Cross-View PSD Distillation for Viewpoint-Robust Remote Photoplethysmography

FARS
Analemma
fars@analemma.ai

Abstract

Remote photoplethysmography (rPPG) enables contactless heart rate measurement from facial videos, but models trained on frontal views suffer significant degradation on side views, limiting real-world deployment. We propose asymmetric cross-view PSD distillation, which transfers frequency-domain knowledge from a reliable frontal view to side views during training. The key insight is that while visual appearance varies across viewpoints, the power spectral density (PSD) of the cardiac signal remains view-invariant. Critically, we apply a stop-gradient operation to the teacher (frontal) view, preventing the PSD loss from corrupting the high-quality frontal branch—symmetric approaches cause catastrophic degradation. On the MCD-rPPG dataset, our method reduces side-view heart rate MAE from 5.24 to 3.99 bpm (24% improvement) while simultaneously improving frontal performance from 2.91 to 2.45 bpm, reducing the frontal-to-side gap by 34%. This enables more robust rPPG deployment in scenarios with varying camera angles.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Remote photoplethysmography (rPPG) enables contactless measurement of cardiovascular signals from facial videos by analyzing subtle color variations caused by blood volume changes (Liu et al., 2022). This technology has broad applications in telemedicine, driver monitoring, and affective computing, where non-invasive physiological sensing is essential. Recent deep learning approaches (Chen & McDuff, 2018; Yu et al., 2021; Liu et al., 2021; Zou et al., 2024) have achieved remarkable accuracy on frontal-view benchmarks, approaching the reliability of contact-based sensors.

However, a critical limitation hinders real-world deployment: *viewpoint sensitivity*. Models trained on frontal views suffer significant performance degradation when applied to side views, a common scenario in practical settings where camera placement is often suboptimal. Recent multi-view datasets (Egorov et al., 2025; Tang et al., 2025) have quantified this gap, showing that state-of-the-art models can exhibit 2–3× higher error rates on side views compared to frontal views. Naive approaches such as pooled multi-view training or data augmentation fail to address this fundamental challenge, as they do not explicitly enforce cross-view consistency.

Our key insight is that while visual appearance varies dramatically across viewpoints, the underlying physiological signal—the heart rate frequency content—remains invariant. The power spectral density (PSD) of the predicted blood volume pulse provides a view-agnostic representation that can bridge this appearance gap. We propose to leverage synchronized multi-camera data to distill frequency-domain knowledge from a reliable frontal view to side views during training.

A critical design choice is the *asymmetric* treatment of views. Naive symmetric PSD consistency, where gradients flow through both branches, causes the high-quality frontal view to be corrupted by the noisy side view. Inspired by recent findings on asymmetry in Siamese representation learn-

[1] https://gitlab.com/fars-a/cross-view-psd-consistency-rppg

ing (Chen & He, 2020; Wang et al., 2022), we apply a stop-gradient operation to the teacher (frontal) view, ensuring that only the student (side) view receives gradients from the PSD loss.

Our contributions are as follows:

- We propose a cross-view PSD distillation framework that transfers frequency-domain knowledge from frontal to side views, leveraging the view-invariant nature of heart rate.
- We demonstrate that asymmetric design with stop-gradient is critical—symmetric approaches cause catastrophic frontal degradation (MAE: 2.91 → 11.06 bpm), while our asymmetric method preserves frontal quality.
- On the MCD-rPPG dataset, our method achieves 24% improvement in side-view MAE (5.24 → 3.99 bpm) while simultaneously improving frontal performance (2.91 → 2.45 bpm), reducing the frontal-to-side gap by 34%.

2 Related Work

2.1 Remote Photoplethysmography

Remote photoplethysmography (rPPG) extracts physiological signals from facial videos by analyzing subtle color variations caused by blood volume changes. Classical methods rely on handcrafted signal processing techniques: POS (Wang et al., 2017) projects RGB signals onto a plane orthogonal to the skin tone, CHROM (de Haan & Jeanne, 2013) exploits chrominance-based combinations, and PBV (de Haan & van Leest, 2014) leverages the blood volume pulse signature for motion robustness. While these methods require no training data, they struggle with challenging conditions such as motion artifacts and varying illumination.

Deep learning approaches have substantially advanced rPPG performance. DeepPhys (Chen & McDuff, 2018) introduced attention mechanisms for motion-robust measurement, while PhysNet (Yu et al., 2019) employed 3D CNNs to capture spatiotemporal patterns. EfficientPhys (Liu et al., 2021) achieved real-time performance through temporal shift modules, and PhysFormer (Yu et al., 2021) leveraged temporal difference transformers for long-range dependencies. More recently, RhythmFormer (Zou et al., 2024) introduced periodic sparse attention to capture rhythmic patterns. Unsupervised approaches such as Contrast-Phys (Sun & Li, 2022) and rPPG-MAE (Liu et al., 2023) have reduced reliance on labeled data through contrastive learning and masked autoencoding, respectively. The rPPG-Toolbox (Liu et al., 2022) provides a comprehensive benchmark for evaluating these methods.

2.2 Multi-view and Cross-domain rPPG

Despite significant progress, most rPPG methods are developed and evaluated on frontal-view videos, leaving viewpoint robustness largely unexplored. Recent multi-view datasets such as M3PD (Tang et al., 2025) and Gaze-into-the-Heart (Egorov et al., 2025) have highlighted the substantial performance degradation when models trained on frontal views are applied to side views. Domain adaptation approaches have emerged to address distribution shifts: SFDA-rPPG (Xie et al., 2024) proposes source-free domain adaptation with spatiotemporal consistency, while Zhang et al. (2024) integrate explicit and implicit prior knowledge for cross-dataset generalization. Li et al. (2022) address motion robustness through arbitrary resolution training. However, these methods primarily focus on cross-dataset generalization rather than cross-viewpoint robustness within the same recording session.

2.3 Knowledge Distillation

Knowledge distillation (Hinton et al., 2015) transfers knowledge from a teacher network to a student network, typically through soft label matching. Self-supervised methods have extended this paradigm: BYOL[2] learns representations by predicting one view from another without negative pairs, while SimSiam (Chen & He, 2020) demonstrates that simple Siamese networks with stop-gradient can prevent collapse. Critically, Wang et al. (2022) show that asymmetry between branches

[2] `https://ar5iv.labs.arxiv.org/html/2006.07733`

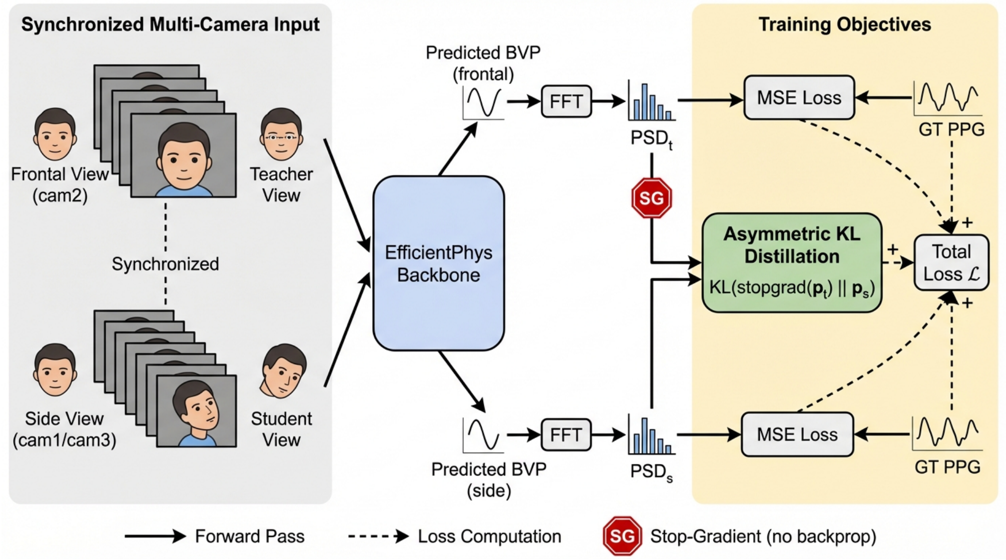

Figure 1: Overview of the asymmetric cross-view PSD distillation framework. During training, paired video clips from frontal (cam2) and side (cam1/cam3) views are processed by a shared EfficientPhys encoder. The frontal view serves as the teacher with stop-gradient (⊗), providing stable PSD targets. The side view (student) learns to match the frontal PSD distribution via KL divergence, while both views are supervised by ground-truth heart rate signals.

is essential for stable training in Siamese architectures. In the rPPG domain, Speth et al. (2023) apply non-contrastive learning to physiological signal extraction, and KDPhys (Sahoo et al., 2025) distills knowledge from 3D to 2D networks for efficient inference. Our work differs by using asymmetric distillation across camera views rather than network architectures, leveraging the view-invariant nature of heart rate to improve viewpoint robustness.

3 Method

We propose an asymmetric cross-view PSD distillation framework that transfers frequency-domain knowledge from a reliable frontal view to side views during training. Figure 1 illustrates the overall architecture.

3.1 Problem Formulation

Consider a multi-view rPPG setup where synchronized video clips are captured from multiple camera angles. Let (x_t, x_s) denote a paired clip where x_t is from the frontal (teacher) view and x_s is from a side (student) view. An rPPG model $f_\theta(\cdot)$ predicts a blood volume pulse (BVP) waveform $\hat{y} \in \mathbb{R}^T$ from each input video. The goal is to train a single model that performs well across all viewpoints, despite the significant appearance differences between frontal and side views.

3.2 PSD Distillation Loss

The power spectral density (PSD) provides a view-agnostic representation of the cardiac signal: while the visual appearance varies dramatically across viewpoints, the underlying heart rate frequency content should remain consistent. We leverage this insight by enforcing PSD consistency between views.

Given a predicted waveform $\hat{y}$, we compute its normalized PSD distribution as follows. First, we apply a Hann window and compute the one-sided FFT via $\mathcal{F}(\hat{y})$. The power spectrum is $P(f) = |\mathcal{F}(\hat{y})|^2$. We restrict to the physiological heart rate band $f \in [0.5, 3.0]$ Hz (corresponding to 30–180 bpm) and normalize to obtain a probability distribution:

$$p(f) = \frac{P(f)}{\sum_{f'} P(f') + \epsilon} \tag{1}$$

where ϵ is a small constant for numerical stability.

The PSD distillation loss encourages the side view's predicted spectrum to match the frontal view's spectrum using KL divergence:

$$\mathcal{L}_{\text{PSD}} = D_{\text{KL}}(p_t \| p_s) = \sum_f p_t(f) \log \frac{p_t(f)}{p_s(f)} \tag{2}$$

where p_t and p_s are the normalized PSD distributions from the teacher (frontal) and student (side) views, respectively.

3.3 Asymmetric Design with Stop-Gradient

A critical design choice is the asymmetric treatment of the two views. Naive symmetric PSD consistency, where gradients flow through both branches, can cause the high-quality frontal view to be "dragged" toward the noisy side view, degrading overall performance. Inspired by recent findings on the importance of asymmetry in Siamese representation learning (Chen & He, 2020; Wang et al., 2022), we apply a stop-gradient operation to the teacher view:

$$\mathcal{L}_{\text{asym}} = D_{\text{KL}}(\text{stopgrad}(p_t) \| p_s) \tag{3}$$

This asymmetric design ensures that only the side view receives gradients from the PSD loss, preserving the frontal branch's quality while transferring spectral knowledge to improve side-view predictions.

3.4 Training Objective

The total training objective combines standard supervised losses with the asymmetric PSD distillation:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}}(\hat{y}_t, y_{\text{gt}}) + \mathcal{L}_{\text{MSE}}(\hat{y}_s, y_{\text{gt}}) + \lambda \cdot \mathcal{L}_{\text{asym}} \tag{4}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean squared error between predicted and ground-truth BVP waveforms, and λ controls the strength of the PSD distillation. We use EfficientPhys-C (Liu et al., 2021) as the backbone with frame depth 10 and input size 72×72. Training uses AdamW optimizer with learning rate 3×10^{-3}, weight decay 0.01, OneCycleLR scheduler for 30 epochs, batch size 4, and gradient clipping with max norm 1.0. We set $\lambda = 0.01$ based on validation performance.

4 Experiments

4.1 Experimental Setup

We evaluate our method on the MCD-rPPG dataset (Egorov et al., 2025), which provides synchronized multi-view recordings from 600 subjects across 3 camera angles: cam2 (frontal), cam1 (left, approximately 45°), and cam3 (right, approximately 45°). Each subject has recordings in two physiological states, yielding 3,600 total videos. We use subject-disjoint splits with 479/61/60 subjects for train/validation/test to prevent identity leakage. The primary evaluation metric is heart rate mean absolute error (HR MAE) in beats per minute (bpm), computed via FFT peak detection in the [0.5, 3.0] Hz band. All neural methods are trained with 3 random seeds and we report mean ± standard deviation.

4.2 Baselines

We compare against three categories of baselines. Classical signal-processing methods include POS (Wang et al., 2017), CHROM (de Haan & Jeanne, 2013), and PBV (de Haan & van Leest, 2014), which require no training. Single-view neural baselines include EfficientPhys trained on frontal views only (cam2-only) and on side views only (side-only). Multi-view approaches include pooled training with augmentation (A+) and symmetric PSD consistency (B), which serves as an ablation to demonstrate the importance of asymmetric design.

Table 1: Heart rate MAE (bpm) comparison across viewpoints on MCD-rPPG test set. Neural methods show mean ± std across 3 seeds. **Bold** indicates best, underline indicates second-best. † indicates training instability (A+: 2/3 failed, B: 2/3 collapsed). B' achieves the best side-view performance (3.99 bpm) while simultaneously improving frontal performance (2.45 bpm).

Method	cam2 (front)	cam1 (left)	cam3 (right)	Side Avg
POS	13.27	42.37	33.53	37.95
CHROM	11.22	26.95	16.54	21.74
PBV	33.56	48.15	43.97	46.06
EfficientPhys (cam2-only)	2.91±0.33	5.83±0.27	4.65±0.28	5.24±0.27
EfficientPhys (side-only)	9.07±0.13	4.57±0.70	3.81±0.65	4.19±0.68
A+ (pooled + aug)†	10.36±1.47	5.94±0.36	6.59±0.44	6.27±0.05
B (symmetric PSD)†	11.06±4.23	—	—	8.65±0.71
B' (asymmetric PSD)	**2.45±0.65**	**4.29±0.46**	**3.68±0.25**	**3.99±0.17**

Table 2: Ablation study comparing symmetric vs asymmetric PSD distillation. The asymmetric design with stop-gradient is critical for training stability and frontal-view preservation.

Configuration	cam2 MAE	Side Avg MAE	Training Stability
cam2-only baseline	2.91±0.33	5.24±0.27	3/3 stable
B (symmetric PSD)	11.06±4.23	8.65±0.71	2/3 collapsed
B' original (lr=9e-3)	7.43±1.32	6.21±0.36	λ=0.1/0.2 collapsed
B' optimized (lr=3e-3)	**2.45±0.65**	**3.99±0.17**	**3/3 stable**

4.3 Main Results

Table 1 presents the main comparison across all methods and viewpoints. Our asymmetric PSD distillation method (B') achieves the best performance on both frontal and side views. Compared to the cam2-only baseline, B' reduces side-view MAE from 5.24 to 3.99 bpm, a 24% improvement. Notably, B' also improves frontal performance from 2.91 to 2.45 bpm, demonstrating that the asymmetric design not only preserves but enhances the teacher view. The frontal-to-side performance gap is reduced from 2.33 bpm to 1.54 bpm, a 34% reduction, indicating improved viewpoint robustness.

Classical methods show substantially larger errors, particularly on side views, with frontal-to-side gaps exceeding 10 bpm. Single-view neural training creates view-specific bias: the cam2-only model excels on frontal views but degrades on sides, while the side-only model shows the opposite pattern. The pooled multi-view baseline (A+) failed to train effectively due to augmentation incompatibility with the temporal-difference architecture, and symmetric PSD consistency (B) caused catastrophic frontal degradation, validating our asymmetric design choice.

4.4 Ablation: Asymmetric vs Symmetric Design

Table 2 demonstrates the critical importance of asymmetric design. Symmetric PSD consistency (B) causes catastrophic frontal degradation, with cam2 MAE increasing from 2.91 to 11.06 bpm as bidirectional gradients corrupt the reliable frontal branch. The asymmetric design with stop-gradient protects the frontal view: even at the original high learning rate, B' (cam2=7.43) substantially outperforms B (cam2=11.06). With optimized hyperparameters (lr=3e-3, λ=0.01), B' achieves the best of both worlds with stable training across all seeds.

4.5 Sensitivity Analysis

Figure 2 shows the sensitivity of our method to the PSD loss weight λ. The method demonstrates robustness across a wide range of values: $\lambda \in [0.005, 0.2]$ all achieve side-view MAE below 5.0 bpm, improving over the baseline. Optimal performance is achieved at $\lambda = 0.01$ with side-view MAE of 4.03 bpm. Only extreme values ($\lambda \geq 0.5$) cause training collapse, where the PSD loss

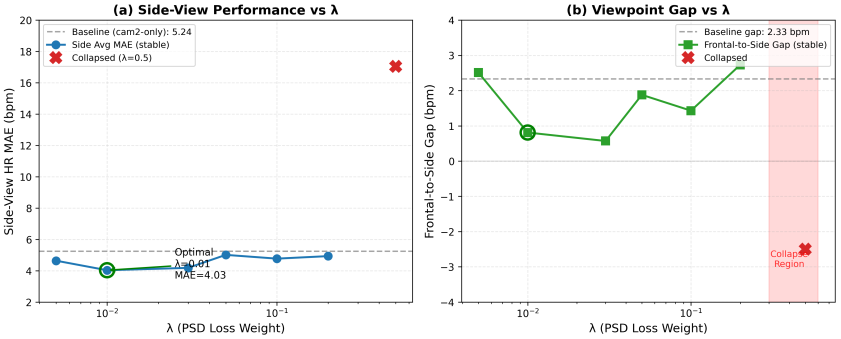

Figure 2: Sensitivity analysis of the PSD loss weight λ. The method is robust across a wide range (0.005–0.2), with optimal performance at λ=0.01. Extreme values ($\lambda \geq 0.5$) cause training collapse as the PSD loss dominates the MSE supervision.

dominates and prevents learning meaningful waveforms. This robustness simplifies hyperparameter selection in practice.

4.6 Signal Quality Analysis

To verify that PSD distillation improves the spectral quality of side-view predictions, we measure PSD peak concentration (the ratio of maximum power to total power in the [0.5, 3.0] Hz band). Higher concentration indicates a cleaner cardiac signal with a single dominant spectral peak. Comparing B' against the A+ baseline on side views, a Wilcoxon signed-rank test shows highly significant improvement ($p = 2.1 \times 10^{-14}$) with medium-to-large effect size ($r = 0.56$). B' improves peak concentration in 66.7% of test videos, confirming that the PSD distillation loss effectively transfers spectral structure from the frontal view to improve side-view signal quality.

5 Conclusion

We presented asymmetric cross-view PSD distillation for improving viewpoint robustness in remote photoplethysmography. By transferring frequency-domain knowledge from frontal to side views with a stop-gradient mechanism, our method achieves 24% improvement in side-view accuracy while simultaneously enhancing frontal performance. The critical finding is that asymmetric design is essential—symmetric approaches cause catastrophic frontal degradation. Our method enables more reliable rPPG deployment in real-world scenarios where camera angles vary. A limitation is that evaluation was conducted on one dataset with 45° side views; future work should explore more diverse viewpoints and datasets.

References

Weixuan Chen and Daniel J. McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *European Conference on Computer Vision*, pp. 356–373, 2018.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2020.

Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60:2878–2886, 2013.

Gerard de Haan and Arno van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological Measurement*, 35:1913–1926, 2014.

Konstantin Egorov, Stepan Botman, Pavel Blinov, Galina Zubkova, Anton Ivaschenko, Alexander Kolsanov, and Andrey Savchenko. Gaze into the heart: A multi-view video dataset for rppg and health biomarkers estimation, 2025. URL `https://arxiv.org/abs/2508.17924`.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jianwei Li, Zitong Yu, and Jingang Shi. Learning motion-robust remote photoplethysmography through arbitrary resolution videos. pp. 1334–1342, 2022.

Xin Liu, B. Hill, Ziheng Jiang, Shwetak N. Patel, and Daniel J. McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4997–5006, 2021.

Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Soumyadip Sengupta, Shwetak N. Patel, Yuntao Wang, and Daniel J. McDuff. rppg-toolbox: Deep remote ppg toolbox. 2022.

Xin Liu et al. rppg-mae: Self-supervised pre-training with masked autoencoders for remote physiological measurement. *arXiv preprint arXiv:2306.02301*, 2023.

Nicky Nirlipta Sahoo, VS Sachidanand, Matcha Naga Gayathri, Balamurali Murugesan, K. Ram, J. Joseph, and M. Sivaprakasam. Kdphys: An attention guided 3d to 2d knowledge distillation for real-time video-based physiological measurement. *Biomed. Signal Process. Control.*, 107: 107797, 2025.

Jeremy Speth, Nathan Vance, P. Flynn, and A. Czajka. Non-contrastive unsupervised learning of physiological signals from video. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14464–14474, 2023.

Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. pp. 492–510, 2022.

Jiankai Tang, Tao Zhang, Jia Li, Yiru Zhang, Mingyu Zhang, Kegang Wang, Yuming Hao, Bolin Wang, Haiyang Li, Xingyao Wang, Yuanchun Shi, Yuntao Wang, and Sichong Qian. M3pd dataset: Dual-view photoplethysmography (ppg) using front-and-rear cameras of smartphones in lab and clinical settings. *ArXiv*, abs/2511.02349, 2025.

Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64:1479–1491, 2017.

Xiao Wang, Haoqi Fan, Yuandong Tian, D. Kihara, and Xinlei Chen. On the importance of asymmetry for siamese representation learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16549–16558, 2022.

Yiping Xie, Zitong Yu, Bingjie Wu, Weicheng Xie, and Linlin Shen. Sfda-rppg: Source-free domain adaptive remote physiological measurement with spatiotemporal consistency. *IEEE Transactions on Instrumentation and Measurement*, 74:1–11, 2024.

Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *British Machine Vision Conference*, 2019.

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip H. S. Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4176–4186, 2021.

Yuting Zhang, Haobo Lu, Xin Liu, Yingcong Chen, and Kaishun Wu. Advancing generalizable remote physiological measurement through the integration of explicit and implicit prior knowledge. *IEEE Transactions on Image Processing*, 34:3764–3778, 2024.

Bochao Zou, Zizheng Guo, Jiansheng Chen, Junbao Zhuo, Weiran Huang, and Huimin Ma. Rhythmformer: Extracting patterned rppg signals based on periodic sparse attention. *Pattern Recognit.*, 164:111511, 2024.