Cluster-Phys: Facial Clues Clustering Towards Efficient Remote Physiological Measurement——Supplementary Material

Wei Qian*
School of Computer Science and Information Engineering, School of Artificial Intelligence, Hefei University of Technology
qianwei.hfut@gmail.com

Kun Li*
CCAI, Zhejiang University
kunli.hfut@gmail.com

Dan Guo†
Hefei University of Technology, Institute of Artificial Intelligence (IAI), Hefei Comprehensive National Science Center
guodan@hfut.edu.cn

Bin Hu
Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University
bh@lzu.edu.cn

Meng Wang†
Hefei University of Technology, Institute of Artificial Intelligence (IAI), Hefei Comprehensive National Science Center
eric.mengwang@gmail.com

ACM Reference Format:
Wei Qian, Kun Li, Dan Guo, Bin Hu, and Meng Wang. 2024. Cluster-Phys: Facial Clues Clustering Towards Efficient Remote Physiological Measurement——Supplementary Material. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3664647.3680670

1 Evaluation Metrics.

For average HR estimation, following the evaluation protocol [1, 2, 4], we report the most commonly used performance metrics, such as *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), *Standard Deviation* (SD), and *Pearson correlation coefficient* (r). For Heart Rate Variability (HRV) and Respiration Frequency (RF) estimation, we follow [3, 5] and report low frequency (LF), high frequency (HF), and LF/HF ratio.

2 Ablation Study of Hyper-parameters

As depicted in Fig. 1, we investigate different hyper-parameters in facial ROI prototypical clustering. For the iteration number M, we test its value of $\{1, 2, 3, 4\}$ and set M to 3 to build prototypes. For the cluster sparse ratio ρ, we evaluate its value from 0.3 to 1.0. The results show that ρ=0.5 performs best, and an excessively high sparsity will lead to the loss of crucial rPPG clues, leading to a decrease in performance. For the depth L of our Cluster-Phys, we search its value from 2 to 5. As shown in Fig. 1 (c), the deeper model will make it difficult for the model to converge, which is harmful to the performance.

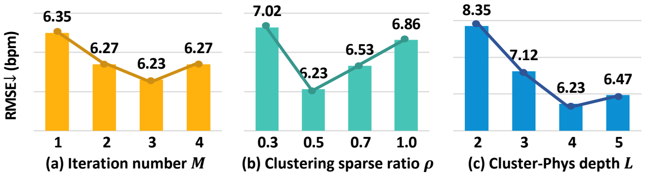

Figure 1: RMSE results of our Facial ROI Prototypical Clustering with different hyper-parameters on the VIPL-HR dataset.

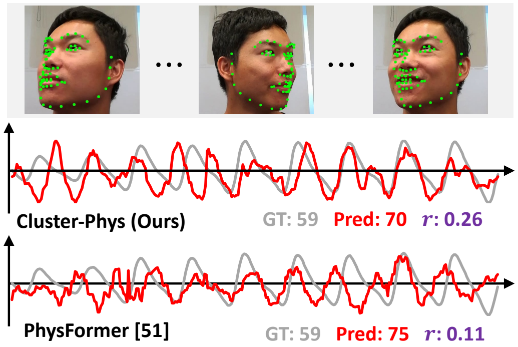

Figure 2: A failure case in VIPL-HR.

3 More Implementation Details.

For each video, we extract the facial ROI regions using landmark detection to generate the MSTmap [4]. The MSTmap and its corresponding Heart Rate (HR) label are sampled at a rate of 30Hz within the video. Each MSTmap is configured to consist of 300 frames, with a sliding window of 15 frames overlapping with the adjacent MSTmap. The model is trained by the Adam optimizer with a learning rate of $1e^{-4}$ and a batch size of 16. We train the

*Wei Qian and Kun Li contributed equally to this research.
†Corresponding authors

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680670

model for 100 epochs on the VIPL-HR dataset and 30 epochs on the other datasets.

4 Failure Case.

We visualize a fail case from VIPL-HR in Fig. 2. *Inaccurate landmark detection* is inevitable with *intense head movements*. Our average Pearson correlation coefficient ($r \uparrow$) on VIPL-HR is 0.84. In this case, our r is 0.26, whereas PhysFormer has a r of 0.11. Although our method performs better than PhysFormer, it is significantly lower than the average r by a large margin. How to exploit occluded areas more effectively is a direction for future research. From a practical perspective, a multi-camera setup is considered a good solution. Additionally, we may need to address new challenges such as multi-camera alignment.

5 Limitation

For the HR estimation task, inaccurate landmarks caused by intense head movements are inevitable to reduce the accuracy of facial ROI; cross-domain learning and cross-ethnic application remain challenges described in Section 4.3 (see experiments in Table 2 of the main paper). These are also some limitations of most current HR estimation methods. For our method, the cluster ratio is a fixed hyperparameter in each step. Developing an adaptive clustering strategy will be our future research.

References

[1] Weixuan Chen and Daniel McDuff. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision*. 349–365.

[2] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. 2014. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4264–4271.

[3] Hao Lu, Hu Han, and S Kevin Zhou. 2021. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12404–12413.

[4] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. 2020. Video-based remote physiological measurement via cross-verified feature disentangling. In *Proceedings of the European Conference on Computer Vision*. 295–310.

[5] Zhaodong Sun and Xiaobai Li. 2024. Contrast-Phys+: Unsupervised and Weakly-supervised Video-based Remote Physiological Measurement via Spatiotemporal Contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).