The 3rd Vision-based Remote Physiological Signal Sensing (RePSS) Challenge & Workshop

Zhaodong Sun[2], Xiaobai Li[1,2,*], Hu Han[3], Jiyang Tang[3], Chenhang Ying[1], Jieyi Ge[1], Antitza Dantcheva[4], Shiguang Shan[3] and Guoying Zhao[2]

[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China
[2]Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland
[3]Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), China
[4]STARS team, INRIA, France

Abstract

The remote measurement of physiological signals from video recordings is a topic of growing interest. Despite its potential, progress in this field is being impeded by the absence of publicly available benchmark databases and a standardized validation platform. To address these issues, the RePSS Challenge is held annually. The 3rd RePSS Challenge is being conducted alongside IJCAI 2024 and features two competition tracks. Track 1 focuses on self-supervised learning for heart rate measurement using unlabeled facial videos, while Track 2 tackles the more complex task of measuring blood pressure from facial videos. This paper provides an overview of the challenge, detailing the data, protocols, analysis of results, and discussions. We highlight the top-performing solutions to offer insights for researchers and outline future directions for this field and the challenge itself.

Keywords

rPPG, physiological signal, facial video, heart rate, blood pressure

1. Introduction

Physiological signals, including heart rate (HR), respiration rate (RR), heart rate variability (HRV), and blood pressure (BP), are crucial indicators of human health. Traditionally, these signals are measured using specialized medical instruments such as electrocardiography (ECG), photoplethysmography (PPG) oximeters, and breathing belts. However, using contact medical sensors is expensive and inconvenient for long-term monitoring. Later, researchers discovered that PPG signals can be captured remotely from human faces under ambient light conditions. For instance, Verkruysse et al. [1] demonstrated the measurement of PPG signals from the forehead. Subsequently, numerous studies have proposed various remote PPG (rPPG) measurement techniques. Early methods relied on empirically designed filters and lacked a training process. Some approaches [2, 3, 4, 5, 6, 7] utilized subtle color changes in facial pixels for rPPG measurement, while others [8, 9, 10] focused on tracking vertical head motions. Most researchers have adopted supervised approaches for rPPG measurement, such as [11], [12], [13], and [14]. Recently, more researchers are developing unsupervised/self-supervised rPPG methods [15, 16, 17, 18, 19, 20] to train rPPG measurement models with only facial videos.

Despite significant research interest, the development of this field is hindered by the lack of publicly available benchmark databases and a standardized validation platform. To address these issues, we organized the 1st RePSS challenge [21] [1] in conjunction with CVPR 2020, followed by the 2nd RePSS challenge [22] [2] with ICCV 2021, aiming to provide benchmark datasets and a fair comparison platform

The 3rd Vision-based Remote Physiological Signal Sensing (RePSS) Challenge & Workshop, August 5, 2024, Jeju, South Korea
*Corresponding author.
zhaodong.sun@oulu.fi (Z. Sun); xiaobai.li@zju.edu.cn (X. Li); hanhu@ict.ac.cn (H. Han); tangjiyang22s@ict.ac.cn (J. Tang); chying@zju.edu.cn (C. Ying); jyge@zju.edu.cn (J. Ge); antitza.dantcheva@inria.fr (A. Dantcheva); sgshan@ict.ac.cn (S. Shan); guoying.zhao@oulu.fi (G. Zhao)
0000-0002-0597-0765 (Z. Sun); 0000-0003-4519-7823 (X. Li); 0000-0001-6010-1792 (H. Han); 0000-0003-0107-7029 (A. Dantcheva); 0000-0002-8348-392X (S. Shan); 0000-0003-3694-206X (G. Zhao)

[1]https://competitions.codalab.org/competitions/22287
[2]https://competitions.codalab.org/competitions/30855

for researchers. The RePSS challenge series is intended to be an annual event with a continuous and evolving theme. The inaugural 1st RePSS challenge focused on the fundamental task of measuring average HR from color facial videos. The 2nd RePSS challenge, held alongside ICCV 2021, introduced two tracks: inter-beat interval (IBI) and respiration measurement. This year, the 3rd RePSS, held in conjunction with IJCAI 2024, introduces two new tracks: self-supervised facial video-based heart rate measurement and blood pressure measurement.

The paper is structured as follows: Section 2 provides an overview of the 3rd RePSS challenge, detailing the tasks, datasets, challenge protocol, and evaluation metrics. Section 3 discusses the approaches proposed by the top-performing teams in the challenge. Section 4 presents the challenge results and discussions, and Section 5 explores future directions in this research area.

2. Challenge Overview

2.1. Challenge tracks

There are two tracks for the 3rd RePSS challenge held on Kaggle. There are 18 teams registered for Track 1 and 15 for Track 2. By the final submission date, valid results were submitted by 13 teams in Track 1 and six teams in Track 2. There are totally 313 result submissions and 58 participants in the track 1, and there are 148 result submissions and 23 participants in the track 2.

Track 1 is self-supervised learning for heart rate measurement using unlabeled facial videos'. Since there are only a few facial videos with HR labels, track 1 mainly focuses on developing self-supervised training methods on large-scale unlabeled facial videos. Track 1 was organized on the Kaggle website [3].

Track 2 is facial video-based blood pressure measurement, which is an emerging topic and more challenging. Blood pressure measurement requires high-quality physiological signals from facial videos, so each participant in this track should design both an accurate remote physiological signal measurement algorithm and a blood pressure estimation algorithm. Track 2 was organized on the Kaggle website [4].

2.2. Data and protocol

Track 1. Since track 1 is about self-supervised training, there are three stages for this track including the pre-training stage, the model fine-tuning stage, and the test stage. For the pre-training stage that focuses on unsupervised pretraining, we have summarized a list of open-source, large-scale facial video datasets including (a) VFHQ [23] [5], (b) FaceForensics++ [24] [6], (c) DeeperForensics [25] [7], (d) CelebV-HQ [26] [8], (e) DISFA [27] [9], and (f) MAHNOB Laughter Database [28] [10]. We have checked each of the datasets to confirm that the video quality is suitable for the task, no ground truth is available, and the data can be easily accessed online. Participants can also use other face video data for pre-training without ground truth. For the model fine-tuning stage, we provide the VIPL-V2 dataset [21, 29, 30] built by the organizers' team. The dataset contains facial videos with ground truth physiological signals from 400 persons. For the test stage, we provide a subset of 200 persons' data from the VIPL-HR-V2 and the OBF datasets as the testing data. The ground truth signals of the test set have never been released in previous challenges. Participants should submit their HR prediction to the Kaggle website to get the evaluation results. Each team has a maximum of 5 submissions per day. The ranking will be based on the RMSE on the test data.

Track 2. Track 2 is facial video-based blood pressure measurement, which contains the training and

[3] https://www.kaggle.com/competitions/the-3rd-repss-t1
[4] https://www.kaggle.com/competitions/the-3rd-repss-t2
[5] https://liangbinxie.github.io/projects/vfhq/
[6] https://github.com/ondyari/FaceForensics
[7] https://github.com/EndlessSora/DeeperForensics-1.0
[8] https://celebv-hq.github.io/
[9] http://mohammadmahoor.com/disfa/
[10] https://mahnob-db.eu/laughter/

test stages. For the training stage, there is a large-scale rPPG dataset called vital videos [31] [11] with facial videos and blood pressure labels from around 880 subjects. The video and labels in the dataset are of good quality. We have made an agreement with the dataset owner that the dataset can be used for the challenge track. Participants can use this labeled dataset to train models for rPPG-based blood pressure measurement. Participants can split part of the training data as the validation set. For the test stage, we will use the OBF dataset [32] including 100 subjects. There are 200 facial videos with blood pressure labels. Only the facial videos will be released, and the blood pressure labels have never been released. Participants should submit their systolic and diastolic BP prediction to the Kaggle website to get the evaluation results. Each team has a maximum of 5 submissions per day. The ranking will be based on the RMSE results on the test data.

2.3. Evaluation metrics

We use root mean squared errors (RMSE) as the evaluation metrics. For Track 1, the RMSE between ground truth heart rates y and submitted heart rates y' is calculated as

$$RMSE_1 = \sqrt{\frac{\Sigma_{i=1}^{N}(y_i - y_i')^2}{N}}. \tag{1}$$

For Track 2, the systolic RMSE between ground truth systolic blood pressure s and submitted systolic blood pressure s' is calculated first, and the diastolic RMSE between ground truth diastolic blood pressure d and submitted diastolic blood pressure d' is calculated. The final RMSE is the mean of systolic RMSE and diastolic RMSE as shown below.

$$RMSE_2 = 0.5\sqrt{\frac{\Sigma_{i=1}^{N}(s_i - s_i')^2}{N}} + 0.5\sqrt{\frac{\Sigma_{i=1}^{N}(d_i - d_i')^2}{N}}. \tag{2}$$

3. Proposed approaches

To ensure fair competition, only pre-registered teams with authorized IDs are included in the final performance evaluation and ranking. The leaderboard in both tracks are displayed in Table 1. We reached out to the top three teams in both tracks, requesting brief descriptions of their methods for inclusion in this review paper. These methods are detailed below.

3.1. Track 1

3.1.1. Team 'Face AI' (Agency for Science, Technology and Research)

The proposed solution includes two stages. In the pre-training stage, they 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 technique is utilized to train the model based on limited number of labeled facial videos. The overall framework is depicted in Fig.1.

They utilize an end-to-end framework based on PhysNet-large 3D-CNN model where a sequence of face frames is fed directly into the deep learning model. They 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 [33] 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 x 128.

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

[11] https://vitalvideos.org/

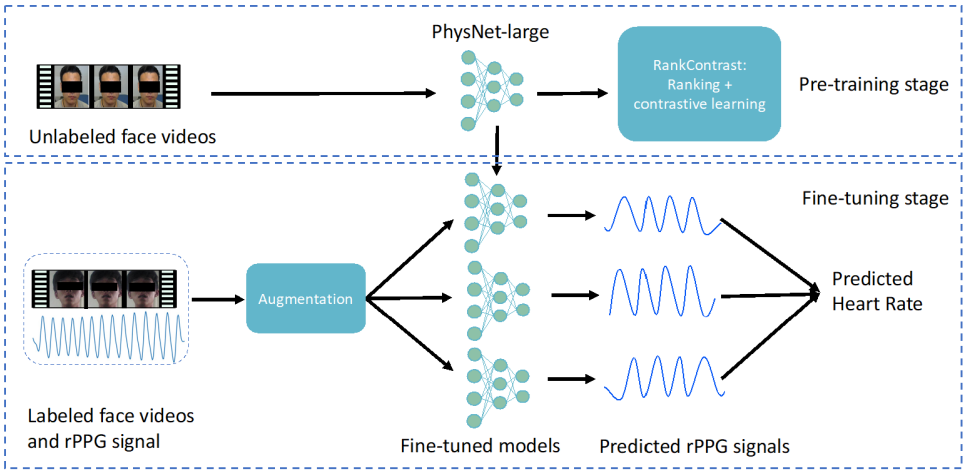

Figure 1: Overall framework for Team Face AI in track 1.

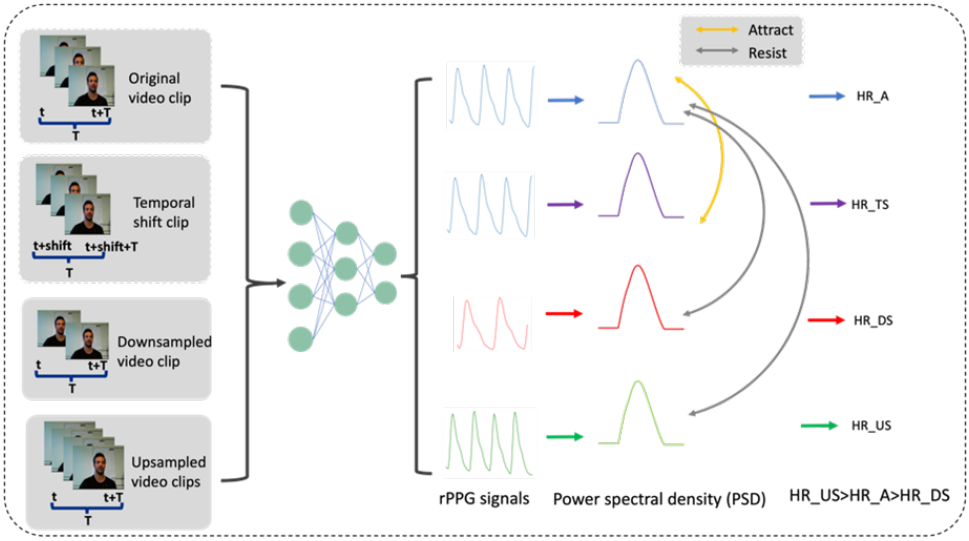

Figure 2: Proposed RankContrat Method for Team Face AI in track 1

the clips will increase the heart rate [34]. According to these characteristics, a ranking loss function is designed to extract features with upsampling and downsampling of the video clips.

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 as negative pairs.

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. They adopt two supervised loss functions: the classification loss and the Pearson loss to guide the learning process.

3.1.2. Team 'HFUT-VUT' (Hefei University of Technology)

The team HFUT-VUT participated in Track 1, and they presented two self-supervised HR estimation solutions that integrate spatial-temporal modeling and contrastive learning, respectively. They first propose a non-end-to-end self-supervised HR measurement framework (solution 1) based on spatial-temporal modeling. Meanwhile, they employ complementaryly an excellent end-to-end solution based

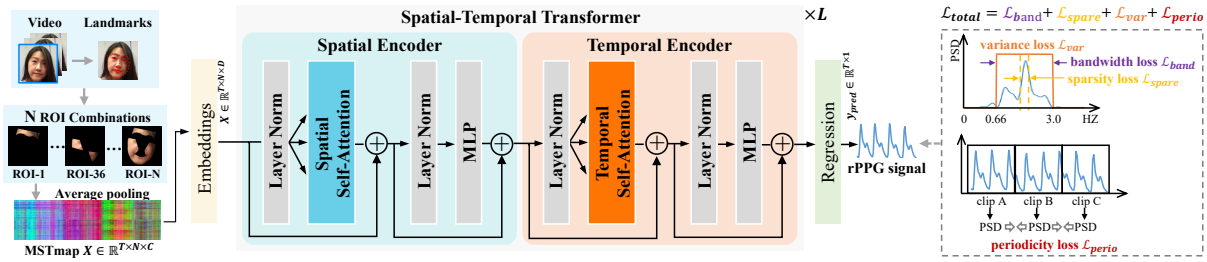

Figure 3: Overview of the proposed solution 1 of team HFUT-VUT in Track 1.

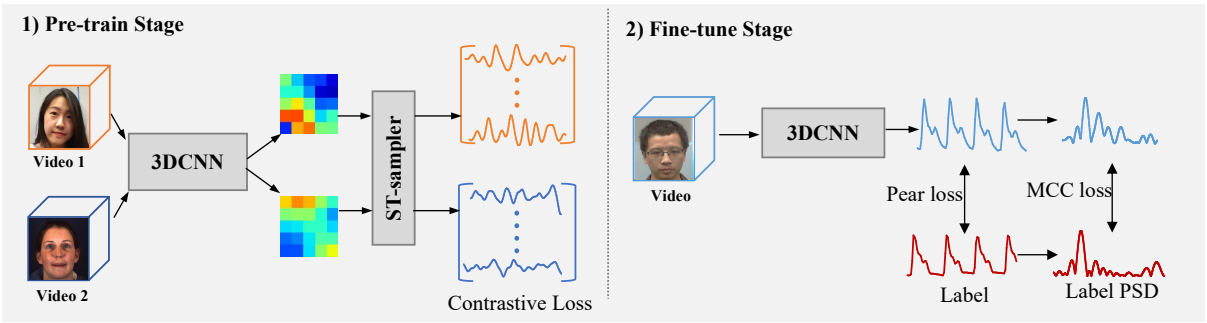

Figure 4: Overview of the solution 2 of team HFUT-VUT in Track 1.

on contrastive learning (solution 2). Finally, they combine the strengths of the above solutions through an ensemble strategy to generate the final predictions.

Solution 1. This solution is a non-end-to-end self-supervised HR measurement framework based on a spatial-temporal Transformer to capture subtle rPPG clues. The overview of this solution is illustrated in Figure 3. The method contains three steps. 1) Data pre-processing: The raw facial video is first transformed into MSTmap to suppress the irrelevant background and noise features while retaining most of the temporal characteristics of the periodic physiological signals. 2) Spatial-Temporal Transformer: Inspired by Dual-TL [35], a spatial-temporal Transformer is proposed to perceive the temporal and spatial correlations. It includes two encoders (spatial encoder and temporal encoder) to refine the ROI representation containing rPPG clues by capturing long-term spatiotemporal contextual information. 3) Self-supervised Loss: In this solution, they employ four self-supervised loss functions by incorporating prior of rPPG bandwidth and periodicity [18]. A bandwidth loss $\mathscr{L}_{band}$ is first adopted to penalize the model for producing signals that exceed the healthy HR bandwidth limits. Then, a sparsity loss $\mathscr{L}_{sparse}$ is adopted to emphasize the periodic heartbeats by suppressing non-heartbeat frequencies. To avoid the model collapsing to a specific frequency, they use a variance loss $\mathscr{L}_{var}$ to spread the variance of the power spectral density into a uniform distribution over the desired frequency band. Besides, a periodicity loss $\mathscr{L}_{perio}$ is proposed to avoid abnormal periodic fluctuations of the predicted signal, thereby ensuring temporal periodicity consistency.

Solution 2. This solution provides the end-to-end self-supervised HR measurement framework the Contrast-Phys+ [20]. The framework is depicted in Figure 4 and consists of three steps. 1) Data pre-processing: Firstly, face detection is performed using MTCNN [36] to obtain the face bounding box. The face video is then cropped to 1.5 times the size of the bounding box and resized to 128×128. Subsequently, each video is segmented into clips, and frame differencing is applied to generate normalized difference frames as input to the model. 2) Pre-training: Following the setup of [20], the 3DCNN-based PhysNet is used to obtain spatiotemporal rPPG (ST-rPPG) block representation. Observing the rPPG spatial and temporal similarity in [20], a contrastive loss is adopted to pull together the rPPG signals from the same ST-rPPG block and push away the signals from different ST-rPPG blocks extracted from different videos. 3) Fine-tuning: With the pre-trained 3DCNN-based PhysNet model, the model is then fine-tuned in a supervised manner. Specifically, given the predicted rPPG signal y_{pred} and the

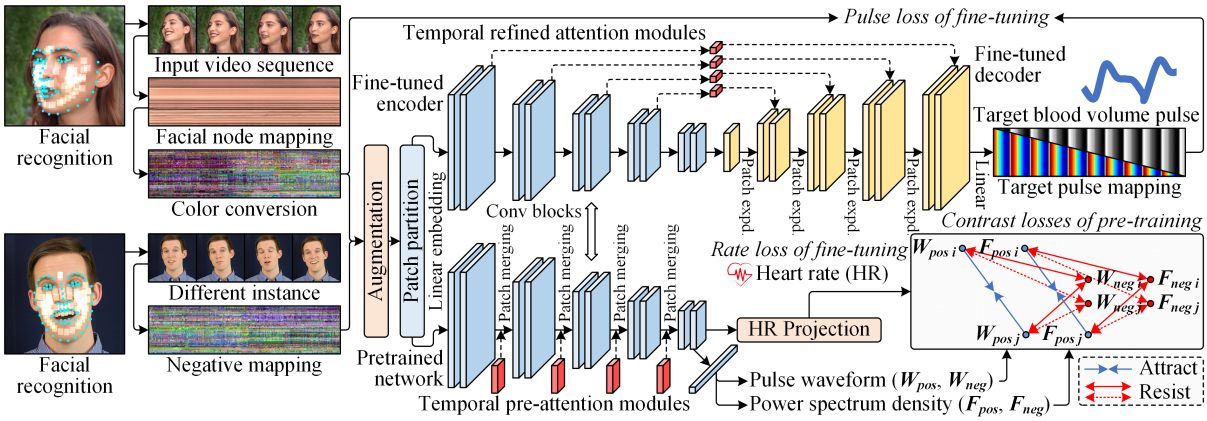

Figure 5: Method Diagram for PCA_Vital team in Track1

ground-truth PPG signal y_{gt}, the popular time domain-based Negative Pearson correlation (Pear) loss and frequency domain-based Negative max cross-correlation (MCC) [16] loss are selected to perform supervised training. The MCC is robust to temporal offsets in the ground truth, which can make up for the Pear loss.

3.1.3. Team 'PCA_Vital' (Nanjing University of Science and Technology)

The team PCA_Vital participated in Track 1 of self-supervised heart rate sensing, and they used a method based on contrastive learning and spatiotemporal reconstruction to learn heart rate from unlabeled facial videos. The framework of the proposed method is shown in Fig. 5.

First, to overcome the redundant skin information, they designed a novel regions of interest extraction method that focuses on facial muscles and capillary-rich areas while ignoring the interference of explicit edges, corners, and textures. They converted the video segment into spatiotemporal mapping, independently performed temporal normalization on each sub-block feature dimension, and then performed YUV color space conversion to mine the subtle color changes of blood volume pulsation feedback in unlabeled facial videos. This process can harvest certain rhythm and color variation characteristics in the preprocessing and enhancement stages without relying on a learning model.

Second, after converting the input video clips into spatiotemporal mappings, they guided inter-instance and intra-instance contrastive learning by enriching positive and negative sample pairs during the pre-training stage. They constructed positive and negative sets between different individual instances, and randomly reorganized and reselected these samples at the feature point level to increase diversity. Then, they constructed an encoder to extract features from the input samples, obtained waveform and frequency features, and respectively calculated the contrast correlation and power spectral density to bring the representation of the same instance closer and different instances farther apart.

Finally, they improved the traditional remote photoplethysmography regression into spatiotemporal reconstruction, and further improved the robustness of the model by focusing on the interaction of temporal features between different sub-regions of the face in the fine-tuning stage. The fine-tuning stage uses a U-shaped network as the backbone to constrain waveform reconstruction, and extracts intermediate layer feature features to construct a mapping of the same scale as the target pulse label. In addition, they embedded a series of temporal attention modules at the skip-layer connections of the U-shaped structure, calculated the global self-attention scores within the encoder features, and concatenated them with the main path features to the decoder.

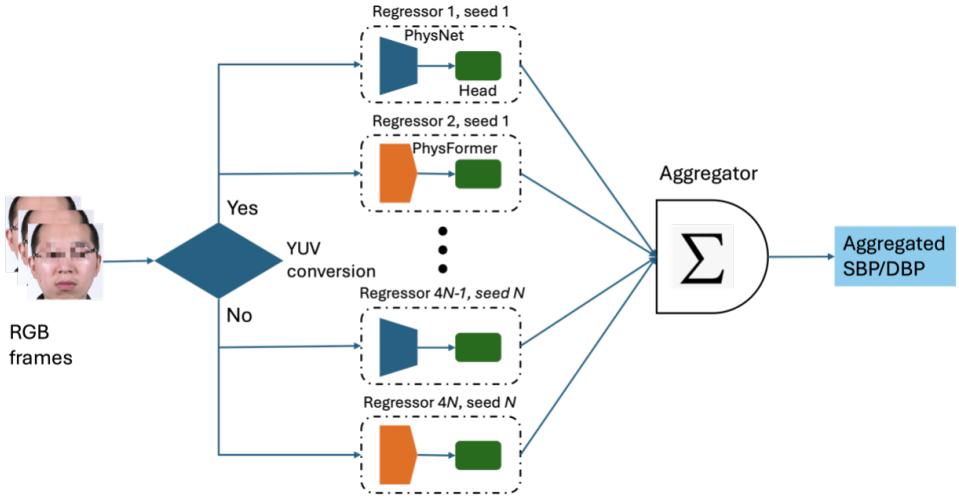

Figure 6: Overall framework for Team Face AI (BP) in Track 2.

3.2. Track 2

3.2.1. Team 'Face AI (BP)' (Agency for Science, Technology and Research)

The overall framework of their ensemble deep learning method is illustrated in Fig. 6, from which we can see that there are multiple regression models. To import diversity, multiple models are trained using different input feature vectors, backbones, or random seeds. The outputs of individual models are then fused with an aggregator.

Data Preprocessing: A short clip is extracted from the original full video and then partitioned into frames. They select the clip closest to the time when blood pressure (BP) is measured to mitigate the impact of BP fluctuation during video taking. If the video is recorded before BP measurement, the last part the video is selected and vice versa for videos taken after BP measurement. The face region of each frame is then cropped and resized to 128×128. To improve model performance in different lighting conditions, data augmentation technique is applied during the training process. As it has been demonstrated in [29], [37] that alternative color spaces derived from RGB videos are beneficial for better representation of HR signal, they also explored using YUV color space for BP estimation other than the original RGB space.

Network Structure: They utilize two state-of-the-art models as the backbone for theyr BP estimation model, including a 3D CNN model named PhysNet [11] and a transformer-based model named PhysFormer [38]. They keep all the layers of the backbones so that the output of the backbone remains as the PPG signal. Then, they stack a regression head with one hidden layer on top of the backbone and the regression head has two output nodes corresponding to systolic BP (SBP) and diastolic BP (DBP), respectively. The average RMSE of SBP and DBP is used as the loss function to train their models.

3.2.2. Team 'PCA_Vital' (Nanjing University of Science and Technology)

The team PCA_Vital participated in Track 2 of facial video-based remote blood pressure measurement, for witch they used a method based on convolutional neural network and random forest feature fusion. The framework of the proposed method is shown in Fig. 7. First, they extract the blood volume pulse signal that changes with optical reflectance from the input visible light facial video clips based on the pixel-level chromaticity transformation information. Then, they combined residual convolution, local and global attention mechanisms to design a convolutional neural network for remote blood pressure measurement, named RBP-CNN, to learn the blood pressure relationship information implicit in the blood volume pulse in spatial and temporal dimensions. At the same time, they also captured the prior information of the participants' body mass index and age from the facial video frames, and calculated the corresponding heart rate value based on the blood volume pulse. In this process, they found that

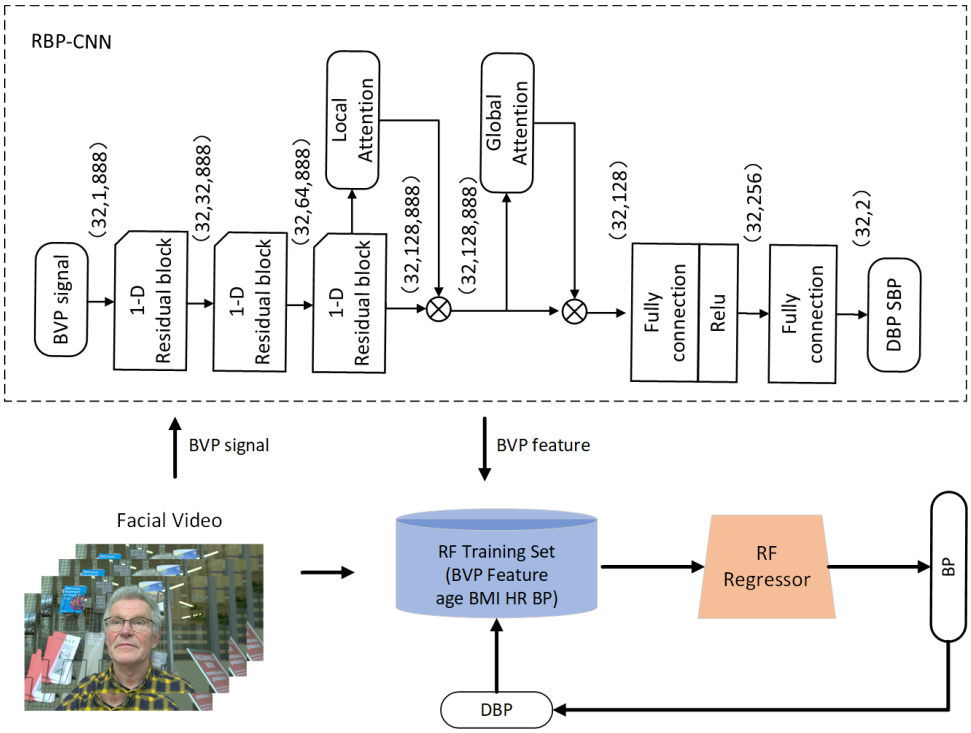

Figure 7: Method Diagram for PCA_Vital team in Track2.

there was a strong correlation between diastolic and systolic blood pressure and utilized diastolic blood pressure for systolic one prediction. Finally, they used an ensemble learning strategy and a random forest manner to fuse multiple features to achieve blood pressure measurement, and employed the feature importance of random forest to verify the rationality of the proposed remote detection approach.

3.2.3. Team 'Rhythm' (University of Science and Technology Beijing)

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. Therefore, they 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 Fig. 8, the overall framework of the proposed method mainly consists of three components: 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. The BVP Network is based on RhythmMamba, which constrains a state space model across multiple temporal scales in both the temporal and frequency domains. This approach maintains linear computational complexity while possessing the capability for long-range dependency modeling. They aim to refine the granularity of pulse wave reconstruction through long-range dependency modeling, thereby improving the accuracy of blood pressure estimation. Finally, the BP Network estimates BP values based on the two BVP waves, primarily utilizing the convolutional neural network.

4. Challenge results and discussion

The main results and ranking in the two competition tracks are summarized and shown in Table 1. In this section we also provide more detail statistics of the results for both tracks.

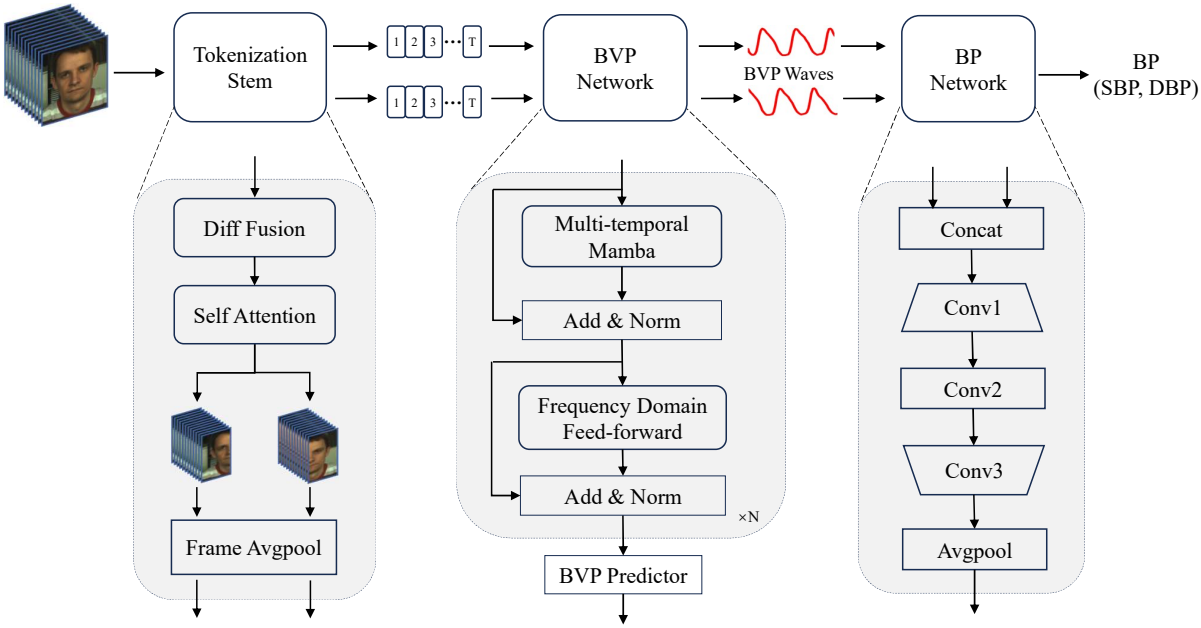

Figure 8: Method Diagram for Rhythm team in Track2

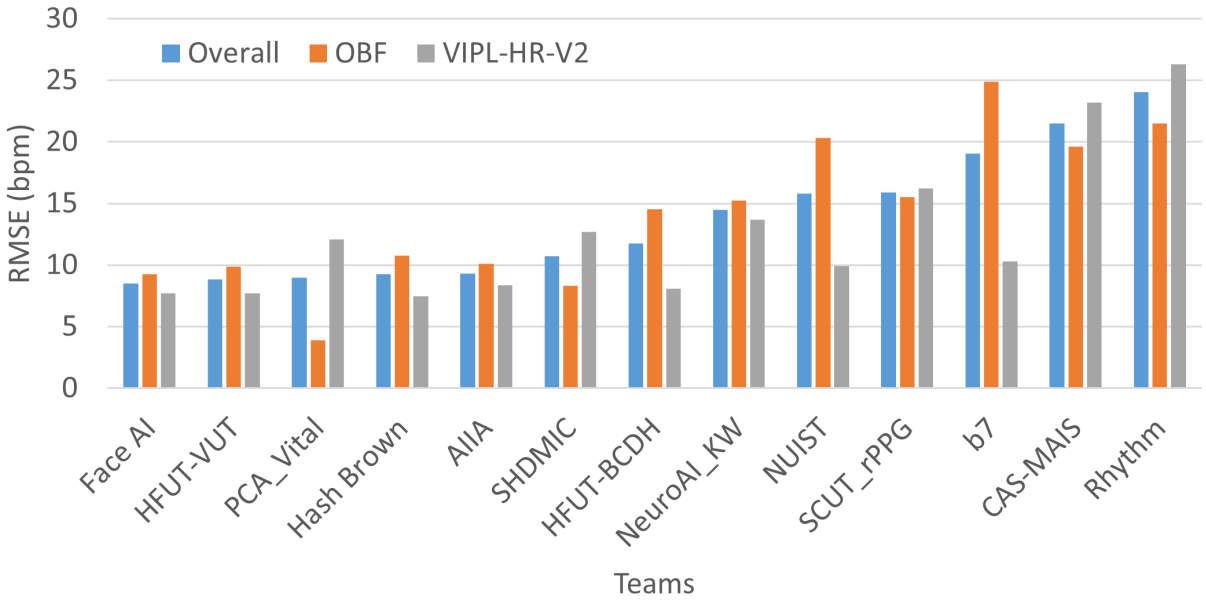

Figure 9: RMSE results on the overall test set, OBF test partition, and VIPL-HR-V2 test partition for all teams in Track 1.

4.1. Track1 result analysis

Fig. 9 showcases the Root Mean Square Error (RMSE) results for various teams in track 1. The results are divided into three categories: overall performance (blue bars), performance on the OBF dataset (orange bars), and performance on the VIPL-HR-V2 dataset (gray bars). This structured approach allows for a detailed analysis of how well different teams performed across diverse datasets.

Examining the overall performance, it is evident that teams with high rankings like ”Face AI” demonstrated consistently low RMSE values, indicating their strong overall performance. On the other hand, teams with lower rankings such as ”Rhythm” and ”CAS-MAIS” exhibited higher RMSE values, suggesting that their models were less accurate in heart rate measurements compared to others in the competition.

Table 1
The final leaderboard of the 3rd challenge of RePSS.

(1) Track 1

Ranking	Team Name	Organization	Score
1	Face AI	Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR)	8.50693
2	HFUT-VUT	Hefei University of Technology	8.85277
3	PCA_Vital	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	10.74201
7	HFUT-BCDH	Hefei University of Technology	11.77657
8	NeuroAI_KW	Kwangwoon University	14.4793
9	NUIST	Nanjing University of Information Science and Technology	15.7968
10	SCUT_rPPG	South China University of Technology	15.88228
11	b7	University of Science and Technology of China	19.06485
12	CAS-MAIS	Institute of Automation, Chinese Academy of Sciences	21.48006
13	Rhythm	University of Science and Technology Beijing	24.0241

(2) Track 2

Ranking	Team Name	Organization	Score
1	Face AI(BP)	Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR)	12.95258
2	PCA_Vital	Nanjing University of Science and Technology	13.48281
3	Rhythm	University of Science and Technology Beijing	13.59307
4	SCUT_rPPG	South China University of Technology	15.06056
5	IAI-USTC	University of Science and Technology of China	16.01179
6	NeuroAI	Kwangwoon University	16.56091

When focusing on the OBF dataset specifically, teams such as ”PCA_Vital” and ”SHDMIC” performed particularly well, with notably low RMSE values. This indicates that their models were highly effective at processing the data characteristics inherent to the OBF dataset. Conversely, other teams had higher RMSE values on the OBF dataset, reflecting challenges in adapting their models to the OBF dataset when fine-tunined on VIPL-HR-V2. This variability points to the importance of dataset-specific tuning and the potential difficulty in developing models that can handle a wide range of input variations.

In terms of performance on the VIPL-HR-V2 dataset, teams like ”HFUT-BCDH” and ”Face AI” excelled, achieving lower RMSE values. Their success suggests effective utilization of the VIPL-HR-V2 dataset’s characteristics for fine-tuning. In contrast, teams like ”CAS-MAIS” and ”Rhythm” exhibited the high RMSE in this category, indicating potential difficulties in leveraging the challenging VIPL-HR-V2 dataset for precise heart rate measurement.

The competition results underscore the importance of consistency and robustness in model performance. Teams with consistently low RMSE across both datasets, such as ”Face AI” and ”HFUT-VUT,” likely developed more robust models capable of generalizing well across different facial video data. This indicates that their pre-training and fine-tuning stages effectively captured the underlying features necessary for accurate heart rate measurement.

Certain teams exhibited strong performance on one dataset but not the other. For example, ”PCA_Vital” showed excellent results on the OBF dataset but struggled significantly with the VIPL-HR-V2 dataset. This disparity could be due to differences in video quality, lighting conditions, or variations in facial expressions and movements between the datasets. Such differences highlight the importance of diverse and comprehensive pre-training data to ensure models can handle various real-world conditions.

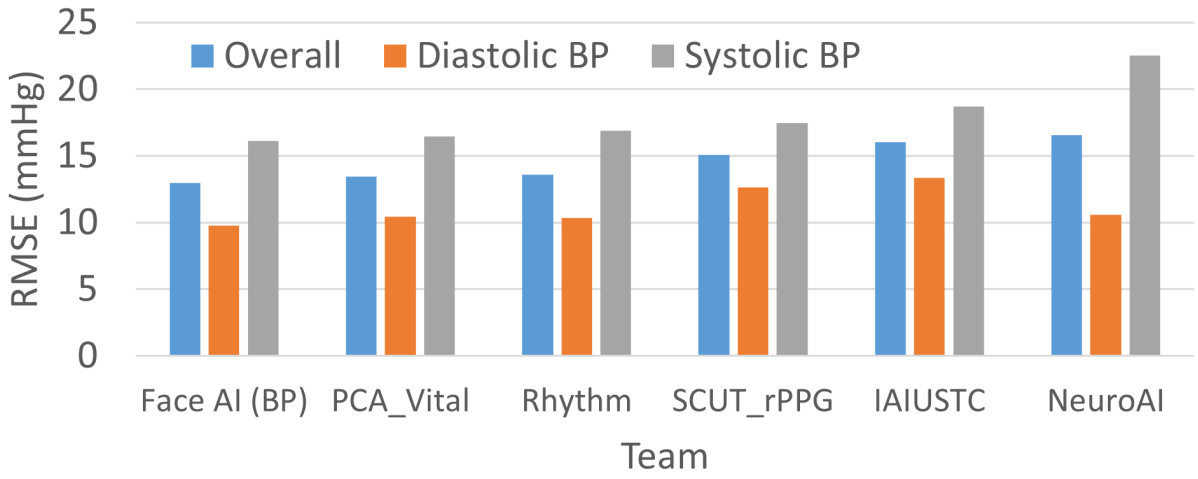

Figure 10: The overall BP RMSE results along with systolic and diastolic BP RMSE for all teams in Track 2.

Table 2
The cummulative percentage of errors (CPE) of diastolic and systolic BP and the corresponding British Hypertension Society (BHS) grade for all teams. The BHS grade standard is shown in Table 3.

Teams	Diastolic BP				Systolic BP			
	CPE5	CPE10	CPE15	BHS Grade	CPE5	CPE10	CPE15	BHS Grade
Face AI (BP)	38%	75.50%	90.50%	D	22.50%	44%	64.50%	D
PCA_Vital	42.50%	73.50%	90.50%	C	21.50%	42.50%	63%	D
Rhythm	41.50%	75.50%	92%	C	23.50%	44.50%	66%	D
SCUT_rPPG	38%	63.50%	84.50%	D	25.50%	44.50%	68.50%	D
IAIUSTC	30.50%	61%	78.50%	D	25%	44%	58.50%	D
NeuroAI	42.50%	74%	90.50%	C	13%	27.50%	44%	D

Table 3
The British Hypertension Society (BHS) grade standard.

	Grade A	Grade B	Grade C	Grade D
CPE5	>=60%	>=50%	>=40%	<40%
CPE10	>=85%	>=75%	>=65%	<65%
CPE15	>=95%	>=90%	>=85%	<85%

4.2. Track2 result analysis

The competition results for facial video-based blood pressure (BP) measurement reveal variations in performance among the participating teams, as shown by their root mean square error (RMSE) and cumulative percentage of errors (CPE) for diastolic and systolic BP presented in Fig. 10 and Table 3.

In terms of overall RMSE for BP measurement, the team Face AI (BP) exhibited the lowest overall RMSE, indicating the most accurate performance among the teams. Focusing on diastolic BP RMSE, Face AI (BP) still achieved the lowest error, underscoring their strong performance in this specific metric, while other teams, such as NeuroAI and IAIUSTC, had relatively higher RMSE values. For systolic BP RMSE, NeuroAI showed the highest error, indicating less accurate performance in systolic BP. Face AI (BP) again performed well, followed by PCA_Vital and Rhythm. When comparing the RMSE between diastolic and systolic BP, diasotlic BP RMSE is always lower than systolic BP RMSE, which was also observed in the contact PPG BP research [39, 40].

The cumulative percentage of errors (CPE) and corresponding British Hypertension Society (BHS)

grades in Table 3 provide further insight into the teams' performance. he CPE5, CPE10, and CPE15 values reflect the percentage of errors within 5, 10, and 15 mmHg, respectively. For diastolic BP, PCA_Vital, Rhythm, and NeuroAI achieves BHS grade C while others achievs the lowest grade D. For systolic BP, all teams fell into the lowest BHS grade D. The results suggest that while there is notable variation in the performance of different teams, all exhibit relatively high errors as evidenced by the BHS grades. Since Grade A and B are recommended for clinical use, the rPPG-based BP estimation from the teams of track 2 still needs performance improvement to achieve clinical use.

The results across all teams, especially in systolic BP measurements, highlight the complexity of accurately estimating BP from facial videos. Enhancements in video preprocessing, feature extraction, and model training could help improve performance. Additionally, incorporating more diverse datasets for training could help models generalize better to the test set.

5. Conclusion and future directions

As a continuous event, the 3rd RePSS challenge advanced beyond the 2nd and 1st RePSS in two key ways: 1) In Track 1, participants utilized self-supervised methods to pre-train models on unlabeled facial videos, unlike previous challenges that relied on supervised methods requiring labeled facial videos. 2) Track 2 introduced a new competition for rPPG-based blood pressure estimation, which necessitates high-quality rPPG signals for accurate blood pressure estimation. The 3rd RePSS challenge attracted more specialized research groups and led to the proposal of interesting approaches from the participating teams, potentially offering valuable insights for future research.

For track 1, the competition results highlight both the potential and the challenges of self-supervised learning for heart rate measurement. While some teams demonstrated impressive accuracy, there remains significant room for improvement, particularly in ensuring models generalize well across diverse datasets. The findings suggest that a focus on dataset diversity, advanced pre-training methods, and the exploration of multi-modal data could drive further advancements in this field. To further improve performance, future work could explore the integration of multi-modal data, such as combining facial video with other modalities like radar and infrared bands. Additionally, enhancing the diversity and quality of pre-training datasets could improve the pre-trained models.

For track 2, these blood pressure results underscore the need for further research and development in rPPG-based blood pressure measurement. While the competition showcases promising advancements in facial video-based BP measurement, the results indicate substantial room for improvement before these methods can be considered reliable for clinical or real-world applications. Future competitions could also focus on rPPG signal waveform evaluation, which is the fundamental to BP estimation .

Acknowledgments

This work was supported by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (grants 336116, 345122), ICT 2023 project TrustFace (grant 345948), the University of Oulu & Research Council of Finland Profi 7 (grant 352788), Infotech Oulu, and National Natural Science Foundation of China (grant 62176249). The authors would like to acknowledge Pieter-Jan Toye for providing data in track 2 of the RePPS challenge. The authors also acknowledge CSC-IT Center for Science, Finland, for providing computational resources.

References

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

[2] M.-Z. Poh, D. J. McDuff, R. W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation., Opt. Express 18 (2010) 10762–10774.

[3] M.-Z. Poh, D. J. McDuff, R. W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam, IEEE Trans. Biomed. Eng. 58 (2011) 7–11.
[4] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE Trans. Biomed. Eng. 60 (2013) 2878–2886.
[5] X. Li, J. Chen, G. Zhao, M. Pietikainen, Remote heart rate measurement from face videos under realistic situations, in: Proc. IEEE CVPR, 2014, pp. 4264–4271.
[6] D. McDuff, S. Gontarek, R. W. Picard, Improvements in remote cardiopulmonary measurement using a five band digital camera, IEEE Trans. Biomed. Eng. 61 (2014) 2593–2601.
[7] W. Wang, A. C. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote ppg, IEEE Trans. Biomed. Eng. 64 (2017) 1479–1491.
[8] G. Balakrishnan, F. Durand, J. Guttag, Detecting pulse from head motions in video, in: Proc. IEEE CVPR, 2013, pp. 3430–3437.
[9] A. V. Moco, S. Sander, G. de Haan., Ballistocardiographic artifacts in ppg imaging, IEEE Trans. Biomed. Eng. 63 (2016).
[10] C. Yang, G. Cheung, V. Stankovic, Estimating heart rate and rhythm via 3D motion tracking in depth video, IEEE Trans. Multimedia 19 (2017) 1625–1636.
[11] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, Proc. BMVC (2019).
[12] Z. Yu, W. Peng, X. Li, X. Hong, G. Zhao, Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement, in: Proc. IEEE ICCV, 2019.
[13] W. Chen, D. Mcduff, Deepphys: Video-based physiological measurement using convolutional attention networks, Proc. ECCV (2018) 356–373.
[14] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, X. Chen, Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography, IEEE Journal of Biomedical and Health Informatics 25 (2021) 1373–1384. doi:10.1109/JBHI.2021.3051176.
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] Y. Yang, X. Liu, J. Wu, S. Borac, D. Katabi, M.-Z. Poh, D. McDuff, Simper: Simple self-supervised learning of periodic targets, in: The Eleventh International Conference on Learning Representations, 2023.
[20] 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).
[21] 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.
[22] 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, 2021, pp. 2404–2413.
[23] L. Xie, X. Wang, H. Zhang, C. Dong, Y. Shan, Vfhq: A high-quality dataset and benchmark for video face super-resolution, in: The IEEE Conference on Computer Vision and Pattern Recognition

Workshops (CVPRW), 2022.
[24] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, FaceForensics++: Learning to detect manipulated facial images, in: International Conference on Computer Vision (ICCV), 2019.
[25] L. Jiang, R. Li, W. Wu, C. Qian, C. C. Loy, DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection, in: CVPR, 2020.
[26] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, C. C. Loy, CelebV-HQ: A large-scale video facial attributes dataset, in: ECCV, 2022.
[27] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, J. F. Cohn, Disfa: A spontaneous facial action intensity database, IEEE Transactions on Affective Computing 4 (2013) 151–160.
[28] S. Petridis, B. Martinez, M. Pantic, The mahnob laughter database, Image and Vision Computing 31 (2013) 186–202.
[29] X. Niu, S. Shan, H. Han, X. Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, IEEE Trans. Image Processing (2019).
[30] 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.
[31] P.-J. Toye, Vital videos: A dataset of videos with ppg and blood pressure ground truths, arXiv preprint arXiv:2306.11891 (2023).
[32] 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: Proc. IEEE FG, 2018, pp. 1–6.
[33] 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.
[34] 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.
[35] W. Qian, D. Guo, K. Li, X. Zhang, X. Tian, X. Yang, M. Wang, Dual-path tokenlearner for remote photoplethysmography-based physiological measurement with facial videos, IEEE Transactions on Computational Social Systems (2024).
[36] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE signal processing letters 23 (2016) 1499–1503.
[37] 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 (2023).
[38] 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.
[39] M. Kachuee, M. M. Kiani, H. Mohammadzade, M. Shabany, Cuffless blood pressure estimation algorithms for continuous health-care monitoring, IEEE Transactions on Biomedical Engineering 64 (2016) 859–869.
[40] Z.-D. Liu, Y. Li, Y.-T. Zhang, J. Zeng, Z.-X. Chen, Z.-W. Cui, J.-K. Liu, F. Miao, Cuffless blood pressure measurement using smartwatches: a large-scale validation study, IEEE Journal of Biomedical and Health Informatics 27 (2023) 4216–4227.