The 4th Vision-based Remote Physiological Signal Sensing (RePSS) Challenge & Workshop

Huiyu Yang[1], Yunchi Zhang[2], Chenhang Ying[3], Youchen Luo[3], Jieyi Ge[3], Antitza Dantcheva[4], Shiguang Shan[2], Guoying Zhao[1], Hu Han[2] and Xiaobai Li[3,1,*]

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

Abstract

Remote photoplethysmography (rPPG) is a non-contact technique for estimating physiological signals—such as heart rate—from subtle color changes in facial videos. While recent advances in rPPG have predominantly relied on RGB videos, these signals are highly susceptible to variations in environmental lighting and individual skin tones, which limits their robustness in real-world scenarios. In contrast, near-infrared (NIR) videos are less affected by illumination changes and dark skin-tones. So, to improve the accuracy and reliability of rPPG measurements, the 4th RePSS challenge is held to promote the novel multimodal fusion strategies which combines the RGB and NIR videos for heart rate (HR) prediction. It is held in conjunction with the International Joint Conference on Artificial Intelligence (IJCAI 2025). This paper provides an overview of the challenge, showing details about the track setting, the data and protocols, the proposed approaches, and the results and discussions. The top-performing solutions are analyzed to provide valuable insights and guide future research directions in the field.

Keywords

rPPG, remote physiological signal measurement, multimodal fusion, facial video, heart rate

1. Introduction

Remote photoplethysmography (rPPG) is a non-contact technique for measuring physiological signals, such as heart rate (HR), by analyzing subtle color changes in facial videos caused by blood volume fluctuations. Unlike traditional contact-based photoplethysmography (PPG) which relies on physical sensors attached to the skin (e.g., fingertip or wrist), rPPG enables vital sign monitoring without direct contact, offering a more comfortable and unobtrusive experience. This makes rPPG particularly suitable for a variety of application scenarios, including remote healthcare and monitoring [1, 2, 3], affective computing [4, 5, 6], and remote education [7, 8].

Over the past decade, rPPG research has evolved significantly. Verkruysse et al. [9] first proposed a rPPG measurement based on the green channel of facial videos. Then, traditional methods primarily relied on hand-crafted signal processing techniques, such as ICA [10], CHROM [11] and POS [12], which apply color space transformations and bandpass filtering to extract pulse signals from RGB video streams. However, these traditional approaches often struggle under uncontrolled conditions due to sensitivity to motion, lighting variations, and differences in skin tone. To overcome these challenges, recent research has increasingly adopted machine learning and deep learning techniques. Models based on 2D-CNN [13, 14, 15], 3D-CNN [16, 17, 18], and transformer [19, 20, 21] architectures

The 4th Vision-based Remote Physiological Signal Sensing (RePSS) Challenge & Workshop, August 2025, Montreal, Canada, Satelite Event in Guangzhou, China
*Corresponding author.
Huiyu.Yang@oulu.fi (H. Yang); zhangyunchi19@mails.ucas.ac.cn (Y. Zhang); chying@zju.edu.cn (C. Ying); youchen.luo@zju.edu.cn (Y. Luo); jyge@zju.edu.cn (J. Ge); antitza.dantcheva@inria.fr (A. Dantcheva); sgshan@ict.ac.cn (S. Shan); guoying.zhao@oulu.fi (G. Zhao); hanhu@ict.ac.cn (H. Han); xiaobai.li@zju.edu.cn (X. Li)
0000-0003-0107-7029 (A. Dantcheva); 0000-0002-8348-392X (S. Shan); 0000-0003-3694-206X (G. Zhao); 0000-0001-6010-1792 (H. Han); 0000-0003-4519-7823 (X. Li)

have been explored to extract visual cues for accurate rPPG estimation. Many data-driven models aim to automatically learn complex spatial-temporal patterns to extract physiological features directly from video data [22, 23, 24], demonstrating improved generalizability and robustness. But, despite the progress, significant challenges remain, particularly in improving the accuracy and robustness under various environments and expanding the applicability of rPPG systems across diverse populations.

To bring together multidisciplinary researchers and further advance rPPG technologies, the Remote Physiological Signal Sensing (RePSS) Challenges have been held annually as a platform to promote the remote physiological signals measurement—such as heart rate (HR), heart rate variability (HRV), and blood pressure (BP)—from facial videos. Since 2020, RePSS challenges have been organized in conjunction with top-tier conferences such as CVPR 2020 [25] [1], ICCV 2021 [26] [2] , and IJCAI 2024 [27] [3] [4]. Through workshops, challenges, and invited talks, the initiative aims to foster technological innovation, resource sharing, and fair competition in the remote physiological sensing field.

This year, the 4th RePSS Challenge and Workshop is held in conjunction with the International Joint Conference on Artificial Intelligence (IJCAI 2025) in August 2025. It is organized on the Kaggle website [5]. The theme centers on "RGB-NIR Fusion for Robust rPPG Measurement", which invites participants to develop innovative data fusion techniques that integrate RGB and Near-Infrared (NIR) facial videos to enhance the accuracy and robustness of rPPG estimation. This focus is motivated by the limitations of conventional RGB-based rPPG methods, which are often sensitive to motion artifacts, lighting changes, and skin tone variations. In contrast, NIR imaging is more robust under such challenging conditions while still retaining meaningful physiological information, which could benefit RGB-based rPPG methods. However, despite the promising potential, multimodal fusion of RGB-NIR videos for rPPG measurement has been rarely explored in existing literature [28, 29]. This gap highlights an important but under-investigated research direction, particularly for real-world applications where single-modality solutions often fall short. By incorporating NIR data, this challenge aims to stimulate the development of advanced multimodal fusion strategies that can overcome current limitations and advance the field of remote physiological monitoring toward greater reliability and broader applicability.

The paper is structured as follows: Section 2 presents an overview of the 4th RePSS Challenge, including details on the track setting, dataset and protocols, and evaluation metrics. Section 3 outlines the top-ranked solutions submitted by participating teams. Section 4 discusses the results and provides analysis and insights. Finally, Section 5 concludes the paper and highlights potential future research directions in this field.

2. Challenge Overview

2.1. Track setting

The 4th RePSS Challenge focuses on the multimodal fusion of RGB and NIR videos for robust heart rate (HR) measurement. Participants are provided with a portion of the VIPL-HR dataset [22, 30] as the training set, which consists of 10-second RGB-NIR video clips from 107 subjects. During the evaluation phase, RGB-NIR video pairs from 100 subjects in the VIPL-HR-v2 dataset and another 100 subjects from the OBF [31] dataset are each segmented into three 10-second clips to form the test set. This task aims to highlight the potential of innovative RGB-NIR fusion strategies to improve the accuracy and reliability of rPPG-based HR estimation, particularly in challenging real-world conditions.

(a) (b)

Figure 1: (a) Illustrations of the VIPL-HR database in terms of illumination condition, head movement, and acquisition device. Image from [30]. (b) A test sample pair from the OBF [31] dataset.

2.2. Data and protocol

2.2.1. Training data

To support the development of multimodal rPPG algorithms, the 4th RePSS Challenge utilizes part of the VIPL-HR database as its training data. VIPL-HR is a large-scale, multimodal database designed for non-contact heart rate estimation from facial videos captured under diverse and less-constrained real-world conditions. The entire dataset comprises 3,130 face videos from 107 subjects, including 2,378 RGB videos and 752 NIR videos. Each video is accompanied by synchronized physiological measurements such as heart rate, blood oxygen saturation (SpO2), and blood volume pulse (BVP) signals.

As shown in Fig. 1(a), VIPL-HR simulates multiple real-world rPPG challenges by recording with different devices-including smartphones, RGB-D cameras, and webcams-under varying illumination conditions (only the ceiling lamp on, both the ceiling and filament lamps on, and both lamps off) and substantial head movements (large rotations, talking). As a publicly available dataset with multi-modality and rich condition diversity, VIPL-HR facilitates the development of practical remote HR detection approaches that generalize well to real-world challenges.

2.2.2. Test data

Data samples from the VIPL-HR-v2 and OBF [31] datasets are used as the test set, which includes paired RGB-NIR videos of 100 subjects from the reserved portion of the VIPL-HR-v2 dataset (different from the training set) and another 100 subjects from the OBF dataset. An example pair of sample images is shown in Fig. 1(b). For each RGB-NIR pair, three 10-second video clips are randomly selected, and the corresponding groun-truth heart rate values are kept for the final evaluation.

It is worth noting that the test data include challenging scenarios: the VIPL-HR-v2 dataset contains head motions and varying lighting conditions, while the OBF dataset includes subjects with diverse skin tones. Due to protocol restrictions, the OBF videos have been anonymized by applying mosaics to sensitive facial regions to safeguard individual privacy. To compensate for this anonymization, facial landmarks of the OBF videos generated by OpenFace are provided to the participants. These landmarks

[1]https://competitions.codalab.org/competitions/22287
[2]https://competitions.codalab.org/competitions/30855
[3]https://www.kaggle.com/competitions/the-3rd-repss-t1/data
[4]https://www.kaggle.com/competitions/the-3rd-repss-t2/data
[5]https://www.kaggle.com/competitions/the-4th-repss-t1/data

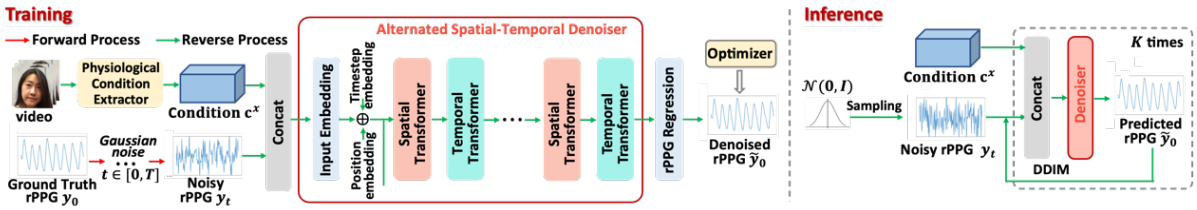

Figure 2: The overall framework of Team HFUT-VUT.

enable precise localization and analysis of relevant facial regions, ensuring effective model training and evaluation while adhering to privacy protection requirements.

2.3. Evaluation metric

The root mean squared errors (RMSE) is used as the evaluation metric. The RMSE value between the ground truth heart rates y and submitted heart rates prediction y' is calculated as follows, where N represents the amount of test samples:

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

3. Proposed approaches

To maintain consistency and fairness in the evaluation process, the final evaluation and ranking only include registered teams. The leaderboard is listed as Table 1. To get a better understanding of the top-ranked methods, brief methodological summaries were gathered from the top three teams and are presented in the subsequent sections.

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

The HFUT-VUT team introduces DiffRePSS, a diffusion-based framework for remote physiological signal sensing. Their method aims to measure remote photoplethysmography (rPPG) signals for heart rate (HR) estimation from facial videos, especially under real-world challenges including head movements and varying illumination. The overall framework of their proposed method is illustrated in Fig. 2.

Their approach is based on a conditional denoising diffusion model (DDIM), which progressively reconstructs clean rPPG signals from Gaussian noise in a step-wise reverse process. Instead of direct signal regression, the model learns to denoise latent representations iteratively, guided by visual physiological cues from the input video. To extract these cues, the team proposes a multi-scale spatial-temporal map (MSTmap) that captures chrominance dynamics across facial regions, together with a temporal difference representation to emphasize fine-grained pulsatile variations. These features are combined and fed into an alternating spatial-temporal Transformer denoiser, which models spatial and temporal dependencies effectively through the attention-based module. Additionally, to address the common issue of heart rate distribution imbalance in physiological datasets, a data augmentation strategy is introduced. It involves upsampling or downsampling MSTmaps along the temporal axis to simulate higher or lower HRs, which effectively expands the HR distribution during training.

In the evaluation phase, DiffRePSS demonstrates strong performance on both the VIPL-HR-v2 and OBF test sets, which include diverse subjects and challenging scenarios. The method achieves the best RMSE and ranks 1st in the competition, highlighting the potential of diffusion-based modeling for robust and accurate rPPG signal estimation.

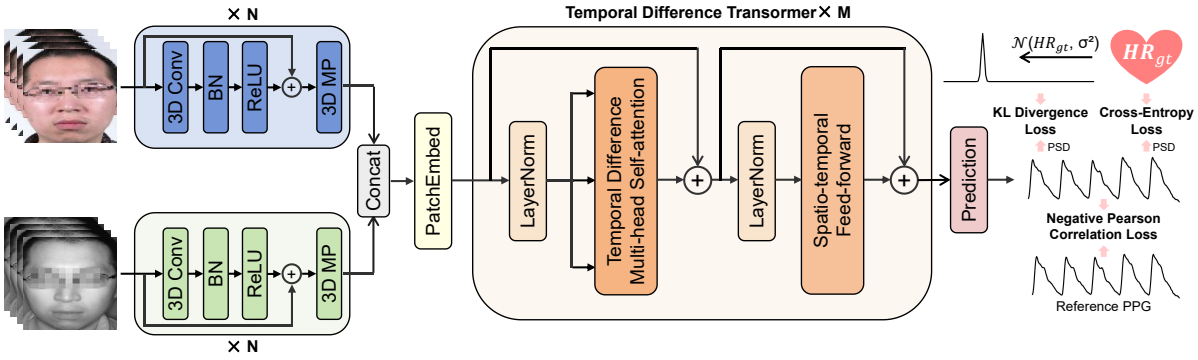

Figure 3: The overall framework of Team IST.

3.2. Team 'IST' (Nanjing University)

The IST team developes a robust multimodal framework for HR estimation based on the temporal difference transformer (TDT) [19], which includes spatio-temporal encoders for feature extraction, a TDT module with temporal difference multi-head self-attention (TD-MHSA) mechanism for multimodal fusion, and the final prediction module. The framework of the proposed method is shown in Fig. 3.

The method starts from processing RGB and NIR video sequences separately through two spatio-temporal encoders. Each encoder consists of multiple residual 3D convolutional blocks, which contains a Conv3D layer, a batch normalization layer, a ReLU activation layer, and a 3D max-pooling layer. These blocks capture both spatial and temporal patterns from the video. After encoding, features from two modalities are aligned in shape but maintain modality-specific semantics.

Subsequently, the TDT module is employed to enhance the quasi-periodic nature of rPPG signals and also to align and fuse the RGB and NIR representations adaptively. The TDT integrates a temporal difference multi-head self-attention (TD-MHSA) mechanism, which enhances standard self-attention with a temporal difference convolution (TDC) [19]. Local temporal changes are explicitly encoded by TDC, making the model more sensitive to periodic variations. For each attention head, query, key, and value matrices are derived from normalized input using TDC (for queries and keys) and standard convolution (for values). The outputs of all attention heads are aggregated and passed through a learnable projection layer, then reshaped back into spatio-temporal form. To further enhance local modeling, the transformer replaces traditional feed-forward layers with a spatio-temporal feed-forward (ST-FF) module. This module uses pointwise and depthwise 3D convolutions to capture fine-grained spatial and temporal patterns, improving representation ability for downstream tasks.

After fusion, a two-stage temporal upsampling module is introduced to increase the temporal resolution while reducing feature dimensionality. Then, a spatial average pooling operation reduces spatial dimensions, and a final regression head outputs a predicted rPPG signal. The HR is obtained by applying Fast Fourier Transform (FFT) to this signal and identifying the dominant frequency peak in the physiological range.

To train the model effectively, a composite loss function combining three objectives is utilized, which includes: (1) a negative Pearson correlation loss aligning the predicted rPPG waveform with the ground truth [17], (2) a KL divergence loss supervising frequency-domain distributions [32], and (3) a cross-entropy loss enhancing frequency-bin classification [33]. This composite loss ensures the model learns accurate and physiologically consistent predictions.

During evaluation, their proposed method achieve outstanding performance, ranking the 2nd on the leaderboard.

3.3. Team 'xjgroupscu' (Sichuan University)

The xjgroupscu team proposed algorithm consists of four steps as shown in Fig. 4, including: (1) ROI segmentation and raw rPPG signal acquisition, (2) Color space transformation, (3) Signal decomposition

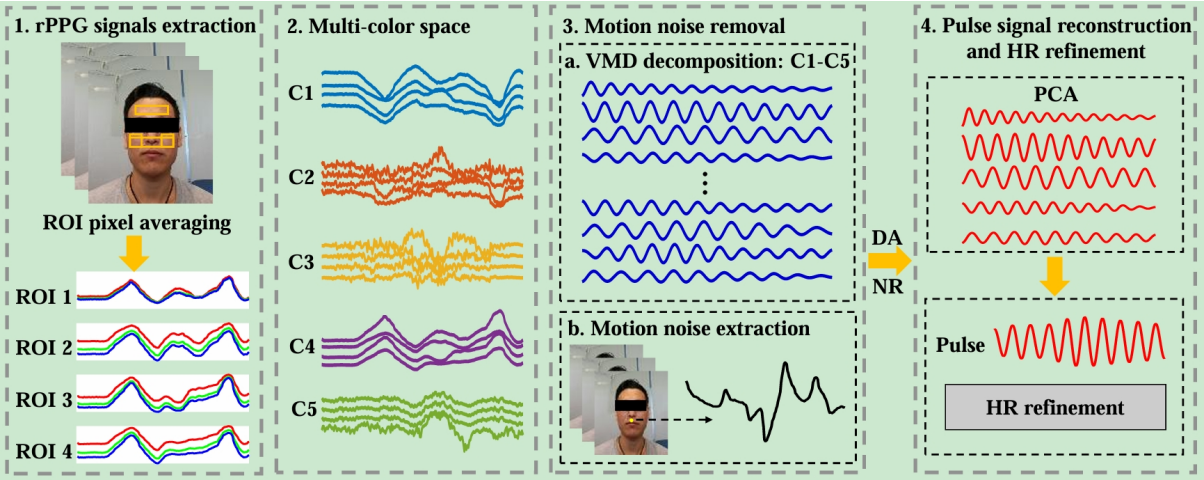

Figure 4: The overall framework of Team xjgroupscu.

and motion noise removal, and (4) Pulse reconstruction and heart rate refinement.

At the beginning of the proposed pipeline, facial landmarks are detected using the MediaPipe Face Mesh [34]. Based on these landmarks, regions of interest (ROIs) are defined, as illustrated in Fig. 4. Spatial averaging is then applied to the pixel values within each ROI to generate temporal signals. While RGB videos provide three-channel data, near-infrared (NIR) videos contain only a single channel. Subsequently, the signals are transformed into multiple channel representations, including: the combined RGB-NIR channel, CHROM-projected channel [11], POS-projected channel [12], the green channel, and the NIR channel. Next, each channel signal is decomposed using Variational Mode Decomposition (VMD) [35] to separate motion noise from pulse-related components. Motion-related noise is estimated based on the positional changes of a key facial point located at the center of the upper lip. The time delay between each decomposed signal component and the motion reference signal is computed. If the delay is below a predefined threshold, the component is considered motion-induced noise and is therefore discarded. Finally, Principal Component Analysis (PCA) [36] is applied to the remaining (cleaned) signals to reconstruct the pulse waveform, and heart rate is estimated from this signal using the Fast Fourier Transform (FFT).

Also, for each sample, three signal segments spaced 0.2 seconds apart are selected. A constraint is imposed to ensure that the heart rate (HR) difference among these segments does not exceed 10 bpm, which ensures the temporal consistency of rPPG signals. Furthermore, a heart rate probability constraint is applied to handle outliers: if the estimated HR falls outside the plausible physiological range of 35.6 to 122.6 bpm, the signal window is shifted forward frame by frame. This process continues until the updated HR estimate deviates by less than 5 bpm from the initial value. The final heart rate estimation is then obtained from this refined signal window.

In the evaluation phase, their proposed work achieves the 3rd place on the final leaderboard, demonstrating its accuracy and robustness under noisy situation.

4. Challenge results and discussion

The final leaderboard of the 4th RePSS Challenge is shown in Table 1. In this section, we provide a detailed analysis and discussion of the final results.

Performance across two datasets. The Root Mean Square Error (RMSE) results for all teams are presented in Fig. 5. The results are divided into three categories: overall performance on the entire test set (blue bars), performance on the OBF test partition (orange bars), and performance on the VIPL-HR-v2 test partition (grey bars). This division facilitates a detailed evaluation across diverse datasets, enabling a comprehensive understanding of the robustness and cross-dataset generalizability of the proposed methods.

Table 1
The final leaderboard of the 4th challenge of RePSS.

Ranking	Team Name	Affiliation	Score
1	HFUT-VUT	Hefei University of Technology	11.89505
2	IST	Nanjing University	12.31846
3	xjgroupscu	Sichuan University	12.70790
4	NJU_TEAM	Nanjing University	14.51449
5	Sgt.Pepper's	Hefei University of Technolog	14.69105
6	xiuxejia	Hefei University of Technology	16.25080
7	NUAA_AI	Nanjing University of Aeronautics and Astronautics	17.68826

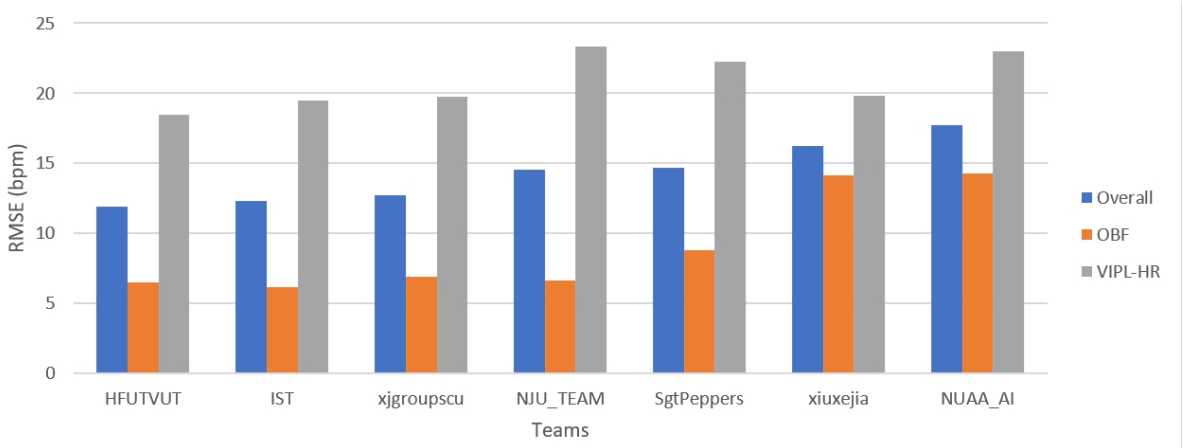

Figure 5: RMSE results of all teams on the overall test set, the OBF test partition, and the VIPL-HR-v2 test partition.

The proposed methods from the top three teams demonstrate relatively low RMSE values across all three settings, including the overall test set, the OBF partition, and the VIPL-HR-v2 partition, indicating their strong performance in video-based remote heart rate estimation. Interestingly, although all models were trained on the VIPL-HR training set, most teams achieved better performance on the OBF test partition than on the VIPL-HR-v2 test partition. This discrepancy may be caused by the substantial gap between the training set from VIPL-HR and the test set from VIPL-HR-v2: VIPL-HR mainly includes videos recorded in a controlled meeting-room environment on participants around twenty years old, whereas VIPL-HR-v2 involves participants in a broader range of ages and was collected in less controllable environments across different regions. This inconsistency may bring challenges to the proposed methods. Moreover, varying lighting conditions and head motions in the VIPL-HR-v2 test data further bring difficulty to video-based remote physiological signal measurement, as varying lighting conditions can hinder the accurate capture of facial color changes, and head movements may cause key facial regions to become partially or fully invisible.

Performance on different skin-tone groups of the OBF test partition. We further divide the OBF test data into three groups according to the participants' skin tones. The sample numbers of each sub-set are: 31 samples for light skin-tone, 41 samples for medium skin-tone, and 28 samples for dark skin-tone. One fact observed from previous rPPG studies is that most RGB-based rPPG approaches work better on lighter skin tone, while darker skin is more challenging. By fusing the NIR data, it is expected to compensate for the challenge. The RMSE results on different skin-tone groups of each participating team are shown in Fig. 6. As expected, the worst performance is observed on the OBF-Dark set across all participating teams. Unexpectedly, the lowest RMSE values were achieved on the OBF-Medium set rather than the OBF-Medium set, which contradicts the common assumption that heart rate estimation is easier for lighter skin tones. We believe this may result from a domain mismatch: the VIPL-HR

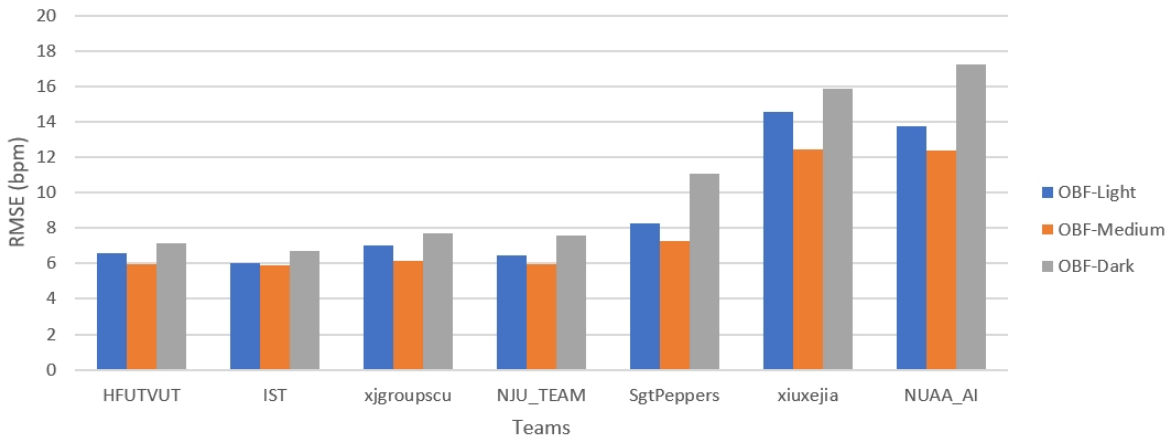

Figure 6: RMSE results of all teams on different skin-tone groups of the OBF test partition.

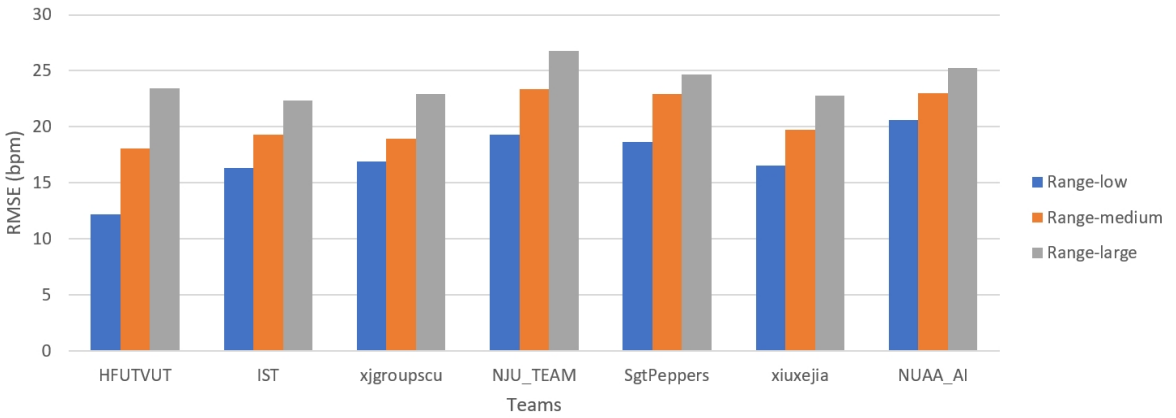

Figure 7: RMSE results of all teams on different rotation range groups of the VIPL-HR-v2 test partition.

training set mainly contains medium skin-tone samples, while the OBF test set is more diverse in terms of skin-tone distribution. Another interesting observation from Fig. 6 is that the performance gaps (i.e., RMSE differences between Medium and Dark sets) of the winning teams (left three groups of bars) are obviously smaller than the rest (right three groups of bars). This may indicate that balanced skin-tone performance (small gap between different skin groups) is an indicator of a well-performing RGB-NIR fusion model.

Performance on different motion levels of the VIPL-HR-v2 test partition. All samples in the VIPL-HR-v2 test partition include head motions. We divided the samples into three sub-sets based on the 'RANGE' and the 'SPEED' of head rotations to investigate their impact on rPPG models. For rotation range, each subset contained 100 samples. For rotation speed, the sample numbers of each sub-set are: 111 samples for low speed, 85 samples for medium speed, 104 samples for high speed.s Fig. 7 shows that a larger rotation range generally results in higher RMSE values. Similarly, Fig. 8 demonstrates that faster head rotation speeds degrade model performance. These findings suggest that both large and rapid head movements interfere with rPPG measurements by making it difficult to consistently track critical facial areas such as the cheeks and forehead. These areas may become occluded or distorted during head movement, and color fluctuations are harder to detect during fast rotations. This highlights important future directions for rPPG research, especially in improving motion-robust signal extraction.

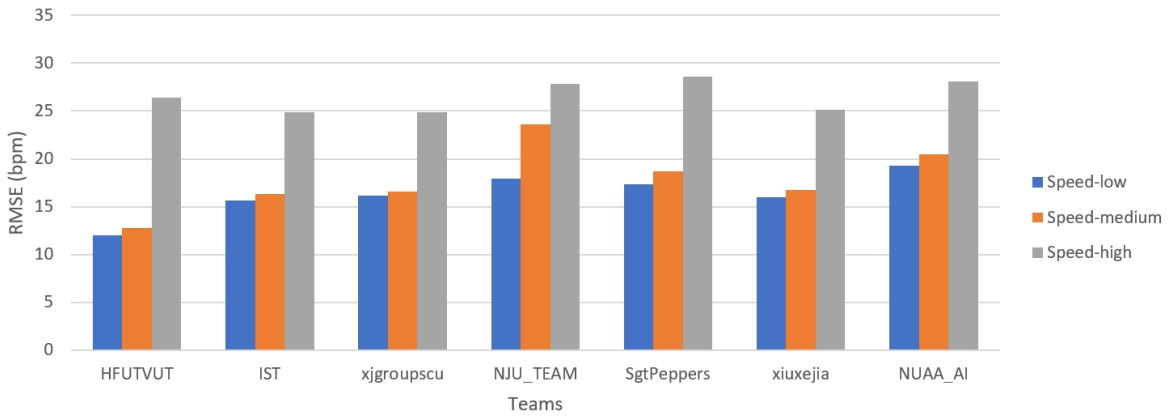

Figure 8: RMSE results of all teams on different rotation speed groups of the VIPL-HR-v2 test partition.

5. Conclusion and future directions

As a continuation of the RePSS series, the 4th RePSS Challenge maintains its focus on remote physiological signal measurement. However, unlike previous editions that only employed RGB modality, this challenge invited participants to explore innovative data fusion techniques by integrating RGB and Near-Infrared (NIR) facial videos to enhance the accuracy and robustness of rPPG estimation.

This challenge highlights an important yet underexplored direction in rPPG research—multimodal fusion for real-world applications where single-modality approaches often fall short. By incorporating NIR data, the challenge encourages the development of advanced fusion strategies that can overcome current limitations and advance the field toward more reliable and generalizable remote physiological monitoring systems.

For evaluation, test data from both the OBF and VIPL-HR-v2 datasets were used. The OBF dataset contains videos of participants from different skin-tone groups, while the VIPL-HR-v2 dataset includes samples with head motions and varying lighting conditions. Although the top-performing teams achieved relatively low RMSE values, all methods experienced performance degradation in challenging conditions such as dark skin tones, large head movements, and fast head rotations. These issues pose challenges to existed rPPG methods and point to future directions for developing more accurate and robust rPPG systems.

Acknowledgments

This work is supported by the Finnish Doctoral Program Network in Artificial Intelligence, AI-DOC (decision number VN/3137/2024-OKM-6), and the National Natural Science Foundation of China under Grant (U2336213 and 62176249).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 for grammar and spelling check. After using the tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Shi, I. Alikhani, X. Li, Z. Yu, T. Seppänen, G. Zhao, Atrial fibrillation detection from face videos by fusing subtle variations, IEEE Transactions on Circuits and Systems for Video Technology 30

(2019) 2781–2795.
[2] Z. Sun, J. Junttila, M. Tulppo, T. Seppänen, X. Li, Non-contact atrial fibrillation detection from face videos by learning systolic peaks, IEEE Journal of Biomedical and Health Informatics 26 (2022) 4587–4598.
[3] F. Ding, Y. Qin, L. Zhang, H. Lyu, Driver drowsiness detection based on facial video non-contact heart rate measurement, Journal of Advanced Computational Intelligence and Intelligent Informatics 29 (2025) 306–315.
[4] W. Yu, S. Ding, Z. Yue, S. Yang, Emotion recognition from facial expressions and contactless heart rate using knowledge graph, in: 2020 IEEE International Conference on Knowledge Graph (ICKG), IEEE, 2020, pp. 64–69.
[5] S. Ziaratnia, T. Laohakangvalvit, M. Sugaya, P. Sripian, Multimodal deep learning for remote stress estimation using cct-lstm, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 8336–8344.
[6] P. Kumar, S. Misra, Z. Shao, B. Zhu, B. Raman, X. Li, Multimodal interpretable depression analysis using visual, physiological, audio and textual data, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 5305–5315.
[7] L. Zhao, X. Zhang, X. Niu, J. Sun, R. Geng, Q. Li, X. Zhu, Z. Dai, Remote photoplethysmography (rppg) based learning fatigue detection, Applied Intelligence 53 (2023) 27951–27965.
[8] K. Wang, Y. Wei, J. Tang, Y. Wang, Z. Li, M. Tong, J. Gao, Y. Ma, Z. Zhao, Camera-based hrv prediction for remote learning environments, in: 2024 IEEE Smart World Congress (SWC), IEEE, 2024, pp. 1165–1173.
[9] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light., Opt. Express 16 (2008) 21434–21445.
[10] 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.
[11] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE Trans. Biomed. Eng. 60 (2013) 2878–2886.
[12] 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.
[13] W. Chen, D. Mcduff, Deepphys: Video-based physiological measurement using convolutional attention networks, Proc. ECCV (2018) 356–373.
[14] X. Liu, J. Fromm, S. Patel, D. McDuff, Multi-task temporal shift attention networks for on-device contactless vitals measurement, Advances in Neural Information Processing Systems 33 (2020) 19400–19411.
[15] E. M. Nowara, D. McDuff, A. Veeraraghavan, The benefit of distraction: Denoising camera-based physiological measurements using inverse attention, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4955–4964.
[16] 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.
[17] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, Proc. BMVC (2019).
[18] 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.
[19] 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.
[20] Z. Yu, Y. Shen, J. Shi, H. Zhao, Y. Cui, J. Zhang, P. Torr, G. Zhao, Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer, International Journal of Computer Vision 131 (2023) 1307–1330.
[21] X. Liu, B. Hill, Z. Jiang, S. Patel, D. McDuff, Efficientphys: Enabling simple, fast and accurate

camera-based cardiac measurement, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, pp. 5008–5017.
[22] 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).
[23] 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.
[24] H. Lu, H. Han, S. K. Zhou, Dual-gan: Joint bvp and noise modeling for remote physiological measurement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12404–12413.
[25] 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.
[26] 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.
[27] Z. Sun, X. Li, H. Han, J. Tang, C. Ying, J. Ge, A. Dantcheva, S. Shan, G. Zhao, The 3rd vision-based remote physiological signal sensing (repss) challenge & workshop, in: CEUR Workshop Proceedings, R. Piskac c/o Redaktion Sun SITE, Informatik V, RWTH Aachen, 2024.
[28] K. Kurihara, D. Sugimura, T. Hamamoto, Non-contact heart rate estimation via adaptive rgb/nir signal fusion, IEEE Transactions on Image Processing 30 (2021) 6528–6543.
[29] S. Park, B.-K. Kim, S.-Y. Dong, Self-supervised rgb-nir fusion video vision transformer framework for rppg estimation, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–10.
[30] X. Niu, H. Han, S. Shan, X. Chen, VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video, in: Proc. ACCV, 2018, pp. 562–576.
[31] 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.
[32] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Transactions on Image Processing 26 (2017) 2825–2838.
[33] 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.
[34] Y. Kartynnik, A. Ablavatski, I. Grishchenko, M. Grundmann, Real-time facial surface geometry from monocular video on mobile gpus, arXiv preprint arXiv:1907.06724 (2019).
[35] K. Dragomiretskiy, D. Zosso, Variational mode decomposition, IEEE transactions on signal processing 62 (2013) 531–544.
[36] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2 (2010) 433–459.