# PulseMamba: An Efficient Framework with Multi-Scale Fusion and Frequency Enhancement for Non-Contact Heart Rate Estimation

IEEE Publication Technology, *Staff, IEEE,*

*Abstract*—**Facial-video based remote photoplethysmography (rPPG) enables the extraction of physiological signals in a non-contact manner, offering great potential in various medical and industrial applications. However, existing methods struggle to balance long-range spatiotemporal modeling with computational efficiency, while also facing difficulties in capturing multi-scale features and accurately modeling the periodicity behavior of rPPG signals under complex scenarios. To address these issues, we propose PulseMamba, a novel Mamba-based end-to-end architecture to enhance the robustness and accuracy of rPPG signal extraction. First, PulseMamba introduces the Residual Spatial Channel Mamba (RSCMamba) Block to efficiently capture long-range spatiotemporal dependencies while maintaining linear computational complexity. Additionally, we employs a Frequency Spatial Enhancement Attention (FSEA) module that enables multi-scale feature extraction and dynamic contextual fusion to enhance rPPG signal extraction. Furthermore, we also integrates a Fourier Residual Network (FRN) module to selectively enhance informative frequency components and refine frequency domain characteristics, ultimately improving the periodicity detection of rPPG signals. Extensive experiments on three benchmark datasets demonstrate that PulseMamba exhibiting strong performance across both intra- and cross-datasets evaluations, achieving notable robustness in complex real-world conditions. Furthermore, PulseMamba offers superior generalization capability and faster computational speed, laying a solid foundation for deploying rPPG networks on resource-constrained Internet of Medical Things devices.**

*Index Terms*—**Remote photoplethysmography (rPPG), remote heart rate measurement, healthcare monitoring, Mamba, facial video analysis.**

## I. Introduction

Heart rate (HR) is one of the most important physiological indicators and is closely linked to various cardiovascular diseases. Accurate HR monitoring plays a vital role in chronic disease prevention and public health management [1]. Recently, the rise of wearable health monitoring technologies has made physiological signal estimation more convenient and intelligent in everyday life [2], [3].

Traditional contact-based HR monitoring methods rely on placing sensors directly on the skin. These methods can be broadly categorized into electrocardiography(ECG)-based [4] and photoplethysmography(PPG)-based approaches [5]. ECG-based methods offer high accuracy by measuring the electrical activity of the heart through electrodes [6], but the need to attach multiple sensors to the body can be very inconvenient. PPG-based methods estimate HR by detecting changes in blood volume through the reflected light, making them easier to integrate into wearable devices [7]. However, contact-based HR estimation methods require direct skin contact, which limits their uses for long-term monitoring or on sensitive groups such as infants and burn patients.

Nowadays, most consumer electronics, especially smartphones and laptops, come with built-in cameras, making visual data more accessible [8]–[10]. To overcome the limitations of contact-based methods, researchers have proposed a new non-contact HR measurement technique, remote photoplethysmography (rPPG) [11]. This method captures facial videos using consumer-grade RGB or near-infrared cameras and analyzes subtle skin color changes caused by blood volume fluctuations to extract pulse signals [12]–[14]. Since Verkruysse et al. [15] first demonstrated its feasibility using a standard RGB camera, rPPG technology has been widely applied in various scenarios, such as emotion recognition, health assessment, and driver state detection. However, accurately extracting HR signals from facial videos remains challenging due to the interference of various noise sources, including illumination variations, facial expressions, and head movements [16].

Early approaches for rPPG signal extraction primarily relied on hand-crafted signal processing techniques to separate HR signals from facial videos, such as color space transformations [17], [18] and blind source separation methods [19], [20]. However, these traditional methods struggle in real-world scenarios due to their reliance on simplified assumptions, like skin reflection models and the linear combination of noise sources, which limit their adaptability and robustness.

With the impressive performance of deep learning, data-driven approaches have proved to be an effective way of overcoming the limitations of traditional methods in remote physiological estimation [21]–[23]. However, these deep learning models still suffer from high computational complexity and large parameter sizes, making them less suitable for deployment on resource-constrained edge devices or mobile platforms [24].

Recently, Mamba [25] architecture has demonstrated impressive performance in both vision and language tasks [26], [27]. Unlike conventional 3D-CNN and Transformer-based models, Mamba leverages a state space modeling (SSM) framework with linear time complexity, allowing efficient sequence modeling while significantly reducing computational complexity. This characteristic makes Mamba a promising candidate for remote HR estimation task, and also well-suited for deployment on resource-constrained mobile and edge devices. However, despite these achievements, several key challenges remain unaddressed in rPPG signal extraction. First, it remains highly challenging to capture long-range

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

spatiotemporal dependencies while preserving the model's lightweight. 3D-CNN is not suitable for modeling long-range temporal dynamics due to its fixed receptive fields [28]. Despite being able to effectively capture long-range dependencies, transformer-based models are limited by their quadratic complexity Transformer-based models can effectively capture long-range and scalability on long sequences, making it difficult for existing models to be flexibly deployed on real-time or edge computing devices.

Second, achieving effective multi-scale feature extraction and dynamic contextual fusion remains a key challenge. Most existing methods extract features at fixed scales and lack mechanisms to integrate multi-scale information [29], [30], limiting their ability to capture both fine-grained fluctuations and global patterns. This restricts the stability and precision of the extracted physiological signals.

Third, precisely modeling the full-period behavior of rPPG signals remains a significant challenge. Most existing methods fail to capture the periodicity of rPPG signals due to the lack of frequency-domain features, making it difficult to distinguish true physiological signals from non-periodic noise under complex real-world conditions. While a few approaches utilize frequency analysis [31], [32], they are typically limited to shallow signal processing techniques and lack effective fusion with spatial-temporal representations.

To address the aforementioned challenges, we propose PulseMamba, an end-to-end HR estimation framework based on Mamba through an encoder-decoder architecture.The PulseMamba consists of three newly designed modules. The first is the *RSCMamba* (*R*esidual *S*patial *C*hannel Mamba) module, which is designed to efficiently capture long-range temporal dependencies. It enables global spatiotemporal modeling over long sequences while maintaining linear computational complexity relative to input tokens. The second component is the *FSEA* (*F*requency-*S*patial *E*nhancement *A*ttention) module, which is designed to perform multi-scale frequency feature extraction and dynamic contextual information fusion, further enhancing the robustness of rPPG signal modeling under real-world scenarios. Finally, we introduce the *FRN* (*F*requency *R*esidual *N*etwork) module, which applies Fourier Transform to enhance frequency-domain representations and improves model's capacity to capture periodicity of rPPG signals. By jointly modeling temporal, spatial, and frequency information, PulseMamba achieves more accurate and efficient HR estimation across diverse and challenging environments.

The main contributions of this paper can be summarized as follows:

- We propose PulseMamba, an novel Mamba-based framework capable of efficiently extracting rPPG signals through an encoder-decoder architecture.
- We propose RSCMamba module that enables global spatiotemporal modeling over long sequences while maintaining linear computational complexity relative to input tokens.
- We introduce the FSEA module, which enables multi-scale feature extraction and dynamic contextual fusion to enhance rPPG signal extraction.
- We introduce the FRN module, which enhances frequency-domain representations and improves model's capacity to capture periodicity of rPPG signals.
- Comprehensive experimental results demonstrate that the proposed PulseMamba achieves superior performance in HR estimation accuracy, robustness, and computational efficiency compared to previous CNN- and Transformer-based methods.

The remainder of this paper is organized as follows. In Section II, we overview related work on vision-based physiological measurements. In Section III, we describe the specific architectural details of the proposed PulseMamba framework. Section IV analyzes the experimental settings and results. Finally, we conclude this work in Section V.

## II. Related Work

### A. Traditional Methods for rPPG Signal Extraction

Traditional signal-based approaches can generally be categorized into two main groups: Blind Source Separation (BSS)-based methods and chrominance-based skin reflectance models.

BSS-based methods aim to separate periodic components from complex RGB signals, identifying them as the underlying rPPG signal. Methods such as Independent Component Analysis (ICA) [19] and Principal Component Analysis (PCA) [33], decompose raw RGB signal over time and extract the dominant pulse-related component. Model-based approaches use color vector information to separate signal components, effectively reducing the dependency of RGB signals on the skin reflectance chrominance channel. For instance, Wang *et al.* proposed POS [18] for pulse extraction in a temporally normalized RGB space to enhance signal extraction. The CHROM method [17] treats pulse signal as linear combination of RGB channels to minimize motion artifacts. PBV [20] distinguishes pulse-related color fluctuations by analyzing variations in blood volume.

Despite their initial success, these traditional methods still face limitations like insufficient accuracy and simplified assumptions, making them less effective in real-world environments.

### B. Deep Learning Methods for rPPG Signal Extraction

Recently, deep learning models have been widely used in HR estimation tasks and achieved significant progress. Early works relied on 2D convolutional neural networks (2D-CNNs), which primarily focused on spatial feature extraction and played a foundational role in early-stage HR estimation networks. For instance, Chen *et al.* proposed DeepPhys [34], jointly models appearance and motion to achieve stable rPPG signal extraction under varying lighting conditions. However, its reliance on 2D-CNNs limits its ability to capture temporal periodic features. To overcome this limitation, Liu *et al.* [35] introduced an improved framework TS-CAN [35], which incorporate temporal shift modules (TSM) to better capture temporal dependencies while reducing inference time.

In 2018, Yu *et al.* proposed PhysNet [36], a 3D-CNN-based model designed to analyze raw video sequences and extract meaningful temporal features for accurate HR estimation. They then successively propose EfficientPhys [37] to further optimize performance by balancing computational efficiency and signal modeling capability. Nevertheless, the limited receptive fields of 3D-CNNs limit their abilities to capture long-term temporal dependencies.

To overcome this challenge, Yu *et al.* [38] employs a temporal difference transformer to capture long-range spatiotemporal dependencies. By focusing on global context and fine-tuning local spatial-temporal patterns, this approach effectively captures quasi-periodic rPPG characteristics and achieves outstanding results. However, Transformer-based models for video processing often suffer from excessive computational redundancy, which commonly requires complex preprocessing, resulting in the loss of subtle yet critical signal details.

### C. Mamba

Recently, Mamba [25] has emerged as a promising architecture for maintaining global context while significantly reducing computational redundancy. This model attracts increasing interests due to its ability to model long sequence data. Initially, Mamba was introduced to natural language processing (NLP) tasks to enable the modeling of long-range dependencies with linear complexity. As research progressed, Mamba has also been extended to vision-related tasks. In 2024, Zhu *et al.* [27] propose the Vision Mamba (ViM) architecture, which converts images into ordered visual sequences using bidirectional scanning, enhancing the model's ability to capture spatial context.

Compared to Transformer-based models, Mamba demonstrates superior efficiency in handling long-range dependencies and offers improved sensitivity to subtle motion changes. These characteristics make it particularly suited for video-based HR estimation tasks.

## III. Method

In this section, we first introduce the essential concepts of the state space models (SSMs). Then we delve into a comprehensive description of PulseMamba, including its overall framework and module design.

### A. Preliminaries

**State Space Models (SSMs)** leverages intermediate state variables to model the linear interaction among system inputs, internal states, and outputs. Inspired by SSM, the Structured State Space for Sequence Model (S4) [39] has been proposed to effectively model long sequence data. In this model, one-dimensional input sequence $x(t) \in \mathbb{R}$ is mapped to $y(t) \in \mathbb{R}$. The process can be mathematically represented as a linear Ordinary Differential Equation (ODE):

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

where $h(t) \in \mathbb{R}^N$ represents the hidden state, $h'(t) \in \mathbb{R}^N$ indicates the time derivative of $h(t)$. $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the state transition matrix, and $N$ denotes the number of states. $\mathbf{B} \in \mathbb{R}^N$ and $\mathbf{C} \in \mathbb{R}^N$ are projection matrices.

To fit discrete signals such as images and text, the continuous-time SSM in the S4 model is discretized using the Zero-Order Hold (ZOH) assumption. Specifically, a timescale parameter $\Delta$ is introduced to convert the continuous parameters $\mathbf{A}$ and $\mathbf{B}$ into their discrete counterparts $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, which can be defined as follows:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \end{aligned} \tag{2}$$

After discretizing $\mathbf{A}$ and $\mathbf{B}$, Equation (1) can be reformulated as follows:

$$\begin{aligned} h(t) &= \overline{\mathbf{A}}h(t-1) + \overline{\mathbf{B}}x(t), \\ y(t) &= \mathbf{C}h(t). \end{aligned} \tag{3}$$

In addition, the model can be trained efficiently and in parallel by using global convolution. Equation (3) can be reformulated as:

$$\begin{aligned} \overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \\ y &= x * \overline{\mathbf{K}}, \end{aligned} \tag{4}$$

where $L$ and $\overline{\mathbf{K}} \in \mathbb{R}^L$ represent the length of the input sequence and the convolution kernel used in S4 model.

### B. Overview

An overview of the PulseMamba is presented in Fig. 1 (a). The model adopts an encoder-decoder architecture that takes facial video sequences as inputs and and generates the pulse signals as outputs. Our proposed approach strengthens weak physiological signal inherent in rPPG data while efficiently capturing multi-scale features and accurately modeling the periodicity behavior of rPPG signals. As shown in Fig.1 (a), the architecture consists of three key modules: RSCMamba, FRN, and FSEA.

The input to the network is a sequence of raw video frames denoted as $X \in \mathbb{R}^{T \times C \times H \times W}$, where $C$ represents the number of channels, $T$ is the video length, and $H$ and $W$ indicate the height and width of each frame, respectively. First, the input video frames are processed by the Fusion stem [40] and Encoder modules to extract low-level spatiotemporal features. These feature maps are then passed to the RSCMamba module to capture long-range spatiotemporal dependencies. Subsequently, the outputs are then processed by the FSEA module, which conducts multi-scale feature extraction and contextual information fusion in the frequency domain. Thereafter, the FRN module further enhances periodic structures and suppresses redundant noise. Finally, the output of the FRN is fed into the Decoder and Predictor module to generate the estimated pulse signal. The network is trained using a negative Pearson correlation loss to ensure the temporal consistency and the waveform similarity between the predicted signal and the ground-truth signal. In the following section, we provide a detailed description of each component.

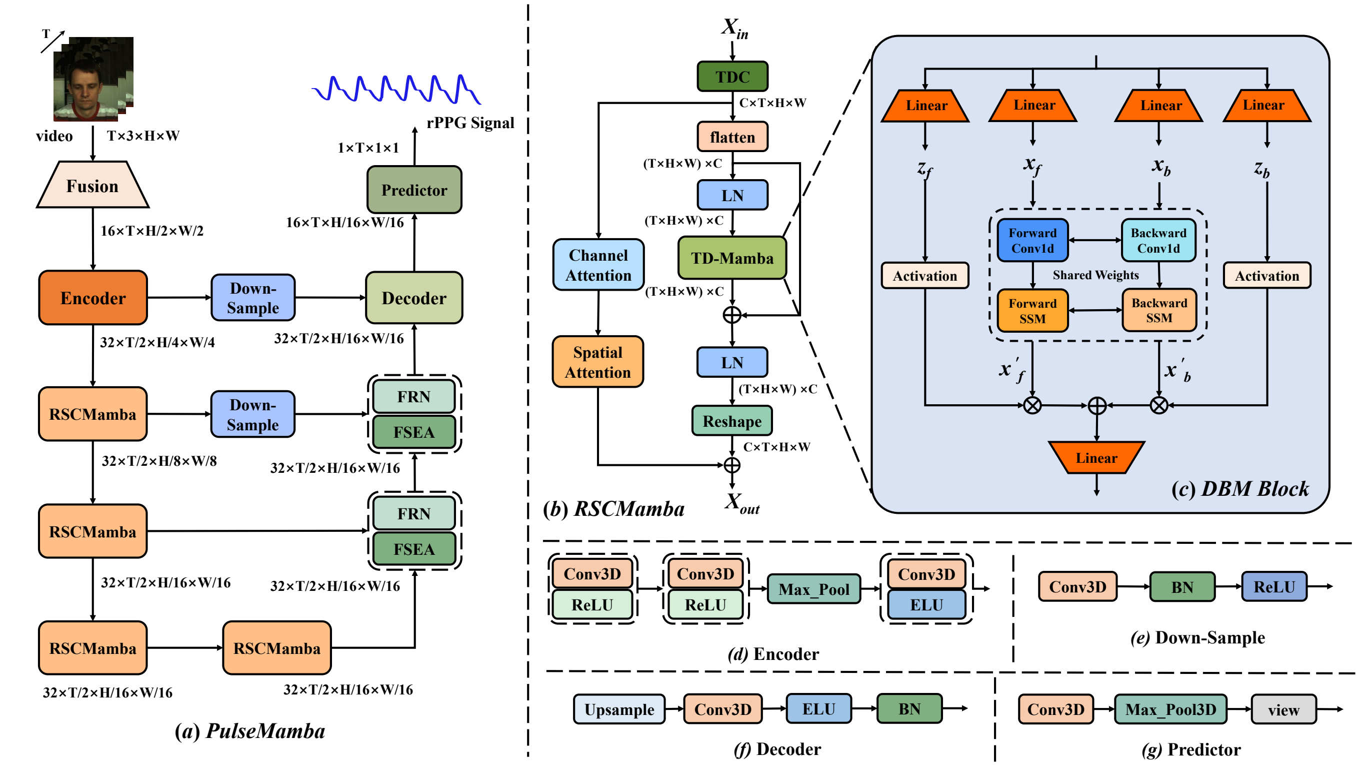


Fig. 1. The overall framework of the proposed PulseMamba.

## C. RSCMamba

*1) The Architecture:* As shown in Fig. 1 (b), RSCMamba module integrates SSMs, 3D attention mechanisms, and temporal difference convolution (TDC) [41] to enhance the modeling capability for spatiotemporal features, providing strong support for more accurate signal estimation.

The input feature $X_{in} \in \mathbb{R}^{B\times C\times T\times H\times W}$ passes through two parallel branches. In the first branch, the input $X_{in}$ first undergoes TDC module to enhance time-normalized frame difference feature representation. Then the spatiotemporal feature maps are flattened into a one-dimensional sequence $f_k \in \mathbb{R}^{B\times L\times C}$, where $L = T \times H \times W$. The flattened sequence is first fed into *LN* (LayerNorm), then passed to *DBM* (Decomposed Bidirectionally Mamba) [42] module for sequence modeling. This module captures long-range dependencies between frames, the output of DBM is added with the flattened sequence $f_k$ through a residual connection, preserving original information and improving model's robustness. After that, another LN is applied to normalize the output of DBM $f_{k+1}$ and reshape it back to $X_1 \in \mathbb{R}^{B\times C\times T\times H\times W}$.

Meanwhile, the input feature $X_{in} \in \mathbb{R}^{B\times C\times T\times H\times W}$ after TDC are integrated with channel and spatial attention, followed by a *ReLU* activation function to generate $X_2 \in \mathbb{R}^{B\times C\times T\times H\times W}$. Afterward, the features from the two parallel branches are summed to obtain the final output $X_{out}$. In summary, RSCMamba module performs computations as follows:

$$\begin{aligned}
f_k &= \text{flatten}(\text{TDC}(X_{\text{in}})),\\
X_1 &= \text{Reshape}(\text{LN}(\text{DBM}(\text{LN}(f_k)) + f_k)),\\
X_2 &= \text{Spatial\_Attn}(\text{Channel\_Attn}(\text{TDC}(X_{\text{in}})))),\\
X_{\text{out}} &= X_1 \oplus X_2,
\end{aligned} \tag{5}$$

where $f_k$ is the flattened sequence, LN represents LayerNorm, and DBM is the 3D SSM module.

*2) DBM Block:* The structure of the DBM module is shown in Fig. 1 (c). The DBM module is an improved bi-directional Mamba module. It refines the bi-directional scanning process by decomposing the input features and sharing parameters, enabling the integration of spatiotemporal information via a structured state-space model (SSM). The process of DBM is defined as follows:

$$f_{k+1} = \text{DBM}(\text{LN}(f_k)) + f_k, \tag{6}$$

where $f_k \in \mathbb{R}^{B\times C\times L}$ is the input sequence of DBM module.

The input feature $f_k$ is first projected into two hidden states, $x$ and $z$, using a linear layer with an expansion factor $E$. To enable bidirectional processing, the hidden state $x$ is reversed along the temporal axis to form backward and forward features $x_b$ and $x_f$. These features are subsequently fed into a shared-weights SSM for bidirectional processing and generate $x_f^{'}$ and $x_b^{'}$. The resulting forward and backward outputs are gated by corresponding $z_f$ and $z_b$, and then fused to generate the next hidden state $f_{k+1}$.

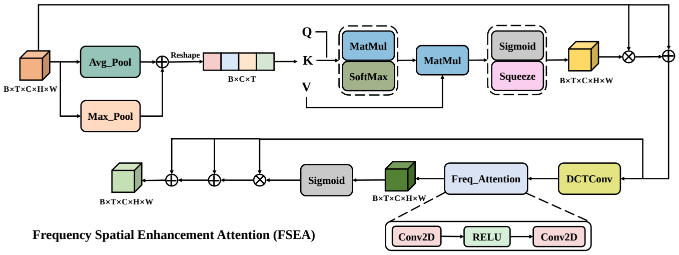


Fig. 2. The detailed structure of the Frequency Spatial Enhancement Attention module, which enables multi-scale feature extraction and dynamic contextual fusion to enhance rPPG signal extraction.

## D. FSEA

Most methods in HR estimation tasks extract features at fixed scales and lack explicit mechanisms to integrate information across different scales, which limits their ability to capture both fine-grained signal fluctuations and broader global patterns, restricting the model's ability to produce stable and precise physiological signals. To address this, we introduce the FSEA module as shown in Fig. 2. The FSEA module expands the input features in the frequency domain using Discrete Cosine Transform (DCT), decomposing them into multiple scales and applies frequency attention mechanism to assign adaptive weights to different frequency bands, thereby dynamically enhancing key frequency information while suppressing redundant components.

Specifically, we first utilize adaptive average pooling and max pooling separately along the spatial dimensions to extract channel-level global features $F_{\text{avg}}$ and $F_{\text{max}}$ from input feature $F \in \mathbb{R}^{B\times C\times T\times H\times W}$. Then we combined a weighted fusion using learnable parameters $\alpha$ and $\beta$ to obtain the temporal feature map $F_{\text{fusion}} \in \mathbb{R}^{B\times C\times T}$, which can be defined as follows:

$$F_{\text{fusion}} = \alpha F_{\text{avg}} + \beta F_{\text{max}}. \tag{7}$$

The fused features $F_{\text{fusion}}$ are subsequently fed into a temporal *Multi-Head Self-Attention* (MHSA) module to effectively capture temporal dependencies among video frames. After that, we apply sigmoid activation and mean calculation to construct temporal attention weights and generate temporally enhanced feature representation $F_t \in \mathbb{R}^{B\times C\times T\times H\times W}$, which can be represented as:

$$F_t = F \times Sigmoid(Mean(\text{MHSA}(F_{fusion})) + F. \tag{8}$$

In the spatial domain, we further introduce a DCT convolution to explicitly transform spatial features into frequency domain, thus generating spatial frequency features $F_{\text{freq}} \in \mathbb{R}^{B\times 1\times H\times W}$ :

$$F_{freq} = W_{\text{DCT}} * \text{Mean}(F_t, \text{dim} = 2), \tag{9}$$

where $W_{\text{DCT}}$ is the fixed kernels, which designed to explicitly capture various spatial frequency components. This characteristic enables the model to decompose features into multiple frequency bands, where low-frequency components reflect global trends such as illumination and head movement, and high-frequency components preserve fine-grained local variations like subtle skin tone changes and facial textures. By leveraging this frequency-aware decomposition, the module is particularly effective in enhancing informative frequency components.

After that, the frequency feature $F_{freq}$ are fed into a frequency convolutional attention module *Freq_Attn*, which identifies salient frequency components and generates an adaptive spatial attention map to facilitate dynamic contextual frequency information fusion, producing the spatially enhanced features $F_s \in \mathbb{R}^{B\times C\times T\times H\times W}$ as follows:

$$F_s = Sigmoid(\text{Freq\_Attn}(F_{freq})). \tag{10}$$

Finally, the module combines the temporal branch $F_t$ and spatial branch $F_s$ outputs through learnable weighting parameters $t$ and $s$, averaging their weighted sum to generate the final output feature $F' \in \mathbb{R}^{B\times C\times T\times H\times W}$.

$$F' = \frac{t \odot F_t + s \odot F_s}{2} \tag{11}$$

## E. FRN

Effectively modeling the non-linear, periodicity of rPPG signals plays a vital role in achieving accurate HR estimation. However, most existing methods fail to capture the periodic characteristics of rPPG signals due to the lack of frequency-domain features, making it difficult to distinguish true physiological signals from non-periodic noise under complex real-world conditions. In this paper, we propose a FRN module to effectively capture periodic information in the frequency domain, which is critical for modeling the subtle physiological signals within facial areas.

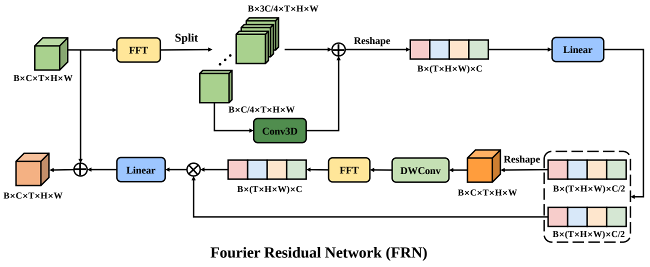


Fig. 3. The detailed structure of the Fourier Residual Network module. This module mainly consists of fourier transformation and partial convolution.

Specifically, We apply Fast Fourier Transform (*FFT*) to transform the temporal feature maps $\tilde{M} \in \mathbb{R}^{B\times C\times T\times H\times W}$ into the frequency domain, and utilize PConv3D [43] to three-quarters of the channels to enhance the informative elements within features. The processed channels and untouched channels are then concatenated along the channel dimension and reshape to $\tilde{M}_f \in \mathbb{R}^{B\times L\times C}$, where $L = T \times H \times W$. Subsequently, a linear transformation with GELU activation function is applied to enhance periodic feature representation:

$$\tilde{M}_f = GELU(Linear(PConv(\tilde{M}))). \tag{12}$$

Subsequently, we divide $\tilde{M}_f$ evenly into two tensors $M_1$ and $M_2$. $M_1$ is reshaped back to $\mathbb{R}^{B\times C\times T\times H\times W}$ and undergoes depth-wise convolution (DWConv) [44] in the frequency domain, which captures subtle frequency patterns along temporal and spatial dimensions without cross-channel interference. Then, we apply *Inverse FFT* (IFFT) to revert frequency-domain features back into the spatial-temporal domain.

After that, this feature is rearranged and modulated through a gating mechanism with another tensor $M_2$ to highlight informative elements. We finally apply linear projection and reshape the feature map to its original dimensions, which can be represented as follows:

$$\begin{aligned} M_r' &= M_2 \otimes \text{iFFT}(\text{DWConv}\left(\text{Reshape}(M_1)\right)), \\ M_{\text{out}} &= \text{Reshape}\left(\text{Linear}(M_r')\right) + \tilde{M}. \end{aligned} \tag{13}$$

## IV. Experiments And Results

To demonstrate the effectiveness of PulseMamba, we perform experiments with intra-dataset and cross-dataset testing, ablation studies, and computational cost on three benchmark datasets.

### A. Datasets

**UBFC-rPPG** [45] consists of 42 videos and each video lasts approximately 1 minute with a resolution of 640 × 480 pixels, captured in uncompressed 8-bit RGB format. Subjects are instructed to play a time-sensitive mathematical game while sit 1 meter in front of the camera to increase HR.

**PURE** [46] consists of videos captured by 10 subjects. Each subject is instructed to perform six types of head movements at a distance of 1.1 meters in front of the camera. These videos are recorded at 30 Hz using with each video lasting approximately 1 minute.

**COHFACE** [47] contains 160 videos and corresponding physiological signals collected from 40 healthy individuals. These videos are recorded under four different conditions involving varying illumination and motion settings. Each video was captured at a resolution of 640 × 480 pixels with a frame rate of 20 fps.

### B. Implementation Details

Our proposed method is performed under the PyTorch framework with an NVIDIA RTX 3090 GPU in AutoDL platform. The entire video and its corresponding ground truth BVP signals are clipped to 120 frames and 30 Hz, respectively. The dataset is split into training and testing sets with a ratio of 7:3. The model is trained for 30 epochs using the Adam optimizer, with an learning rate of 0.003 and a batch size of 2, using the negative Pearson correlation loss as the loss function.

### C. Evaluation Metrics

To perform a quantitative analysis of the HR estimation model proposed in this paper, four quantitative evaluation metrics are employed: Mean Absolute Error (*MAE*) [48], Root Mean Square Error (*RMSE*) [49], Pearson Correlation Coefficient (*r*) [50].

### D. Main Comparsion

*1) Intra-dataset Comparsion:* We conducted intra-dataset comparsions on three datasets and compared our results with both traditional signal-based methods (CHROM [17], POS [18], ICA [19]) and learning-based approaches (PhysNet [36], DeepPhys [34], TS-CAN [35], PhysFormer [38], and EfficientPhys [37]). The best results are highlighted in bold, and the second-best results are underlined. Traditional signal-based algorithms perform relatively well on controlled environment like UBFC-rPPG dataset. However, their performances degrade significantly in more complex scenarios (PURE and COHFACE datasets) due to their reliance on simplified assumptions.

**Performance on UBFC-rPPG Dataset.** As shown in Table I, on UBFC-rPPG dataset, the proposed PulseMamba outperforms the other methods, with MAE of 0.25 bpm and RMSE of 0.61 bpm, indicating stronger robustness in handling fine-grained spatiotemporal features. PhysFormer is the second-best model, as its Transformer-based architecture allows it to effectively capture global spatiotemporal dependencies, showing strong temporal modeling capabilities. However, its reliance on global dependencies limits its ability to capture subtle changes in rPPG signals. 2D-CNN methods like DeepPhys and TS-CAN have limitations in capturing long-term dependencies and accuracy, leading to relatively higher MAE and RMSE results.

**Performance on PURE Dataset.** The PURE dataset includes videos with significant head movements, making rPPG estimation particularly challenging. Approaches based on 2D-CNN, such as DeepPhys, processes video frames independently, failing to fully capture the temporal dependencies between frames. PhysNet and EfficientPhys struggle to capture global temporal dynamics due to their fixed receptive fields and strong reliance on local features. While PhysFormer excels in modeling long-term dependencies through global self-attention, it struggles with robustness when faced with complex backgrounds large head movements. In contrast, our model effectively improves spatiotemporal modeling, achieving an MAE of 0.34 bpm and an RMSE of 0.87 bpm, surpassing the performance of other methods.

**Performance on COHFACE Dataset.** To evaluate performance in challenging scenarios like different lighting conditions and diverse head movements, we use the most challenging dataset COHFACE. While PhysNet and DeepPhys are able to capture spatiotemporal features using CNNs, their limited receptive fields hinder their ability to effectively model complex spatiotemporal dependencies, leading to poor performance on physiological singal extraction. The proposed PulseMamba addresses this by incorporating the FSEA module, which expands the receptive field, and integrates fine-grained

TABLE I
INTRA-DATASET RESULTS ON UBFC-RPPG, PURE AND COHFACE DATASETS.

| Methods | UBFC-rPPG | | | PURE | | | COHFACE | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE↓ | RMSE↓ | r↑ | MAE↓ | RMSE↓ | r↑ | MAE↓ | RMSE↓ | r↑ |
| ***Signal-based Methods*** | | | | | | | | | |
| CHROM | 6.69 | 8.82 | 0.82 | 4.17 | 6.26 | 0.92 | 12.08 | 16.30 | 0.26 |
| POS | 4.05 | 8.05 | 0.78 | 3.14 | 10.57 | 0.95 | 13.43 | 17.05 | 0.07 |
| ICA | 5.17 | 11.76 | 0.65 | 2.59 | 4.23 | 0.94 | 8.89 | 14.55 | 0.42 |
| ***Learning-based Methods*** | | | | | | | | | |
| PhysNet | 2.95 | 3.57 | 0.77 | 2.10 | 2.60 | **0.99** | 8.63 | 9.36 | 0.54 |
| DeepPhys | 6.27 | 10.82 | 0.65 | 3.77 | 12.14 | 0.33 | 8.25 | 14.71 | 0.28 |
| TS-CAN | 1.70 | 2.72 | **0.99** | 2.48 | 9.01 | 0.92 | - | - | - |
| PhysFormer | 0.50 | 0.71 | **0.99** | 1.10 | 1.75 | **0.99** | 2.49 | 7.79 | 0.83 |
| EfficientPhys | 1.14 | 1.81 | **0.99** | 1.33 | 5.99 | 0.97 | - | - | - |
| **Ours** | **0.25** | **0.61** | **0.99** | **0.34** | **0.87** | **0.99** | **1.03** | **1.87** | **0.99** |

temporal differences, allowing the network to better extract rPPG signals from compressed videos. Our approach achieves the best performance with an MAE of 0.89 bpm and an RMSE of 1.03 bpm.

*2) Cross-dataset Comparsion:* We also perform cross-dataset validation to demonstrate the generalization ability of the proposed method. Specifically, the model trained on one dataset is directly validated on another dataset. Table II presents the comparison results on UBFC-rPPG and PURE datasets with other existing methods. It is clear that the proposed PulseMamba achieves the best performance across all metrics, demonstrating its strong generalization ability in diverse scenarios.

This can be attributed to the integration of the RSCMamba module for global spatiotemporal modeling, along with the FSEA and FRN modules, which enhance multi-scale feature extraction and frequency-domain representations. These components enable the model to effectively capture rPPG signals across different datasets, demonstrating its strong generalization ability.

TABLE II
CROSS-DATASET RESULTS ON UBFC-RPPG AND PURE DATASETS.

| Methods | UBFC-rPPG → PURE | | | PURE → UBFC-rPPG | | |
|---|---|---|---|---|---|---|
| | MAE↓ | RMSE↓ | r↓ | MAE↑ | RMSE↓ | r↑ |
| PhysNet | 8.06 | 19.71 | 0.61 | 0.98 | 2.48 | **0.99** |
| DeepPhys | 5.54 | 18.51 | 0.66 | 1.21 | 2.90 | **0.99** |
| TS-CAN | 3.69 | 13.80 | 0.82 | 1.30 | 2.87 | **0.99** |
| PhysFormer | 1.99 | 3.28 | **0.99** | 1.44 | 3.77 | 0.98 |
| EfficientPhys | 5.47 | 17.04 | 0.71 | 2.07 | 6.32 | 0.94 |
| **Ours** | **1.12** | **2.39** | **0.99** | **0.57** | **1.40** | **0.99** |

### *E. Testing Under Specific Conditions*

To evaluate the robustness of PulseMamba against variations in illumination and head movements, we performed experiments using videos from the COHFACE dataset (with different lighting conditions) and the PURE dataset (with different types of head movements). In addition, we generate saliency maps for PulseMamba under different head movements and lighting scenarios.

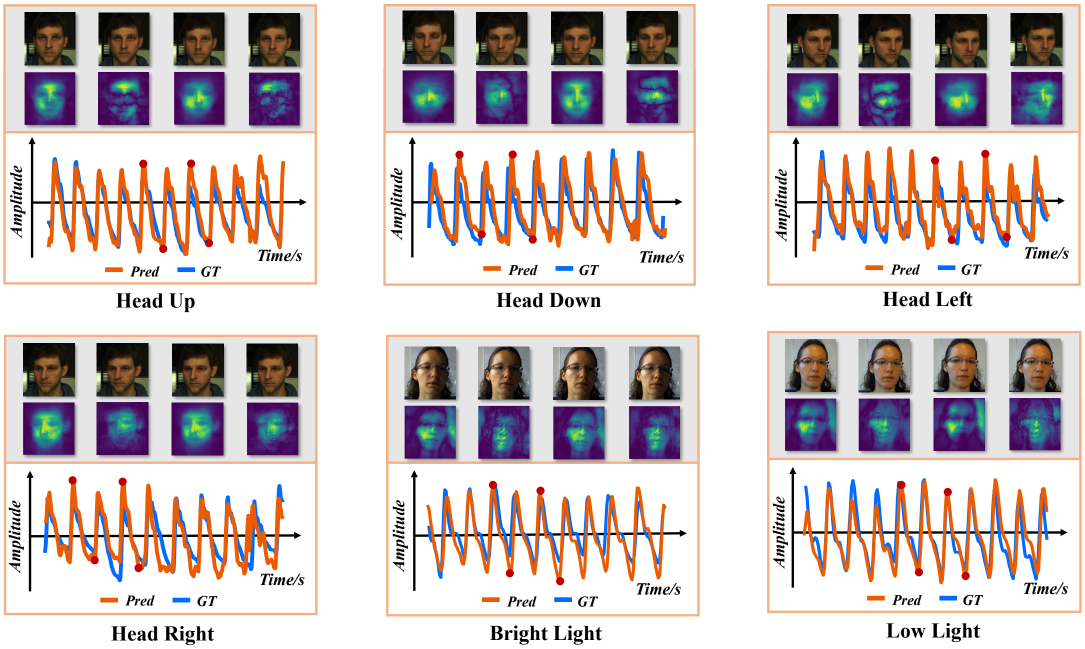
The results are shown in Fig. 4, we can see that the predicted rPPG (*Pred*) signal aligns with the ground-truth (*GT*) signal well in challenging scenairos. For each video sample, we project the feature maps into the spatial dimension and obtain a sequence of feature maps. The feature maps of the two examples have a consistent visualization law, namely weak visual responses at peaks of rPPG signals and strong visual responses at troughs. Despite the changes in lighting and head movements, PulseMamba can still focus on spatial information, particularly in the facial skin regions. Notably, even under highly compressed and poor light conditions in the COHFACE dataset, PulseMamba is still able to capture critical facial regions and extract reliable rPPG signals, demonstrating strong robustness against noise and complex real-world environments.

### *F. Ablation Studies*

We conducted an ablation study on the UBFC-rPPG and PURE datasets to analyze the impact of each proposed method's component on experimental performance. The results are presented in Table III.

*1) Impact of RSCMamba:* As shown in Table III, the results reveal that incorporating the RSCMamba module yields a significant improvement across all metrics, most notably on the PURE dataset, where the MAE decrased from 1.58 to 0.81 bpm. Traditional CNNs, constrained by their local receptive fields, struggle to capture the subtle pulse signal within the video sequence. While Transformers possess global modeling capabilities, their self-attention mechanism lacks an inductive bias for spatiotemporal structure. The RSCMamba module leverages the Mamba architecture's linear complexity advantage, efficiently captures the dynamic changes of facial pixels over long temporal sequences. This allows the model to fundamentally distinguish the periodic rPPG signal caused by blood volume changes from environmental noise such as head movements and facial expressions, thereby enhancing the ability for rPPG signal extraction and analysis.

*2) Impact of FSEA:* In addition, when the FSEA module was removed, the model's ability to capture the periodic characteristic of physiological signals significantly weakened, leading to a noticeable performance decline.



Head Up

Head Down

Head Left

Head Right

Bright Light

Low Light

Fig. 4. Visualization of HR prediction results in challenging scenarios. We show the predicted rPPG signals (Pred) and corresponding ground-truth PPG signals (GT) in orange and blue curves, respectively. Some representative video frames are marked in the waveform with dots and placed on the top of the signal diagram.

After integrating the FSEA module, the model demonstrates significantly improved accuracy, especially in UBFC-rPPG dataset, with the MAE dropping from 1.17 to 0.42 bpm. The FSEA module extracts features from multiple spatial scales, enabling the model to focus on both large facial areas—such as the forehead and cheeks—and smaller skin patches with stronger signal quality. More importantly, its dynamic contextual fusion mechanism adaptively weighs features across different regions and scales, significantly enhancing the model's ability to extract subtle rPPG signals under complex visual conditions.

*3) Impact of FRN:* After adding the FRN module, the MAE on the UBFC dataset drops from 1.17 to 0.92 bpm, and on the PURE dataset from 0.81 to 0.50 bpm. FRN enhances the frequency-domain representation of rPPG signals by amplifying the dominant frequency peaks and suppressing noise components that are hard to remove in the time domain. This design specifically strengthens the periodicity of the signal, helping the model focus more on physiologically plausible HR ranges.

TABLE III
ABLATION RESULTS OF THE PULSEMAMBA MODEL ON UBFC-RPPG DATASET.

| RSCMamba | FSEA | FRN | UBFC-rPPG | | | PURE | | |
|---|---|---|---|---|---|---|---|---|
| | | | MAE↓ | RMSE↓ | r↑ | MAE↓ | RMSE↓ | r↑ |
| - | - | - | 1.29 | 2.75 | 0.93 | 1.58 | 3.26 | 0.87 |
| ✓ | | | 1.17 | 2.50 | 0.98 | 0.81 | 1.69 | 0.91 |
| ✓ | ✓ | | 0.42 | 1.02 | 0.99 | 0.46 | 1.09 | 0.95 |
| ✓ | | ✓ | 0.92 | 1.92 | 0.99 | 0.50 | 1.22 | 0.98 |
| ✓ | ✓ | ✓ | **0.25** | **0.61** | **0.99** | **0.34** | **0.87** | **0.99** |

## G. *Computaional Cost*

Table IV presents the comparison of our proposed PulseMamba with other approaches in terms of memory footprint (Memory), multiply-accumulate operations (MACs), and the number of parameters (Para). The results demonstrate that our proposed method achieves the lowest parameter count, smallest model size, and minimal memory footprint, significantly reducing computational cost, making it particularly promising for real-world rPPG applications.

TABLE IV
COMPARISON OF COMPUTATIONAL COST.

| Methods | MACs ($10^9$) | Para ($10^6$) | Memory (MB) |
|---|---|---|---|
| PhysNet | 438.24 | 0.77 | 11.43 |
| DeepPhys | 744.45 | 1.98 | 37.28 |
| TS-CAN | 744.45 | 1.98 | 38.91 |
| PhysFormer | 316.29 | 7.38 | 28.63 |
| EfficientPhys | 373.72 | 1.91 | 26.68 |
| Ours | **146.14** | **0.37** | **11.07** |

# V. CONCLUSION

In this study, we propose PulseMamba, a novel Mamba-based model for remote physiological prediction and HR estimation. Our method contains three key modules: the RSC-Mamba module for global spatiotemporal modeling over long sequences, the FSEA module for multi-scale feature extraction and dynamic contexual fusion, and the FRN module to model periodicity of rPPG signals. Extensive experiments on UBFC-rPPG, PURE, and COHFACE datasets have demonstrated that the proposed approach outperforms the other methods with lower MAE and RMSE, smaller model size and less Parameters. This highlights the effectiveness of PulseMamba in addressing the limitations of existing CNN- and Transformer-based models, particularly their challenges in modeling fine-grained physiological signals and maintaining efficiency in long-sequence processing.

Despite these advantages, the proposed method still face some limitations. First, although PulseMamba generalizes well on the three datasets, its reliance on supervised learning may still restrict its adaptability to unconstrained environments. Second, while our method can effectively capture spatial and frequency-domain information, it currently does not leverage multi-modal inputs such as NIR or depth data, which may contribute to better performance in complex real-world scenarios with variable lighting and motion artifacts. Third, although PulseMamba demonstrates computational efficiency in terms of parameters and MACs, optimizing its real-time processing capabilities on edge devices remains crucial for broader adoption in resource-constrained environments.

In future work, we plan to extend our research beyond HR estimation to other rPPG-based physiological measurements, such as respiration rate or blood pressure monitoring. Additionally, we seek to incorporate multi-modal data sources and investigate lightweight variants of PulseMamba to support real-time deployment in mobile and wearable devices.

## References

[1] Y. Benezeth, P. Li, R. Macwan, K. Nakamura, R. Gomez, and F. Yang, "Remote heart rate variability for emotional state monitoring," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2018, pp. 153–156.

[2] D. Li, X. Chen, Q. Li, F. Zhu, X. Lu, S. Routray, U. Ghosh, and M. Al-Numay, "Intelligent biomedical photoplethysmography signal cycle division with digital twin in metaverse for consumer health," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2116–2128, 2024.

[3] J. Chen, S. Sun, L.-b. Zhang, B. Yang, and W. Wang, "Compressed sensing framework for heart sound acquisition in internet of medical things," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 2000–2009, 2022.

[4] blueYu, Zhaocheng and Chen, Junxin and Liu, Yu and Chen, Yongyong and Wang, Tingting and Nowak, Robert and Lv, Zhihan, "Ddcnn: A deep learning model for af detection from a single-lead short ecg signal," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 4987–4995, 2022black.

[5] blueZhang, Zhilin, "Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 8, pp. 1902–1910, 2015black.

[6] blueZhang, Qingxue and Zeng, Xuan and Hu, Wenchuang and Zhou, Dian, "A machine learning-empowered system for long-term motion-tolerant wearable monitoring of blood pressure and heart rate with ear-ecg/ppg," *IEEE Access*, vol. 5, pp. 10 547–10 561, 2017black.

[7] S. Puranik and A. W. Morales, "Heart rate estimation of ppg signals with simultaneous accelerometry using adaptive neural network filtering," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 1, pp. 69–76, 2020.

[8] D. McDuff, S. Gontarek, and R. W. Picard, "Improvements in remote cardiopulmonary measurement using a five band digital camera," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2593–2601, 2014.

[9] Y. Zhang, W. Liang, X. Yuan, S. Zhang, G. Yang, and Z. Zeng, "Deep learning-based abnormal behavior detection for elderly healthcare using consumer network cameras," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2414–2422, 2024.

[10] C. Savur, R. Dautov, K. Bukum, X. Xia, J.-P. Couderc, and G. R. Tsouri, "Monitoring pulse rate in the background using front facing cameras of mobile devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 5, pp. 2208–2218, 2023.

[11] Z. Yue, S. Ding, S. Yang, H. Yang, Z. Li, Y. Zhang, and Y. Li, "Deep super-resolution network for rppg information recovery and noncontact heart rate estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[12] S.-N. Yu, C.-S. Wang, and Y. P. Chang, "Heart rate estimation from remote photoplethysmography based on light-weight u-net and attention modules," *IEEE Access*, vol. 11, pp. 54 058–54 069, 2023.

[13] M. Hu, D. Guo, M. Jiang, F. Qian, X. Wang, and F. Ren, "rppg-based heart rate estimation using spatial-temporal attention network," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1630–1641, 2022.

[14] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, and X. Chen, "Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7411–7421, 2020.

[15] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[16] J. He and N. Jiang, "Fast multilevel mental stress identification from bispectrum-based heart rate variability feature," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 2, pp. 1124–1133, 2024.

[17] 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.

[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, 2017.

[19] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in non-contact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[20] de G Gerard Haan and van Aj Arno Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, pp. 1913 – 1926, 2014.

[21] Y. Zhang, W. Liang, X. Yuan, S. Zhang, G. Yang, and Z. Zeng, "Deep learning-based abnormal behavior detection for elderly healthcare using consumer network cameras," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2414–2422, 2024.

[22] Z. Guo, J. Chen, T. He, W. Wang, H. Abbas, and Z. Lv, "Ds-cnn: Dual-stream convolutional neural networks-based heart sound classification for wearable devices," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 4, pp. 1186–1194, 2023.

[23] J. Dai, J. Wang, W. Huang, J. Shi, and Z. Zhu, "Machinery health monitoring based on unsupervised feature learning via generative adversarial networks," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 5, pp. 2252–2263, 2020.

[24] C. Zhao, P. Cao, M. Hu, B. Huang, H. Chen, and J. Li, "Wtc3d: An efficient neural network for noncontact pulse acquisition in internet of medical things," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 2, pp. 1547–1556, 2025.

[25] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[26] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[27] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," in *Forty-first International Conference on Machine Learning*.

[28] S.-Q. Liu and P. C. Yuen, "Robust remote photoplethysmography estimation with environmental noise disentanglement," *IEEE Transactions on Image Processing*, vol. 33, pp. 27–41, 2024.

[29] Q. Li, D. Guo, W. Qian, X. Tian, X. Sun, H. Zhao, and M. Wang, "Channel-wise interactive learning for remote heart rate estimation from facial video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 4542–4555, 2024.

[30] S. Zhao, F. Wang, X. Huang, X. Yang, N. Jiang, J. Peng, and Y. Ban, "Mamba-unet: Dual-branch mamba fusion u-net with multiscale spatio-temporal attention for precipitation nowcasting," *IEEE Transactions on Industrial Informatics*, pp. 1–10, 2025.

[31] H.-Y. Chang, Y.-Y. Chen, and W.-H. Chung, "Two-stage zoom fft-enhanced deep learning-aided weighted scheme for wireless vital sign

estimation using mmwave fmcw radar," *IEEE Sensors Letters*, vol. 8, no. 7, pp. 1–4, 2024.

[32] Y. Wei, J. Liu, L. Hu, B. W.-K. Ling, and Q. Liu, "Time frequency analysis-based averaging and fusion of features for wearable non-invasive blood glucose estimation," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 510–521, 2023.

[33] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam — a non-contact method for evaluating cardiac activity," in *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2011, pp. 405–410.

[34] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 349–365.

[35] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 19 400–19 411.

[36] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *Procedings of the British Machine Vision Conference*, 2019, pp. 3–6.

[37] 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. 4997–5006.

[38] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, 2019.

[39] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *The International Conference on Learning Representations (ICLR)*, 2022.

[40] 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, 2024.

[41] Z. Yu, B. Zhou, J. Wan, P. Wang, H. Chen, X. Liu, S. Z. Li, and G. Zhao, "Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition," *IEEE Transactions on Image Processing (TIP)*, 2021.

[42] J. X. B. P. Z. C. Z. L. J. W. K. L. T. L. L. W. Guo Chen, Yifei Huang, "Video mamba suite: State space model as a versatile alternative for video understanding," 2024.

[43] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," *arXiv preprint arXiv:2303.03667*, 2023.

[44] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv e-prints*, p. arXiv:1704.04861, Apr. 2017.

[45] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, 2019, award Winning Papers from the 23rd International Conference on Pattern Recognition (ICPR).

[46] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 1056–1062.

[47] G. Heusch, A. Anjos, and S. Marcel, "A Reproducible Study on Remote Heart Rate Measurement," *arXiv e-prints*, p. arXiv:1709.00962, Sep. 2017.

[48] C. Zhao, H. Wang, H. Chen, W. Shi, and Y. Feng, "Jamsnet: A remote pulse extraction network based on joint attention and multi-scale fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2783–2797, 2023.

[49] C. Zhao, M. Zhou, W. Han, and Y. Feng, "Anti-motion remote measurement of heart rate based on region proposal generation and multi-scale roi fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[50] Y. Sun and N. Thakor, "Photoplethysmography revisited: From contact to noncontact, from point to imaging," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 463–477, 2016.