EXPLOITING NEAR-INFRARED PRIORS FOR ILLUMINATION-ROBUST RPPG ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote photoplethysmography (rPPG) enables contact-free physiological measurement from facial videos, but remains vulnerable to illumination variations that can easily overwhelm subtle pulse-related appearance changes. RGB-NIR sensing provides a promising solution by introducing near-infrared observations that are relatively more stable than visible-light imaging. However, raw RGB-NIR features still contain illumination-related disturbance, and direct fusion may propagate it into the final prediction. To address this problem, we propose the Near-Infrared Prior Network (NIRP-Net), a prior-guided framework for illumination-robust RGB-NIR rPPG estimation. NIRP-Net first performs prior-guided feature disentanglement to derive disturbance-suppressed RGB and NIR physiological representations by exploiting the 940-nm low-irradiance and physiology-irrelevant background priors. It then introduces gated cross-attention aggregation to selectively retrieve and aggregate complementary NIR cues for temporal prediction. We further collect FaceIVL, a new RGB-NIR facial rPPG dataset under controlled illumination variations. Experiments on public datasets and FaceIVL demonstrate state-of-the-art accuracy and greater computational efficiency than existing RGB-NIR approaches.

1 INTRODUCTION

Remote photoplethysmography (rPPG) enables contact-free measurement of physiological signals from facial videos (Verkruysse et al., 2008) and has attracted growing attention in healthcare (Yan et al., 2018; Jiang et al., 2025), affective computing (Yu et al., 2021), and driver monitoring (Savic & Zhao, 2025). Benefiting from recent advances in deep neural networks, rPPG methods have achieved promising performance under controlled laboratory conditions. However, reliable physiological measurement in real-world scenarios is still far from solved, because the target pulsatile signal is extremely subtle and can be affected by various factors (Yu et al., 2019b; Zhao et al., 2022; Zhang et al., 2024). Among these factors, illumination variation is particularly challenging, as it directly corrupts facial appearance dynamics and easily overwhelms subtle physiological cues (Nowara et al., 2020).

To improve the robustness of physiological measurement under complex illumination, RGB-NIR sensing provides a practical multimodal solution by offering complementary observations of facial dynamics. RGB videos contain rich color variations related to blood volume changes, but these variations are highly sensitive to visible-light fluctuations. In contrast, NIR observations have been shown to be relatively more stable in many scenes and can therefore serve as a reliable auxiliary modality for physiological analysis (Shao & Hu, 2025; Nowara et al., 2020; Magdalena Nowara et al., 2018), rather than being regarded as an illumination-invariant source. In addition, non-facial background regions do not contain target pulsatile information and can provide useful references for characterizing environmental disturbance. Together, these priors suggest that RGB and NIR should not be treated as equivalent inputs for naive fusion, but should instead be exploited according to their distinct roles in physiological representation learning and disturbance suppression.

Despite the promise of RGB-NIR sensing, robust rPPG estimation under complex illumination remains challenging. Existing RGB-NIR methods mainly rely on direct feature fusion or prior-guided feature-difference modeling (Liu et al., 2023a; Ying et al., 2025; Shao & Hu, 2025). As illustrated in Fig. 1, direct fusion lacks explicit disturbance-aware feature refinement, whereas feature-difference

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.

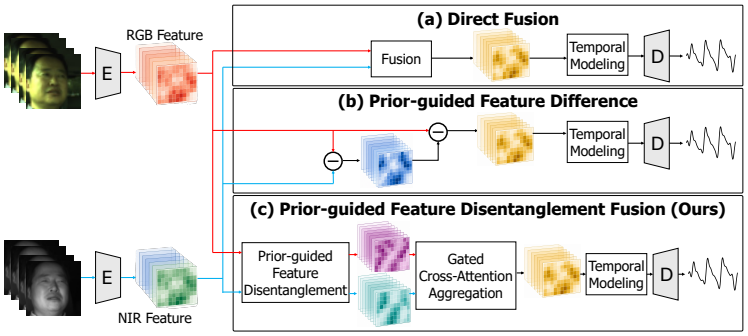

Figure 1: Comparison of three representative RGB-NIR fusion paradigms for robust rPPG measurement: (a) Direct Fusion, which directly merges RGB and NIR features; (b) Prior-guided Feature Difference, which exploits prior knowledge through feature difference between the RGB and NIR features; and (c) Prior-guided Feature Disentanglement Fusion (ours), which first disentangles modality-specific physiological representations from disturbance-related cues and then performs gated cross-attention aggregation for robust rPPG estimation. "E" and "D" denote the encoder and rPPG decoder, respectively.

modeling uses NIR primarily as a subtraction reference. These limitations leave two fundamental challenges unresolved. The first challenge is feature entanglement: physiological variations are tightly entangled with illumination-induced appearance changes in raw spatiotemporal features. Although NIR and background priors can provide useful references for disturbance suppression, existing strategies do not explicitly construct disturbance-suppressed modality-specific physiological representations, allowing residual illumination-related components to affect the final prediction. The second challenge is cross-modal complementarity: RGB and NIR observations exhibit different reliability and disturbance characteristics under complex lighting conditions. Treating them as equivalent inputs or using NIR merely as a subtraction reference limits the adaptive exploitation of complementary physiological cues. These challenges therefore call for both disturbance-aware feature disentanglement and reliability-aware complementary fusion.

To address these challenges, we propose the Near-Infrared Prior Network (NIRP-Net), a prior-guided framework for robust RGB-NIR rPPG estimation under complex illumination. Instead of treating NIR merely as an additional input channel, NIRP-Net incorporates near-infrared and background-derived disturbance priors into the modeling process to construct disturbance-aware physiological representations before cross-modal aggregation. Specifically, it first performs prior-guided feature disentanglement to derive disturbance-suppressed RGB and NIR physiological features, and then introduces gated cross-attention aggregation to retrieve complementary NIR cues through local cross-attention and adaptively aggregate them for rPPG waveform prediction. In addition to evaluations on public RGB-NIR rPPG datasets, we collect FaceIVL to evaluate RGB-NIR physiological measurement under controlled normal and low-light indoor conditions. Experimental results on both public datasets and FaceIVL demonstrate the effectiveness of the proposed framework. Our main contributions are summarized as follows:

- We propose NIRP-Net, a prior-guided RGB-NIR rPPG framework that combines prior-guided feature disentanglement (PFD) and gated cross-attention aggregation (GCA) to suppress illumination-related disturbance and exploit cross-modal complementarity
- We collect FaceIVL, a new RGB-NIR facial rPPG dataset for controlled evaluation under controlled illumination variations.
- Extensive experiments on public datasets and FaceIVL demonstrate state-of-the-art accuracy, while ablation and computational analyses verify the proposed design and its efficiency relative to existing RGB-NIR methods.

2 Related Work

Early rPPG methods primarily rely on traditional signal processing techniques to extract physiological signals from facial videos (Verkruysse et al., 2008; Poh et al., 2010; De Haan & Jeanne, 2013; De Haan & Van Leest, 2014; Wang et al., 2016). With the advancement of deep learning, data-driven approaches have become the dominant paradigm. Early deep-learning methods, including HR-CNN (Špetlík et al., 2018), DeepPhys (Chen & McDuff, 2018), TS-CAN (Liu et al., 2020), and EfficientPhys (Liu et al., 2023b), are mainly built upon 2D convolutional networks for RGB-based physiological representation learning. Subsequent methods improve temporal modeling using 3D convolutional networks (Yu et al., 2019a;b), spatiotemporal attention mechanisms, and Transformer-based architectures (Yu et al., 2022; Zou et al., 2025a; Liu et al., 2024). More recently, several works explore improved temporal modeling strategies, including fast-slow dual-branch designs (Yu et al., 2023), factorized attention mechanisms (Joshi et al., 2024), and state-space sequence modeling (Zou et al., 2025b). Despite these advances, RGB-based methods fundamentally rely on subtle visible-light appearance variations, making them vulnerable to illumination changes, shadows, and brightness fluctuations in unconstrained environments. This limitation motivates the exploration of complementary spectral information for more robust physiological measurement.

To reduce the sensitivity of RGB-based rPPG to visible-light variations, NIR sensing has been explored as a complementary modality for physiological measurement. Early RGB-NIR studies introduced paired RGB/NIR facial datasets and showed the potential of narrow-band NIR imaging, especially in driving and illumination-varying scenarios (Magdalena Nowara et al., 2018; Nowara et al., 2020). Recent methods further exploit NIR information for robust heart-rate estimation, including NIR-prior-based disturbance suppression (Shao & Hu, 2025), information-enhanced RGB-NIR fusion (Liu et al., 2023a), and flexible multimodal fusion across visible, NIR, and other sensing modalities (Ying et al., 2025). Although these methods demonstrate the value of multimodal sensing, they usually treat NIR as an auxiliary input, a difference reference, or a generic complementary modality. In contrast, NIRP-Net explicitly incorporates near-infrared prior and background-derived disturbance cues to refine modality-specific features before adaptive cross-modal aggregation.

3 Method

Robust RGB-NIR rPPG estimation requires disentangling weak physiological cues from illumination-induced appearance changes and adaptively exploiting cross-modal complementarity. To address these challenges, we propose NIRP-Net, which first refines modality-specific physiological features through prior-guided feature disentanglement and then selectively aggregates complementary NIR cues through gated cross-attention aggregation for rPPG prediction.

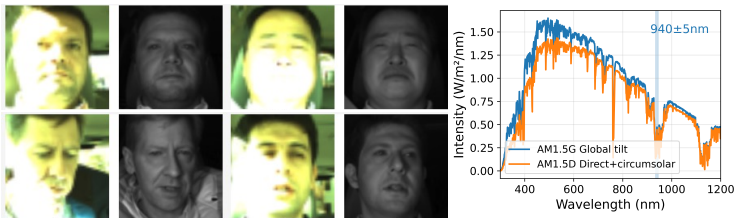

Figure 2: RGB-NIR observations and the spectral motivation for 940-nm imaging. Left: paired RGB and NIR samples from MR-NIRP under challenging illumination. Right: reference solar spectral irradiance (ASTM International, 2020), highlighting local attenuation around the 940 $\pm$ 5 nm band.

3.1 Preliminary and Motivation

Given synchronized RGB and NIR facial videos, rPPG estimation aims to recover the underlying pulse waveform from subtle facial appearance variations. Under complex illumination, these weak physiological variations are often entangled with non-physiological changes induced by ambient light and shadows, making direct RGB-NIR fusion unreliable. This motivates the use of two priors for disturbance-aware representation learning: the relative stability of 940-nm NIR observations and the disturbance-reference property of non-facial background regions.

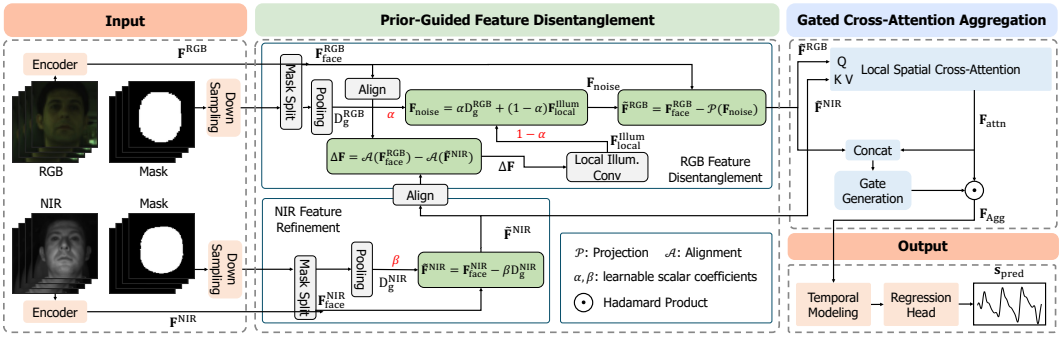

Figure 3: Overview of the proposed NIRP-Net. Encoded RGB-NIR features are refined by prior-guided feature disentanglement to suppress disturbance-related components, aggregated by gated cross-attention aggregation to incorporate complementary NIR cues, and processed by temporal modeling for rPPG waveform prediction.

940-nm low-irradiance prior. Fig. 2 compares paired RGB/NIR samples from the MR-NIRP dataset and the solar spectral irradiance around the 940-nm band. RGB facial appearance varies substantially under challenging illumination, whereas the corresponding NIR observations remain relatively more stable. According to the ASTM G173 reference spectrum, terrestrial solar irradiance exhibits local attenuation around 940 nm due to atmospheric water-vapor absorption (ASTM International, 2020; Gueymard et al., 2002). Under active narrow-band 940-nm illumination, this reduced ambient spectral component provides a physical basis for treating NIR observations as a relatively stable reference, rather than an illumination-invariant signal, for disturbance-aware physiological representation learning.

Physiology-irrelevant background prior. Non-facial background regions are irrelevant to the target physiological signal, so their appearance variations mainly reflect illumination changes and sensor noise. They can therefore serve as disturbance references for separating physiological cues from illumination-related interference. NIRP-Net explicitly incorporates this prior together with the 940-nm low-irradiance prior to improve disturbance-aware physiological representation learning and robust RGB-NIR complementarity modeling under complex illumination.

3.2 Framework Overview

Fig. 3 illustrates the overall pipeline of NIRP-Net. The framework follows a four-stage design: spatiotemporal feature extraction, prior-guided feature disentanglement, gated cross-attention aggregation, and temporal regression for rPPG prediction.

Given synchronized RGB and NIR facial clips $\mathbf{V}^{\mathrm{RGB}} \in \mathbb{R}^{3\times T\times H\times W}$ and $\mathbf{V}^{\mathrm{NIR}} \in \mathbb{R}^{1\times T\times H\times W}$, where T, H, and W denote the temporal length, frame height, and frame width, respectively, two 3DCNN encoders are used to extract spatiotemporal features $\mathbf{F}^{\mathrm{RGB}}, \mathbf{F}^{\mathrm{NIR}} \in \mathbb{R}^{C\times T\times H'\times W'}$, where C denotes the number of channels and $H' \times W'$ denotes the downsampled spatial resolution. The extracted features, together with the corresponding RGB and NIR masks, are then fed into the prior-guided feature disentanglement (PFD) module to obtain disturbance-suppressed modality-specific physiological representations $\widetilde{\mathbf{F}}^{\mathrm{RGB}}, \widetilde{\mathbf{F}}^{\mathrm{NIR}} \in \mathbb{R}^{C\times T\times H'\times W'}$. These two disentangled features are further processed by the gated cross-attention aggregation (GCA) module, which adaptively exploits RGB-NIR complementarity and produces the aggregated feature $\mathbf{F}_{\mathrm{agg}} \in \mathbb{R}^{C\times T\times H'\times W'}$. Finally, $\mathbf{F}_{\mathrm{agg}}$ is spatially pooled into a temporal feature sequence, which is then processed by a Transformer-based temporal modeling module and a regression head to predict the rPPG waveform $\mathbf{s}_{\mathrm{pred}} \in \mathbb{R}^{T}$.

3.3 Prior-Guided Feature Disentanglement

The RGB and NIR encoders provide spatiotemporal representations from synchronized facial clips, but these features still contain illumination-induced appearance variations and other non-physiological components. Directly aggregating such features may propagate disturbance-related components into the rPPG waveform prediction stage, especially when visible-light variations dom-

inate subtle pulse-related changes. Instead of enforcing feature separation through direct distance-based constraints, which may be unreliable without ground-truth illumination noise supervision, we introduce the prior-guided feature disentanglement (PFD) module for disturbance-aware feature refinement. PFD exploits the 940-nm low-irradiance prior to provide a relatively stable NIR reference and uses the physiology-irrelevant background prior to construct global disturbance references, which are further used to refine RGB and NIR features before cross-modal aggregation.

Given the encoded RGB and NIR features $\mathbf{F}^{\mathrm{RGB}}, \mathbf{F}^{\mathrm{NIR}}$, we downsample the corresponding RGB and NIR masks to the feature resolution and use them to separate facial regions from non-facial background regions. The masked facial features are denoted as $\mathbf{F}^{\mathrm{RGB}}_{\mathrm{face}}$ and $\mathbf{F}^{\mathrm{NIR}}_{\mathrm{face}}$, global pooling over the masked background regions yields the disturbance references $\mathbf{D}^{\mathrm{RGB}}_{\mathrm{g}}$ and $\mathbf{D}^{\mathrm{NIR}}_{\mathrm{g}}$.

NIR Feature Refinement. Although the 940-nm NIR observation is relatively more stable than visible-light imaging, the encoded NIR feature may still contain background-related disturbance components. Therefore, we use the NIR background disturbance reference to refine the NIR facial representation. Specifically, the refined NIR feature is computed as

$$\widetilde{\mathbf{F}}^{\mathrm{NIR}} = \mathbf{F}^{\mathrm{NIR}}_{\mathrm{face}} - \beta \mathbf{D}^{\mathrm{NIR}}_{\mathrm{g}}, \tag{1}$$

where β is a learnable scalar coefficient that controls the strength of NIR disturbance suppression. The global reference $\mathbf{D}^{\mathrm{NIR}}_{\mathrm{g}}$ is broadcast to the spatial resolution of $\mathbf{F}^{\mathrm{NIR}}_{\mathrm{face}}$ before subtraction. This background-guided refinement mitigates disturbance components while preserving the relatively stable NIR representation for subsequent RGB feature disentanglement.

RGB Feature Disentanglement. Compared with NIR observations, RGB features are more sensitive to light variations. We therefore use the refined NIR representation as a relatively stable reference to estimate the local disturbance contained in the RGB feature. Specifically, the RGB facial feature and the refined NIR feature are first aligned by lightweight convolutional layers and then subtracted to obtain the local RGB-NIR discrepancy,

$$\Delta \mathbf{F} = \mathcal{A}_R(\mathbf{F}^{\mathrm{RGB}}_{\mathrm{face}}) - \mathcal{A}_N(\widetilde{\mathbf{F}}^{\mathrm{NIR}}), \tag{2}$$

where $\mathcal{A}_R(\cdot)$ and $\mathcal{A}_N(\cdot)$ denote alignment layers for RGB and NIR features, respectively. The discrepancy feature $\Delta\mathbf{F}$ is then processed by a lightweight 3D convolutional block to obtain the local illumination-related representation $\mathbf{F}^{\mathrm{Illum}}_{\mathrm{local}} \in \mathbb{R}^{B\times C\times T\times H'\times W'}$. To incorporate both local RGB-NIR discrepancy and global background disturbance, we combine $\mathbf{F}^{\mathrm{Illum}}_{\mathrm{local}}$ with the RGB background disturbance reference $\mathbf{D}^{\mathrm{RGB}}_{\mathrm{g}}$,

$$\mathbf{F}_{\mathrm{noise}} = \alpha \mathbf{D}^{\mathrm{RGB}}_{\mathrm{g}} + (1-\alpha)\mathbf{F}^{\mathrm{Illum}}_{\mathrm{local}}, \tag{3}$$

where α is a learnable scalar coefficient that balances the global and local disturbance references. The global reference $\mathbf{D}^{\mathrm{RGB}}_{\mathrm{g}}$ is broadcast to the spatial resolution of $\mathbf{F}^{\mathrm{Illum}}_{\mathrm{local}}$ before fusion. Finally, the RGB physiological representation is obtained by subtracting the projected disturbance feature from the RGB facial feature:

$$\widetilde{\mathbf{F}}^{\mathrm{RGB}} = \mathbf{F}^{\mathrm{RGB}}_{\mathrm{face}} - \mathcal{P}(\mathbf{F}_{\mathrm{noise}}), \tag{4}$$

where $\mathcal{P}(\cdot)$ denotes a projection layer. This process mitigates disturbance-related components in the RGB representation by jointly exploiting the local RGB-NIR discrepancy and the background-derived global disturbance reference. Through the above prior-guided process, PFD outputs the refined NIR representation $\widetilde{\mathbf{F}}^{\mathrm{NIR}}$ and the refined RGB representation $\widetilde{\mathbf{F}}^{\mathrm{RGB}}$, which are then fed into the subsequent GCA module for cross-modal aggregation.

3.4 Gated Cross-Attention Aggregation

For rPPG estimation, cross-modal aggregation should selectively exploit complementary cues rather than indiscriminately merging all modality information. Since pulse-related variations are subtle and easily affected by residual non-physiological components, direct summation or concatenation may propagate redundant or disturbance-related information into the rPPG waveform prediction stage. We design the gated cross-attention aggregation (GCA) module to use the RGB representation as the aggregation anchor, retrieve locally relevant NIR cues through cross-attention, and control their residual contribution through gate modeling.

Given the disturbance-suppressed RGB representation $\widetilde{\mathbf{F}}^{\mathrm{RGB}}$ and the refined NIR representation $\widetilde{\mathbf{F}}^{\mathrm{NIR}}$, GCA first performs local spatial cross-attention to retrieve NIR cues relevant to the RGB representation. We partition both features into non-overlapping local windows along the spatial dimensions,

$$\mathbf{X}^{\mathrm{RGB}} = \mathcal{W}_w(\widetilde{\mathbf{F}}^{\mathrm{RGB}}), \quad \mathbf{X}^{\mathrm{NIR}} = \mathcal{W}_w(\widetilde{\mathbf{F}}^{\mathrm{NIR}}), \tag{5}$$

where $\mathcal{W}_w(\cdot)$ denotes the $w \times w$ local window partitioning operation. The partitioned features satisfy $\mathbf{X}^{\mathrm{RGB}}, \mathbf{X}^{\mathrm{NIR}} \in \mathbb{R}^{BTN_w \times w^2 \times C}$, where $N_w = H'W'/w^2$ is the number of local windows in each frame. Within each local window, the RGB feature is used as the query, while the NIR feature is used as the key and value. The attended NIR cue in each local window is then computed as

$$\mathbf{F}_w = \mathrm{Softmax}\left(\frac{\mathbf{Q}^{\mathrm{RGB}}(\mathbf{K}^{\mathrm{NIR}})^\top}{\sqrt{d}}\right)\mathbf{V}^{\mathrm{NIR}}, \tag{6}$$

where $\mathbf{Q}^{\mathrm{RGB}} = \mathbf{X}^{\mathrm{RGB}}\mathbf{W}_Q$, $\mathbf{K}^{\mathrm{NIR}} = \mathbf{X}^{\mathrm{NIR}}\mathbf{W}_K$, and $\mathbf{V}^{\mathrm{NIR}} = \mathbf{X}^{\mathrm{NIR}}\mathbf{W}_V$. Here, d denotes the attention feature dimension. The window-level output $\mathbf{F}_w$ is then reshaped back to the original spatiotemporal feature layout, yielding the RGB-guided attended NIR cue $\mathbf{F}_{\text{attn}} \in \mathbb{R}^{B\times C\times T\times H'\times W'}$. This local attention design preserves the spatial correspondence between synchronized RGB and NIR facial regions while reducing unnecessary interactions with distant or irrelevant areas.

To avoid unconditional injection of the attended NIR cue, we generate an adaptive gate from the RGB representation and the attended NIR feature and compute the aggregated feature as

$$\mathbf{F}_{\mathrm{agg}} = \widetilde{\mathbf{F}}^{\mathrm{RGB}} + \sigma\left(\phi_g\left([\widetilde{\mathbf{F}}^{\mathrm{RGB}}, \mathbf{F}_{\mathrm{attn}}]\right)\right) \odot \mathbf{F}_{\mathrm{attn}}, \tag{7}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation and $\phi_g(\cdot)$ denotes the gate generation layer. In this way, GCA selectively incorporates complementary NIR cues for robust rPPG waveform prediction.

4 Experiments

4.1 Datasets

Public RGB-NIR rPPG datasets remain limited, especially those acquired with active 940-nm near-infrared illumination. Moreover, severe underexposure or facial occlusion in the NIR recordings of some existing datasets can cause commonly used face detection procedures to fail, limiting their use for controlled RGB-NIR rPPG evaluation. To complement existing resources, we collect FaceIVL, a new RGB-NIR facial video dataset with controlled illumination variations. The overall statistics and acquisition settings of FaceIVL are summarized in Table 1.

Table 1: Statistics and acquisition settings of the FaceIVL dataset.

Subjects	60 (30 male, 30 female)
Total videos	120
Duration	∼1 min / video
Resolution	1920 × 1080
Frame rate	30 fps
Modalities	RGB + NIR
NIR illumination	940 nm
Lighting conditions	Normal / low light

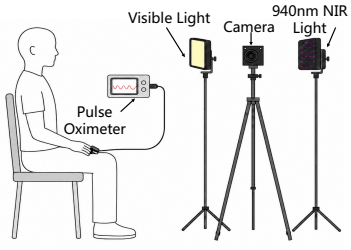

Figure 4: FaceIVL acquisition with synchronized RGB-NIR imaging and contact PPG reference.

FaceIVL is designed to evaluate whether RGB-NIR methods can maintain robust physiological estimation when visible-light conditions change. Each subject is recorded under normal and low-light indoor conditions, while active 940-nm near-infrared illumination is used for the NIR modality. Subjects are asked to avoid large head movements, while slight facial motions and speaking are not strictly constrained. The data are captured using a JAI FS-3200D-10GE multi-spectral camera, which simultaneously captures visible color and near-infrared images through the same optical path. A contact finger PPG sensor is used to record the ground-truth physiological signal synchronized

with the facial videos. The overall acquisition setup is illustrated in Fig. 4. Additional examples of synchronized RGB-NIR observations under both illumination conditions are provided in Appendix A. The data collection protocol was approved by the ethics review committee of the relevant institution, and all participants signed informed consent forms before data collection.

4.2 Comparison Results

Intra-Dataset Evaluation. We evaluate NIRP-Net on three RGB-NIR rPPG datasets, including MR-NIRP Driving (Nowara et al., 2020), MR-NIRP Indoor (Magdalena Nowara et al., 2018), and FaceIVL. For MR-NIRP Indoor, we use 8-fold cross-validation. For MR-NIRP Driving, the data are divided into training and testing subsets with a ratio of 7:3. For FaceIVL, we adopt a subject-independent 6:4 split, where the gender distribution is balanced in both the training and testing sets. As shown in Table 2, NIRP-Net achieves state-of-the-art performance across all three RGB-NIR datasets. Fig. 5 presents representative examples of the predicted and ground-truth signals from the MR-NIRP Driving and FaceIVL datasets. Fig. 6 further shows low bias and strong agreement between the estimated and ground-truth heart rates. The stronger performance of RGB-NIR methods over RGB-based baselines confirms the value of NIR observations under illumination variations, while the consistent advantage of NIRP-Net over existing RGB-NIR fusion methods shows that naive fusion or difference-based NIR modeling remains insufficient. These results demonstrate the effectiveness of NIRP-Net for robust RGB-NIR rPPG estimation, the contribution of each component is examined in Sec. 4.3.

Table 2: Intra-dataset evaluation results on several datasets. The best, the second-best and the third-best results are marked in **red**, blue and green respectively.

Method	MR-NIRP Driving				MR-NIRP Indoor				FaceIVL			
	MAE	RMSE	MAPE	ρ	MAE	RMSE	MAPE	ρ	MAE	RMSE	MAPE	ρ
POS (Wang et al., 2016)	12.46	17.79	17.96	0.18	6.03	12.39	9.90	0.34	7.35	14.96	6.66	0.60
CHROM (De Haan & Jeanne, 2013)	12.39	17.87	17.86	0.16	6.49	12.98	10.74	0.23	10.11	18.87	9.19	0.42
OMIT (Casado & López, 2023)	12.87	17.93	18.18	0.14	6.25	11.76	9.48	0.44	18.30	26.64	17.08	0.12
DeepPhys (Chen & McDuff, 2018)	3.10	5.05	4.27	0.72	0.88	1.13	1.22	**0.99**	2.42	8.67	2.09	0.83
TS-CAN (Liu et al., 2020)	4.33	8.06	5.54	0.38	1.21	2.05	1.58	0.96	6.02	14.38	5.46	0.44
EfficientPhys (Liu et al., 2023b)	3.69	6.82	4.80	0.58	0.92	1.13	1.25	**0.99**	5.12	12.65	4.73	0.59
PhysNet (Yu et al., 2019a)	2.87	5.63	4.10	0.66	0.83	1.31	1.19	**0.99**	1.18	3.43	1.06	0.98
PhysFormer (Yu et al., 2022)	2.76	4.91	3.77	0.74	0.62	0.75	0.87	**0.99**	2.93	7.50	2.51	0.88
RhythmFormer (Zou et al., 2025a)	2.65	4.54	3.65	0.78	0.74	0.98	1.05	**0.99**	4.65	10.42	4.09	0.76
RhythmMamba (Zou et al., 2025b)	2.90	5.52	4.13	0.67	0.67	**0.99**	0.93	0.99	1.27	2.98	1.31	0.98
IENet (Liu et al., 2023a)	2.72	5.63	3.95	0.75	0.72	1.12	1.05	**0.99**	1.13	2.35	1.12	**0.99**
FusionPhys (Ying et al., 2025)	2.33	5.80	3.51	0.70	0.58	0.86	0.84	**0.99**	1.13	2.67	1.14	**0.99**
Ours	**1.73**	**4.17**	**2.51**	**0.83**	**0.51**	**0.66**	**0.73**	**0.99**	**0.95**	**2.26**	**0.92**	**0.99**

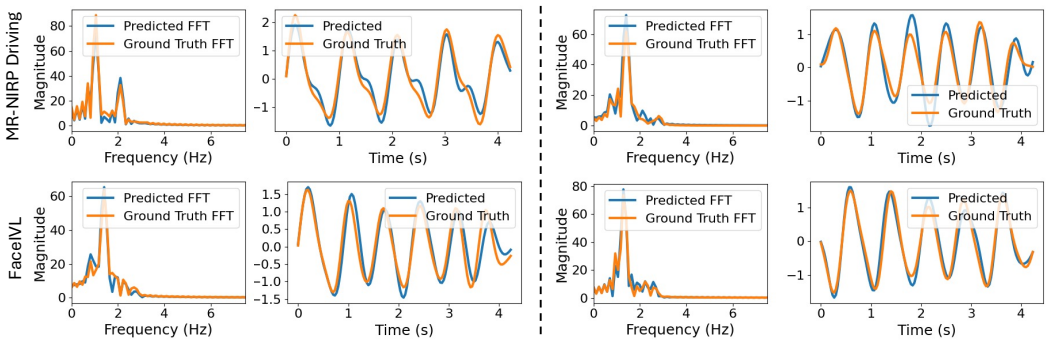

Figure 5: Representative frequency spectra and predicted/ground-truth waveforms on MR-NIRP Driving (top) and FaceIVL (bottom).

Cross-Dataset Evaluation. We further evaluate the generalization capability of NIRP-Net by training on MR-NIRP Driving and testing on MR-NIRP Indoor. Specifically, MR-NIRP Driving is divided into training and validation subsets with a ratio of 8:2, while the entire MR-NIRP Indoor

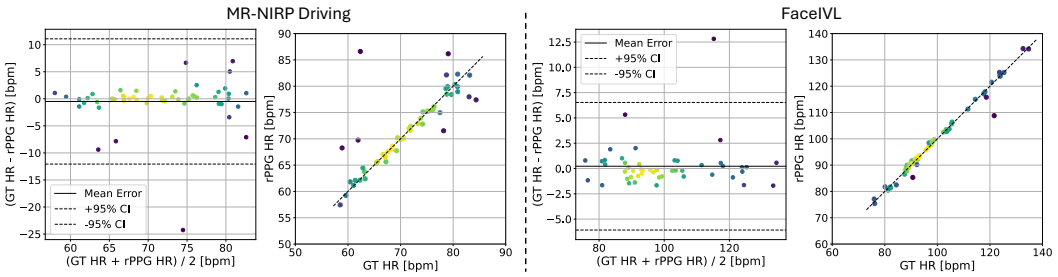

Figure 6: Bland-Altman plots and scatter plots from intra-dataset evaluation on the MR-NIRP Driving and FaceIVL datasets. "CI" denotes the confidence interval.

dataset is used for testing. No samples from MR-NIRP Indoor are used during training or validation. As shown in Table 3, NIRP-Net achieves the best performance across all error metrics and maintains a high correlation with the ground truth. Compared with RGB-based baselines, RGB-NIR methods generally show stronger cross-dataset performance, indicating that NIR observations provide useful complementary information under domain shifts. Meanwhile, the consistent advantage of NIRP-Net over existing RGB-NIR fusion methods further suggests that naive fusion or difference-based NIR modeling is insufficient for robust cross-dataset rPPG estimation.

Table 3: Cross-dataset heart-rate estimation. Models are trained on MR-NIRP Driving dataset and evaluated on MR-NIRP Indoor dataset. The best and second-best ones are in bold and underlined, respectively.

Method	MAE	RMSE	MAPE	ρ
DeepPhys	1.70	2.47	2.61	0.92
TS-CAN	3.70	6.62	5.51	0.70
EfficientPhys	2.16	3.46	2.90	0.92
PhysNet	1.43	2.52	2.28	<u>0.98</u>
PhysFormer	1.23	1.68	1.71	<u>0.98</u>
RhythmFormer	1.08	1.51	1.52	**0.99**
RhythmMamba	2.12	4.96	3.52	0.83
IENet	0.81	1.16	1.14	**0.99**
FusionPhys	<u>0.77</u>	<u>1.02</u>	<u>1.12</u>	**0.99**
Ours	**0.66**	**0.99**	**0.91**	**0.99**

Table 4: Ablation study on the overall pipeline of the proposed method on MR-NIRP Driving dataset. "Diff", "NIR" and "RGB" denote the prior-guided feature difference, NIR feature refinement and RGB feature disentanglement in PFD, respectively.

Setting	Disentanglement			Aggregation		MR-NIRP Driving			
	Diff	PFD		Direct Fusion	GCA	MAE	RMSE	MAPE	ρ
		NIR	RGB						
1	-	-	-	✓	-	3.25	6.54	4.79	0.68
2	✓	-	-	-	-	3.11	6.23	4.57	0.64
3	-	-	✓	-	-	3.27	6.82	4.87	0.64
4	-	✓	✓	-	-	3.16	6.12	4.59	0.62
5	-	-	-	-	✓	2.48	5.86	3.69	0.70
6	-	-	✓	-	✓	2.49	6.36	3.88	0.62
7	-	✓	-	-	✓	2.26	4.89	3.21	0.80
8	-	✓	✓	-	✓	1.73	4.17	2.51	0.83

4.3 Ablation Study

We conduct ablation studies on the MR-NIRP Driving dataset to assess the contributions of the proposed components.

Overall Pipeline. We first evaluate the overall design of NIRP-Net. As shown in Table 4, Setting 1 with direct fusion gives limited performance, indicating that simply merging RGB and NIR features is insufficient under illumination variations. Setting 2 improves over direct fusion by introducing prior-guided feature difference, but still remains clearly inferior to the full model, suggesting that using NIR only as a difference reference cannot fully exploit RGB-NIR complementarity. Comparing Setting 4 and Setting 8 further shows that PFD alone is not enough, and the refined modality-specific features need to be adaptively aggregated by GCA. In addition, comparing Settings 5–7 shows that adding NIR feature refinement to GCA consistently improves all metrics, whereas adding RGB feature disentanglement alone provides no performance gain. This result highlights the importance of refining NIR cues before cross-modal aggregation. Overall, Setting 8 performs best, demonstrating the complementary benefits of PFD and GCA for robust RGB-NIR rPPG estimation.

Fusion Mechanism. We further evaluate different aggregation strategies for combining RGB and NIR representations. As shown in Table 5, simple summation and concatenation give limited performance, indicating that directly merging the two modalities cannot adapt to their different reliability

under illumination variations. Cross-attention improves the results by retrieving NIR cues according to the RGB representation, but it still performs worse than the full GCA. The superior performance of GCA demonstrates the benefit of adaptive gated aggregation over the evaluated fusion alternatives.

Temporal Modeling. We also study the influence of temporal modeling. As shown in Table 6, removing temporal modeling leads to clear performance degradation, confirming the importance of temporal dependency modeling for rPPG estimation. Temporal convolution and lightweight temporal attention both improve the results, showing that temporal context is beneficial after RGB-NIR feature refinement. Among the evaluated alternatives, the Transformer encoder achieves the best results and is therefore adopted in NIRP-Net.

Table 5: Ablation study on the fusion mechanism of the proposed method on MR-NIRP Driving dataset.

Setting	MAE	RMSE	MAPE	ρ
Sum	2.99	6.78	4.55	0.61
Concat	2.97	6.43	4.46	0.61
Cross-Attention	2.92	6.22	4.30	0.67
Ours	1.73	4.17	2.51	0.83

Table 6: Ablation study on temporal modeling in the proposed method on MR-NIRP Driving dataset.

Setting	MAE	RMSE	MAPE	ρ
w/o Temporal Modeling	3.10	6.80	4.62	0.61
Conv Block	2.66	6.18	4.02	0.66
Attention Block	2.66	6.10	4.06	0.67
Ours	1.73	4.17	2.51	0.83

4.4 Computational Cost

We further compare the computational cost and inference efficiency of different methods on the MR-NIRP Driving dataset. As shown in Table 7, NIRP-Net achieves the lowest MAE and RMSE while maintaining a moderate computational cost. Compared with lightweight RGB-based methods, NIRP-Net requires more computation due to the additional NIR branch and cross-modal aggregation, but it provides substantially better estimation accuracy. Compared with existing RGB-NIR methods, NIRP-Net is more efficient in terms of MACs and throughput, while achieving better performance. NIRP-Net incurs additional cost over lightweight RGB-only models, but roughly halves the MACs of FusionPhys and IENet while providing higher throughput and lower estimation error.

Table 7: Computational cost and inference efficiency for 160-frame 128×128 clips on an RTX A6000. MAE and RMSE are evaluated on MR-NIRP Driving dataset.

Method	Param. (M)	MACs (G)	Throughput (FPS)	MAE (bpm)	RMSE (bpm)
DeepPhys	1.975	95.290	987	3.10	5.05
TS-CAN	1.975	95.290	916	4.33	8.06
EfficientPhys	1.909	47.837	1167	3.69	6.82
PhysNet	0.769	56.095	1092	2.87	5.63
PhysFormer	7.381	40.485	77	2.76	4.91
RhythmFormer	3.251	30.795	79	2.65	4.54
RhythmMamba	1.065	10.355	40	2.90	5.52
FusionPhys	3.450	183.208	191	2.72	5.63
IENet	18.072	179.354	35	2.33	5.80
Ours	4.749	89.095	280	1.73	4.17

5 Conclusions

In this work, we propose NIRP-Net for illumination-robust RGB-NIR rPPG estimation. The proposed framework combines prior-guided feature disentanglement with gated cross-attention aggregation to refine modality-specific physiological representations and selectively exploit complementary NIR cues. We also collect FaceIVL, a new RGB-NIR facial rPPG dataset under controlled illumination variations. Experiments on public datasets and FaceIVL demonstrate that NIRP-Net achieves state-of-the-art performance in both intra-dataset and cross-dataset evaluations. Ablation studies demonstrate the effectiveness of the proposed design, while computational analysis shows its efficiency relative to existing RGB-NIR methods.

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

ASTM International. ASTM G173-03(2020): Standard Tables for Reference Solar Spectral Irradiances: Direct Normal and Hemispherical on 37 Degree Tilted Surface. ASTM International, West Conshohocken, PA, USA, 2020. Available: https://www.astm.org/g0173-03r20.html.

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.

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.

Christian A Gueymard, David Myers, and Keith Emery. Proposed reference irradiance spectra for solar energy systems testing. *Solar energy*, 73(6):443–467, 2002.

Titong Jiang, Yuan Ma, Jiaqi Li, Qing Dong, Xuewu Ji, and Yahui Liu. LSTS: Periodicity learning via long short-term temporal shift for remote physiological measurement. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):6452–6465, 2025.

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

Lili Liu, Zhaoqiang Xia, Xiaobiao Zhang, Jinye Peng, Xiaoyi Feng, and Guoying Zhao. Information-enhanced network for noncontact heart rate estimation from facial videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2136–2150, 2023a.

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, 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.

Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. SparsePPG: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1272–1281, 2018.

Ewa M Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE transactions on intelligent transportation systems*, 23(4):3589–3600, 2020.

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.

Marko Savic and Guoying Zhao. Rs+rppg: Robust strongly self-supervised learning for rppg. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Hang Shao and Chuanfei Hu. Remote Heart Rate Measurement Based on Near Infrared Prior Information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 538–545, 2025.

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.

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.

Chenhang Ying, Huiyu Yang, Jieyi Ge, Zhaodong Sun, Xu Cheng, Kui Ren, and Xiaobai Li. FusionPhys: A Flexible Framework for Fusing Complementary Sensing Modalities in Remote Physiological Measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9363–9373, 2025.

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

Yizhu Zhang, Jingang Shi, Jiayin Wang, Yuan Zong, Wenming Zheng, and Guoying Zhao. MaskFusionNet: A dual-stream fusion model with masked pre-training mechanism for rPPG measurement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11521–11534, 2024.

Changchen Zhao, Hongsheng Wang, Huiling Chen, Weiwei Shi, and Yuanjing Feng. JAMSNet: A remote pulse extraction network based on joint attention and multi-scale fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2783–2797, 2022.

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 VISUALIZATION OF FACEIVL

To provide a qualitative overview of FaceIVL, Fig. 7 presents paired RGB and NIR facial samples from representative subjects under normal and low-light conditions. For each illumination condition, the RGB and NIR frames are synchronously captured through the same optical path. The visible-light facial appearance changes substantially across the two conditions, whereas the corresponding NIR observations generally preserve discernible facial structures under active 940-nm illumination. These examples illustrate the controlled illumination variations in FaceIVL and complement the dataset statistics and acquisition settings reported in the main paper.

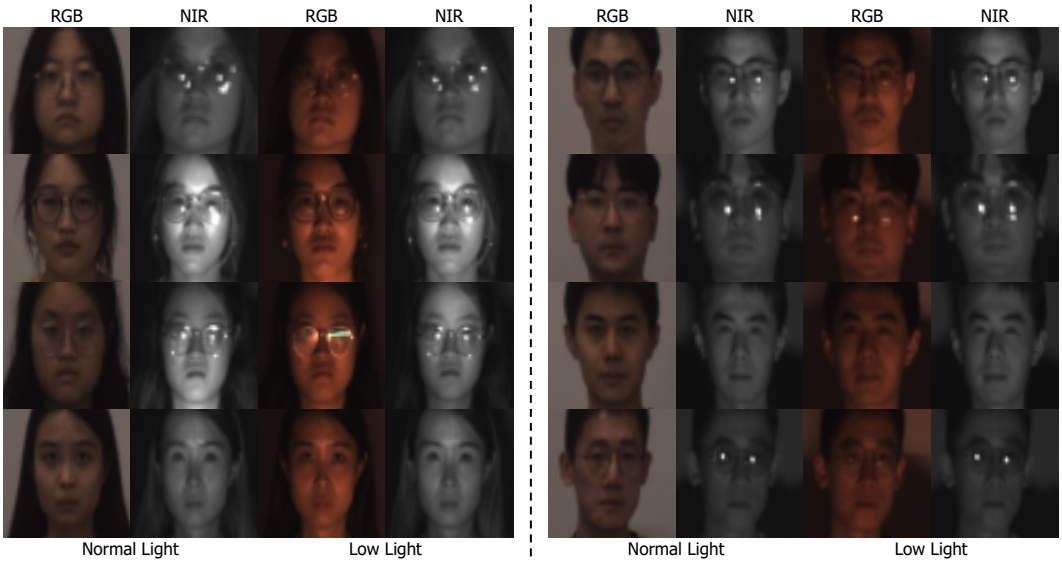

Figure 7: Representative paired RGB and NIR facial samples from FaceIVL under normal and low-light conditions. Each row shows the same subject across the two illumination conditions, with synchronized RGB and NIR observations shown for each condition.

B APPENDIX: IMPLEMENTATION DETAILS

Our method is implemented using PyTorch 2.4.0. We train NIRP-Net for 30 epochs using the AdamW optimizer with a weight decay of 1×10^{-4} and an initial learning rate of 2×10^{-4}. A OneCycle learning rate schedule is adopted, and the batch size is set to 32. We use a random seed of 100 for all experiments. All experiments are conducted on an NVIDIA RTX A6000 GPU. For preprocessing, synchronized RGB-NIR videos are divided into 128-frame clips. The facial regions are detected, cropped, and resized to 128×128 pixels. The RGB clips, NIR clips, and corresponding facial masks are temporally aligned before being fed into the network. During training, both predicted and ground-truth rPPG signals are temporally normalized before loss computation. The model is optimized with a negative Pearson loss, together with auxiliary regularization terms for stable global-local disturbance balancing and local illumination consistency. The gradient norm is clipped to 1.0. For postprocessing, the predicted rPPG waveform is filtered by a second-order Butterworth bandpass filter with cutoff frequencies of 0.75–2.5 Hz. The heart rate is estimated from the

power spectral density computed by the Welch algorithm. In NIRP-Net, the local spatial window size in GCA is set to $w = 4$, and the default temporal modeling module is implemented with a Transformer encoder.

C Appendix: Network Configuration

Table 8 summarizes the detailed architecture and tensor dimensions of NIRP-Net for a 128-frame RGB-NIR clip with a spatial resolution of 128×128. The RGB and NIR encoders have identical architectures but are independently parameterized.

Table 8: Configuration of the main learnable layers in NIRP-Net.

Module	Layer	Kernel / Window	Configuration	Output Size
RGB encoder	Conv3D + 6×BasicBlock3D	$1 \times 7 \times 7$; $3 \times 3 \times 3$	Channels: $3 \to 32 \to 64 \to 128$	$128 \times 128 \times 8 \times 8$
NIR encoder	Conv3D + 6×BasicBlock3D	$1 \times 7 \times 7$; $3 \times 3 \times 3$	Channels: $1 \to 32 \to 64 \to 128$	$128 \times 128 \times 8 \times 8$
Feature alignment	2×(Conv3D + BN)	$1 \times 1 \times 1$	$128 \to 128$ for each modality	$128 \times 128 \times 8 \times 8$
Local disturbance encoder	Conv3D + BN + Relu + Conv3D	$3 \times 3 \times 3$ $1 \times 1 \times 1$	$128 \to 128 \to 128$	$128 \times 128 \times 8 \times 8$
Disturbance projection	Conv3D	$1 \times 1 \times 1$	$128 \to 128$	$128 \times 128 \times 8 \times 8$
Local cross-attention	Multi-head cross-attention + Group normalization	4×4 spatial window	$D = 128$, 8 heads	$128 \times 128 \times 8 \times 8$
Gating network	Conv3D + BN + ReLU + Conv3D + Sigmoid	$1 \times 1 \times 1$	$256 \to 128 \to 128$	$128 \times 128 \times 8 \times 8$
Temporal encoder	4×TransformerEncoderLayer	–	$D = 128$, 8 heads FFN $= 512$, dropout $= 0.1$	128×128 $(T \times D)$
Prediction head	Linear	–	$128 \to 1$	128