Supplementary Materials for Remote Photoplethysmography in Real-World and Extreme Lighting Scenarios

Hang Shao[1], Lei Luo[1*], Jianjun Qian[1], Mengkai Yan[1], Shuo Chen[2], Jian Yang[1*]
[1]PCA Lab,[†] Nanjing University of Science and Technology, [2]Nanjing University, China
{shaohang,cslluo,csjqian,ymk,csjyang}@njust.edu.cn, shuo.chen.ya@foxmail.com

1. Comparison with transformer-based rPPGs

Admittedly, we propose the remote photoplethysmography (rPPG) framework based on the Swin vision transformer, in order to validate that our performance and innovation come from the improvement of the learning manner and the disentanglement strategy, rather than relying solely on the self-attention and the temporal linear complexity mechanisms of the transformer structure, we compare our approach with two recent rPPG models (PhysFormer++ [1] and RhythmFormer [2]) based on the transformer architecture, and draw their scatter plots of the heart rate (HR) estimation results in Fig. 1. Specifically, we report the results from HR measurements of 2,000 random time periods, the training data is the joint MR-NIRP-IND [3] and MR-NIRP-DRV [4] datasets (MR-NIRP), and the test set is the outdoor driving condition of the MR-NIRP-DRV dataset to prove our improvements in the more challenging environment. At can be seen that the remote perceptual results of our algorithm are closer to the ground truth (GT) values, and more values are in the GT$\pm$6 beats per minute (bpm) space. This serve as a useful illustration of our paradigm and its ability to mine high-quality physiological cues.

2. Visualization of pulse and HR predictions

As a complement to HR prediction and ablation studies to verify the scientific rationality of our architecture and settings, we discuss the performance of the core component of our method, that is the interference disentanglement module. Specifically, we compare the full model with the VIPL-HR dataset [5] learning results without the disentanglement module (w/o ND), and plot visualizations of one-minute HR sensing, as shown in Fig. 2. At the same time, to fully discuss the impact of the time window length on the prediction, we set the raw 320-frame time dimension of the spatiotemporal map (STMap) of the undisentangled input to 300, 280, 260, 240, 220, 200, 180, 160, and 140 frames, respectively

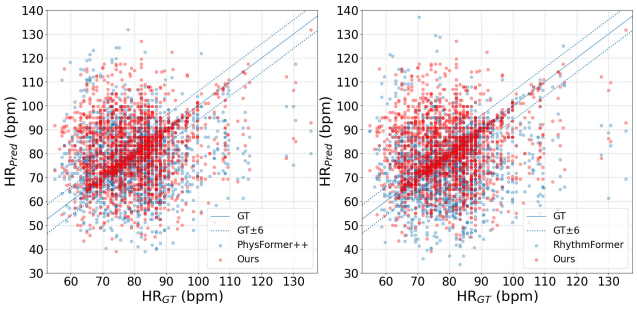

Figure 1. Our framework improves upon the state-of-the-art vision transformer-based rPPG models.

(as shown by the corresponding light green lines). In addition, we also draw our blood volume pulse (BVP) prediction results in Fig. 3 to better reflect the tracking ability of our algorithm for physiological signals. It can be observed that although increasing the input time length can improve the model's fit to BVP, thereby improving the ability of perceive biosignals and further optimization remote HR recognition. However, without the introduction of the disentanglement architecture, the results under insufficient time dimension still show similar trends, and they have difficult in properly mining certain complete heartbeat cycles and peaks, so their HR estimation performances are also generally poor.

3. Computational performance analysis

As a supplement to the effectiveness analysis, we comprehensively compare all the latest representative deep rPPG technologies we have discussed in our main paper, including convolutional network-based methods (DeepPhys [6], TS-CAN [7], Dual-GAN [8], PFE-TFA [9], NEST [10], and ND-DeeprPPG [11]) and vision transformer-based methods (EfficientPhys [12], PhysFormer++ [1], and RhythmFormer [2]). The model parameters, the floating point operations per second (FLOPs), the computing time on the RTX 4090 GPU device, and the root mean square error (RMSE) prediction results of the VIPL-HR dataset [5] are listed in Tab. 1. Among them, for the approaches that take the facial

*Corresponding authors.
†Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering.

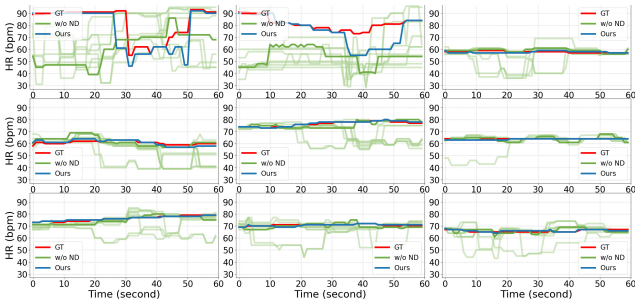

Figure 2. HR estimation within a one-minute facial video segment and comparison with and without the disentanglement module.

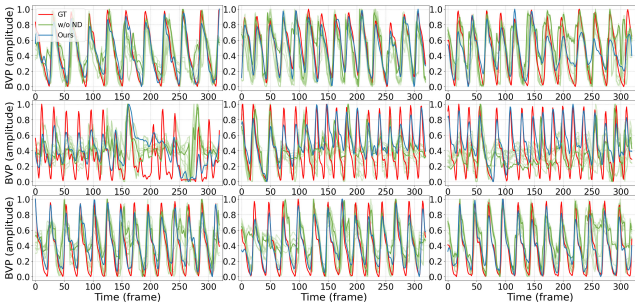

Figure 3. BVP prediction within a 320-frame time sequences and comparison with and without the disentanglement module.

video segment as input, the input scale is 3×320×64×64 (color channel, input video segment length, frame height, and frame width), and for the methods that take the STMap as input, the input scale is 3×64×320 (color channel, map height, and video segment length). In addition, we also list the relevant results of the convolutional architecture based on our paradigm (Ours-Conv). It can be seen that, benefiting from our framework layout, while effectively improving performance, our network is more lightweight and efficient for the practical deployment of diversity downstream tasks of remote physiological perception.

4. Ablation study on boosting disentanglement

Our method is based on the boosting algorithm design idea, and proposes corresponding elimination strategies for general noise and extreme interference. In order to verify the scientificity and rationality of each of our main modules separately, we follow the training and verification in Sec. 1, and calculate our full model and compare it with the models without the spatiotemporal biological prior-based sliding window enhancement (w/o ST) and the interference disentanglement (w/o ND). The HR estimation scatter plots of 2,000 random complex scenes under outdoor conditions in the MR-NIRP-DRV dataset [4] are shown in Fig. 4. On the left side is the ablation of spatiotemporal denoising and reduction for general noise and motion, and on the right side

Table 1. Computational cost and performance comparison, where ↓ means the smaller the better, and the best result is **bolded** (the same below).

rPPG Methods	Parameter	FLOPs	RTX 4090	RMSE↓
DeepPhys [6]	1.46 M	64.86 G	6.27 ms	13.80
TS-CAN [7]	3.91 M	110.15 G	5.52 ms	14.59
Dual-GAN [8]	6.17 M	302.14 G	24.65 ms	7.68
PFE-TFA [9]	1.31 M	79.60 G	36.48 ms	8.65
NEST [10]	13.82 M	3.11 G	11.27 ms	7.96
ND-DeeprPPG [11]	6.05 M	320.08 G	29.87 ms	7.52
EfficientPhys-T1 [12]	15.73 M	345.20 G	26.64 ms	8.25
PhysFormer++ [1]	9.79 M	49.85 G	217.07 ms	7.62
RhythmFormer [2]	3.25 M	39.49 G	29.49 ms	7.49
Ours-Conv	2.85 M	32.11 G	25.43 ms	7.50
Ours	6.03 M	53.04 G	24.64 ms	**7.09**

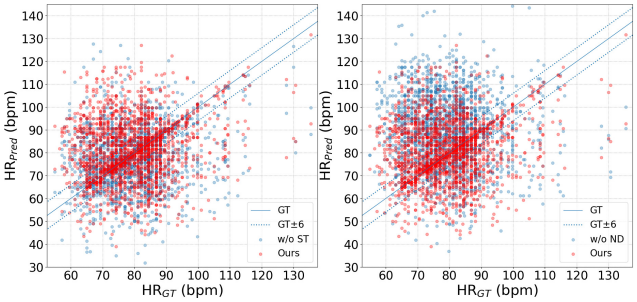

Figure 4. Illustration of a comparative study of our boost design.

is the self-supervised disentanglement of extreme interference. It can be seen that our solution can not only effectively improve the detection performance, but also has more obvious advantages in processing marginal HR estimation points (such as very low and very high GT values) to ensure that HR measurements are captured correctly and the network doesn't overfit. Meanwhile, it can also be found that the improvement of model performance by the interference disentanglement paradigm is more intuitive.

5. Ablation study on loss hyperparameters

To illustrate the rationality of our settings, Fig. 5 shows a set of control experiments on the hyperparameters of the overall loss function $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{total}} = \alpha\,\mathcal{L}_{\text{r}} + \beta\,\mathcal{L}_{\text{c}} + \gamma\,\mathcal{L}_{\text{p}}, \tag{1}$$

We report the RMSE results on the VIPL-HR dataset [5] trained models with different combinations of α, β, and γ of $\mathcal{L}_{\text{r}}$, $\mathcal{L}_{\text{c}}$, and $\mathcal{L}_{\text{p}}$ for the whole loss function in the main paper. We fix one certain hyperparameter value to verify the performance of the remaining values to select the best combination. It is worth mentioning that when β is 0, our model can be regarded as a network that does not using the interference disentanglement module and is adapted to the current general rPPG framework. Furthermore, the verification of

Table 2. Comparison of our rPPG method with the facial skin reflectance baseline under multiple illumination modalities and quantifiable patterns.

Models	VIPL-HR [5]			BUAA-MIHR [13]											
	Studio illumination			Lux											
	Lamp	Bright	Dark	$10^{0.0}$	$10^{0.2}$	$10^{0.4}$	$10^{0.6}$	$10^{0.8}$	$10^{1.0}$	$10^{1.2}$	$10^{1.4}$	$10^{1.6}$	$10^{1.8}$	$10^{2.0}$
POS [14]	16.59	17.38	17.64	74.31	57.33	40.31	16.17	8.49	4.88	8.46	9.96	8.03	13.04	4.42
Ours w/o ND	7.57	8.84	9.08	15.78	13.74	9.33	7.39	2.61	1.43	2.38	1.50	1.37	1.71	**1.15**
Ours	**6.53**	**7.83**	**8.11**	**5.04**	**3.97**	**2.68**	**2.35**	**1.28**	**1.09**	**1.36**	**1.16**	**1.23**	**1.12**	1.22

α can illustrate the importance of our spatiotemporal reconstruction module in optimizing the overall learning process, even though it does not participate in the final prediction.

6. Ablation study on diversity illuminations

In order to verify the impact of illumination of our measured results and our insensitivity to lighting conditions, we split two datasets, VIPL-HR [5] and BUAA-MIHR [13] , which contain multiple illumination variations, according to the subjects and test them under different illuminations. Among them, the illumination design of the VIPL-HR dataset is to qualitatively distinguish modes, namely general illumination (the studio light source, "Lamp"), bright scene (the filament lamp is turned on, "Bright"), and dark situation (the ceiling lamp of the room is turned off, "Dark"). The illumination of the BUAA-MIHR dataset is quantifiable, and it subdivides the captured videos into eleven modes ranging from $10^{0.0}$ to $10^{2.0}$ lux in intervals of $10^{0.2}$. We report the RMSE results of our full model and the model without the disentanglement module on their remote measurements in Tab. 2. At the same time, as a benchmark for verifying illumination, we introduce the plan-orthogonal-to-skin (POS) method [14] followed by Xi et al. [13] as a quantitative reference for feedback of skin light reflectance under different illuminations. It can be seen that our method is insensitive to illumination and contrast under dynamic lighting.

7. Implementation datasets

We construct the proposed network's training, testing, and validation of time-varying interference disentanglement on four of the most popular publicly available remote physiological monitoring datasets, which are BUAA-MIHR [13], VIPL-HR [5], MR-NIRP-IND [3], and MR-NIRP-DRV [4]. Specifically, the BUAA-MIHR dataset has 165 one-minute facial videos of thirteen participants, divided into eleven illumination modes ($10^{0.0}$, $10^{0.2}$, $10^{0.4}$, $10^{0.6}$, $10^{0.8}$, $10^{1.0}$, $10^{1.2}$, $10^{1.4}$, $10^{1.6}$, $10^{1.8}$, and $10^{2.0}$ lux, respectively) with precise lighting span. Its frame rate is 30 frames per second (fps), resolution is 640×480 pixels, and has corresponding BVP waveform and same frequency HR value labels. The videos are collected by a Logitech HD pro webcam C930E

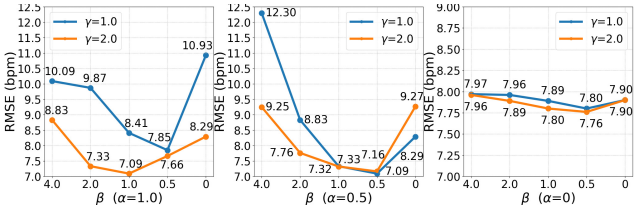

Figure 5. Illustration of the loss function hyperparameter settings.

color camera, and the physiological signals are obtained by a CONTEC CMS50E finger clip sensor.

The VIPL-HR dataset covers 2,379 visible light facial videos of 107 subjects, divided into three acquisition modes and nine dynamic conditions, with corresponding BVP and HR labels to meet the situations encountered in real-world daily life. Among them, the three video acquisition patterns are recorded by a Logitech HD C310 web-camera (25 fps and 960×720 resolution), a HUAWEI P9 frontal camera (30 fps and 1920×1080 resolution), and a RealSense F200 camera (30 fps and 1920×1080 resolution), respectively. The nine dynamic modes include stable scenario, motion scenario, talking scenario, dark scenario, bright scenario, long distance scenario, exercise scenario, phone stable scenario, and phone motion scenario. Although all videos are shot indoors, it strives to simulate real scenes and is currently the largest and most important rPPG dataset.

The MR-NIRP-IND dataset contains 1,914 seconds of still and moving visible light facial data (RGB) and corresponding near-infrared imaging data (NIR) of eight subjects, with the frame rate of 30 fps, the frame resolution of 640×640 pixels, and corresponding BVP signal labels. Furthermore, the RGB camera is a FLIR Blackfly BFLY-U3-23S6C-C, and the NIR camera is a Grey Grasshopper GS3-U3-41C6NIR-C equipped with a narrow-band 940 nm bandpass filter.

The MR-NIRP-DRV dataset extends the data scale of the former to eighteen subjects and 30,198 seconds, covering outdoor driving, day and night changes, with corresponding NIR maps. Its ground truth label, RGB camera, frame rate, and resolution are consistent with the MR-NIRP-IND dataset, while its NIR additionally introduces 975 nm bandpass. The scene settings of the MR-NIRP-DRV dataset are

Figure 6. Illustration of sample variations from the four datasets.

very diverse, and each subject recorded six types of video: driving large motion, driving small motion, driving still, garage large motion, garage small motion, and garage still. The lighting changes in driving are complex, time-varying, diverse, and extreme. To the best of our knowledge, it is the largest dataset currently available for pure outdoor scenes and dynamic lighting. Representative participant examples and modalities of these datasets are shown in Fig. 6.

References

[1] Z. Yu, Y. Shen, J. Shi, H. Zhao, Y. Cui, J. Zhang, ... and G. Zhao. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *IJCV*, 2023. 1, 2

[2] B. Zou, Z. Guo, J. Chen, and H. Ma. Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer. *arXiv:2402.12788*, 2024. 1, 2

[3] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan. Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *CVPR Workshops*, 2018. 1, 3

[4] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE TITS*, 2022. 1, 2, 3

[5] X. Niu, S. Shan, H. Han, and X. Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE TIP*, 2020. 1, 2, 3

[6] W. Chen, and D. McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *ECCV*, 2018. 1, 2

[7] X. Liu, J. Fromm, S. Patel, and D. McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In *NeurIPS*, 2020. 1, 2

[8] H. Lu, H. Han, and S. K. Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *CVPR*, 2021. 1, 2

[9] J. Li, Z. Yu, and J. Shi. Learning motion-robust remote photoplethysmography through arbitrary resolution videos. In *AAAI*, 2023. 1, 2

[10] H. Lu, Z. Yu, X. Niu, and Y. C. Chen. Neuron structure modeling for generalizable remote physiological measurement. In *CVPR*, 2023. 1, 2

[11] S. Q. Liu, and P. C. Yuen. Robust remote photoplethysmography estimation with environmental noise disentanglement. *IEEE TIP*, 2024. 1, 2

[12] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *WACV*, 2023. 1, 2

[13] L. Xi, W. Chen, C. Zhao, X. Wu, and J. Wang. Image enhancement for remote photoplethysmography in a low-light environment. In *IEEE FG*, 2020. 3

[14] W. Wang, A. C. D. Brinker, S. Stuijk, and G. D. Haan. Algorithmic principles of remote ppg. *IEEE TBME*, 2017. 3