Supplementary Material for BeatFormer: Efficient motion-robust remote heart rate estimation through unsupervised spectral zoomed attention filters

Joaquim Comas and Federico Sukno
Department of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain
{joaquim.comas,federico.sukno}@upf.edu

1. Datasets

The **Multi-domain Mobile Video Physiology Dataset (MMPD) [12]** comprises 660 one-minute videos from 33 subjects, totaling 11 hours of recordings from mobile phones. Each subject contains twenty trials to capture variations in body motion, and lighting conditions. These trials included stationary, rotation, talking, and walking tasks performed under four different lighting conditions. The videos were recorded at 30 FPS with a resolution of 1280x720 pixels, while PPG signals were simultaneously recorded using an HKG-07C+ oximeter at 200Hz and downsampled to 30Hz. Additionally, each subject was provided with eight descriptive labels, including the Fitzpatrick scale for skin color.

The **UBFC-rPPG** dataset [1] includes 42 RGB videos from 42 subjects. The subjects were asked to play a time-sensitive mathematical game, emulating a standard human-computer interaction scenario, to obtain varied HR during the experiment. The recorded facial videos were acquired indoors with varying sunlight and indoor illumination at 30 FPS with a webcam (Logitech C920 HD Pro) at a resolution of 640x480 in uncompressed 8-bit RGB format. The bio-signals ground truth was acquired using a CMS50E transmissive pulse oximeter to record the PPG signal and heart rate.

The **Pulse Rate Detection (PURE)** dataset [10] contains 60 videos from 10 subjects performing six different head motion tasks: steady, talking, slow translation, fast translation, small rotation, and medium rotation. The facial videos were recorded using an ECO274CVGE camera with a resolution of 640 x 480 pixels and 30 FPS. Each video is about 1 minute long and stored in uncompressed PNG format. The gold-standard measures of BVP and SpO2 were collected using a finger pulse oximeter.

2. Additional experiments

In this section, we provide relevant experiments done to evaluate the BeatFormer performance, but not included in the main paper due to space constraints. These experiments include an intra-dataset evaluation on PURE and UBFC-rPPG datasets, an extended comparison between supervised and unsupervised BeatFormer training and additional ablation studies regarding the BeatFormer configuration.

2.1. Intra-dataset evaluation

For intra-dataset experiments, we follow the protocols of [5, 9] for PURE and [5] for UBFC-rPPG. Table 1 compares the proposed BeatFormer with state-of-the-art methods, including three handcrafted approaches, five supervised, and three unsupervised data-driven models. The results show that supervised methods outperform both handcrafted and unsupervised approaches, likely due to the similar conditions between training and testing, highlighting the advantage of data-driven models in this setting. Additionally, since PURE and UBFC-rPPG are uncompressed datasets recorded under controlled conditions, rPPG extraction is facilitated. Notably, BeatFormer achieves state-of-the-art

Table 1. Pulse rate intra-dataset results on PURE and UBFC-rPPG datasets (in BPMs).

Method	PURE			UBFC		
	MAE ↓	RMSE ↓	ρ ↑	MAE ↓	RMSE ↓	ρ ↑
ICA [7]	3.76	12.60	0.85	5.17	11.76	0.65
CHROM [2]	0.75	2.23	1.00	2.36	9.23	0.87
POS [13]	0.80	4.11	0.98	2.11	9.11	0.87
HR-CNN [9]	1.84	2.37	0.98	4.90	5.89	0.64
Dual-GAN [5]	0.82	1.31	0.99	0.44	0.67	0.99
ETA-rPPGNet [4]	0.34	0.77	0.99	1.46	3.97	0.93
PhysMamba [6]	0.25	0.40	0.99	0.54	0.76	0.99
RhythmFormer [14]	0.27	0.47	0.99	0.50	0.78	0.99
BeatFormer-SL (ours)	**0.17**	**0.36**	**1.00**	**0.15**	**0.37**	**1.00**
Gideon2021 [3]	2.30	2.90	0.99	1.85	4.28	0.93
Contrast-Phys [11]	1.00	1.40	0.99	0.64	1.00	0.99
SiNC [8]	0.61	1.84	**1.00**	0.59	1.83	0.99
BeatFormer-SCL (ours)	0.29	0.60	**1.00**	0.44	1.29	0.99

Table 2. Comparison between supervised (SL) and spectral contrastive learning (SCL) in MMPD cross dataset evaluation (in BPMs).

Scenario	Training	MAE ↓	RMSE ↓	MAPE ↓	ρ ↑
Stationary	SL	6.23 ± 0.88	11.87 ± 31.18	9.08 ± 1.47	0.51 ± 0.08
	SCL	**6.03 ± 0.85**	**11.47 ± 27.67**	**8.71 ± 1.37**	**0.54 ± 0.07**
Stationary (after exercise)	SL	11.99 ± 1.26	18.79 ± 61.13	12.15 ± 1.13	0.29 ± 0.08
	SCL	**11.66 ± 1.22**	**18.26 ± 56.74**	**12.02 ± 1.14**	**0.31 ± 0.08**
Rotation	SL	6.07 ± 0.89	11.89 ± 38.66	8.97 ± 1.51	0.45 ± 0.08
	SCL	**5.88 ± 0.76**	**10.47 ± 24.79**	**8.19 ± 1.23**	**0.53 ± 0.07**
Talking	SL	**6.25 ± 0.83**	**11.40 ± 29.00**	**8.04 ± 1.16**	**0.47 ± 0.08**
	SCL	7.14 ± 0.89	12.47 ± 31.01	9.21 ± 1.25	0.36 ± 0.08
Walking	SL	**13.66 ± 1.16**	**19.09 ± 52.22**	**14.41 ± 1.19**	0.06 ± 0.09
	SCL	14.97 ± 1.19	20.30 ± 56.41	15.71 ± 1.17	**0.07 ± 0.09**
All	SL	**8.85 ± 0.47**	**15.04 ± 20.24**	**10.54 ± 0.59**	0.39 ± 0.04
	SCL	9.14 ± 0.47	15.13 ± 19.32	10.78 ± 0.56	**0.40 ± 0.04**

performance on both PURE and UBFC-rPPG, surpassing PhysMamba and RhythmFormer, which obtain similar performance. Regarding unsupervised learning, BeatFormer achieves the best results for both datasets and all the performance metrics, except the RMSE in UBFC-rPPG, where Contrast-Phys yields 1 BPM of error, the best performance. Despite the superiority of supervised methods, we highlight that recent unsupervised methods are obtaining close performance without requiring PPG or HR labels.

2.2. Comparison supervised against SCL training

Table 2 provides a detailed comparison between the supervised and unsupervised versions of BeatFormer in MMPD cross-dataset scenarios, including both stationary splits. The results show that SCL training outperforms the supervised version in stationary, stationary after exercise, and rotation scenarios, while the supervised model performs better in talking and walking.

Although SCL achieves superior results in three scenarios compared to two, the overall performance slightly favors the supervised version, primarily due to a major difference in the last two splits, approximately around 1 BPM error. Nonetheless, both versions deliver strong results, with the unsupervised model achieving competitive performance without requiring any labeled data, neither in terms of PPG or HR information.

(a) ID 1 (b) ID 2 (c) ID 3 (d) ID 4

Figure 1. Example of tested video processings for the same subject.

2.3. Additional ablation studies

This subsection incorporate additional ablations studies regarding the impact of the video preprocessing, window size influence and BeatFormer configuration. In addition, we include Tables 5 and 6 which show the complete results discussed in the main paper during the ablation studies in the ZOCA and CZT impact, as well as, the impact of video transformations in SCL training.

Impact of video processing framework. Table 3 highlights the impact of video preprocessing on BeatFormer's performance. This experiment evaluates four different preprocessing pipelines, illustrated in Fig. 1. The first two (ID 1 and ID 2) use static facial detection with a 1.5-box spacing, while the latter two (ID 3 and ID 4) employ tracking-based facial detection with a 1.2-box spacing. Additionally, we assess the effect of skin masking in the dynamic and tracking-based versions (ID 2 and ID 4).

The results show that tracking-based facial detection improves HR estimation by 1–3 BPM. However, the most significant enhancement comes from skin masking. Comparing ID 1 with ID 2, or ID 3 with ID 4, reveals an HR error reduction of approximately 4–5 BPM, demonstrating its substantial impact on BeatFormer's performance. While averaging spatially RGB values over frames helps maintain robustness to local motions, it relies on accurate skin region detection for optimal performance. As discussed, future work should focus on reducing BeatFormer's dependence on specific preprocessing frameworks.

Impact of blocks and heads configuration. Fig. 2 illustrates the ablation study regarding the number of blocks and heads. Regarding the ZOCA blocks, while maintaining fixed the number of heads, we test BeatFormer with 2, 4 and 6 blocks, where the configuration with the best performance using 4 blocks. On the other hand, for the number of heads, we test between 1 and 3 heads since our frequential embedding is equal to three which represents the filtered RGB

Table 3. Influence of video processing framework in cross-dataset evaluation on MMPD, trained with PURE dataset (in BPMs).

ID	Video Processing	MAE ↓	RMSE ↓	MAPE ↓	ρ ↑
1	Crop_{static} ($\times$1.5-Box) + Resize	14.00 ± 0.51	19.16 ± 24.23	16.89 ± 0.63	0.13 ± 0.04
2	Crop_{static} ($\times$1.5-Box) + Resize + Skin mask	9.80 ± 0.49	16.04 ± 21.00	11.64 ± 0.61	0.42 ± 0.04
3	$\text{Crop}_{Dynamic}$ ($\times$1.2-Box) + Resize	11.75 ± 0.51	17.52 ± 22.87	14.01 ± 0.63	0.30 ± 0.04
4	$\text{Crop}_{Dynamic}$ ($\times$1.2-Box) + Resize + Skin mask	**8.85 ± 0.47**	**15.04 ± 20.24**	**10.54 ± 0.59**	**0.39 ± 0.04**

channels. For this comparison, we use four ZOCA blocks which are the best configuration found previously. Unlike the block ablation study, where the performance varies more, in terms of heads the performance is very similar, being the configuration with three heads slightly better.

Impact of window size. Table 4 examines the effect of window size L on performance. We set a minimum window size of two seconds to ensure at least one detectable heartbeat, assuming a minimum HR frequency of 0.66 Hz (one beat every 1.51 seconds). For comparison, we also evaluate window sizes of 4 and 6 seconds.

Our results show that in the MMPD cross-dataset evaluation, a 2-second window yields significantly better performance than larger windows. This may be due to the high motion variability in MMPD, where shorter windows are more effective at capturing rapid changes. Future work should further investigate the impact of window size across different datasets and explore integrating a temporal multi-resolution framework.

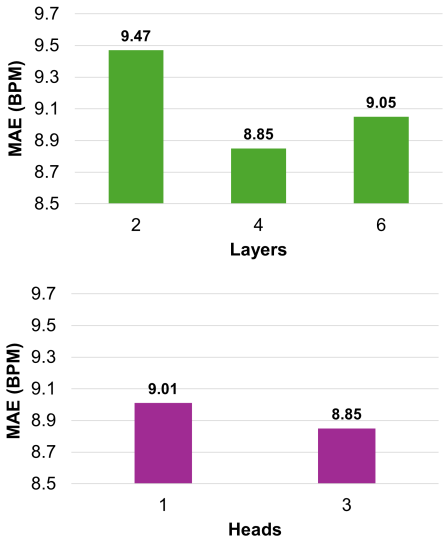


Figure 2. BeatFormer blocks and heads configuration ablation study (in BPMs).

For this study, the final BeatFormer configuration uses a 2-second window size, four blocks, and three heads.

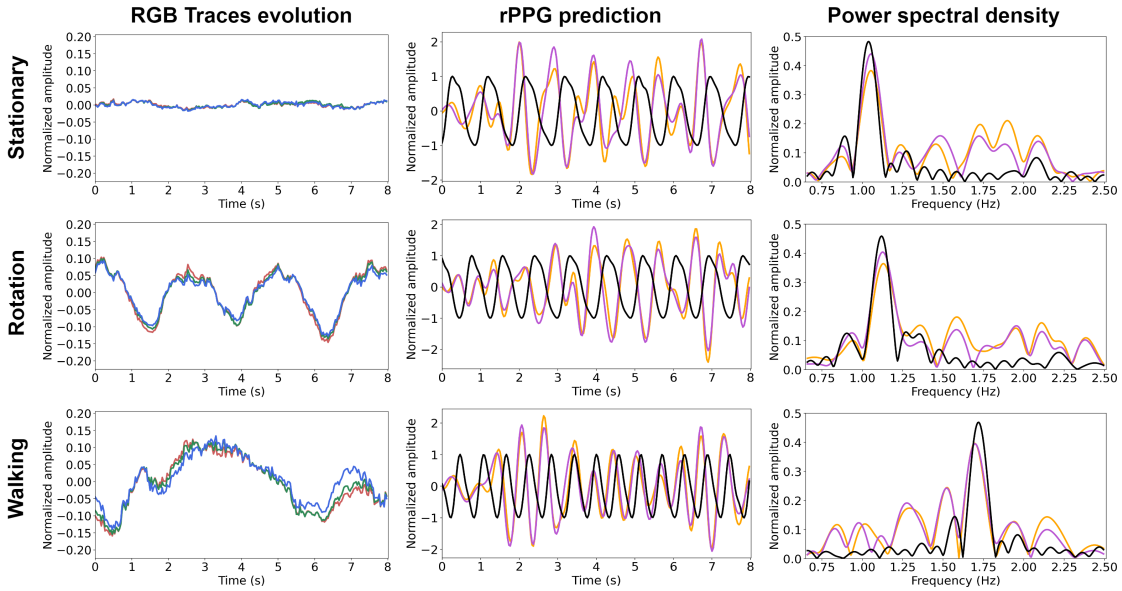
2.4. Qualitative results

Fig. 3 presents two qualitative examples from BeatFormer using MMPD cross-dataset samples from different scenarios. The models are trained on PURE (top) and UBFC-rPPG (bottom), with inference results for BeatFormer-SL (yellow) and BeatFormer-SCL (magenta) shown for each example. The RGB trace evolution highlights significant amplitude changes when comparing stationary scenarios to rotation and walking. Despite these noticeable differences in the RGB traces, the rPPG prediction and its spectrum, compared with the ground truth, show significant robustness on these motion splits such as in the PURE rotation example, where horizontal head motions are quite noticeable. As in the main paper, we can appreciate the temporal offset between our rPPG prediction and the PPG-finger ground truth signal. This is because using frequency optimization and physiological constraints the BeatFormer is able to be invariant to the Pulse Transit Time (PTT) between the pulse information extracted from the face and the finger.

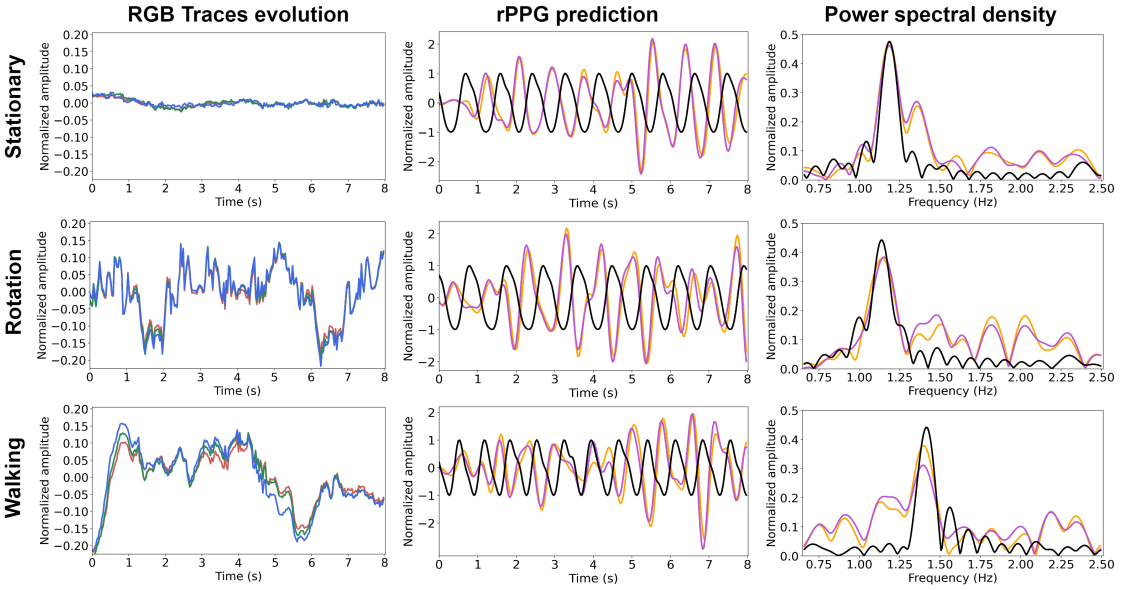
Importantly, despite this temporal offset, the maximum spectral peaks of both signals align, leading to the same heart rate estimation. Additionally, as noted in our discussion, some prominent peaks in both the spectral and temporal domains may result from second harmonics or residual motion artifacts, particularly in the walking scenarios. For this reason, future work should focus on the enhancement of pulsatile frequency filtering.

Table 4. Temporal window size influence in cross-dataset performance on MMPD trained on PURE across different motion scenarios (in BPMs).

Window size	MAE ↓	RMSE ↓	MAPE ↓	ρ ↑
2 sec	**8.85 ± 0.47**	**15.04 ± 20.24**	**10.54 ± 0.59**	**0.39 ± 0.04**
4 sec	10.07 ± 0.47	15.75 ± 20.05	12.02 ± 0.57	0.30 ± 0.04
6 sec	10.05 ± 0.47	15.75 ± 20.07	12.13 ± 0.60	0.30 ± 0.04



(a) Sample from subject P2 in the MMPD dataset, trained on the PURE dataset.



(b) Sample from subject P9 in the MMPD dataset, trained on the UBFC-rPPG dataset.

Figure 3. Inference examples for cross-dataset MMPD subjects trained on PURE and UBFC-rPPG. The left column shows RGB trace evolution across three motion scenarios: stationary, rotation, and walking. The middle column depicts rPPG predictions for BeatFormer-SL (yellow), BeatFormer-SCL (magenta), and the PPG ground truth (**black**). The right column presents the corresponding rPPG spectra for each signal.

Table 5. Ablation study of ZOCA and CZT influence (in BPMs).

CZT	*FFT*	*ZOCA*	MAE ↓	RMSE ↓	MAPE ↓	ρ ↑
✗	✓	✗	14.50 ± 0.61	21.40 ± 32.15	16.07 ± 0.62	0.06 ± 0.04
✓	✗	✗	9.15 ± 0.48	15.28 ± 19.91	10.85 ± 0.60	0.37 ± 0.04
✗	✓	✓	13.07 ± 0.53	18.88 ± 24.71	14.85 ± 0.57	0.16 ± 0.04
✓	✗	✓	$\mathbf{8.85 \pm 0.47}$	$\mathbf{15.04 \pm 20.24}$	$\mathbf{10.54 \pm 0.59}$	$\mathbf{0.39 \pm 0.04}$

Table 6. Impact of each SCL video transformation in MMPD cross dataset evaluation across different scenarios, trained on PURE dataset (in BPMs).

Scenario	Transformation	MAE ↓	RMSE ↓	MAPE ↓	ρ ↑
Stationary	HSV	7.11 +/- 0.91	12.62 +/- 29.34	10.14 +/- 1.45	0.45 +/- 0.08
	LAB	6.52 +/- 0.90	12.25 +/- 29.88	9.63 +/- 1.48	0.48 +/- 0.08
	FLIP	6.60 +/- 0.89	12.20 +/- 32.50	9.53 +/- 1.46	0.47 +/- 0.08
	OCC	7.06 +/- 0.88	12.33 +/- 29.14	9.90 +/- 1.38	0.43 +/- 0.08
	FUSION	**6.03 +/- 0.85**	**11.47 +/- 27.67**	**8.71 +/- 1.37**	**0.54 +/- 0.07**
Stationary (after exercise)	HSV	11.93 +/- 1.21	18.31 +/- 55.32	12.42 +/- 1.15	0.28 +/- 0.08
	LAB	**11.46 +/- 1.23**	**18.23 +/- 58.11**	**11.84 +/- 1.15**	**0.29 +/- 0.08**
	FLIP	12.05 +/- 1.25	18.72 +/- 58.62	12.31 +/- 1.16	0.27 +/- 0.08
	OCC	12.44 +/- 1.24	18.93 +/- 56.78	12.57 +/- 1.11	0.22 +/- 0.09
	FUSION	11.66 +/- 1.22	18.26 +/- 56.74	12.02 +/- 1.14	**0.31 +/- 0.08**
Rotation	HSV	6.31 +/- 0.81	11.20 +/- 25.24	9.01 +/- 1.32	0.45 +/- 0.08
	LAB	5.94 +/- 0.79	10.77 +/- 25.56	8.34 +/- 1.24	0.50 +/- 0.08
	FLIP	5.91 +/- 0.77	10.63 +/- 25.65	8.62 +/- 1.31	0.52 +/- 0.08
	OCC	5.89 +/- 0.78	10.66 +/- 26.42	8.66 +/- 1.33	0.52 +/- 0.08
	FUSION	**5.88 +/- 0.76**	**10.47 +/- 24.79**	**8.19 +/- 1.23**	**0.53 +/- 0.07**
Talking	HSV	8.23 +/- 0.93	13.43 +/- 31.69	10.52 +/- 1.27	0.26 +/- 0.08
	LAB	7.13 +/- 0.90	12.57 +/- 31.54	9.34 +/- 1.27	0.35 +/- 0.08
	FLIP	**6.57 +/- 0.89**	**12.15 +/- 33.88**	**8.57 +/- 1.31**	**0.41 +/- 0.08**
	OCC	7.01 +/- 0.84	11.89 +/- 28.99	9.09 +/- 1.18	0.43 +/- 0.08
	FUSION	7.14 +/- 0.89	12.47 +/- 31.01	9.21 +/- 1.25	0.36 +/- 0.08
Walking	HSV	15.74 +/- 1.15	20.55 +/- 51.01	16.57 +/- 1.14	0.07 +/- 0.09
	LAB	14.91 +/- 1.18	20.16 +/- 53.61	15.58 +/- 1.16	**0.10 +/- 0.09**
	FLIP	14.26 +/- 1.22	**19.99 +/- 55.74**	**14.80 +/- 1.19**	0.00 +/- 0.09
	OCC	**13.99 +/- 1.25**	20.02 +/- 59.24	14.83 +/- 1.38	-0.08 +/- 0.09
	FUSION	14.97 +/- 1.19	20.30 +/- 56.41	15.71 +/- 1.17	0.07 +/- 0.09
All	HSV	9.87 +/- 0.47	15.65 +/- 18.64	11.74 +/- 0.58	0.35 +/- 0.04
	LAB	9.20 +/- 0.47	15.26 +/- 19.31	10.95 +/- 0.58	0.38 +/- 0.04
	FLIP	**9.09 +/- 0.48**	15.24 +/- 19.97	**10.77 +/- 0.59**	0.37 +/- 0.04
	OCC	9.28 +/- 0.47	15.28 +/- 19.71	11.01 +/- 0.58	0.35 +/- 0.04
	FUSION	9.14 +/- 0.47	**15.13 +/- 19.32**	10.78 +/- 0.56	**0.40 +/- 0.04**

References

[1] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.*, 124:82–90, 2019. 1

[2] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Trans. Biomed. Eng.*, 60(10): 2878–2886, 2013. 1

[3] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *ICCV*, pages 3995–4004, 2021. 1

[4] Min Hu, Fei Qian, Dong Guo, Xiaohua Wang, Lei He, and Fuji Ren. Eta-rppgnet: Effective time-domain attention network for remote heart rate measurement. *IEEE Trans. Instrum. Meas.*, 70:1–12, 2021. 1

[5] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *CVPR*, pages 12404–12413, 2021. 1

[6] Chaoqi Luo, Yiping Xie, and Zitong Yu. Physmamba: Efficient remote physiological measurement with slowfast temporal difference mamba. *arXiv preprint arXiv:2409.12031*, 2024. 1

[7] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 1

[8] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. Non-contrastive unsupervised learning of physiological signals from video. In *CVPR*, 2023. 1

[9] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *BMVC*, 2018. 1

[10] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mo-

bile service robot. In *RO-MAN*, pages 1056–1062. IEEE, 2014. 1

[11] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *ECCV*, pages 492–510. Springer, 2022. 1

[12] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: Multi-domain mobile video physiology dataset. *arXiv preprint arXiv:2302.03840*, 2023. 1

[13] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Trans. Biomed. Eng.*, 64(7):1479–1491, 2016. 1

[14] Bochao Zou, Zizheng Guo, Jiansheng Chen, and Huimin Ma. Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer. *arXiv preprint arXiv:2402.12788*, 2024. 1