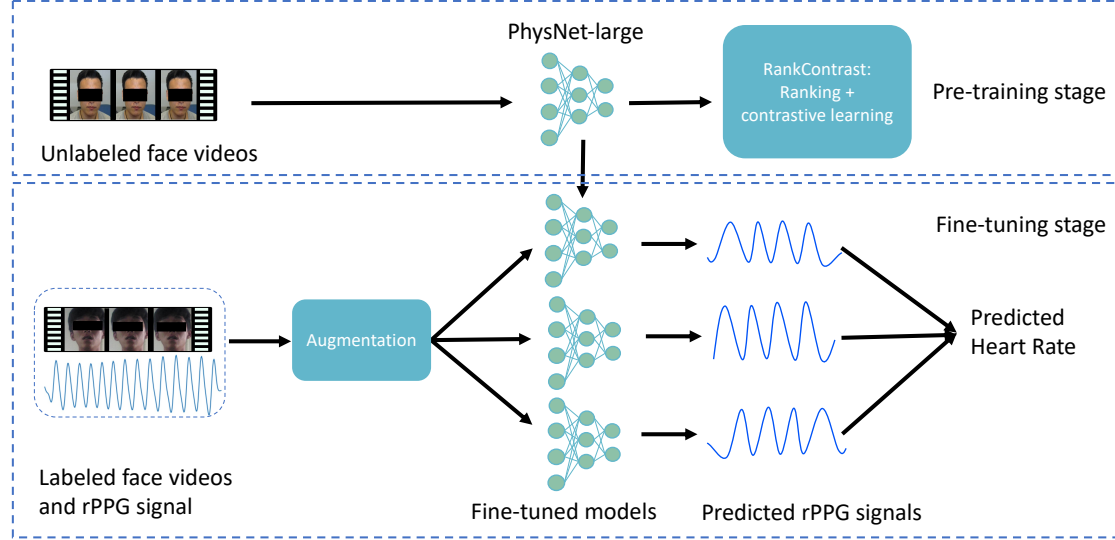

Figure 1: The framework of our proposed solution. It includes two stages: the pre-training stage and the fine-tuning stage. In the pre-training stage, the model is trained with a self-supervised RankContrast method. In the fine-tuning stage, the pre-trained model is fine-tuned with labeled VIPL-HR-V2 dataset with data augmentation and ensemble learning techniques.

3.1. Data Pre-processing

Many existing rPPG learning methods convert the face clips into ST-maps [17, 18, 19] for further feature extraction. In contrast, our solution presents an end-to-end framework where a sequence of face frames is fed directly into the deep learning model. We use multiple datasets with highly complex backgrounds to train the model during the pre-training stage. To minimize noise, only the face area reflecting the rPPG signal is cropped for training. The human faces are detected by MTCNN [20] on the first frame, and then the whole video is cropped by a larger bounding box based on the detected face with a scale factor of 1.3. The cropped image frames are resized to 128×128.

3.2. Pre-training with Self-supervised Learning

A RankContrast self-supervised learning method that integrates the ranking loss and the contrastive learning loss is proposed in this work, as shown in Fig. 2. Since the rPPG signal is periodic, the heart rate varies by resampling the video clips. Upsample the clips will reduce the heart rate and downsample the clips will increase the heart rate [21]. According to these
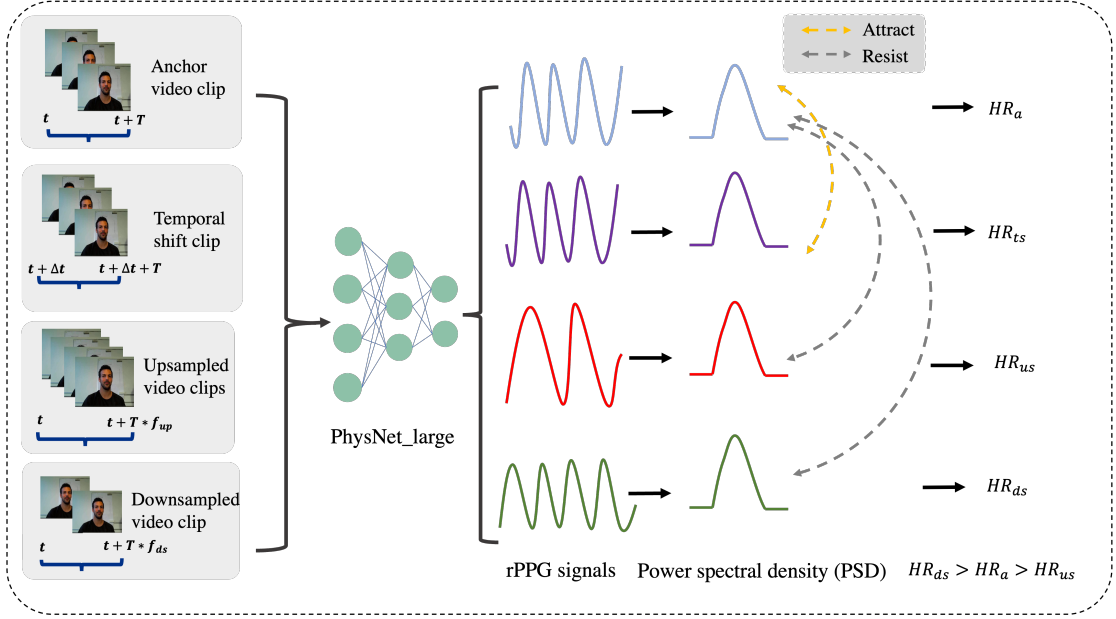
Figure 2: The overview of our proposed RankContrast self-supervised learning method. t represents the t-th frame, Δt is the shifted frame number, T is the temporal length of the trained video clip, f_{up} is the upsampling factor, f_{ds} is the downsampling factor.

characteristics, a ranking loss function is designed to extract features with upsampling and downsampling video clips. A random upsampling factor is selected between 1 and 1.1 and a downsampling factor is selected between 0.9 and 1.0. The ranking loss L_r is defined as:

$$L_r = max(HR_{us} - HR_a, 0) + max(HR_a - HR_{ds}, 0) \tag{1}$$

where HR_{us}, HR_{ds}, and HR_a are the heart rate of the upsampled clip, downsampled clip, and anchor clip, respectively.

The heart rate is calculated by multiplying the one-hot vector of the power spectral density (PSD) with the vector of frequency bins.

The contrastive learning loss is to compare similar (positive) clips and dissimilar (negative) clips with the anchor clips through the attracting and resisting strategy. As the heart rate is relatively stable for an individual in a short time, the positive pairs are constructed by shifting the training clip for some frames in the same video. The resampled samples from the anchor sample are considered negative pairs. Based on this idea, the contrastive learning loss L_c is formulated as:

$$L_c = MSE(PSD_{ts} - PSD_a) - MSE(PSD_{us} - PSD_a) - MSE(PSD_{ds} - PSD_a) \tag{2}$$

where PSD_{us}, PSD_{ds}, PSD_a, PSD_{ts} are the power spectral density of the upsampled clip, downsampled clip, anchor clip, and temporal shift clip, respectively. MSE denotes the mean squared error.

The total loss function for our self-supervised learning L_{un} is:

$$L_{un} = \lambda L_r + L_c \tag{3}$$

where λ is a hyperparameter to trade off the contributions of the two terms.

3.3. Fine-tuning

The pre-trained model is then fine-tuned on the VIPL-HR-V2 dataset that consists of 400 subjects in a supervised learning manner. The ground truth of blood volume pulse (BVP) wave and heart rate are provided in the VIPL-V2 dataset. We adopt two supervised loss functions: the classification loss L_{ce} and the Pearson loss L_p to guide the learning process. Given that the typical resting heart rate for humans falls within the range of 40-180 beats per minute, each heart rate is mapped to a class label ranging from 0 to 140. The classification loss is defined as:

$$L_{ce} = CE(PSD(\hat{y}), \mathbf{e}_{HR}) \tag{4}$$

where CE is the cross-entropy loss, $\hat{y}$ is the predicted BVP, $PSD(\hat{y})$ represents the power spectral density of $\hat{y}$, and $\mathbf{e}_{HR}$ denotes the one-hot encoding of the class of the ground-truth heart rate HR.

Since the ground-truth BVP signal is available, we impose a Negative Pearson correlation coefficient loss L_p to constrain the model's output as:

$$L_p = 1 - \frac{\sum_{i=1}^{T}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{T}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{T}(y_i - \bar{y})^2}} \tag{5}$$

where y is the ground-truth BVP, $\hat{y}$ is predicted BVP, $\bar{\hat{y}}$ is the average value of the predicted BVP, $\bar{y}$ is average value of ground-truth BVP, T is the length of the BVP signal, and y_i and $\hat{y}_i$ are the i-th element of y and $\hat{y}$, respectively.

The total loss for the fine-tuning stage L_s is:

$$L_s = L_{ce} + \mu L_p \tag{6}$$

where μ is a hyperparameter to trade off the contributions of the two terms.

4. Experiments

4.1. Dataset

In the pre-training phase, the model is trained using datasets from VIPL-HR-V2, VIPL-HR, PURE, UBFC-rPPG, COHFACE, and VSIGN without labels. Following this, the model undergoes fine-tuning on VIPL-HR-V2, which includes data from 400 participants. The test set comprises data from 200 individuals drawn from the VIPL-HR-V2 and OBF datasets.

VIPL-HR-V2 [22] offers 2500 RGB videos featuring 500 subjects. Each subject has five clips of 10 seconds which are cut from a thirty-second long video with a five-second stride.

VIPL-HR [23] is a database with face videos of 107 subjects and the corresponding heart rate, SpO2, and BVP wave. The data is collected under 9 scenarios under various illumination and motion conditions with three camera devices. This database has 2,378 visible light videos (VIS) and 752 near-infrared (NIR) videos. We only use visible light videos for training.

PURE [24] comprises recordings of 10 subjects, each recorded for one minute across six different scenarios. These videos are captured at 30 fps with a resolution of 640×480 pixels. Pulse rate waveforms and SpO2 readings, sampled at 60 Hz, are recorded alongside the videos.

UBFC-rPPG [25] includes 42 subjects, recorded at 30 fps with image frames of 640×480 pixels in uncompressed RGB format for each subject. Simultaneously, the ground-truth PPG waveform is collected.

COHFACE [26] features 160 one-minute videos from 40 subjects, with synchronized heart rate and breathing rate data. These videos are recorded with image frames of 640×480 pixels and a frame rate of 20 Hz.

The VSIGN dataset is a facial video dataset collected by our team (Face AI) at A*STAR for research purposes. It encompasses signals including BVP, blood pressure, respiratory rate, and SpO2. In this work, facial videos from 90 subjects have been amassed. Each subject is captured using six RGB cameras across six distinct scenarios. The video frame rate is around 30 fps.

4.2. Implementation Details

In this work, a 3D-CNN model of PhysNet-large is utilized as the backbone, which is modified from the PhsNet [14]. A sequence of T frames is selected as input for the model. T is set as 280 in this study. Each image frame is revised to a size of 128×128. Since the common heart rate falls between 40-180 bpm, the signals are filtered with a cutoff frequency of [0.66, 3]Hz to filter out irrelevant noises. The Adam optimizer [27] with a learning rate of 1×10^{-5} is employed. The number of training epochs is set as 50 for both pre-training and fine-tuning. The frame rates of all the clips and BVP signals are resampled to 30 fps. The facial videos of subject no. 351-400 from VIPL-HR-V2 are used as the validation dataset.

4.3. Evaluation Metrics

The experimental results have been reported in terms of heart rate. The root mean square error (RMSE) is utilized to evaluate the performance of the tested measurement methods as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{HR}_i - HR_i)^2}{N}} \tag{7}$$

where HR_i is ground-truth heart rate, $\hat{HR}_i$ is predicted heart rate of the i-th sample, and N is the total number of samples.

4.4. Experimental Results

The model is pre-trained on the merged dataset of VIPL-HR-V2, VIPL-HR, PURE, UBFC-rPPG, COHFACE, and VSIGN. The results on the validation dataset of VIPL-HR-V2 are shown in Fig. 3. The RMSE fluctuates initially and decreases drastically around the number of 40 epochs. The

minimum RMSE on the validation data with the RankContrast method is 19.84. The model with the minimum RMSE is selected for further fine-tuning in the next stage.
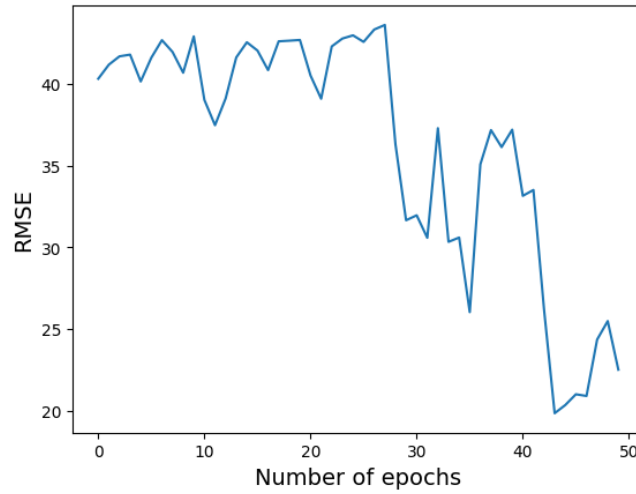
Figure 3: The results of the pre-trained model from RankContrast. The model is pre-trained on the merged dataset and evaluated on VIPL-HR-V2

The RankContrast method is compared with two self-supervised learning methods, i.e., Contrast-Phys [14] and Gideon2021 [15]. In this experiment, only the VIPL-HR-V2 dataset is utilized for pre-training by considering the computational time costs. The results are reported in Table 1, from which we can see that RankContrast demonstrates superior performance compared to the other two self-supervised learning methods with a significant margin of 7.81.

Table 1
Comparsion of Self-supervised Methods Pre-trained and Evaluated on VIPL-HR-V2

Index	Method	RMSE
1	Contrast-Phys[14]	33.55
2	Gideon2021[15]	30.08
3	RankContrast	22.27

In the second stage, the pre-trained model is fine-tuned with labeled data from VIPL-HR-V2 using Equation 6. With data augmentation and ensemble learning techniques, our solution achieves the RMSE of 8.50693 on the test dataset in the 3rd RePPS Challenge, as shown in Table 2. It outperforms the results from other teams by a large margin, indicating the effectiveness of our solution.

4.5. Discussion

The Contrast-Phys and Gideon2021 methods perform well on small and simple datasets, such as PURE and UBFC-rPPG, as reported in their original papers. However, VIPL-HR-V2 is much

Table 2
The Public Leaderboard of 3rd RePSS Challenge Track 1

Rank	Team name	RMSE
1	Face AI	8.50693
2	HFUT_VUT	8.85277
3	PCA_VItal	8.96941
4	Hash Brown	9.26198
5	AIIA	9.28902

more challenging due to complex backgrounds, diverse illumination conditions, and substantial motions. Besides, the heart rates for VIPL-HR-V2 have a wide range of distribution. It is more difficult to identify the specific category of the heart rate for a video clip with self-supervised learning methods. RankContrast provides an additional ranking loss on heart rates besides the contrastive learning loss to constrain the model learning, making it more effective on VIPL-HR-V2.

5. Conclusion

In this paper, we introduced a self-supervised learning method called RankContrast, which leverages periodic signal characteristics. RankContrast employs both a ranking loss and a contrastive learning loss to extract physiological signals through the resampling of video clips. We compared our method with two other peer methods on VIPL-HR-V2. Our results show that RankContrast achieves the best performance. The pre-trained model using RankContrast was then fine-tuned on VIPL-HR-V2 in a supervised learning manner. The final fine-tuned model achieved an RMSE of 8.51 on the test dataset of the 3rd RePSS Challenge, significantly outperforming other solutions.

References

[1] D. N. Tran, H. Lee, C. Kim, A robust real time system for remote heart rate measurement via camera, in: 2015 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2015, pp. 1–6.

[2] M. Hu, F. Qian, X. Wang, L. He, D. Guo, F. Ren, Robust heart rate estimation with spatial–temporal attention network from facial videos, IEEE Transactions on Cognitive and Developmental Systems 14 (2021) 639–647.

[3] J. Du, S.-Q. Liu, B. Zhang, P. C. Yuen, Weakly supervised rppg estimation for respiratory rate estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2391–2397.

[4] M. Alnaggar, A. I. Siam, M. Handosa, T. Medhat, M. Rashad, Video-based real-time monitoring for heart rate and respiration rate, Expert Systems with Applications 225 (2023) 120135.

[5] B.-F. Wu, B.-J. Wu, B.-R. Tsai, C.-P. Hsu, A facial-image-based blood pressure measurement system without calibration, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–13.

[6] F. Schrumpf, P. Frenzel, C. Aust, G. Osterhoff, M. Fuchs, Assessment of deep learning based blood pressure prediction from ppg and rppg signals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3820–3830.

[7] 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.

[8] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE transactions on biomedical engineering 60 (2013) 2878–2886.

[9] M.-Z. Poh, D. J. McDuff, R. W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam, IEEE transactions on biomedical engineering 58 (2010) 7–11.

[10] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, arXiv preprint arXiv:1905.02419 (2019).

[11] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, 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.

[12] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light., Optics express 16 (2008) 21434–21445.

[13] 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, volume 36, 2022, pp. 2431–2439.

[14] Z. Sun, X. Li, Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast, in: European Conference on Computer Vision, Springer, 2022, pp. 492–510.

[15] J. Gideon, S. Stent, The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3995–4004.

[16] J. Speth, N. Vance, P. Flynn, A. Czajka, Non-contrastive unsupervised learning of physiological signals from video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14464–14474.

[17] 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 (2019) 2409–2423.

[18] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, G. Zhao, Video-based remote physiological measurement via cross-verified feature disentangling, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer, 2020, pp. 295–310.

[19] A. Das, H. Lu, H. Han, A. Dantcheva, S. Shan, X. Chen, Bvpnet: Video-to-bvp signal prediction for remote heart rate estimation, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, 2021, pp. 01–08.

[20] J. Xiang, G. Zhu, Joint face detection and facial expression recognition with mtcnn, in: 2017 4th international conference on information science and control engineering (ICISCE),

IEEE, 2017, pp. 424–427.

[21] Z. Li, L. Yin, Contactless pulse estimation leveraging pseudo labels and self-supervision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20588–20597.

[22] X. Li, H. Han, H. Lu, X. Niu, Z. Yu, A. Dantcheva, G. Zhao, S. Shan, The 1st challenge on remote physiological signal sensing (repss), in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 314–315.

[23] X. Niu, H. Han, S. Shan, X. Chen, Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video, in: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14, Springer, 2019, pp. 562–576.

[24] R. Stricker, S. Müller, 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, IEEE, 2014, pp. 1056–1062.

[25] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters 124 (2019) 82–90.

[26] G. Heusch, A. Anjos, S. Marcel, A reproducible study on remote heart rate measurement, arXiv preprint arXiv:1709.00962 (2017).

[27] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).