

MS-rPPG: Multi-spectral State Space Model for Remote Photoplethysmography in Driver Monitoring Systems

Jiho Choi¹ and Sang Jun Lee^{1,*}

Abstract—Remote photoplethysmography (rPPG) is a camera-based technique for measuring physiological signals, particularly cardiac activity. From the remotely measured signals, heart rate can be estimated, which is crucial for health monitoring. In this study, we investigate a driver health monitoring system based on remote heart rate estimation. However, driving environments represent uncontrolled settings where videos are subject to varying illumination conditions and frequent head movements. We introduce MS-rPPG, a multi-spectral framework that combines RGB with near-infrared (NIR) face video to alleviate rPPG estimation under challenging driving conditions. To combine the complementary features from two spectral videos, we propose a cross-spectral linear modulation (CSLM) strategy based on frequency-domain analysis. Moreover, we introduce MS-Mamba, a novel state space model designed to effectively model long-range temporal dependencies while jointly capturing cross-channel interactions between multi-spectral features. We collected a real-world dataset called MS-Drive, which was recorded from 50 participants while driving the vehicle. The proposed method was evaluated on the MR-NIRP Car dataset and MS-Drive datasets. The experimental results indicate that MS-rPPG shows better robustness and heart rate estimation accuracy than previous methods, highlighting its promise for driver health monitoring. The codes are available at github.com/ziiho08/MS-rPPG.

I. INTRODUCTION

Blood volume pulse (BVP) is a fundamental physiological signal used to derive key indicators of human health, including heart rate (HR) and respiration rate. Photoplethysmography (PPG) is a widely adopted non-invasive technique that measures BVP by detecting variations in light absorption and reflection caused by blood flow driven by cardiac activity. However, PPG sensors require direct skin contact, which restricts user movements and limits their suitability for long-term health monitoring. To overcome these limitations, remote photoplethysmography (rPPG) has emerged as a camera-based alternative that does not require physical contact. Similar to conventional PPG sensors, rPPG captures subtle variations in light reflected from the facial skin in video recordings, enabling contactless physiological monitoring. Owing to its non-contact characteristics, rPPG holds significant potential for diverse applications, including healthcare [1], [2], affective computing [3], [4], and driver monitoring systems [5], [6].

*The corresponding author of this paper.

¹Jiho Choi and Sang Jun Lee are with the Division of Electronics and Information Engineering, Jeonbuk National University, Republic of Korea. Emails: {jihochoi, sj.lee}@jbnu.ac.kr

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2025-RS-2024-00439292).

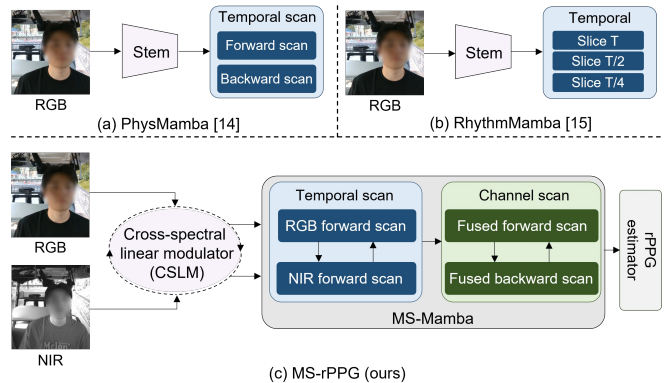


Fig. 1. Comparison with recent Mamba-based rPPG frameworks

Early rPPG estimation approaches mainly utilized conventional signal processing methods such as chrominance [7] and principal component analysis [8]. Substantial improvements in remote physiological measurement have been achieved with deep learning methods, particularly the convolutional neural network (CNN) and the vision transformer (ViT). While CNN-based algorithms [9], [10] are effective in extracting spatio-temporal features, they face limitations in capturing long-range temporal dependencies due to their limited receptive fields. Although ViT-based methods [11], [12] can effectively capture local and global long-range features, the quadratic computational cost of attention restricts their practicality for long video sequences. These limitations are important to consider when applying rPPG to real-world systems such as driver health monitoring.

Recently, Mamba [13] has been introduced as a state space model (SSM) that is effective for sequence modeling with reduced computational complexity. Mamba achieves linear-time complexity while effectively capturing long-term temporal dependencies, facilitating its use in long-sequence tasks such as rPPG estimation. As shown in Figure 1(a), PhysMamba [14] is a lightweight model for real-time rPPG estimation, leveraging linear complexity for efficient long-range temporal modeling. RhythmMamba [15] scans features along the time axis with multiple temporal lengths, enabling effective modeling of both short-term and long-range temporal dependencies in rPPG signals. However, existing rPPG methods have been evaluated under controlled laboratory conditions, with limited evidence of robustness in unconstrained environments.

This paper focuses on driver heart rate estimation among rPPG applications and addresses the problem of uncontrolled

environmental illumination variations. Varying illumination poses significant challenges for rPPG estimation accuracy, especially for light-susceptible RGB video. To alleviate the effects of real-world lighting variations, recent studies have explored the use of radio-frequency (RF) signals and near-infrared (NIR) video. RF signals are robust to lighting variations but remain vulnerable to motion-induced distortions, which are common in dynamic driving scenarios [16]. NIR video is more robust under low-light or rapidly varying illumination, as longer wavelengths are less susceptible to ambient visible light sources [5]. Therefore, we utilize NIR videos along with RGB videos to build a robust model for remote physiological estimation.

As discussed above, this study introduces MS-rPPG, which leverages an SSM with linear computational complexity to capture long-range temporal dependencies from RGB and NIR facial videos in uncontrolled driving environments. To mitigate drastic illumination changes during driving, we propose a cross-spectral linear modulator (CSLM) based on frequency-domain analysis guided by physiological priors, effectively integrating the complementary information from RGB and NIR images. Moreover, to enhance long-range temporal dependencies, the multi-spectral Mamba (MS-Mamba) performs two key operations. First, it conducts a temporal scan across the embeddings for each spectral feature, and it utilizes bidirectional channel scanning to effectively model inter-channel dependencies.

Additionally, real-world driving scenario datasets are limited. To address this gap, we collected the MS-Drive dataset, which contains RGB and NIR videos with reliable ground truth signals. We evaluated the proposed framework on both the publicly available MR-NIRP Car dataset and the collected MS-Drive dataset. Experimental results demonstrate that our approach achieves state-of-the-art performance with HR estimation task under real-world driving conditions, underscoring its potential for deployment in driver monitoring systems. The main contributions of this paper are:

- We propose MS-rPPG, a novel multi-spectral framework that integrates RGB and NIR video to enhance the robustness of rPPG estimation in driving environments.
- We introduce the CSLM based on the frequency domain analysis that effectively fuses the pulse signal from complementary spectral images.
- We develop MS-Mamba, a state space model designed to capture long-range temporal dependencies with linear computational complexity. MS-mamba scans each spectral video along the temporal axis and performs bidirectional scanning along the channel axis.
- We collected the real-world driving dataset MS-Drive and conducted extensive experiments on the MR-NIRP Car and MS-Drive datasets, demonstrating that MS-Mamba outperforms previous approaches in terms of remote HR estimation performance under unconstrained driving environments.

II. RELATED WORK

A. Remote photoplethysmography estimation

Early rPPG research primarily employed traditional signal processing methods based on hand-crafted features and color space transformations from RGB video [8], [17], [18], [7]. Recently, deep learning models have established 2D and 3D CNN and ViT-based architectures as the dominant paradigms for rPPG tasks [19], [11], [12]. However, relying solely on RGB facial video remains susceptible to external perturbations, including lighting fluctuations and head movement.

To address these limitations, several studies have explored rPPG estimation from NIR images, which exhibit less sensitivity to varying light. Nowara et al. [5] investigated the characteristics of the NIR spectrum, revealing that this spectral band contains minimal energy from sunlight and ambient light. They proposed a robust PCA algorithm and optimized a low-rank matrix to extract the periodic pulse signal, demonstrating effectiveness under varying illumination conditions using the collected real-world MR-NIRP dataset. However, using only NIR can be challenging because NIR yields a lower signal-to-noise ratio (SNR) than visible light for rPPG due to weaker hemoglobin absorption and reduced sensor sensitivity, leading to much weaker pulsatile signals than RGB. Park et al. [20] proposed a self-supervised video-based rPPG framework that integrates RGB and NIR modalities using a dual-branch ViT architecture, achieving enhanced robustness to illumination changes. Despite using multi-spectral images, ViT inherits substantial computational overhead and has primarily been evaluated under controlled laboratory conditions.

B. State space model in rPPG

Mamba represents a selective state space model that achieves efficient processing of long-range sequences through linear computational complexity [13]. Within computer vision applications, vision Mamba has demonstrated competitive performance across video understanding and image classification tasks, providing an alternative to ViTs with reduced memory requirements and computational latency.

Recent advances in rPPG have increasingly adopted Mamba-based architectures. PhysMamba [14] advances Mamba-based rPPG modeling by introducing a temporal difference Mamba block. It employs a bidirectional Mamba as depicted in Figure 1(a), achieving superior performance over CNN and Transformer baselines. As shown in Figure 1(b), RhythmMamba [15] established a temporal modeling framework specifically designed for rPPG applications through multi-temporal constraints and frequency-domain feed-forward modules, achieving state-of-the-art performance while preserving computational efficiency. ME-rPPG [21] further advanced the field by enabling single-frame, low-latency inference with minimal memory consumption, addressing critical real-time deployment requirements. Unlike existing approaches, our framework models temporal state transitions and simultaneously captures inter-channel dependencies via a bidirectional channel SSM. This

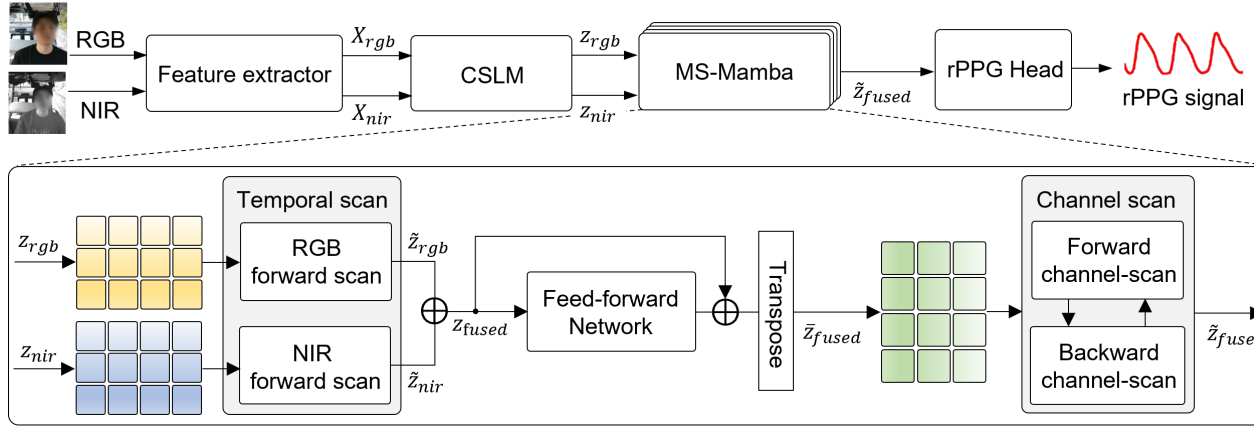


Fig. 2. Overview of proposed framework. Our method integrates RGB and NIR face videos for robust remote physiological measurement under real-world driving conditions. MS-rPPG includes the CSLM module for multi-spectral fusion and the MS-Mamba block with temporal and bidirectional channel scanning.

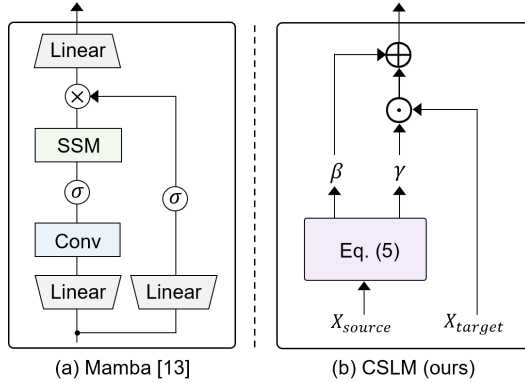


Fig. 3. Architecture of the Mamba [13] and the proposed CSLM.

design is an efficient and dynamic channel scanning mechanism.

III. METHOD

A. Preliminaries of Mamba

To model the sequential physiological data, we adopt the state space model, which describes continuous linear time-invariant (LTI) systems. An LTI system maps a continuous input signal $x(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$ through a latent state $h(t) \in \mathbb{R}^N$. To this end, the system uses a state transition matrix $A \in \mathbb{R}^{N \times N}$, projection matrices $B \in \mathbb{R}^{N \times 1}$ and $C \in \mathbb{R}^{1 \times N}$, and skip connection $D \in \mathbb{R}$. The system is described by the following ordinary differential equation.

$$\begin{cases} h(t)' = Ah(t) + Bx(t), \\ y(t) = Ch(t) + Dx(t). \end{cases} \quad (1)$$

To integrate continuous-time SSM into deep learning frameworks, these equations are discretized using the Zero-Order Hold (ZOH) principle. Specifically, a time scale parameter Δ is introduced, and the continuous parameters A and B are converted into their discrete counterparts \bar{A} and \bar{B} as follows:

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I)\Delta B, \quad (2)$$

where I is the identity matrix of compatible dimensions. The discrete state transition and observation equations can then be written as:

$$\begin{cases} h_k = \bar{A}h_{k-1} + \bar{B}x_k, \\ y_k = Ch_k + Dx_k, \end{cases} \quad (3)$$

where k denotes the discrete time step index.

Discrete SSM can be further interpreted as resembling both CNNs and recurrent neural networks. Specifically, by unrolling the hidden state h_k over the previous time steps and collecting terms of A and B , a structured convolutional kernel \bar{K} can be defined. This enables the following 1D convolution formulation:

$$\bar{K} = (CB, CAB, \dots, CA^{L-1}B), \quad y = x * \bar{K}, \quad (4)$$

where $*$ is a 1D convolution and L is sequence length. This equation formulation allows for efficient parallelization on modern hardware computation, enabling scalable training and inference for long-sequence modeling tasks.

B. General framework of proposed method

The proposed framework takes RGB and NIR images as inputs and predicts the rPPG signal. The proposed model is composed of three primary components: a feature extractor, CSLM, and MS-Mamba. First, the feature extractor is composed of 3D convolutions and pooling operations to extract spatio-temporal features, which are crucial for capturing quasi-periodic signals from local spatial regions. After the feature extractor, we obtain $X_{rgb} \in \mathbb{R}^{T \times C}$ and $X_{nir} \in \mathbb{R}^{T \times C}$, where T and C denotes sequence length and channel dimension, respectively.

Since RGB and NIR signals capture distinct spectral properties, leveraging their respective advantages is essential. To this end, we introduce the CSLM, which performs feature-wise linear modulation by conditioning one spectral representation on the other. This process allows each modality to be adaptively refined based on complementary spectral information. The modulated features are then fed into the MS-Mamba module to model long-term temporal dependencies. Each modality undergoes sequential modeling along the

temporal axis through a state space mechanism, followed by a feed-forward network that enhances frequency-domain representations relevant to physiological rhythms. We apply a bidirectional channel scan to exploit dependencies across feature channels and further strengthen cross-channel representations. Lastly, the estimated PPG signal is regressed by the rPPG head module.

C. Frequency domain cross-spectral linear modulation

The intensity changes in skin regions due to blood flow are subtle compared to the absolute intensity values. As reported in [5], NIR has a lower SNR than RGB because its lower hemoglobin absorption results in much weaker pulse signals. While RGB provides stronger periodic signals with a higher SNR in stable lighting conditions, NIR is more beneficial for acquiring physiological signals in low-light environments or under severe illumination changes. Our goal is to leverage the complementary information between different spectral modalities. To this end, we introduce CSLM, a fusion method that adopts feature-wise linear modulation [22] rather than simple summation or concatenation. In our approach, CSLM computes scale γ and shift β parameters from one set of features to adjust and modulate the features from the other modality.

The generator selectively extracts cardiac-related frequency information from an input signal $X_{\text{source}} \in \mathbb{R}^{T \times C}$ to compute the modulation parameters γ and β . Specifically, the signal is first transformed into the frequency domain via fast Fourier transform (FFT), after which a band-pass mask M is applied to retain only the frequency components corresponding to a predefined physiological range of 40 bpm to 200 bpm. The filtered signal is then passed through a lightweight neural network G , composed of 1D depthwise and pointwise convolutions, which outputs the parameters $\gamma, \beta \in \mathbb{R}^{C \times T}$.

$$\gamma, \beta = G(F^{-1}(F(X_{\text{source}}) \odot M)), \quad (5)$$

where F and F^{-1} denote the FFT and inverse FFT, respectively, and \odot represents element-wise multiplication. The generated parameters are constrained to a stable range using the \tanh activation function. Finally, these parameters are used to modulate X_{target} through an affine transformation, as defined as follows:

$$\text{CSLM}(X_{\text{target}} | \gamma, \beta) = X_{\text{target}} \odot (1 + \gamma') + \beta', \quad (6)$$

where γ' and β' are the parameters transposed and broadcast to match the dimensions of X_{target} . The modulated features obtained through this process are denoted as z_{rgb} and z_{nir} . Figure 3(b) depicts the operation of the CSLM.

D. Multi-spectral Mamba

We propose MS-Mamba, a model for multi-spectral inputs that enhances feature representation by jointly modeling RGB and NIR information. Mamba has demonstrated strong performance in efficiently capturing long-range dependencies, such as sequential relationships in time-series data.

In our framework, Mamba is leveraged to capture temporal dependencies between the two spectral modalities and facilitate their interaction through a shared state transition matrix. Initially, the two input streams z_{rgb} and z_{nir} are processed through independent Mamba blocks to learn modality-specific characteristics. Each block utilizes a shared Mamba layer across both RGB and NIR channels, thereby capturing long-range temporal dependencies while maintaining spectral consistency.

$$\tilde{z}_m = \text{Mamba}(z_m), \quad m \in \{\text{rgb}, \text{nir}\}. \quad (7)$$

The resulting features, denoted as \tilde{z}_{rgb} and \tilde{z}_{nir} , represent the embeddings of each modality. These features are then fused into z_{fused} and further refined by a frequency-domain feed-forward network (FFN) [15]. FFN is to amplify physiologically relevant periodic components and suppress noise artifacts, yielding clearer pulse representations.

While Mamba effectively models long-range temporal dependencies, its architecture applies the SSM independently to each channel, overlooking cross-channel interactions and thus limiting representational capacity. To address this limitation, we reformulate the channel dimension as a sequence by transposing the fused representation z_{fused} to $\bar{z}_{\text{fused}} \in \mathbb{R}^{C \times T}$. The transposition allows the state space model to operate along the channel axis, enabling bidirectional scanning across feature channels. Consequently, the model effectively captures inter-channel dependencies missed when channels are treated independently, enriching the fused representation with complementary spectral information. Bidirectional state-space modeling involves both forward and backward scans, as illustrated in Figure 2. This operation highlights informative channels and simultaneously models their interactions with other feature channels.

$$\bar{z}_{\text{fused}} = \text{BiMamba}(\bar{z}_{\text{fused}}). \quad (8)$$

The output of the channel scanning, denoted as \tilde{z}_{fused} , is added to the z_{rgb} and z_{nir} through skip connections, ensuring that each stream retains its unique modality-specific information while also incorporating the enriched contextual representation obtained from MS-Mamba. The outputs from the two branches are then averaged to yield the final fused representation, which is fed into the rPPG head to regress the estimated pulse signal. Our architecture includes 4 sequential MS-Mamba blocks.

E. Loss function

We employed the negative Pearson correlation loss, a temporal constraint calculated between the estimated rPPG and ground truth signals.

$$L_{\text{time}} = 1 - \frac{T \sum_1^T y y' - \sum_1^T y \sum_1^T y'}{\sqrt{(T \sum_1^T y^2 - (\sum_1^T y)^2)(T \sum_1^T y'^2 - (\sum_1^T y')^2)}}, \quad (9)$$

where y and y' are the ground truth pulse signal and predicted rPPG signal, respectively. L_{time} aligns the trend and peak values of the predicted signal with those of the sensor signal.

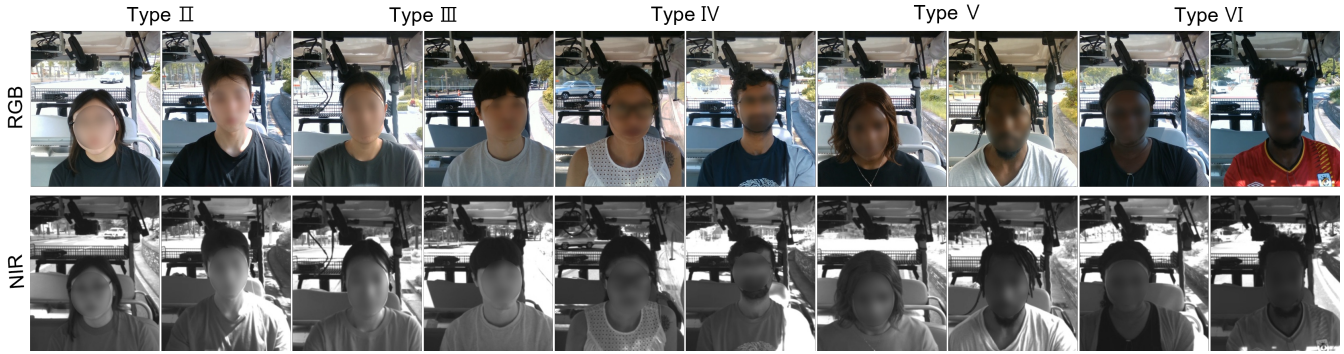


Fig. 4. Example images from the MS-Drive dataset, which contains simultaneously captured RGB, NIR, and reliable ECG sensor data.

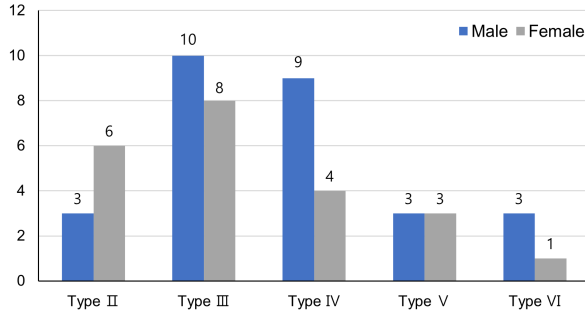


Fig. 5. Number of subjects per Fitzpatrick skin type in the MS-Drive dataset.

For the frequency domain constraint, the loss L_{freq} is defined using cross-entropy. It is calculated between the heart rate derived from the ground truth signal y and the power spectral density (PSD) of the predicted signal y' , as formulated below:

$$L_{freq} = CE(maxIdx(PSD(y)), PSD(y')), \quad (10)$$

where $maxIdx$ represents the index of the maximum value. The overall loss function is $\alpha \cdot L_{time} + \beta \cdot L_{freq}$, where α and β are set to 0.2 and 1.0, respectively.

IV. EXPERIMENTAL RESULTS

A. Datasets

The MR-NIR-Car dataset is a publicly available rPPG dataset that contains facial videos and ground truth pulse signals collected during driving. The experiments were conducted in both driving and parked conditions inside a garage, and videos were simultaneously recorded using RGB and narrow-band NIR cameras from 18 subjects at 30 FPS with a resolution of 640×640 . The NIR video was captured using narrow-band filters at 940 nm and 975 nm. Participants were seated in the passenger seat and instructed to remain still, move their heads naturally, and occasionally perform large movements. The PPG signals were collected using a CMS 50D+ finger pulse oximeter recording at 60 FPS, which was then downsampled to match the video frame rate and synchronized with the video recordings. Small motion refers to natural head movements that drivers typically make, such

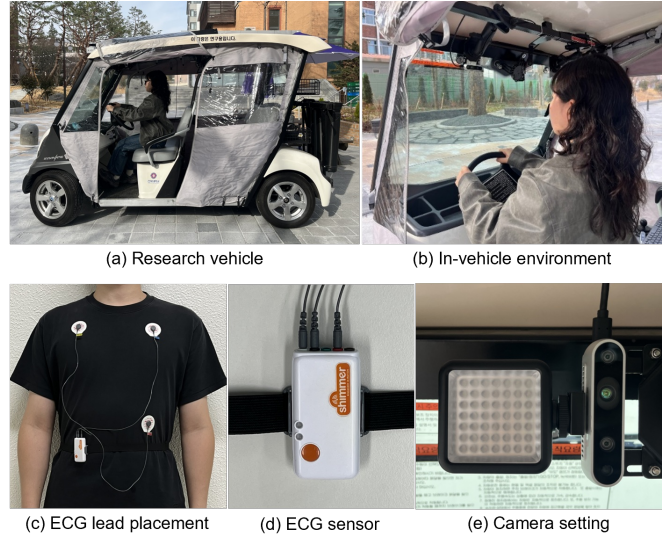


Fig. 6. The data acquisition setup for the MS-Drive dataset collection.

as glancing at rear-view and side-view mirrors or talking, without large or sudden movements, while the large motion condition required participants to significantly and abruptly move their heads out of plane.

We present MS-Drive, a real-world driving dataset designed to benchmark remote physiological estimation algorithms. To ensure diversity in the dataset, we recruited 50 participants (28 male, 22 female) aged 20-34, representing various ethnic backgrounds and distributed across the Fitzpatrick [23] skin types. The distribution is shown in Figure 5. The dataset includes participants with Fitzpatrick skin types 2-6, accounting for 18%, 36%, 26%, 12%, and 8%, respectively. The inclusion of five distinct skin types shows a reasonable diversity in skin tone representation across the dataset.

In contrast to the MR-NIRP Car dataset, participants operated the research vehicle directly during driving sessions. They were instructed only to avoid looking backward, while all other movements were performed naturally. For data acquisition, we used an Intel RealSense D435i camera with an external 850 nm NIR illuminator to record synchronized RGB videos at a resolution of 1920×1080 and NIR videos

TABLE I

INTRA-DATASET EVALUATION RESULTS ON THE MR-NIRP CAR DATASET WITH THE 975 NM NIR SETTING UNDER SMALL AND LARGE MOTION SCENARIOS. BEST RESULTS ARE MARKED IN BOLD.

Methods	Small motion					Large motion				
	MAE↓	RMSE↓	MAPE↓	Pearson↑	SNR↑	MAE↓	RMSE↓	MAPE↓	Pearson↑	SNR↑
DeepPhys [9]	11.746	13.165	16.753	-0.621	-11.486	16.336	17.443	22.675	-0.003	-11.715
EfficientPhys [10]	12.016	15.448	15.665	-0.464	-11.581	13.771	16.915	18.377	0.043	-10.091
PhysNet [19]	6.616	10.257	10.123	-0.118	-5.329	6.211	7.817	9.013	0.006	-5.512
PhysFormer [11]	5.363	7.671	7.473	0.464	-10.777	34.833	35.738	49.78	0.229	-12.479
RhythmFormer [12]	6.211	9.077	9.358	0.034	-1.780	7.021	9.125	9.991	0.581	-7.854
PhysMamba [14]	8.101	10.834	11.696	-0.362	-6.135	26.462	27.391	36.206	NaN	NaN
RhythmMamba [15]	7.561	10.240	11.151	-0.344	-6.181	9.991	13.517	14.828	0.254	-11.088
MS-rPPG w/o RGB	8.100	10.437	11.648	-0.397	-4.959	6.886	8.635	9.400	-0.058	-5.686
MS-rPPG w/o NIR	4.590	7.656	7.152	0.483	-3.668	5.671	9.132	8.981	0.054	-4.858
MS-rPPG (ours)	4.320	6.038	6.470	0.728	-1.606	4.320	6.447	6.546	0.589	-5.228

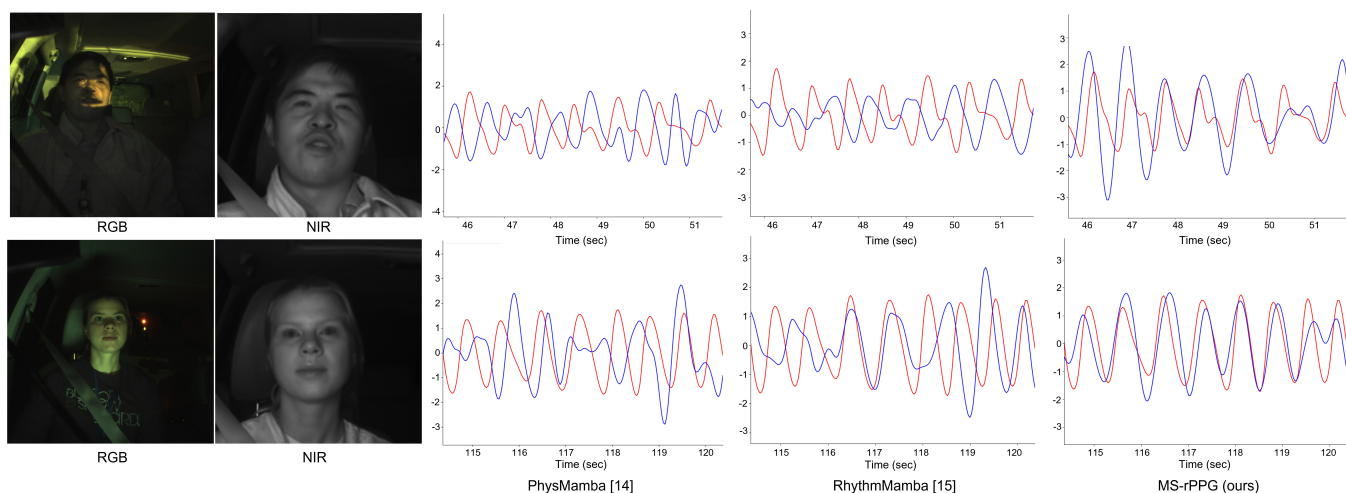


Fig. 7. Visualization of test results on the MR-NIRP dataset, showing the ground-truth and predicted PPG signals in red and blue, respectively.

at 848×480 , both at 30 FPS. Simultaneously, ground truth electrocardiogram (ECG) signals were recorded at 128 Hz using a 3-lead Shimmer3 ECG unit, as shown in Figure 6. For model training, the physiological signals were synchronized with the video frames. Example images from the dataset are shown in Figure 4. All data collection procedures were conducted under the approval of the Institutional Review Board (IRB Number: JBNU 2025-04-035-002).

B. Implementation Details

For preprocessing, the RGB and NIR video inputs were first divided into 160 frame segments. Facial recognition was applied to the first frame of each segment to detect the facial region, which was then cropped and resized to 128×128 , consistent with existing methods. We applied the DiffNormalized method [9] to input frames, first calculating the normalized difference between consecutive frames. The resulting sequence is then standardized by its standard deviation. The proposed method was implemented using an open-source rPPG toolbox [24] for fair comparisons with existing methods, and the implementation was based on PyTorch.

The models were trained using a batch size of 4 and a learning rate of $3e-4$ for 30 epochs. All experiments were performed with NVIDIA RTX 6000 Ada. To evaluate the performance of the remote HR estimation task, we employed standard metrics including mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), Pearson correlation coefficient (ρ), and SNR.

C. Intra-dataset Evaluation

Table I summarizes the intra-dataset evaluation results on the MR-NIRP dataset under both small- and large-motion conditions using the 975 nm NIR band. The qualitative results are illustrated in Figure 7. Under the small-motion condition, our model achieves an MAE of 4.320 and a Pearson correlation of 0.728, outperforming EfficientPhys with a 7.696 reduction in MAE. PhysFormer and PhysMamba exhibit unstable performance in the large-motion setting, suggesting that ViT and Mamba-based models are sensitive to such variations. PhysFormer achieves competitive performance in the small-motion case, with an MAE of 5.363 and an RMSE of 7.671. However, its performance degrades

TABLE II

CROSS-DATASET EVALUATION RESULTS. WE TRAINED THE MODEL ON MR-NIRP SMALL MOTION 975NM AND LARGE MOTION 975NM, AND TESTED ON OUR COLLECTED DATASET. BEST RESULTS ARE MARKED IN BOLD.

Methods	Testing on MS-Drive					
	Small motion			Large motion		
	MAE↓	RMSE↓	MAPE↓	MAE↓	RMSE↓	MAPE↓
EfficientPhys [10]	25.659	28.690	29.368	31.479	33.814	36.360
PhysNet [19]	13.189	26.397	15.073	13.345	15.939	15.449
PhysFormer [11]	14.905	17.794	17.190	22.142	25.338	29.090
RhythmFormer [12]	13.201	15.732	15.286	11.245	14.347	13.139
PhysMamba [14]	14.370	17.509	16.549	38.152	39.844	44.560
RhythmMamba [15]	12.373	15.867	14.273	11.785	14.257	14.414
MS-rPPG (ours)	11.797	14.246	13.927	9.601	12.745	11.406

TABLE III

ABLATION STUDY OF KEY CONTRIBUTIONS ON THE MR-NIRP DATASET.

CSLM	MS-Mamba	MAE↓	RMSE↓	MAPE↓	Pearson↑	SNR↑
-	✓	4.860	7.365	7.366	0.488	-1.059
1D Conv	✓	4.455	6.128	6.635	0.759	-2.529
✓	w/o channel scan	6.346	8.388	8.767	0.163	-3.880
✓	w/o BiMamba	6.076	7.896	8.676	0.260	-2.179
✓	✓	4.320	6.038	6.470	0.728	-1.606

significantly under large head motion, with the MAE and RMSE increasing to 34.833 and 35.748, respectively. In real-world scenarios, ambient lighting is often distributed unevenly across the face. Under such conditions, significant head movements can induce drastic temporal variations in illumination for the same spatial region, which explains the performance degradation. MS-Mamba showed significant improvement under large head movements, achieving a 30.513 reduction in MAE and a 0.360 increase in Pearson correlation.

Further analysis highlights the effectiveness of using two spectral modalities, showing that multi-spectral inputs are particularly beneficial under varying lighting conditions. Removing the RGB input led to a significant drop in rPPG estimation performance. In contrast, excluding the NIR input resulted in a comparatively minor degradation, especially in the small-motion case. This suggests that the RGB signal contains high-SNR pulse information, a finding consistent with [5]. While the RGB modality struggled with illumination changes from large head motion, the NIR modality proved more robust under these conditions. However, surpassing both single-modality approaches, the proposed multi-spectral model consistently achieved superior performance with the lowest errors and highest correlations. This demonstrates the success of our design in leveraging the complementary advantages of RGB and NIR to achieve consistent improvements over prior methods.

D. Cross-dataset Evaluation

Table II presents the cross-dataset evaluation results, where models trained on the MR-NIRP dataset were tested on MS-Drive. For the heart rate estimation task, the ground truth heart rate was derived from the collected ECG signal.

Many existing methods, such as EfficientPhys and PhysMamba, struggle to learn robust pulse signal representations, leading to poor generalization in cross-dataset evaluations. While 3D CNN architecture from PhysNet offered a solid approach for extracting spatio-temporal features, especially under large motion, the MS-rPPG model achieves a promising performance. Specifically, our proposed method improves upon PhysNet, lowering the MAE by 1.392 in small-motion and 3.744 in large-motion scenarios.

Although ViT-based models such as RhythmFormer demonstrate strong generalization capability, their quadratic scaling in self-attention is computationally expensive, thereby limiting their practicality for long video sequences. Our proposed MS-rPPG achieves state-of-the-art performance in cross-dataset rPPG estimation for driving scenarios, with MAE scores of 11.797 under small motion and 9.601 under large motion conditions. This result represents a notable improvement over recent SSM-based methods such as RhythmMamba, reducing MAE by 0.576 and 2.184, respectively. Figure 8 presents the cross-dataset qualitative results of our model. These findings confirm the effectiveness of our multi-spectral algorithm for robust rPPG estimation in challenging driving environments.

E. Ablation study

To demonstrate the effectiveness of the proposed CSLM and MS-Mamba, we conducted an ablation study on small motion 975 nm data from the MR-NIRP dataset. The results are shown in Table III. We designed CSLM to modulate each feature with complementary spectral information by leveraging physiological frequency band information obtained from FFT-based analysis. While using a simple 1D convolution for calculating CSLM parameters yielded a high Pearson correlation, it resulted in lower performance across all error metrics and the SNR. Moreover, when features directly extracted from the feature extractor were fed into MS-Mamba without

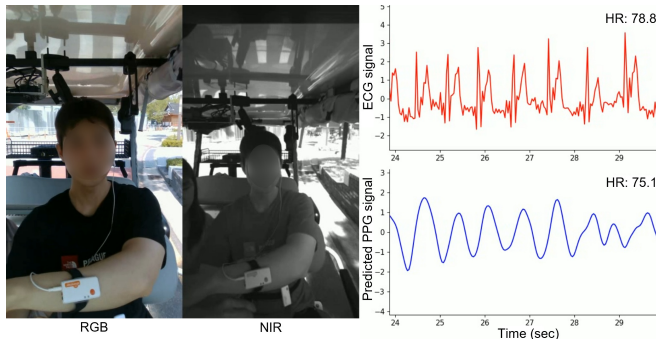


Fig. 8. Cross-dataset evaluation results of MS-rPPG on the MS-Drive dataset.

CSLM, the overall performance declined. However, the drop was not severe, which highlights the inherent robustness of the MS-Mamba architecture.

We also conducted an ablation study on the channel scanning mechanism in MS-Mamba. The key feature of our proposed framework is its bidirectional channel-axis scan, which is designed to effectively capture inter-channel dependencies. As shown in the experimental results, unidirectional scanning leads to noticeable performance degradation compared to the proposed method. This result confirms that bidirectional scanning is crucial for learning robust inter-channel dependencies. The performance degrades even more significantly when channel scanning is omitted entirely, which further underscores its effectiveness. The proposed MS-Mamba and CSLM have proven to be effective for video-based physiological estimation, exhibiting strong robustness against interferences in driving scenarios.

V. CONCLUSION

This work proposes MS-rPPG, a novel multi-spectral framework for rPPG estimation in challenging driving environments. The framework integrates RGB and NIR videos through the CSLM, which adaptively modulates spectral features using a physiological prior derived from FFT analysis. Additionally, the paper introduces MS-Mamba, a new architecture employing SSM. MS-Mamba models long-range temporal information with a time-axis SSM and captures inter-channel dependencies via a bidirectional channel SSM. Comprehensive experiments on the MR-NIRP and MS-Drive datasets demonstrate the effectiveness of the proposed method. These results indicate the potential of MS-rPPG for real-world driver monitoring systems, even under challenging conditions with significant motion and illumination variations.

REFERENCES

- [1] A. Hajr, B. Tarvirdizadeh, K. Alipour, and M. Ghamari, "Contactless health monitoring: An overview of video-based techniques utilising machine/deep learning," *IET Wireless Sensor Systems*, vol. 15, no. 1, p. e70009, 2025.
- [2] Q. Zhu, C.-W. Wong, Z. M. Lazri, M. Chen, C.-H. Fu, and M. Wu, "A comparative study of principled rppg-based pulse rate tracking algorithms for fitness activities," *IEEE Transactions on Biomedical Engineering*, 2024.

- [3] J. Li and J. Peng, "End-to-end multimodal emotion recognition based on facial expressions and remote photoplethysmography signals," *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [4] Z. Yu, X. Li, and G. Zhao, "Facial-video-based physiological signal measurement: Recent advances and affective applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 50–58, 2021.
- [5] E. Magdalena Nowara, T. K. Marks, H. Mansour, and A. Veeraghavan, "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*, 2018, pp. 1272–1281.
- [6] S. G. Ahmed, K. Verbert, N. Siedahmed, A. Khalil, H. Aljassmi, and F. Alnajjar, "Ai innovations in rppg systems for driver monitoring: Comprehensive systematic review and future prospects," *Ieee Access*, 2025.
- [7] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE transactions on biomedical engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.
- [8] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3430–3437.
- [9] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 349–365.
- [10] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 5008–5017.
- [11] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Phys-former: Facial video-based physiological measurement with temporal difference transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4186–4196.
- [12] B. Zou, Z. Guo, J. Chen, and H. Ma, "Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer," *CoRR*, 2024.
- [13] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
- [14] C. Luo, Y. Xie, and Z. Yu, "Phymamba: Efficient remote physiological measurement with slowfast temporal difference mamba," in *Chinese Conference on Biometric Recognition*. Springer, 2024, pp. 248–259.
- [15] B. Zou, Z. Guo, X. Hu, and H. Ma, "Rhythmamba: Fast, lightweight, and accurate remote physiological measurement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 11 077–11 085.
- [16] A. Vilesov, P. Chari, A. Armouti, A. B. Harish, K. Kulkarni, A. Deoghare, L. Jalilian, and A. Kadambi, "Blending camera and 77 ghz radar sensing for equitable, robust plethysmography," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 36–1, 2022.
- [17] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.
- [18] W. Wang, A. C. Den Brinker, S. Stuijk, and G. De Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.
- [19] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," *arXiv preprint arXiv:1905.02419*, 2019.
- [20] S. Park, B.-K. Kim, and S.-Y. Dong, "Self-supervised rgb-nir fusion video vision transformer framework for rppg estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.
- [21] K. Wang, J. Tang, Y. Fan, J. Ji, Y. Shi, and Y. Wang, "Memory-efficient low-latency remote photoplethysmography through temporal-spatial state space duality," *arXiv preprint arXiv:2504.01774*, 2025.
- [22] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [23] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.
- [24] X. Liu, G. Narayanswamy, A. Paruchuri, X. Zhang, J. Tang, Y. Zhang, R. Sengupta, S. Patel, Y. Wang, and D. McDuff, "rppg-toolbox: Deep remote ppg toolbox," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68 485–68 510, 2023.