KDPPG: PRIOR-GUIDED KNOWLEDGE DISTILLATION FOR ROBUST STREAMING RPPG ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote photoplethysmography (rPPG) enables contactless physiological monitoring, but robust low-latency deployment remains challenging. High-performance methods typically rely on buffered video clips, whereas streaming-compatible models often provide weaker temporal modeling and robustness. We propose KDPPG, a prior-guided knowledge distillation framework that transfers spatial and geometric knowledge from a clip-wise teacher to a causal student. KDPPG-T uses facial masks and projected normalized coordinate code (PNCC) in a dual-prior refinement module to suppress background illumination artifacts and improve robustness to motion. KDPPG-S processes one frame per time step, maintains a recurrent state, and incorporates a parallel harmonic resonance module (PHRM) to model periodic dynamics without fixed clip buffering or inference-time priors. Extensive experiments across multiple datasets demonstrate that KDPPG achieves state-of-the-art accuracy, while KDPPG-S offers high inference efficiency and low computational cost, making it suitable for practical streaming deployment.

1 INTRODUCTION

Blood volume pulse (BVP) is a vital physiological signal that enables the non-invasive estimation of key parameters such as heart rate (HR) and respiration. Photoplethysmography (PPG) is the standard optical technique used to detect these volumetric changes in subcutaneous blood vessels by measuring light absorption and scattering. However, conventional contact-based PPG requires sensors to be physically attached to the skin, which inevitably causes discomfort and restricts its usability in sensitive scenarios such as neonatal monitoring and long-term health tracking (Sun & Thakor, 2015). To overcome these physical constraints, remote photoplethysmography (rPPG) has emerged as a promising method for extracting pulse signals from facial videos by tracking imperceptible skin color variations driven by the cardiac cycle (Verkruysse et al., 2008). Although rPPG offers a safer and more convenient solution for applications such as telemedicine (Yan et al., 2018) and affective computing (Yu et al., 2021), its practical deployment still faces a fundamental dilemma: accurate and robust estimation benefits from long-range temporal modeling and auxiliary priors, yet these are difficult to reconcile with low-latency inference. Therefore, achieving both accurate HR estimation under disturbances and efficient real-time inference remains a major challenge.

Accurate rPPG signal extraction remains inherently challenging, as the subtle skin color variations induced by blood flow are easily corrupted by dynamic illumination changes and motion artifacts. Although traditional signal processing techniques provided foundational solutions, their reliance on handcrafted rules limits their robustness to ambient light fluctuations and head motion (Verkruysse et al., 2008; Poh et al., 2010; De Haan & Jeanne, 2013; Pilz et al., 2018; De Haan & Van Leest, 2014; Wang et al., 2016), thereby motivating the transition to deep learning approaches. From the perspective of the inference mechanism, existing deep learning models for rPPG can be broadly categorized into frame-wise and clip-wise architectures. Frame-wise approaches process each frame independently (Špetlík et al., 2018; Chen & McDuff, 2018; Liu et al., 2020; 2023a), whereas clip-wise approaches take a buffered video clip as input and rely on non-causal temporal operations, such as 3D convolutions or global temporal self-attention (Yu et al., 2019a;b; 2022; 2023; Zou et al., 2025a;b), to model multi-frame temporal dependencies. These two paradigms exhibit different limitations in practical deployment: frame-wise approaches are naturally compatible with low-latency

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

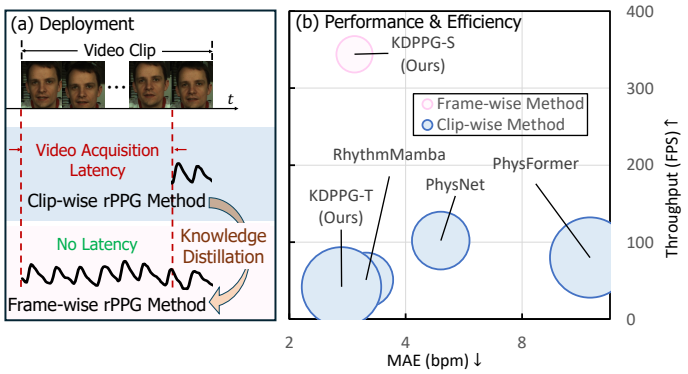

Figure 1: Latency and efficiency characteristics of KDPPG. (a) Inference latency in the deployment scenario. (b) Performance and efficiency metrics on the MMPD dataset. The diameter of each circle represents the peak GPU memory during a single forward inference. While the clip-wise KDPPG-T achieves the lowest MAE on MMPD, it offers no clear efficiency advantage over other clip-wise approaches. In contrast, the distilled frame-wise KDPPG-S attains state-of-the-art accuracy among practical streaming models, while providing substantially higher throughput and lower peak GPU memory than the evaluated clip-wise approaches.

inference, but their restricted temporal receptive field limits robustness and accuracy, whereas clip-wise approaches achieve stronger temporal modeling at the cost of buffering latency.

The practical deployment of rPPG in real-world scenarios is hindered by two fundamental challenges: severe performance degradation under complex interference and insufficient real-time capability caused by video acquisition latency. First, rPPG signals are inherently susceptible to dynamic illumination changes and head motion. Frame-wise approaches remain vulnerable due to their limited temporal modeling capability. Although clip-wise approaches exploit multi-frame context, they often compromise efficiency by restricting model capacity and avoiding available auxiliary priors, which still constrains their robustness in challenging environments. Second, the reliance on clip-wise inference limits real-time performance. Capturing long-period physiological rhythms typically requires buffering 4–5 seconds of video (Zou et al., 2025b; Braun et al., 2024; Cheng et al., 2020), as illustrated in Fig. 1(a). This video acquisition latency cannot be eliminated through model compression and significantly degrades the user experience, as subjects are required to remain relatively still during video acquisition.

To address severe performance degradation under complex interference and the limited real-time capability of existing approaches, inspired by knowledge distillation (Gou et al., 2021; Gu et al., 2026), we propose KDPPG, a knowledge distillation framework tailored to robust frame-wise rPPG prediction. Instead of sacrificing informative priors for deployment efficiency, the proposed framework fully exploits multimodal priors during training and transfers the resulting robustness from a clip-wise teacher model, KDPPG-T, to a frame-wise student model, KDPPG-S. Specifically, KDPPG-T explicitly exploits spatial priors, including facial masks and projected normalized coordinate code (PNCC). Through the proposed dual-prior refinement module (DPRM), KDPPG-T effectively mitigates the adverse effects of dynamic illumination changes and head motion. KDPPG-S adopts a frame-wise architecture to eliminate the latency bottleneck caused by long-video buffering. To model periodic physiological dynamics under single-frame input, KDPPG-S incorporates a parallel harmonic resonance module (PHRM) for sequential temporal modeling. Fig. 1(b) shows that KDPPG-S outperforms existing clip-wise baselines in terms of the accuracy-efficiency trade-off. Moreover, both KDPPG-T and KDPPG-S achieve state-of-the-art performance on several datasets. This demonstrates that KDPPG achieves an effective balance between efficiency and accuracy, thereby enabling real-time rPPG prediction in real-world scenarios. Our main contributions are summarized as follows:

- We propose KDPPG, a knowledge distillation framework for robust frame-wise rPPG prediction. By distilling rich prior-guided representations from a clip-wise teacher into an

efficient frame-wise student, KDPPG achieves high estimation accuracy while enabling real-time streaming inference without video acquisition latency.

- We design KDPPG-T, a clip-wise teacher network that leverages facial masks and PNCC as spatial priors. Through the dual-prior refinement module (DPRM), KDPPG-T effectively mitigates the adverse effects of dynamic illumination changes and head motion in complex scenarios.
- We develop KDPPG-S, a frame-wise student network that incorporates a parallel harmonic resonance module (PHRM) to enhance periodic signal modeling. This design enables the model to capture long-range physiological dynamics while processing frames causally, alleviating the temporal modeling limitations typically associated with frame-wise architectures.
- Extensive experiments across multiple datasets demonstrate that KDPPG achieves state-of-the-art accuracy, while KDPPG-S maintains high inference efficiency and low computational cost.

2 Related Work

Early research on rPPG predominantly relies on traditional signal processing techniques to extract weak physiological rhythms from facial videos (Verkruysse et al., 2008; Poh et al., 2010). Subsequent studies (De Haan & Jeanne, 2013; De Haan & Van Leest, 2014; Wang et al., 2016) leverage the spectral properties of skin reflections to improve measurement robustness. However, these handcrafted methods remain highly vulnerable to head motions and illumination variation in unconstrained scenarios.

To overcome the limitations of traditional approaches, data-driven methods have been widely adopted. Early deep learning-based approaches, including HR-CNN (Špetlík et al., 2018), TS-CAN (Liu et al., 2020), DeepPhys (Chen & McDuff, 2018), and EfficientPhys (Liu et al., 2023a), primarily perform frame-wise estimation using 2D spatial convolutions. Although these architectures offer high computational efficiency, extracting robust physiological signals from isolated frames remains fundamentally challenging. To explicitly capture spatiotemporal dependencies, most current state-of-the-art approaches rely on clip-wise architectures that require long video clips (e.g., 128 to 160 frames) as input. Architectures leveraging 3D convolutions (Yu et al., 2019a;b) or spatiotemporal Transformers (Yu et al., 2022; Liu et al., 2024) have been explored for modeling periodic physiological dynamics. This clip-wise paradigm has been further advanced through various architectural innovations, including fast-slow temporal branches (Yu et al., 2023) and matrix-factorized attention mechanisms (Joshi et al., 2024). More recently, selective state-space models have also been introduced (Zou et al., 2025b) to improve the efficiency of long-range temporal modeling while reducing computational overhead.

Nevertheless, existing clip-wise models still struggle to maintain high accuracy in complex scenarios and inevitably suffer from video acquisition latency, whereas frame-wise architectures lack sufficient temporal modeling capability. These limitations motivate us to develop a knowledge distillation framework that transfers prior-guided robust representations to an efficient frame-wise student for robust streaming inference.

3 Method

Conventional clip-wise rPPG models typically rely on multi-frame inputs and auxiliary priors, which substantially hinder low-latency deployment. To address this limitation, we propose KDPPG, a novel framework that distills rich prior-guided representations from a clip-wise teacher into an efficient frame-wise student. Section 3.1 first presents an overview of the proposed knowledge distillation pipeline. Section 3.2 details the clip-wise teacher network, KDPPG-T, which is equipped with the dual-prior refinement module (DPRM). Section 3.3 then introduces the lightweight frame-wise student network, KDPPG-S, featuring the parallel harmonic resonance module (PHRM).

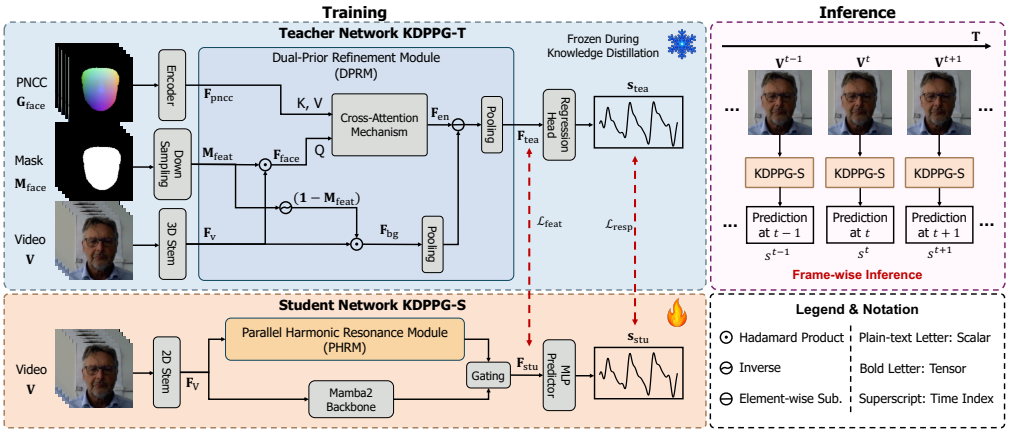

Figure 2: Overview of the KDPPG knowledge distillation framework. In the first training stage, we train a teacher network KDPPG-T utilizing PNCC and facial mask as auxiliary inputs. In the second stage, KDPPG-T is frozen to train a student network KDPPG-S that accepts only video input. During inference, KDPPG-S enables efficient frame-wise streaming inference.

3.1 Overview of KDPPG Knowledge Distillation Framework

The proposed KDPPG framework achieves accurate rPPG estimation while maintaining low-latency inference. By leveraging knowledge distillation, it transfers the robust prior-guided capabilities of a complex teacher model to an efficient frame-wise architecture, enabling real-time streaming inference without clip buffering or auxiliary prior computation.

Fig. 2 illustrates the overall pipeline of the KDPPG framework, which consists of a two-stage training phase and a lightweight inference phase. In the first training stage, we optimize the teacher network, KDPPG-T. It takes the video clip $\mathbf{V} \in \mathbb{R}^{3\times T\times H\times W}$, the projected normalized coordinate code (Zhu et al., 2016) $\mathbf{G}_{\text{pncc}} \in \mathbb{R}^{3\times T\times H\times W}$, and the facial mask $\mathbf{M}_{\text{face}} \in \mathbb{R}^{T\times H\times W}$ as inputs. KDPPG-T fuses these geometric and physiological priors to explicitly suppress ambient light fluctuations and motion artifacts, producing the teacher's feature representation $\mathbf{F}_{\text{tea}} \in \mathbb{R}^{D\times T}$ and the rPPG prediction $\mathbf{s}_{\text{tea}} \in \mathbb{R}^{T}$. In the second stage, the parameters of the well-trained KDPPG-T are frozen. The student network, KDPPG-S, is then trained using only the video clip $\mathbf{V}$ to produce the student's feature representation $\mathbf{F}_{\text{stu}} \in \mathbb{R}^{D\times T}$ and rPPG prediction $\mathbf{s}_{\text{stu}} \in \mathbb{R}^{T}$. Here D, T, H and W represent channels, length, height and width of input, respectively. A knowledge distillation loss $\mathcal{L}_{\text{KD}}$ is then employed at both feature and response levels to enforce consistency between the student and teacher networks, as detailed in Appendix A.

During inference, KDPPG-S can operate in a strictly frame-wise streaming manner. At time step t, given only the current frame $\mathbf{V}^t \in \mathbb{R}^{3\times H\times W}$, KDPPG-S yields the prediction $s^t \in \mathbb{R}$.

3.2 Teacher Network: KDPPG-T with Prior Knowledge

In unconstrained scenarios, abrupt lighting fluctuations and motion artifacts can easily overwhelm the subtle skin color variations, leading to substantial accuracy degradation. The teacher network, KDPPG-T, is designed to learn robust representations in unconstrained scenarios by jointly exploiting clip-wise spatiotemporal context and multimodal priors. Structurally, it consists of a set of encoders, a dual-prior refinement module (DPRM), and a regression head. Given the input video clip $\mathbf{V}$, a spatiotemporal stem first extracts the visual features $\mathbf{F}_{\text{v}} \in \mathbb{R}^{D\times T\times H'\times W'}$, where H' and W' represent height and width of feature, respectively. Meanwhile, the geometric prior $\mathbf{G}_{\text{pncc}}$ is processed by a 3D-CNN encoder to obtain the geometric features $\mathbf{F}_{\text{pncc}} \in \mathbb{R}^{D\times T\times H'\times W'}$, while the facial mask $\mathbf{M}_{\text{face}}$ is downsampled to $\mathbf{M}_{\text{feat}} \in \mathbb{R}^{T\times H'\times W'}$ to match the spatial dimensions of the extracted features.

To deeply integrate these priors, the extracted features are fed into the DPRM. Within this module, we first utilize mask-guided artifact suppression to decompose the visual features $\mathbf{F}_{\text{v}}$

into face-specific representations $\mathbf{F}_{\text{face}} \in \mathbb{R}^{D\times T\times H'\times W'}$ and background representations $\mathbf{F}_{\text{bg}} \in \mathbb{R}^{D\times T\times H'\times W'}$. Next, we use geometry-guided cross-attention to aggregate the face representations $\mathbf{F}_{\text{face}}$ with the geometric features $\mathbf{F}_{\text{pncc}}$, yielding the enhanced features $\mathbf{F}_{\text{en}} \in \mathbb{R}^{D\times T\times H'\times W'}$. Finally, the estimated background illumination component $\mathbf{F}_{\text{bg}}$ is explicitly subtracted from $\mathbf{F}_{\text{en}}$ to produce the teacher's feature representation $\mathbf{F}_{\text{tea}}$. $\mathbf{F}_{\text{tea}}$ is subsequently mapped to the rPPG prediction $\mathbf{s}_{\text{tea}}$ by the regression head. The specific designs of mask-guided artifact suppression and geometry-guided cross-attention within the DPRM are detailed in the following sections.

3.2.1 Mask-guided Artifact Suppression

Environmental illumination variations can severely corrupt subtle blood volume pulse signals. To mitigate this effect, we introduce a mask-guided artifact suppression mechanism that explicitly decomposes the visual features $\mathbf{F}_{\text{v}}$ into face and background components according to spatial semantics, where the background component primarily captures illumination-related cues from non-facial regions.

Specifically, utilizing the downsampled facial mask $\mathbf{M}_{\text{feat}}$, the face-specific representations $\mathbf{F}_{\text{face}}$ and the background representations $\mathbf{F}_{\text{bg}}$ are formulated as

$$\mathbf{F}_{\text{face}} = \mathbf{F}_{\text{v}} \odot \mathbf{M}_{\text{feat}}, \quad \mathbf{F}_{\text{bg}} = \mathbf{F}_{\text{v}} \odot (1 - \mathbf{M}_{\text{feat}}), \tag{1}$$

where $\odot$ denotes the Hadamard product. During this operation, $\mathbf{M}_{\text{feat}}$ is broadcast along the channel dimension D of $\mathbf{F}_{\text{v}}$. Then we apply spatial average pooling on the background representation $\mathbf{F}_{\text{bg}}$. This pooled noise feature is then explicitly subtracted from the motion-aligned enhanced feature $\mathbf{F}_{\text{en}}$ (to be introduced in Section 3.2.2). Finally, group normalization is applied to yield the feature representation $\mathbf{F}_{\text{tea}}$,

$$\mathbf{F}_{\text{tea}} = \text{GN}(\mathbf{F}_{\text{en}} - \text{AvgPool}(\mathbf{F}_{\text{bg}})), \tag{2}$$

where $\text{GN}(\cdot)$ and $\text{AvgPool}(\cdot)$ denote group normalization and spatial average pooling, respectively.

3.2.2 Geometry-guided Cross-Attention

To mitigate spatial misalignment caused by head motion, we design a geometry-guided cross-attention mechanism that leverages PNCC to dynamically align the facial representation $\mathbf{F}_{\text{face}}$. The PNCC is derived using a pre-trained 3DDFA-V2 model (Guo et al., 2020), which provides a dense 3D topological prior by mapping facial structures into a normalized coordinate space. This makes the extracted geometric representation inherently less sensitive to head motion. By employing the encoded PNCC feature $\mathbf{F}_{\text{pncc}}$ as an explicit spatial anchor, this mechanism aligns features based on robust 3D geometric correspondences rather than solely RGB appearance.

Specifically, $\mathbf{F}_{\text{face}}$ and $\mathbf{F}_{\text{pncc}}$ are flattened and then linearly projected into $\mathbf{H}^q_{\text{face}}$, $\mathbf{H}^k_{\text{pncc}}$, and $\mathbf{H}^v_{\text{pncc}}$, which serve as the query, key, and value, respectively. The motion-aligned enhanced feature $\mathbf{F}_{\text{en}}$ is computed as

$$\mathbf{F}_{\text{en}} = \mathbf{F}_{\text{face}} + \text{Softmax}\left(\frac{\mathbf{H}^q_{\text{face}}(\mathbf{H}^k_{\text{pncc}})^\top}{\sqrt{d_k}}\right)\mathbf{H}^v_{\text{pncc}}, \tag{3}$$

where d_k is the feature dimension used for scaling, and the attention output is reshaped back to its original spatial dimensions before the residual addition.

3.3 Student Network: Efficient Frame-wise KDPPG-S

Although the clip-wise teacher model achieves superior accuracy, its computational overhead and reliance on clip-level temporal contexts hinder low-latency deployment. To bridge this gap, we design KDPPG-S, a lightweight student network that achieves high throughput and low memory consumption by performing rPPG estimation in a strictly frame-wise manner.

During training, given an input video clip $\mathbf{V}$, a 2D spatial stem first extracts the visual feature maps $\mathbf{F}_{\text{v}} \in \mathbb{R}^{D\times T\times H'\times W'}$. These feature maps are then flattened along the spatial dimensions and processed by an attention pooling layer. Unlike standard spatial averaging, this mechanism adaptively focuses on physiologically salient regions and removes the spatial dimensions to form a sequence of frame-level representations $\mathbf{x}^t \in \mathbb{R}^D$. The resulting sequence is then fed into the

temporal modeling module to produce the student's deep temporal features $\mathbf{F}_{\text{stu}}$, which are mapped by an MLP predictor to the estimated signal sequence $\mathbf{s}_{\text{stu}}$. During inference, KDPPG-S operates in a frame-wise manner. At time step t, it processes only the current frame $\mathbf{V}^t$ while maintaining the temporal states accumulated from preceding frames, and directly outputs the corresponding rPPG value s^t.

To explicitly model the periodic nature of rPPG signals, the temporal modeling module consists of a Mamba-2 block and the proposed parallel harmonic resonance module (PHRM) connected in parallel and fused through a learnable gating mechanism. A standard Mamba-2 block updates its hidden state $\mathbf{h}^t$ given the input $\mathbf{x}^t$ through a selective state-space mechanism,

$$\mathbf{h}^t = \mathbf{A}^t\mathbf{h}^{t-1} + \mathbf{B}^t\mathbf{x}^t, \;\; \mathbf{y}_m^t = \mathbf{C}^t\mathbf{h}^t + \mathbf{D}\mathbf{x}^t. \tag{4}$$

In common parameterizations, the state transition tensor $\mathbf{A}^t$ is constrained to a diagonal matrix representing stable decays, formulated as $\mathbf{A}^t \approx \text{diag}(\exp(-\boldsymbol{\Delta}^t \odot \mathbf{a}))$ with $\mathbf{a} > 0$. Consequently, the temporal dynamics are dominated by exponential memory. However, modeling sustained physiological oscillations requires complex-conjugate eigenmodes, i.e., explicit rotational dynamics, which are structurally absent in this decay-dominated Mamba formulation.

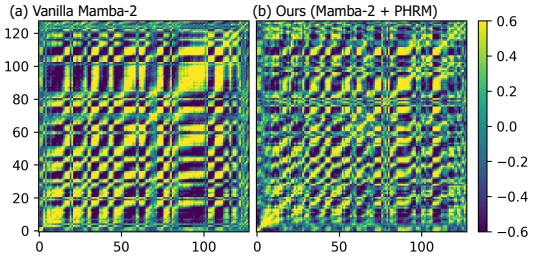

Figure 3: Visualization of feature self-similarity matrices. (a) Vanilla Mamba-2 is dominated by blocky local similarities. While it captures coarse periodicities, it suffers from temporal oversmoothing and struggles to preserve fine-grained rhythmic continuity due to its exponential decay formulation. (b) Ours (Mamba-2 + PHRM) demonstrates a distinct, sustained periodic lattice structure with clear off-diagonal bands.

To bridge this representational gap, the PHRM explicitly models frequency-aware oscillations using K parallel harmonic oscillators. Given the input feature $\mathbf{x}^t$, we first project it into a $2K$-dimensional driving-signal space,

$$\mathbf{u}^t = \tanh(\mathbf{W}_B\mathbf{x}^t + \mathbf{b}_B), \tag{5}$$

where $\mathbf{W}_B$ and $\mathbf{b}_B$ are learnable projection parameters. The output $\mathbf{u}^t \in \mathbb{R}^{2K}$ is partitioned into K independent 2D state-driving signals $\mathbf{u}_k^t \in \mathbb{R}^2$. Each oscillator k maintains a 2D state $\mathbf{h}_k^t \in \mathbb{R}^2$, corresponding to its real and imaginary components. Its internal state is updated iteratively using a damped rotation,

$$\mathbf{h}_k^t = \lambda_k\mathbf{R}(\omega_k)\mathbf{h}_k^{t-1} + (1-\lambda_k)\mathbf{u}_k^t, \tag{6}$$

where ω_k is the learnable frequency, $\lambda_k = \sigma(\alpha_k) \in (0,1)$ is the damping factor, and $\mathbf{R}(\omega_k)$ denotes the 2D rotation matrix,

$$\mathbf{R}(\omega_k) = \begin{bmatrix} \cos\omega_k & -\sin\omega_k \\ \sin\omega_k & \cos\omega_k \end{bmatrix}. \tag{7}$$

The transition matrix $\lambda_k\mathbf{R}(\omega_k)$ provides explicit complex-conjugate eigenmodes $\lambda_k e^{\pm j\omega_k}$, enabling the network to support sustained periodic oscillations at specific frequencies.

After the state update, the updated states $\mathbf{h}_k^t$ across all K oscillators are concatenated into a flat vector $\mathbf{z}^t \in \mathbb{R}^{2K}$ and linearly projected to obtain the harmonic representation $\mathbf{y}_h^t$,

$$\mathbf{y}_h^t = \mathbf{W}_C\mathbf{z}^t + \mathbf{b}_C. \tag{8}$$

To combine the exponential memory of Mamba-2 with the periodic modeling capability of the PHRM, we aggregate the outputs of the two parallel branches using a ReZero gate,

$$\mathbf{y}^t = \mathbf{y}_m^t + \boldsymbol{\gamma} \odot \mathbf{y}_h^t, \tag{9}$$

where $\gamma \in \mathbb{R}^D$ is a learnable gating parameter initialized to zero. This aggregated feature $\mathbf{y}^t$ serves as the refined sequence representation for subsequent physiological signal extraction. Fig. 3 further illustrates the effectiveness of the PHRM. While features from vanilla Mamba-2 exhibit blocky local similarity due to temporal over-smoothing, our aggregated features $\mathbf{y}^t$ preserve a distinct sustained periodic lattice structure.

Please refer to the Appendix A for the loss function design.

4 Experiments

4.1 Comparison Results

We evaluate intra-dataset performance on four benchmarks using established data split protocols described in Appendix B.1. As shown in Table 1, the proposed clip-wise teacher model KDPPG-T achieves state-of-the-art performance across most metrics on PURE, UBFC, and COHFACE. On the MMPD dataset, both proposed models demonstrate leading performance, KDPPG-T obtains the lowest MAE, while the distilled frame-wise student KDPPG-S obtains the lowest RMSE and highest ρ. These findings indicate that our knowledge distillation framework enables the efficient frame-wise KDPPG-S to achieve performance comparable to that of clip-wise architectures.

Table 1: Intra-dataset evaluation results on several datasets. The best, the second-best and the third-best results are in **red**, blue and green respectively. "F" and "C" in the Type column denote frame-wise and clip-wise rPPG methods, respectively. "-" denotes results that are either not reported in the original paper or unavailable under the unified evaluation protocol.

Method	Type	PURE			UBFC			COHFACE			MMPD		
		MAE	RMSE	ρ	MAE	RMSE	ρ	MAE	RMSE	ρ	MAE	RMSE	ρ
Face2PPG$_{best}$(Casado & López, 2023)	F	1.20	1.80	-	0.80	1.10	-	7.50	9.80	-	-	-	-
DeepPhys(Chen & McDuff, 2018)	F	0.83	1.54	**0.99**	6.27	10.82	0.65	6.89	13.89	0.34	22.00	28.13	-0.07
TS-CAN(Liu et al., 2020)	F	2.48	9.01	0.92	1.70	2.72	**0.99**	3.74	9.68	0.71	12.03	20.38	0.17
EfficientPhys(Liu et al., 2023a)	F	-	-	-	1.14	1.81	**0.99**	31.27	33.53	0.11	13.31	20.73	0.28
PhysNet(Yu et al., 2019a)	C	2.10	2.60	**0.99**	2.95	3.67	0.97	5.38	10.76	-	4.93	11.85	0.62
ETA-rPPGNet(Hu et al., 2021)	C	0.34	0.77	**0.99**	1.46	3.97	0.93	4.67	6.65	0.77	-	-	-
PhysFormer(Yu et al., 2022)	C	1.10	1.75	**0.99**	0.50	0.71	**0.99**	11.70	13.83	0.07	11.99	18.41	0.18
Li et al.(Li & Yin, 2023)	C	0.64	1.16	**0.99**	0.48	0.64	**0.99**	-	-	-	-	-	-
PFE-TFA(Li et al., 2023)	C	-	-	-	-	-	-	1.31	3.92	-	-	-	-
SSPD-L3(Zhang et al., 2024)	C	0.53	1.16	**0.99**	0.50	0.71	**0.99**	-	-	-	-	-	-
Contrast-Phys+(Sun & Li, 2024)	C	0.48	0.98	**0.99**	0.21	0.80	**0.99**	-	-	-	-	-	-
FactorizePhys(Joshi et al., 2024)	C	3.08	13.84	0.91	2.08	3.63	0.98	5.25	17.68	0.53	5.93	12.74	0.62
RhythmMamba(Zou et al., 2025b)	C	0.23	0.34	**0.99**	0.50	0.75	**0.99**	11.44	15.98	0.15	3.16	7.27	0.84
KDPPG-T	C	**0.16**	**0.33**	**0.99**	**0.14**	**0.27**	**0.99**	**0.93**	**2.41**	**0.98**	**2.73**	6.86	0.85
KDPPG-S	F	0.23	0.47	**0.99**	0.18	0.31	**0.99**	1.30	3.68	0.95	2.95	**6.83**	**0.86**

We further conduct cross-dataset evaluations to assess the generalization capability of our method. Following the protocol described in Appendix B.1, Table 2 reports the generalization performance on the UBFC dataset. Both KDPPG-T and KDPPG-S outperform the competing baselines, with KDPPG-S maintaining performance comparable to that of the teacher model. These results indicate that the proposed knowledge distillation strategy effectively preserves domain generalization capability in the frame-wise student network.

Table 2 further reports the evaluation results on the MMPD dataset. Under this setting, our clip-wise teacher model KDPPG-T achieves state-of-the-art performance across all metrics, while the distilled frame-wise student model KDPPG-S attains the second-best overall performance.

4.2 Ablation Study

We conduct several ablation studies on the MMPD dataset to evaluate the contribution of the KDPPG framework.

Ablation Study of KDPPG-T Components. We first evaluate the impact of DPRM in KDPPG-T. Table 3 evaluates the mask-guided artifact suppression mechanism under varying illumination conditions on the MMPD dataset. Overall, incorporating the facial mask reduces the MAE from 3.611 to 2.731. Under constant illumination (100 lm, 200 lm, and 300 lm), this mechanism consistently

Table 2: Cross-dataset evaluation results. Models are trained on the PURE dataset and evaluated on the UBFC and MMPD datasets. The first 7 rows are traditional methods, others are deep-learning methods. The best, the second-best and the third-best results are in **red**, blue and green respectively. "F" and "C" in the Type column denote frame-wise and clip-wise rPPG methods, respectively.

Method	Type	PURE → UBFC				PURE → MMPD			
		MAE	RMSE	MAPE	ρ	MAE	RMSE	MAPE	ρ
Green(Verkruysse et al., 2008)	F	19.73	31.00	18.72	0.37	21.68	27.69	24.39	-0.01
ICA(Poh et al., 2010)	F	16.00	25.65	15.35	0.44	18.60	24.30	20.88	0.01
CHROM(De Haan & Jeanne, 2013)	F	4.06	8.83	3.84	0.89	13.66	18.76	16.00	0.08
LGI(Pilz et al., 2018)	F	15.80	28.55	14.70	0.36	13.66	18.76	16.00	0.08
PBV(De Haan & Van Leest, 2014)	F	15.90	26.40	15.17	0.48	17.95	23.58	20.18	0.09
POS(Wang et al., 2016)	F	4.08	7.72	3.93	0.92	12.36	17.71	14.43	0.18
OMIT(Casado & López, 2023)	F	15.96	28.57	14.94	0.36	17.18	23.07	19.23	0.07
DeepPhys(Chen & McDuff, 2018)	F	1.21	2.90	1.42	**0.99**	16.92	24.61	18.54	0.05
TS-CAN(Liu et al., 2020)	F	1.30	2.87	1.50	**0.99**	13.94	21.61	15.15	0.20
EfficientPhys(Liu et al., 2023a)	F	2.07	6.32	2.10	0.94	14.03	21.62	15.32	0.17
PhysNet(Yu et al., 2019a)	C	0.98	2.48	1.12	**0.99**	13.93	20.29	15.61	0.17
PhysFormer(Yu et al., 2022)	C	1.44	3.77	1.66	0.98	14.57	20.71	16.73	0.15
Spiking-Phys(Liu et al., 2025)	C	2.80	-	2.81	0.95	14.57	-	16.55	0.14
FactorizePhys(Joshi et al., 2024)	C	1.04	2.44	1.23	**0.99**	-	-	-	-
RhythmMamba(Zou et al., 2025b)	C	0.95	1.83	1.04	**0.99**	10.44	16.70	12.25	0.36
KDPPG-T	C	**0.40**	**0.60**	**0.40**	**0.99**	**8.05**	**13.93**	**9.42**	**0.48**
KDPPG-S	F	**0.40**	**0.60**	0.41	**0.99**	9.60	16.07	11.19	0.33

improves all evaluation metrics. This indicates that explicitly utilizing the facial mask effectively isolates skin pixels from non-informative regions, thereby mitigating the impact of static specular reflections. Even under dynamic lighting condition ranging from 300-800 lm, this strategy reduces both MAE and MAPE, demonstrating its effectiveness in reducing estimation errors under severe illumination shifts.

Table 3: Ablation study of the mask-guided artifact suppression under different light conditions on the MMPD dataset. The categories "100 lm", "200 lm", and "300 lm" denote constant facial illumination at specific intensity levels, whereas "300-800 lm" represents dynamically changing illumination within a single video clip. The "All Types" category evaluates the overall performance across all evaluated lighting conditions.

Setting	Light	MAE	RMSE	MAPE	ρ
w/o Mask	100 lm	3.842	8.248	4.033	0.778
	200 lm	2.052	5.680	2.082	0.887
	300 lm	3.549	7.495	3.941	0.855
	300-800 lm	4.999	10.367	4.930	0.704
	All Types	3.611	8.123	3.746	0.805
w/ Mask	100 lm	2.299	5.017	2.405	0.910
	200 lm	1.481	3.188	1.558	0.971
	300 lm	3.024	6.435	3.175	0.885
	300-800 lm	4.120	10.569	3.899	0.658
	All Types	2.731	6.864	2.759	0.854

Table 4: Ablation study of the geometry-guided cross-attention under different motion conditions on the MMPD dataset. In the MMPD dataset, motions involving head movement include "Rotation" and "Walking", whereas "Stationary" and "Talking" represent stationary head states. The "All Types" category evaluates the overall performance across all four of these conditions.

Setting	Motion	MAE	RMSE	MAPE	ρ
w/o PNCC	Rotation	3.240	5.305	3.783	0.832
	Walking	11.399	16.802	10.650	-0.250
	All Types	3.414	8.143	3.397	0.793
w/ PNCC	Rotation	2.623	4.413	3.081	0.890
	Walking	9.007	14.464	8.512	0.173
	All Types	2.731	6.864	2.759	0.854

Table 4 presents the ablation of the geometry-guided cross-attention mechanism to evaluate network's robustness against head movements. Incorporating the PNCC spatial prior yields consistent performance gains under both the "Rotation" and "Walking" conditions. Notably, the "Walking" scenario introduces severe rigid head translation and scale variation, causing substantial performance degradation in the baseline network ($\rho = -0.250$, MAE = 11.399). The geometry-guided cross-attention mechanism reduces the MAE to 9.007 and improves ρ to 0.173. These results suggest that the 3D facial structural priors provided by PNCC improve motion robustness by facilitating spatial alignment under large motions.

Ablation Study of KDPPG-S Components. Table 5 evaluates the impact of the PHRM in the student model on the MMPD dataset. Incorporating the PHRM reduces the MAE from 3.45 to 2.95 and improves ρ from 0.79 to 0.86. These consistent improvements across all metrics validate the effectiveness of the PHRM design. By explicitly introducing periodic modeling into the Mamba-2 architecture, the PHRM enables the network to better capture the inherent rhythmic patterns of rPPG signals. Consequently, the frame-wise student model KDPPG-S can reconstruct more accurate and robust rPPG estimation.

Table 5: Ablation study of the PHRM on the MMPD dataset.

Setting	MMPD Dataset			
	MAE	RMSE	MAPE	ρ
w/o PHRM	3.45	8.31	3.58	0.79
w/ PHRM	2.95	6.83	3.13	0.86

4.3 Computational Cost

We use the protocol described in Appendix B.2 to evaluate the computational cost. Table 6 compares the computational cost and inference efficiency of various rPPG methods. Existing frame-wise models are substantially more lightweight than clip-wise architectures but suffer from severe accuracy degradation. To improve accuracy, the proposed KDPPG-T incorporates explicit prior knowledge within a computationally heavy architecture. As a result, although KDPPG-T achieves the lowest MAE, it incurs considerable computational overhead and provides no efficiency advantage. However, through our knowledge distillation framework, the student model KDPPG-S achieves a favorable accuracy-efficiency trade-off. Compared with clip-wise approaches, it substantially reduces computational cost and memory usage while providing higher throughput, and achieves state-of-the-art accuracy that is second only to the teacher model.

Table 6: Computational cost and inference time. MAE is tested on MMPD dataset. "F" and "C" in the Type column denote frame-wise and clip-wise rPPG methods, respectively.

Method	Type	Param. (M)	MACs (G)	Throughput (FPS)	Memory (MB)	MAE (bpm)
DeepPhys	F	1.98	0.74	1798	178.04	22.00
TS-CAN	F	1.98	0.74	1356	186.60	12.03
EfficientPhys	F	1.91	0.37	1550	124.61	13.31
PhysNet	C	0.77	70.12	102	432.96	4.92
PhysFormer	C	7.38	50.61	80	598.53	11.99
RhythmMamba	C	1.07	12.94	51	408.95	3.16
KDPPG-T	C	9.46	237.30	42	593.33	2.73
KDPPG-S	F	2.97	2.35	344	276.54	2.95

Please refer to Appendix C and Appendix D for detailed description of datasets, performance metrics and implementation details. More experiment results can be found in Appendix E.

5 Conclusions

In this work, we propose KDPPG, a knowledge distillation framework for frame-wise rPPG estimation, addressing two key challenges in real-time rPPG deployment: severe performance degradation under complex interference and insufficient real-time capability caused by video acquisition latency. Extensive experiments demonstrate that the teacher model KDPPG-T, together with the DPRM, effectively mitigates the adverse effects of illumination variation and head motion by exploiting explicit priors, while the student model KDPPG-S, equipped with the PHRM, efficiently models temporal and periodic dynamics during frame-wise inference. The comparison results and computational cost analysis further demonstrate that KDPPG achieves state-of-the-art performance across multiple benchmarks, while KDPPG-S maintains high inference efficiency and low computational cost, highlighting the potential of the proposed framework for real-time contactless physiological monitoring.

6 AI USAGE DISCLOSURE

In this work, generative AI tools were used to check code correctness, improve manuscript readability, and suggest manuscript organization. They were not used to design the methodology or experiments, formulate hypotheses, implement the proposed method, or interpret results. All AI-assisted code-review and writing suggestions were manually examined and verified by the authors. The authors take responsibility for the final content, including all text, claims, code, and other AI-assisted artifacts.

REFERENCES

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

Björn Braun, Daniel McDuff, and Christian Holz. How suboptimal is training rPPG models with videos and targets from different body sites? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 410–418, 2024.

Constantino Alvarez Casado and Miguel Bordallo López. Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces. *IEEE Journal of Biomedical and Health Informatics*, 27(11):5530–5541, 2023.

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

Juan Cheng, Ping Wang, Rencheng Song, Yu Liu, Chang Li, Yong Liu, and Xun Chen. Remote heart rate measurement from near-infrared videos based on joint blind source separation with delay-coordinate transformation. *IEEE Transactions on Instrumentation and Measurement*, 70: 1–13, 2020.

Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):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(9):1913, 2014.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Zejun Gu, Zhong-Qiu Zhao, Henghui Ding, Hao Shen, Zhao Zhang, and De-Shuang Huang. Cross-domain knowledge distillation for low-resolution human pose estimation. *IEEE Transactions on Multimedia*, 2026.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Guillaume Heusch, André Anjos, and Sébastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017.

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 Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

Jitesh Joshi, Sos S Agaian, and Youngjun Cho. Factorizephys: Matrix factorization for multidimensional attention in remote physiological sensing. *arXiv preprint arXiv:2411.01542*, 2024.

Jianwei Li, Zitong Yu, and Jingang Shi. Learning motion-robust remote photoplethysmography through arbitrary resolution videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1334–1342, 2023.

Zhihua Li and Lijun Yin. Contactless pulse estimation leveraging pseudo labels and self-supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20588–20597, 2023.

Mingxuan Liu, Jiankai Tang, Yongli Chen, Haoxiang Li, Jiahao Qi, Siwei Li, Kegang Wang, Jie Gan, Yuntao Wang, and Hong Chen. Spiking-physformer: Camera-based remote photoplethysmography with parallel spike-driven transformer. *Neural Networks*, 185:107128, 2025.

Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020.

Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5008–5017, 2023a.

Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. *Advances in Neural Information Processing Systems*, 36:68485–68510, 2023b.

Xin Liu, Yuting Zhang, Zitong Yu, Hao Lu, Huanjing Yue, and Jingyu Yang. rPPG-MAE: Self-supervised pretraining with masked autoencoders for remote physiological measurements. *IEEE Transactions on Multimedia*, 26:7278–7293, 2024.

Hao Lu, Hu Han, and S Kevin 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*, pp. 12404–12413, 2021.

Christian S Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1254–1262, 2018.

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.

Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021.

Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference*, pp. 3–6, 2018.

Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1056–1062. IEEE, 2014.

Yu Sun and Nitish Thakor. Photoplethysmography revisited: from contact to noncontact, from point to imaging. *IEEE Transactions on Biomedical Engineering*, 63(3):463–477, 2015.

Z. Sun and X. Li. Contrast-Phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2024.

Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. MMPD: Multi-domain mobile video physiology dataset. In *Proceedings of the International Conference of the IEEE Engineering in Medicine & Biology Society*, pp. 1–5, 2023. doi: 10.1109/EMBC40787.2023.10340857.

Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008.

Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

Bryan P Yan, William HS Lai, Christy KY Chan, Stephen Chun-Hin Chan, Lok-Hei Chan, Ka-Ming Lam, Ho-Wang Lau, Chak-Ming Ng, Lok-Yin Tai, Kin-Wai Yip, et al. Contact-free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals. *Journal of the American Heart Association*, 7(8):e008585, 2018.

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

Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 151–160, 2019b.

Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38(6):50–58, 2021.

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4186–4196, 2022.

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Yawen Cui, Jiehua Zhang, Philip Torr, and Guoying Zhao. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, 131(6):1307–1330, 2023.

Xinyu Zhang, Weiyu Sun, Hao Lu, Ying Chen, Yun Ge, Xiaolin Huang, Jie Yuan, and Yingcong Chen. Self-similarity prior distillation for unsupervised remote physiological measurement. *IEEE Transactions on Multimedia*, 26:10290–10305, 2024.

Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 146–155, 2016.

Bochao Zou, Zizheng Guo, Jiansheng Chen, Junbao Zhuo, Weiran Huang, and Huimin Ma. RhythmFormer: Extracting patterned rPPG signals based on periodic sparse attention. *Pattern Recognition*, 164:111511, 2025a.

Bochao Zou, Zizheng Guo, Xiaocheng Hu, and Huimin Ma. RhythmMamba: Fast, Lightweight, and Accurate Remote Physiological Measurement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11077–11085, 2025b.

A Appendix: Loss Function

Our training framework consists of two phases: teacher pre-training for KDPPG-T and knowledge distillation training for KDPPG-S. In both phases, standard rPPG measurement losses are used for basic task supervision, including a negative Pearson correlation loss L_{time} to constrain waveform morphology with respect to the ground truth and a frequency-domain Kullback–Leibler divergence loss L_{freq} to supervise spectral consistency.

A.1 Appendix: Teacher Network Pre-training

To encourage the teacher network to learn highly disentangled features robust to motion artifacts, we further introduce a disentanglement loss $\mathcal{L}_{\text{dis}}$. This loss enforces orthogonality between the spatially aggregated teacher features $\mathbf{F}_{\text{tea}}$ and the estimated background noise features $\mathbf{F}_{\text{bg}}$ by minimizing their squared cosine similarity,

$$\mathcal{L}_{\text{dis}} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbf{F}_{\text{tea}}^{t}\cdot\mathbf{F}_{\text{bg}}^{t}}{\|\mathbf{F}_{\text{tea}}^{t}\|\|\mathbf{F}_{\text{bg}}^{t}\|}\right)^{2}. \tag{10}$$

The overall teacher objective is formulated as

$$\mathcal{L}_{\text{teacher}} = \mathcal{L}_{\text{time}} + \alpha_1 \mathcal{L}_{\text{freq}} + \alpha_2 \mathcal{L}_{\text{dis}}, \tag{11}$$

where α_1 and α_2 are balancing weights.

A.2 Appendix: Student Network Knowledge Distillation

During knowledge distillation, the frozen teacher supervises the student network through a dual-level alignment strategy, applied together with the task-specific losses $\mathcal{L}_{\text{time}}$ and $\mathcal{L}_{\text{freq}}$. At the response level, we transfer soft morphological knowledge by computing the Pearson correlation between the student and teacher rPPG estimations,

$$\mathcal{L}_{\text{resp}} = 1 - \text{Pearson}(\mathbf{s}_{\text{stu}}, \mathbf{s}_{\text{tea}}). \tag{12}$$

At the feature level, the student features $\mathbf{F}_{\text{stu}}$ are mapped to the teacher feature space defined by $\mathbf{F}_{\text{tea}}$. To bridge the channel-dimension discrepancy, we employ an MLP-based projector $\phi(\cdot)$ and maximize their cosine similarity across time steps,

$$\mathcal{L}_{\text{feat}} = 1 - \frac{1}{T} \sum_{t=1}^{T} \cos(\phi(\mathbf{F}_{\text{stu}}^t), \mathbf{F}_{\text{tea}}^t). \tag{13}$$

Finally, the total student loss is defined as

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{time}} + \beta_1 \mathcal{L}_{\text{freq}} + \beta_2 \mathcal{L}_{\text{resp}} + \beta_3 \mathcal{L}_{\text{feat}}, \tag{14}$$

where β_1, β_2, and β_3 are balancing weights.

B Appendix: Experiment Protocols

B.1 Appendix: Comparison Experiment Protocol

We follow the established intra- and cross-dataset evaluation protocols used in prior rPPG studies.

Intra-dataset evaluation. PURE dataset: 60% of the data is used for training and 40% for testing (Lu et al., 2021; Song et al., 2021). COHFACE dataset: 60% of the data is used for training and 40% for testing (Heusch et al., 2017). UBFC dataset: the first 30 subjects are used for training and the remaining 12 subjects are used for testing (Lu et al., 2021; Song et al., 2021). MMPD dataset: the dataset is divided into training, validation, and testing subsets containing 70%, 10%, and 20% of the data, respectively (Tang et al., 2023).

Cross-dataset evaluation. The training dataset PURE is split into 80% for training and 20% for validation, and the trained models are subsequently evaluated on the UBFC and MMPD datasets (Liu et al., 2023b).

B.2 Appendix: Computational Cost Evaluation Protocol

To ensure a fair and practically meaningful evaluation, particularly for real-time edge deployment, we redefine the protocol for measuring computational cost. Standard protocols often average the metrics over a long video sequence, which can obscure the true instantaneous inference burden of clip-wise methods. Specifically, our metrics are defined as follows: **MACs** are measured for a single forward pass strictly according to the required input shape (a single frame for frame-wise methods and a 160-frame sequence for clip-wise methods), rather than averaging the total MACs of a long video by its frame count. **Throughput** is evaluated on a 1000-frame video stream to simulate continuous real-time processing. Frame-wise models execute 1000 sequential forward passes, whereas clip-wise models execute 841 passes using a sliding-window scheme (window size of 160, stride of 1). **Memory** is defined as the absolute peak VRAM footprint during a single forward pass—accounting for model parameters, input tensors, and intermediate activations—without any temporal averaging.

C Appendix: Datasets and Performance Metrics

We conduct our experiments on the PURE (Stricker et al., 2014), UBFC-rPPG (Bobbia et al., 2019), COHFACE (Heusch et al., 2017), and MMPD (Tang et al., 2023) datasets.

PURE. The PURE dataset (Stricker et al., 2014) contains 60 one-minute facial video sequences from 10 subjects, recorded under natural lighting. The subjects performed six head motion tasks. The videos have a resolution of 640×480 pixels and a frame rate of 30 Hz.

UBFC-rPPG. The UBFC-rPPG dataset (Bobbia et al., 2019) contains 42 videos, each about 2 minutes long. The videos have a resolution of 640×480 pixels and a frame rate of 30 Hz. The ground-truth bio-signals are captured using a pulse oximeter with a 60 Hz sampling rate.

COHFACE. The COHFACE dataset (Heusch et al., 2017) has 160 video sequences, with each lasting one minute, collected from 40 subjects. The videos are recorded at a resolution of 640×480 pixels and 20 Hz using a Logitech HD C525 webcam. The corresponding blood volume pulse (BVP) and respiration signals are simultaneously recorded using Thought Technologies sensors and the BioGraph Infiniti software.

MMPD. The MMPD dataset (Tang et al., 2023) contains 660 one-minute videos recorded using a Samsung Galaxy S22 Ultra, with a resolution of 320×240 pixels at 30 FPS. It includes 33 participants with Fitzpatrick skin types 3-6, who perform four activities under four lighting conditions. Ground truth PPG signals are captured using a HKG-07C+ oximeter at 200 Hz, downsampled to 30 Hz.

We use Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Pearson Correlation Coefficient (ρ) to evaluate the performance of our proposed method.

D Appendix: Implementation Details

Our method is implemented using PyTorch 2.4.0. We employ the AdamW optimizer with a weight decay of 1×10^{-4} and adopt a CosineAnnealing learning rate schedule with a linear warm-up with warm-up ratio of 0.1 and minimum learning rate ratio of 1×10^{-2}, updated per iteration. For the teacher model, the initial learning rate is 1×10^{-3} and it is trained for 30 epochs. For the student model, the initial learning rate is 3×10^{-4} and it is trained for 50 epochs. Both models are trained with a batch size of 48. In our implementation, the hyperparameter weights for the loss functions are empirically set as follows: $\alpha_1 = 0.1$, $\alpha_2 = 10$, $\beta_1 = 0.1$, $\beta_2 = 0.1$, and $\beta_3 = 1.0$. All experiments are performed on an NVIDIA RTX A6000 GPU.

For all methods, we adopt the same preprocessing and postprocessing steps. Video inputs are divided into 128-frame segments for training. During inference, each method follows its native inference protocol, while KDPPG-S operates in a frame-wise streaming manner. The face region is cropped and resized to 128×128 pixels. For postprocessing, a second-order Butterworth filter with cutoff frequencies of 0.75-2.5 Hz is applied to filter the rPPG waveform. We use the Welch algorithm to compute the power spectral density.

E Appendix: Visualization

E.1 Appendix: Comparison Results

Fig. 4 present a representative example of the predicted and ground-truth signals on the MMPD dataset, illustrating that the proposed models not only capture the correct heart-rate frequencies but also reconstruct physiologically consistent temporal waveforms. Fig. 5 further shows that both KDPPG-T and KDPPG-S exhibit low bias and strong correlation with the ground truth in the intra-dataset experiment on the MMPD dataset.

Fig. 6 illustrates the Bland-Altman and scatter plots of the models trained on PURE and evaluated on UBFC. The results show a tightly bounded 95% limit of agreement, and the predictions closely follow the identity line, indicating excellent cross-domain consistency. This generalization capability is attributed to the higher domain diversity of the PURE dataset than UBFC dataset. The resulting

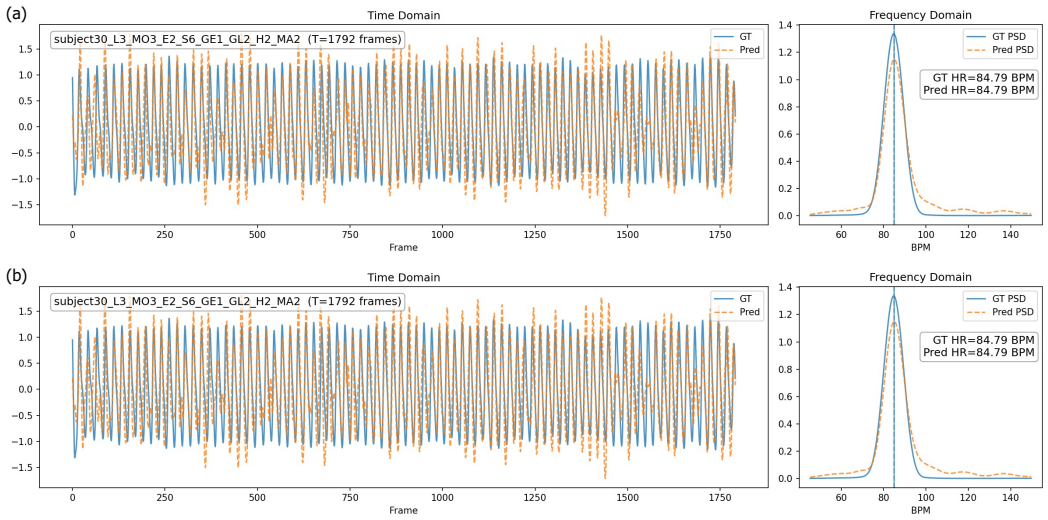

Figure 4: Visual comparison of the rPPG signals predicted by (a) KDPPG-T and (b) KDPPG-S on the MMPD dataset. For visual clarity, all signals are bandpass-filtered (0.75–2.5 Hz) to isolate the HR-related components. The left column displays the time-domain waveforms, while the right column presents the corresponding power spectral density (PSD), where the dominant peaks indicate the estimated and ground-truth heart rates.

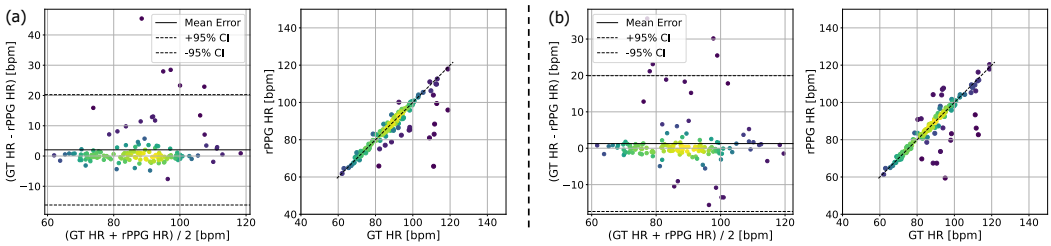

Figure 5: Bland-Altman plots and scatter plots of (a) KDPPG-T and (b) KDPPG-S from intra-dataset evaluation on MMPD. "CI" denotes the confidence interval.

broader feature space can better encompass the data distribution of the UBFC dataset, facilitating the extraction of domain-invariant physiological representations.

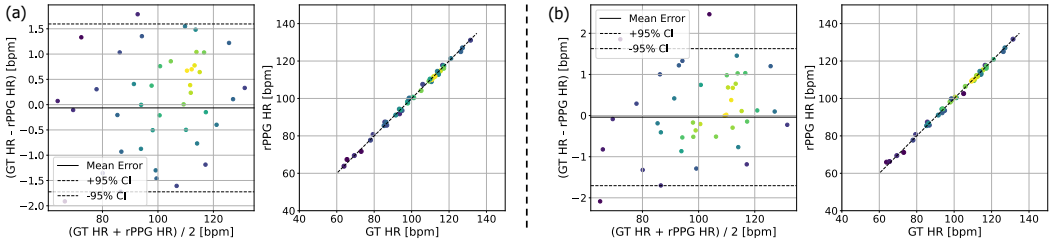

Figure 6: Bland-Altman and scatter plots of (a) KDPPG-T and (b) KDPPG-S under the cross-dataset setting, where the models are trained on PURE and evaluated on UBFC. "CI" denotes the confidence interval.

Figure 7 depicts the corresponding Bland-Altman and scatter plots for the MMPD evaluation. Compared with the UBFC results, the Bland-Altman plot reveals a relatively wider 95% confidence interval, and the scatter plot captures increased deviations from the identity line. This performance difference can be attributed to the domain gap between the two datasets. Specifically, training on

PURE provides limited exposure to the broader environmental variations present in the larger-scale MMPD dataset. Consequently, the model generalizes less effectively when evaluated on a target distribution with higher domain complexity. Despite this domain shift, most predictions remain consistent with the ground truth, which further demonstrates the robustness of the proposed framework.

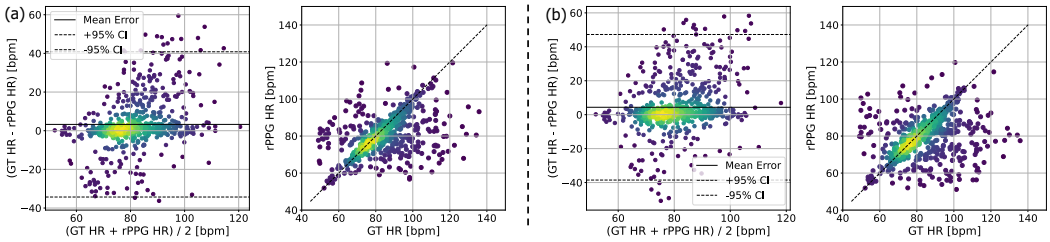

Figure 7: Bland-Altman and scatter plots of (a) KDPPG-T and (b) KDPPG-S under the cross-dataset setting, where the models are trained on PURE and evaluated on MMPD. "CI" denotes the confidence interval.

E.2 Appendix: Ablation Study

Fig. 8 investigates the influence of different pre-trained teacher models on the distillation process. Using the proposed KDPPG-T achieves the best performance (MAE = 2.95, ρ = 0.86), validating its architectural advantages over existing baselines and confirming its effectiveness in providing high-quality supervisory signals. Moreover, the progressive performance variations across the three settings indicate that the accuracy of the student model is positively correlated with the capability of the source teacher. When distilling from other representative clip-wise architectures, namely PhysNet and RhythmMamba, the student network still achieves competitive MAEs of 3.74 and 3.24, respectively. These results demonstrate the robustness of the proposed KDPPG-S architecture and confirm the generality of our knowledge distillation strategy across diverse teacher models.

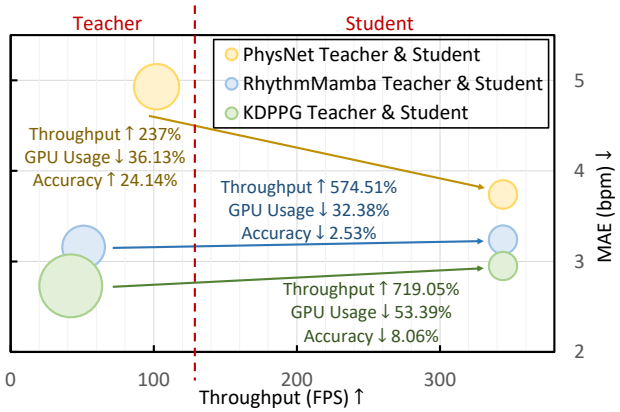

Figure 8: Ablation study on different teacher models distilled into the KDPPG-S student model, evaluated on the MMPD dataset. The diameter of each circle represents the peak GPU memory during a single forward inference.