RhythmMamba: Fast, Lightweight, and Accurate Remote Physiological Measurement

Bochao Zou[1], Zizheng Guo[1], Xiaocheng Hu[2], Huimin Ma[1]*

[1]University of Science and Technology Beijing, Beijing, China
[2]China Academy of Electronics and Information Technology, Beijing, China
zoubochao@ustb.edu.cn, guozizheng@xs.ustb.edu.cn, 675342900@qq.com, mhmpub@ustb.edu.cn

Abstract

Remote photoplethysmography (rPPG) is a method for non-contact measurement of physiological signals from facial videos, holding great potential in various applications such as healthcare, affective computing, and anti-spoofing. Existing deep learning methods struggle to address two core issues of rPPG simultaneously: understanding the periodic pattern of rPPG among long contexts and addressing large spatiotemporal redundancy in video segments. These represent a trade-off between computational complexity and the ability to capture long-range dependencies. In this paper, we introduce RhythmMamba, a state space model-based method that captures long-range dependencies while maintaining linear complexity. By viewing rPPG as a time series task through the proposed frame stem, the periodic variations in pulse waves are modeled as state transitions. Additionally, we design multi-temporal constraint and frequency domain feed-forward, both aligned with the characteristics of rPPG time series, to improve the learning capacity of Mamba for rPPG signals. Extensive experiments show that RhythmMamba achieves state-of-the-art performance with 319% throughput and 23% peak GPU memory.

Code — https://github.com/zizheng-guo/RhythmMamba

1 Introduction

Blood Volume Pulse (BVP) is a vital physiological signal, further enabling the extraction of key signs such as heart rate (HR) and heart rate variability (HRV). Photoplethysmography (PPG) is a non-invasive monitoring method that utilizes optical means to measure changes in blood volume within living tissues. The physiological mechanism of PPG stems from variations in blood volume during cardiac contraction and relaxation in subcutaneous blood vessels, leading to changes in light absorption and scattering. These changes result in periodic color signal variations on imaging sensors, which are imperceptible to the human eye (Verkruysse, Svaasand, and Nelson 2008; Chen and McDuff 2018). Traditionally, BVP extraction requires the use of contact sensors, which brings inconvenience and limitations. In recent years, non-contact methods for obtaining BVP, particularly rPPG, have garnered increasing attention (McDuff 2023; Li, Yu, and Shi 2023; Choi, Kang, and Kim 2024).

Early rPPG research primarily relied on traditional signal processing methods to recover weak rPPG signals from facial videos, which are susceptible to interference from environmental light, motion, and other noises. In complex environments, relying solely on signal processing methods often struggles to achieve satisfactory accuracy. In recent years, data-driven methods have become mainstream, represented by convolutional neural networks (CNNs) and transformers. However, CNNs have limited receptive fields and transformer-based architectures exhibit mediocre performance in capturing long-term dependencies from the computational complexity perspective, especially when dealing with long video sequences.

Recently, Mamba (Dao and Gu 2024) has emerged with its selective state space model, striking a balance between maintaining linear complexity and facilitating long-term dependency modeling. It has been successfully applied to various artificial intelligence tasks such as video understanding (Li et al. 2025). For rPPG tasks that typically require long-term monitoring and are suitable for deployment on mobile devices, the linear complexity and ability to capture long-term dependencies give Mamba an advantage.

However, the direct application of Mamba to rPPG tasks performs poorly. Our experiments reveal that embedding spatiotemporal information into token sequences through patch embedding leads to spatial information significantly disrupting Mamba's comprehension of temporal information (see Section 4.5). This phenomenon may stem from Mamba's linear modeling characteristics, where the states are associated with the temporal phases of the rPPG signals. The incorporation of spatial information increases the dimensionality of the state transition process, thereby adding complexity and impeding the model's learning efficacy.

Motivated by the aforementioned discussion, we propose RhythmMamba, a state space model-based architecture for remote physiological measurement. The proposed frame stem embeds spatial information from a single frame into the channels, thereby viewing the rPPG task as a time series task, allowing the periodic variations in pulse waves to be modeled as state transitions. As shown in Fig. 1, considering the linear modeling characteristics of Mamba, the phase shifts of rPPG signals can be viewed as state tran-

*Corresponding author.

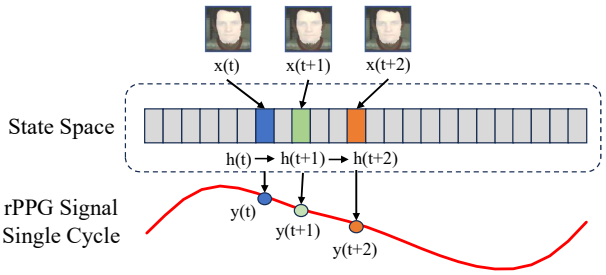

Figure 1: A schematic diagram of state transitions. Considering the periodic nature of rPPG, the rPPG signal can be represented using a finite number of states. Where $h(t)$ represents the state vector, $x(t)$ represents the input vector, and $y(t)$ represents the output vector.

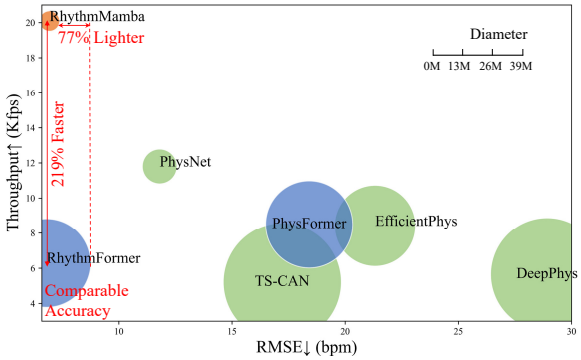

Figure 2: Performance and efficiency evaluation for intra-dataset testing on MMPD. The diameter of the circle indicates the peak GPU memory. The proposed RhythmMamba is faster, lighter, and achieves comparable accuracy, with these advantages becoming more pronounced as the scale increases due to its linear complexity.

sitions within the state space. The periodic nature of rPPG allows the signal to be represented using a finite set of states.

Additionally, we design multi-temporal constraint and frequency domain feed-forward, both aligned with the characteristics of rPPG time series, to improve the learning capacity of Mamba for rPPG signals. By learning from the same sequences with varying temporal lengths, a single Mamba block can simultaneously be constrained by the periodicity of long sequences and the trends of short sequences. Subsequently, through frequency domain feed-forward, the learned temporal features by Mamba undergo inter-channel spatial interaction in the frequency domain, enabling a better discernment of the periodic nature of rPPG signals.

The main contributions are as follows:

• We propose RhythmMamba, which leverages state space models to model periodic variations as state transitions, combining multi-temporal constraints Mamba and frequency domain feed-forward to learn the quasi-periodic patterns of rPPG. To the best of our knowledge, this is the first work to investigate state space models in the rPPG domain.

• In response to the observed phenomenon where spatial information interferes with Mamba's understanding of temporal sequences, we design the frame stem to embed spatial information into channels, reducing the dimensionality of state transitions to boost Mamba's learning.

• We conduct extensive experiments on intra-dataset and cross-dataset scenarios. The results demonstrate that RhythmMamba achieves state-of-the-art performance with 319% throughput and 23% GPU memory, as illustrated in Fig. 2.

2 Related Work

2.1 Remote Physiological Measurement

Early research on rPPG primarily relied on traditional signal processing methods to recover weak rPPG signals from facial videos (Verkruysse, Svaasand, and Nelson 2008; Poh, McDuff, and Picard 2010; De Haan and Jeanne 2013; Wang et al. 2016). In recent years, data-driven approaches have dominated due to their remarkable performance, showcasing a trend in the transition of backbone from 2D CNNs (Špetlík, Franc, and Matas 2018; Niu et al. 2018; Chen and McDuff 2018; Niu et al. 2020; Liu et al. 2020) to 3D CNNs (Yu, Li, and Zhao 2019; Yu et al. 2019; Zhao et al. 2021; Li, Yu, and Shi 2023) and further to transformers (Yu et al. 2022, 2023; Liu et al. 2023a; Shao et al. 2023; Liu et al. 2024; Zou et al. 2024). However, none of them have been able to effectively address the two core issues of rPPG: understanding the periodic pattern of rPPG among long contexts and addressing large spatiotemporal redundancy in video segments. This dilemma underscores a trade-off between computational complexity and the ability to capture long-range dependencies, thereby presenting a barrier to deploying rPPG solutions on mobile devices. Although previously dominant 3D CNNs and video transformers have effectively tackled one of the above issues by utilizing local convolutions or long-range attention, they fail to address both problems simultaneously. Unlike them, the proposed RhythmMamba can capture long-range dependencies while maintaining linear complexity, making it fast, lightweight, and accurate for remote physiological measurement.

2.2 Vision Mamba

Recently, Mamba has distinguished itself with a data-dependent state space model (SSM) and a selection mechanism utilizing parallel scanning, striking a balance between maintaining linear complexity and facilitating long-term dependency modeling. Compared to transformers with quadratic complexity attention (Vaswani et al. 2017; Arnab et al. 2021), Mamba excels at handling long sequences with linear complexity. Subsequently, the immense potential of Mamba has sparked a series of works (Zhu et al. 2024; Patro and Agneeswaran 2024; Li et al. 2025), demonstrating superior performance and higher GPU efficiency of Mamba over Transformers on downstream vision tasks. However, unlike other video tasks, rPPG signals are particularly weak and highly susceptible to noise from factors such as lighting and motion, making the direct application of the traditional Mamba architecture to rPPG tasks perform poorly. In contrast to previous works, we view the rPPG task as the time series task, fully integrating spatial information into the channels and designing multiple modules tailored for time series to boost Mamba's learning.

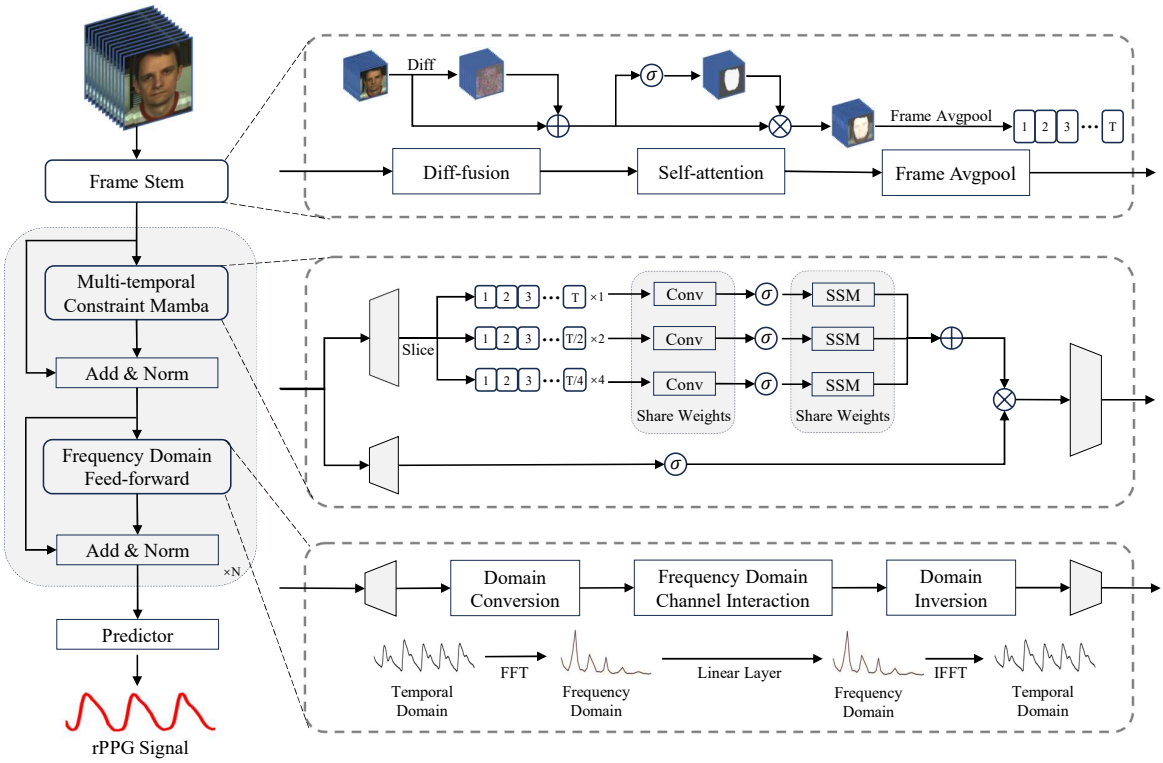

Figure 3: The framework of RhythmMamba. It consists of frame stem, multi-temporal constraint Mamba, frequency domain feed-forward, and rPPG predictor head. Where "+" represents addition, "×" represents multiplication, "σ" represents the activation layer, and trapezoid represents the linear layer.

3 Methodology

Section 3.1 introduces the general framework of RhythmMamba, followed by the presentation of its main components: the frame stem in Section 3.2, the multi-temporal constraint Mamba in Section 3.3, and lastly, the frequency domain feed-forward in Section 3.4.

3.1 The General Framework of RhythmMamba

As shown in Figure 3, RhythmMamba consists of frame stem, multi-temporal constraint Mamba, frequency domain feed-forward, and rPPG predictor head. The frame stem utilizes diff-fusion, self-attention, and frame average pooling to extract rPPG features and embed all spatial information into channels. Specifically, given an RGB video input $X \in \mathbb{R}^{3\times T\times H\times W}$, $X_{stem} = frame\_stem(X)$, where $X_{stem} \in \mathbb{R}^{T\times C}$, and C, T, W, H indicate channel, sequence length, width, and height, respectively.

Subsequently, the output of the frame stem will be fed into the multi-temporal constraint Mamba. The tokens will be sliced into sequences of varying temporal lengths, followed by the processing of hidden information between tokens with the SSM. Then, the output of Mamba will be passed into the frequency domain feed-forward, facilitating the interaction of information across multiple channels. The outputs of Mamba and feed-forward network (FFN) will undergo normalization and residual connections. The two outputs have dimensions identical to the output of the frame stem $X_{stem} \in \mathbb{R}^{T\times C}$. Finally, the rPPG features will be projected into PPG waves through the predictor head.

3.2 Frame Stem

In the field of video understanding, existing transformer-based and Mamba-based methods typically embed spatiotemporal information into token sequences through patch embedding (Arnab et al. 2021; Zhu et al. 2024; Dosovitskiy et al. 2020). Previous works on rPPG have also been based on such foundational models for improvements. For transformer-based methods, spatiotemporal token sequences can inspire long-range spatiotemporal attention both within frames and across frames. However, we found that for linear Mamba, spatial information may interfere with Mamba's understanding of temporal sequences [see section 4.5]. The frame stem is utilized to initially extract rPPG features and embed spatial information fully into the channels, thereby boosting the learning of state transitions in the multi-temporal constraint Mamba and the channel interactions in the frequency domain feed-forward.

Firstly, the diff-fusion module (Zou et al. 2024) integrates frame differences into the raw frames, enabling frame-level representation awareness of BVP wave variations. This effectively enhances the features of rPPG with a small additional computational cost. Additionally, for rPPG, high-frequency information across frames and low-frequency information within frames are required. Therefore, relatively

large convolutional kernels are used to obtain low-frequency information within frames, ensuring that spatial information is fully incorporated into the channels. Here, 'relatively large' refers to the size relative to the image resolution, enabling a large receptive field.

Specifically, for an RGB video input $X \in \mathbb{R}^{3\times T\times H\times W}$, temporal shift is initially applied to obtain X_{t-2}, X_{t-1}, X_t, X_{t+1} and X_{t+2}. Subsequently, frame differences between consecutive frames are computed in reverse chronological order, yielding D_{t-2}, D_{t-1}, D_{t+1}, and D_{t+2}. The frame differences and the raw frames are then passed through $Stem_1$ for feature extraction. $Stem_1$ consists of a 2D convolution layer with (7×7) kernel, followed by batch normalization (BN), ReLU, and MaxPool. The input dimension is 3 when taking raw frames as input, and 12 when taking the concatenation of frame differences as input.

$$\begin{aligned} X_{diff} &= Stem_1(Concat\,(D_{t-2}, D_{t-1}, D_{t+1}, D_{t+2})), \\ X_{raw} &= Stem_1(X_t). \end{aligned} \tag{1}$$

Then frame differences X_{diff} and raw frames X_{raw} are merged and the feature representation is further enhanced through $Stem_2$, which consists of a 2D convolution layer with (7×7) kernel, followed by BN, ReLU and MaxPool.

$$X_{fusion} = Stem_2(X_{raw} + X_{diff}) + Stem_2(X_{diff}). \tag{2}$$

Subsequently, at the resolution of (16×16), $Stem_3$ utilizes a convolution layer with (5×5) kernel to fully integrate spatial information into the channels, followed by BN. Before frame-level global average pooling, a self-attention module is employed to enhance skin regions with rPPG signals in the spatial domain. This self-attention module utilizes sigmoid activation followed by L1 normalization, which is softer than softmax and generates fewer masks (Liu et al. 2023a). The attention mask can be computed as:

$$Mask = \frac{(H/8)(W/8) \cdot \sigma(Stem_3(X_{fusion}))}{2||\sigma(Stem_3(X_{fusion}))||_1}. \tag{3}$$

Finally, the attention output $X_{attn} \in \mathbb{R}^{C\times T\times H/8\times W/8}$, undergoes global average pooling within each frame, resulting in the stem output $X_{stem} \in \mathbb{R}^{T\times C}$.

3.3 Multi-temporal Constraint Mamba

Previous studies (Hu et al. 2022; Kong, Bian, and Jiang 2022; Dai et al. 2022) have shown the effectiveness of modeling periodic tasks with multi-temporal scales, primarily achieved by the extraction and fusion of features from different temporal scales. Unlike these studies, we replace multi-temporal fusion with multi-temporal constraint, which better aligns with the characteristics of the Mamba. Specifically, we slice a video segment into numerous sub-segments of varying lengths to constrain a single Mamba block, rather than downsampling the video segment to different resolutions and using multiple Mamba blocks to extract and fuse features. The aim is to subject a Mamba block to both the periodic constraints of long sequences and the trend constraints of short sequences simultaneously, instead of extracting different features from multi-temporal scales.

After the frame stem, the token sequence can be regarded as a time series, and the state transition of Mamba can be interpreted as the temporal phase shift. Due to the quasi-periodicity of rPPG signals, the signal can be represented using a finite set of states. Specifically, as illustrated by the multi-temporal constraint Mamba in Figure 3, the input X_{stem} is first linearly projected and then processed through three weight-shared paths. Along these paths, the input is sliced into temporal sequences of varying lengths. For the i_{th} path, the sequence is divided into 2^{i-1} sub-sequences, each of which undergoes sequential processing through a convolution layer, activation layer, and selective state space model (see supplementary material A for details). Subsequently, they are recombined into a sequence of the original length, forming the output of the i_{th} path, denoted as X_{path_i}. The output before projection can be represented as follows:

$$X_{mamba} = \sum_{i=1}^{3} X_{path_i} \times \sigma(Proj(X_{stem})). \tag{4}$$

3.4 Frequency Domain Feed-forward

The FFN employs linear transformations to project data into a higher-dimensional space before mapping it back into a lower-dimensional space. Through this channel interaction, deeper features are extracted. In previous rPPG studies, spatio-temporal FFN was frequently applied, which introduced a depthwise 3D convolution layer between the two linear layers of the vanilla FFN, to refine the local inconsistency and provide relative positional cues. (Yu et al. 2022).

In our study, due to the input sequences being solely time-dependent, channel interaction in the frequency domain enables a better discernment of the periodic nature of rPPG signals. So we introduce frequency domain feed-forward, which adds a frequency domain linear layer between the two linear layers of the vanilla FFN. The frequency domain linear layer consists of three stages: domain conversion, frequency domain channel interaction, and domain inversion.

Domain Conversion/Inversion. Domain conversion and inversion utilize fast Fourier transform and inverse Fourier transform, respectively. The utilization of the Fourier transform enables the decomposition of rPPG signals into their constituent frequencies, facilitating the recognition of periodic patterns in the rPPG signals. We transform the input $H(t)$ to the frequency domain $H(f)$ as follows:

$$\begin{aligned} H(f) &= \int_{-\infty}^{\infty} H(t) e^{-j2\pi ft} dt \\ &= \int_{-\infty}^{\infty} H(t)\cos(2\pi ft)dt + j\int_{-\infty}^{\infty} H(t)\sin(2\pi ft)dt \\ &= H(f)_{re} + jH(f)_{im}. \end{aligned} \tag{5}$$

Where f represents frequency, t represents time, and the subscripts re and im denote the real and imaginary components of the corresponding complex data, respectively. After channel Interaction in the frequency domain, inverse Fourier transform is employed to revert to the temporal domain:

$$\begin{aligned} H(t) &= \int_{-\infty}^{\infty} H(f) e^{j2\pi ft} df \\ &= \int_{-\infty}^{\infty} (H(f)_{re} + jH(f)_{im}) e^{j2\pi ft} df. \end{aligned} \tag{6}$$

Frequency Domain Channel Interaction. Through the frame stem, spatial information is embedded into the channels, each of which is treated as a time series. Consequently,

Method	PURE			UBFC			VIPL-HR			MMPD		
	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑
HR-CNN	1.84	2.37	0.98	4.90	5.89	0.64	-	-	-	-	-	-
DeepPhys	0.83	1.54	0.99	6.27	10.82	0.65	11.00	13.80	0.11	22.27	28.92	-0.03
PhysNet	2.10	2.60	0.99	2.95	3.67	0.97	10.80	14.80	0.20	4.80	11.80	0.60
TS-CAN	2.48	9.01	0.92	1.70	2.72	0.99	-	-	-	9.71	17.22	0.44
Gideon et al.	2.30	2.90	0.99	1.85	4.28	0.93	9.01	14.02	0.58	-	-	-
Dual-GAN	0.82	1.31	0.99	0.44	0.67	0.99	4.93	7.68	0.81	-	-	-
PhysFormer	1.10	1.75	0.99	0.50	0.71	0.99	4.97	7.79	0.78	11.99	18.41	0.18
EfficientPhys	-	-	-	1.14	1.81	0.99	-	-	-	13.47	21.32	0.21
TFA-PFE	1.44	2.50	-	0.76	1.62	-	-	-	-	-	-	-
NEST	-	-	-	-	-	-	4.76	7.51	**0.84**	-	-	-
Li et al.	0.64	1.16	0.99	0.48	**0.64**	0.99	5.19	8.26	0.78	-	-	-
Yue et al.	1.23	2.01	0.99	0.58	0.94	0.99	-	-	-	-	-	-
PhysFormer++	-	-	-	-	-	-	4.88	7.62	0.80	-	-	-
Contrast-Phys+	0.48	0.98	0.99	**0.21**	0.80	0.99	-	-	-	-	-	-
RhythmFormer	0.27	0.47	0.99	0.50	0.78	0.99	-	-	-	**3.07**	**6.81**	**0.86**
Ours	**0.23**	**0.34**	0.99	0.50	0.75	0.99	**4.30**	**7.49**	0.81	3.16	7.27	0.84

Table 1: Intra-dataset evaluation. Best results are marked in **bold** and second best in underline.

the frequency domain features obtained after domain conversion can clearly represent the signal's frequency composition. This allows channel interactions in the frequency domain to refine noise interference and more easily focus on critical channels. Specifically, channel interaction is implemented through a linear layer, the theoretical feasibility of which has been demonstrated by (Yi et al. 2024). For complex input $H \in \mathbb{R}^{T\times C}$, given complex weight matrix $W \in \mathbb{R}^{C\times C}$ and complex bias $B \in \mathbb{R}^{C}$, according to the rules of complex multiplication, it can be expressed as:

$$\begin{aligned} H'_{re} &= H_{re}W_{re} - H_{im}W_{im} + B_{re}, \\ H'_{im} &= H_{re}W_{im} + H_{im}W_{re} + B_{im}, \\ H' &= H'_{re} + j \cdot H'_{im}. \end{aligned} \tag{7}$$

After inverse FFT transform, the frequency domain feedforward outputs through linear projection, resulting in $X_{FFN} \in \mathbb{R}^{T\times C}$.

4 Experiment

4.1 Dataset and Performance Metric

The experiments of remote physiological measurement were conducted on four publicly available datasets: PURE (Stricker, Müller, and Gross 2014), UBFC-rPPG (Bobbia et al. 2019), VIPL-HR(Niu et al. 2019), and MMPD (Tang et al. 2023). **PURE** comprises 59 1-minute videos, documenting records of 10 subjects, each engaging in six different activities. **UBFC-rPPG** consists of 42 videos, recording 42 subjects. These videos were derived from a setting where subjects participated in a time-limited digital game. **VIPL-HR** includes 2,378 RGB videos from 107 participants, captured using three RGB cameras, with an unstable fps. **MMPD** includes 660 1-minute videos, documenting records of 33 subjects. Participants engaged in four different activities under four distinct lighting conditions. **Metircs.** The evaluation was conducted using five metrics for video-level heart rate estimations: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Pearson Correlation Coefficient (ρ), and Signal-to-Noise Ratio (SNR).

4.2 Implementation Details

The proposed RhythmMamba was implemented based on PyTorch, and we utilized an open-source rPPG toolbox (Liu et al. 2023b) to conduct a fair comparison against several state-of-the-art methods. In the pre-processing, video inputs were divided into segments of 160 frames. Facial recognition was applied on the first frame of each segment, followed by cropping and resizing of the facial region. These adjustments were then maintained throughout the subsequent frames. In the post-processing, a second-order Butterworth filter (cutoff frequencies: 0.75 and 2.5 Hz) was applied to filter the rPPG waveform, and power spectral density was computed by the Welch algorithm for further heart rate estimation. Following the protocol outlined in (Yu et al. 2020), random upsampling, downsampling, and horizontal flipping were applied for data augmentation. The experiment was conducted on NVIDIA RTX 3090.

Loss. We employed a loss function that integrates constraints from both the temporal and frequency domains (Yu et al. 2020). The negative Pearson correlation coefficient is utilized as temporal constraint $\mathcal{L}_{Time}$, while cross-entropy between the power spectral density of prediction and the HR derived from the power spectral density of ground truth, is employed as frequency constraint $\mathcal{L}_{Freq}$. $\mathcal{L}_{Freq} = CE(maxIndex(PSD(PPG_{gt})), PSD(PPG_{pred}))$, where PSD represents Power Spectral Density and $maxIndex$ represents the index of the maximum value. The overall loss is expressed by: $\mathcal{L}_{overall} = a\cdot\mathcal{L}_{Time} + b\cdot\mathcal{L}_{Freq}$.

Comparison. We compare our method with state-of-the-art approaches in intra-dataset testing (Špetlík, Franc, and Matas 2018; Chen and McDuff 2018; Yu, Li, and Zhao 2019; Liu et al. 2020; Gideon and Stent 2021; Lu, Han, and Zhou 2021; Yu et al. 2022; Liu et al. 2023a; Li, Yu, and Shi 2023; Lu et al. 2023; Li and Yin 2023; Yue, Shi, and Ding 2023; Yu et al. 2023; Sun and Li 2024; Zou et al. 2024). Building on this, additional comparisons with (Verkruysse, Svaasand, and Nelson 2008; Poh, McDuff, and Picard 2010; De Haan and Jeanne 2013; Pilz et al. 2018; De Haan and Van Leest

Method	TrainSet	Test Set														
		PURE					UBFC					MMPD				
		MAE	RMSE	MAPE	ρ	SNR	MAE	RMSE	MAPE	ρ	SNR	MAE	RMSE	MAPE	ρ	SNR
GREEN	-	10.09	23.85	10.28	0.34	-2.66	19.73	31.00	18.72	0.37	-11.18	21.68	27.69	24.39	-0.01	-14.34
ICA	-	4.77	16.07	4.47	0.72	5.24	16.00	25.65	15.35	0.44	-9.91	18.60	24.30	20.88	0.01	-13.84
CHROM	-	5.77	14.93	11.52	0.81	4.58	4.06	8.83	3.84	0.89	-2.96	13.66	18.76	16.00	0.08	-11.74
LGI	-	4.61	15.38	4.96	0.77	4.50	15.80	28.55	14.70	0.36	-8.15	17.08	23.32	18.98	0.04	-13.15
PBV	-	3.92	12.99	4.84	0.84	2.30	15.90	26.40	15.17	0.48	-9.16	17.95	23.58	20.18	0.09	-13.88
POS	-	3.67	11.82	7.25	0.88	6.87	4.08	7.72	3.93	0.92	-2.39	12.36	17.71	14.43	0.18	-11.53
OMIT	-	4.66	15.82	4.97	8.76	4.37	16.99	29.54	15.91	0.34	-7.29	17.02	23.23	18.89	0.04	-12.77
DeepPhys	UBFC	5.54	18.51	5.32	0.66	4.40	-	-	-	-	-	17.50	25.00	19.27	0.06	-11.72
	PURE	-	-	-	-	-	1.21	2.90	1.42	0.99	1.74	16.92	24.61	18.54	0.05	-11.53
PhysNet	UBFC	8.06	19.71	13.67	0.61	6.68	-	-	-	-	-	9.47	16.01	**11.11**	0.31	-8.15
	PURE	-	-	-	-	-	0.98	2.48	1.12	0.99	1.49	13.94	21.61	15.15	0.20	-9.94
TS-CAN	UBFC	3.69	13.80	3.39	0.82	5.26	-	-	-	-	-	14.01	21.04	15.48	0.24	-10.18
	PURE	-	-	-	-	-	1.30	2.87	1.50	0.99	1.49	13.94	21.61	15.15	0.20	-9.94
PhysFormer	UBFC	12.92	24.36	23.92	0.47	2.16	-	-	-	-	-	12.10	17.79	15.41	0.17	-10.53
	PURE	-	-	-	-	-	1.44	3.77	1.66	0.98	0.18	14.57	20.71	16.73	0.15	-12.15
EfficientPhys	UBFC	5.47	17.04	5.40	0.71	4.09	-	-	-	-	-	13.78	22.25	15.15	0.09	-9.13
	PURE	-	-	-	-	-	2.07	6.32	2.10	0.94	-0.12	14.03	21.62	15.32	0.17	-9.95
Spiking-Phys.	UBFC	3.83	-	5.70	0.83	-	-	-	-	-	-	14.15	-	16.22	0.15	-
	PURE	-	-	-	-	-	2.80	-	2.81	0.95	-	14.57	-	16.55	0.14	-
RhythmFormer	UBFC	**0.97**	**3.36**	**1.60**	**0.99**	**12.01**	-	-	-	-	-	**9.08**	**15.07**	11.17	**0.53**	**-7.73**
	PURE	-	-	-	-	-	**0.89**	**1.83**	**0.97**	**0.99**	6.05	**8.98**	**14.85**	**11.11**	**0.51**	-8.39
Ours	UBFC	1.98	6.51	3.59	0.96	8.94	-	-	-	-	-	10.63	17.14	12.14	0.34	-8.28
	PURE	-	-	-	-	-	0.95	**1.83**	1.04	**0.99**	**6.35**	10.44	16.70	12.25	0.36	**-8.18**

Table 2: Cross-dataset evaluation. Best results are marked in **bold** and second best in underline.

2014; Wang et al. 2016; Casado and López 2023; Liu et al. 2024) are presented in cross-dataset testing.

4.3 Intra-Dataset Evaluation

We conducted intra-dataset evaluation on the PURE and UBFC datasets to validate the feasibility of the Mamba architecture. For the evaluation of the PURE dataset, we followed the protocols outlined in (Lu, Han, and Zhou 2021), splitting the dataset sequentially into training and testing sets with a ratio of 6:4. Similarly, for the evaluation of the UBFC dataset, we followed the protocols in (Lu, Han, and Zhou 2021), selecting the first 30 samples as the training set and the remaining 12 samples as the testing set. Due to the absence of the validation set, we selected the checkpoint from the last epoch for testing and compared them with the reported results from previous methods. As shown in Table 1, on the PURE dataset, our method outperformed all state-of-the-art methods across all metrics, achieving the minimum MAE (0.23) and RMSE (0.34). On the UBFC dataset, our method also achieved comparable performance to others.

Due to the relative simplicity of PURE and UBFC, the performance of state-of-the-art methods on these datasets is nearing saturation. To further evaluate the performance, we employed the more challenging dataset. For the VIPL-HR dataset, we followed the subject-exclusive 5-fold cross-validation protocol, as outlined in (Niu et al. 2019; Yu et al. 2022). For the MMPD dataset, following the protocols outlined in (Zou et al. 2024), the dataset was sequentially split into training, validation, and testing sets with a ratio of 7:1:2. As shown in Table 1, our method achieved comparable performance to the previous methods. This indicates that RhythmMamba can accurately extract weak rPPG signals and understand their periodic nature, which provides ample empirical evidence for the feasibility of the Mamba architecture in the rPPG task.

4.4 Cross-Dataset Evaluation

To objectively evaluate the generalization capability to out-of-distribution data, we followed the protocols outlined in (Liu et al. 2023b) for cross-dataset evaluation. The models were trained on either the PURE or UBFC datasets and tested on the PURE, UBFC, and MMPD datasets. The training dataset was sequentially split into training and validation sets with a ratio of 8:2. All comparative methods were implemented based on the rPPG toolbox (Liu et al. 2023b). As shown in Table 2, the proposed RhythmMamba also achieved SOTA performance, demonstrating its capability in modeling domain-invariant features and generalizing to unseen domains. Based on the comparisons in Tables 1 and 2, the improvement in cross-dataset results is less significant compared to intra-dataset results, possibly due to the fine-grained token-wise self-attention, which may have an advantage in capturing domain-invariant features. Nevertheless, founded on fewer parameters and lower computational complexity, RhythmMamba showcases its potential in real-world applications through its robustness and generalization in complex environments. Additional visualization results can be found in Supplementary Material B.

4.5 Ablation Study

Ablation studies were conducted on the MMPD dataset to assess the impact of different modules.

Diff-fusion	Self-attention	Large Kernel	Multi-temporal	FFN	MAE↓	RMSE↓	MAPE↓	ρ ↑	SNR↑
×	✓	✓	✓	Frequency	5.71	10.00	5.97	0.68	-0.29
✓	×	✓	✓	Frequency	4.02	8.43	4.15	0.79	2.98
✓	✓	×	✓	Frequency	3.51	7.53	3.81	0.83	4.22
✓	✓	✓	×	Frequency	3.60	7.78	3.83	0.81	4.38
✓	✓	✓	✓	×	3.53	7.65	3.67	0.83	2.89
✓	✓	✓	✓	Vanilla	3.54	7.68	3.72	0.82	2.88
✓	✓	✓	✓	Spatio-Temporal	3.82	8.06	3.95	0.81	4.06
✓	✓	✓	✓	Frequency	**3.16**	**7.27**	**3.37**	**0.84**	**4.74**

Table 3: Impact of key modules.

Spatial Token numbers	MAE↓	RMSE↓	MAPE↓	ρ ↑	SNR↑
8×8	4.90	10.14	5.15	0.71	-2.10
4×4	4.62	8.87	4.83	0.77	-0.88
4×4 (Temporal Embed)	4.92	10.04	5.07	0.69	-1.43
4×4 (Position Embed)	5.47	10.34	5.77	0.71	-3.03
2×2	4.69	9.97	4.80	0.70	-0.98
1×1 (Avgpool)	**3.54**	**7.68**	**3.72**	**0.82**	**2.88**

Table 4: Impact of spatial information (with vanilla FFN).

Method	Para.	MACs	Throughput	Memory
DeepPhys	1.98	744.45	5.65	37.28
PhysNet	0.77	438.24	11.80	11.43
TS-CAN	1.98	744.45	5.21	38.91
PhysFormer	7.38	316.29	8.50	28.63
EfficientPhys	1.91	373.72	8.42	26.68
RhythmFormer	3.25	240.55	6.30	29.06
Ours	1.07	80.90	20.09	6.66

Table 5: Computational cost. The horizontal axis represents Parameters (M), MACs (M), Throughput (Kfps), and Peak GPU Memory Usage (M).

Impact of Spatial Information. As shown in Table 4, the ablation study of spatial information is presented, where both position embedding and temporal embedding were implemented using learnable parameters. For a fair comparison, the vanilla FFN was used, as the frequency domain FFN might have an advantage with purely temporal token sequences. It is evident that tokenized spatiotemporal information performs poorly, even with temporal embedding or position embedding. The integration of spatial information increases the dimensionality of the state transition process, thereby elevating complexity and making the model more difficult to train. We view the rPPG task as a time series task, embedding spatial information into the channels, with each channel being treated as a purely temporal sequence. This ensures that the state transition process occurs purely along the temporal dimension, while spatial information interactions are facilitated through subsequent channel interactions, effectively resolving this issue.

Impact of Key Modules. As illustrated in Table 3, the comparison between the first four rows and the last row indicates the significant roles played by these modules. The diff-fusion module enables frame-level representation awareness of BVP wave variations, effectively enhancing rPPG features with a small additional computational cost. The use of relatively large convolution kernels and self-attention allows for the integration of spatial information into channels effectively, thereby providing sufficient information for subsequent processing. The multi-temporal constraint Mamba constrains a single Mamba block simultaneously to short-term trends and periodic patterns, facilitating the accurate comprehension of rPPG features.

Impact of Frequency Domain Feed-forward. As shown in Table 3, the last four rows show that the frequency domain FFN plays an important role. Among them, vanilla FFN refers to using two linear layers to compose the FFN, and Spatio-Temporal FFN refers to the addition of a depth-wise convolution layer between the linear layers (Yu et al. 2022). Frequency domain FFN adds a frequency domain linear layer between the linear layers, effectively extracting the most critical frequency domain features in rPPG signals.

4.6 Computational Cost

We conducted a 30-second inference test at a resolution of 128×128, reporting the parameters, average MACs per frame, average throughput per frame, and average peak GPU memory usage per frame. As shown in Table 5 and Figure 2, RhythmMamba achieved 319% throughput and 23% peak GPU memory, demonstrating the potential for effective mobile-level rPPG applications. Building on this advantage, we also found that the proposed method can accept inputs of arbitrary length during inference without any performance degradation (see Supplementary Material C for details).

5 Conclusion

We approach the rPPG task as a time series task, designing multiple modules that align with the temporal characteristics of rPPG signals to boost state space model learning. This approach boasts strong long-range dependency modeling capabilities while maintaining linear complexity. It achieves state-of-the-art performance both within and across datasets with a faster and more lightweight design. However, since Mamba's state transitions align closely with the periodic variations of rPPG signals, we believe that Mamba's potential in rPPG extends beyond the current results. Our utilization of periodic priors is currently limited and we would like to delve into this more deeply in the future.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62206015, 62227801, U21B2048), the National Science and Technology Major Project (2022ZD0117901), and the Fundamental Research Funds for the Central Universities (FRF-TP-22-043A1).

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6836–6846.

Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; and Dubois, J. 2019. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124: 82–90.

Casado, C. A.; and López, M. B. 2023. Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces. *IEEE Journal of Biomedical and Health Informatics*.

Chen, W.; and McDuff, D. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 349–365.

Choi, J.-H.; Kang, K.-B.; and Kim, K.-T. 2024. Fusion-Vital: Video-RF Fusion Transformer for Advanced Remote Physiological Measurement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1344–1352.

Dai, R.; Das, S.; Kahatapitiya, K.; Ryoo, M. S.; and Brémond, F. 2022. MS-TCT: multi-scale temporal convtransformer for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20041–20051.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *International Conference on Machine Learning (ICML)*.

De Haan, G.; and Jeanne, V. 2013. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10): 2878–2886.

De Haan, G.; and Van Leest, A. 2014. Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physiological measurement*, 35(9): 1913.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gideon, J.; and Stent, S. 2021. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3995–4004.

Hu, H.; Dong, S.; Zhao, Y.; Lian, D.; Li, Z.; and Gao, S. 2022. Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19013–19022.

Kong, J.; Bian, Y.; and Jiang, M. 2022. MTT: Multi-scale temporal transformer for skeleton-based action recognition. *IEEE Signal Processing Letters*, 29: 528–532.

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

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2025. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, 237–255. Springer.

Li, Z.; and Yin, L. 2023. Contactless Pulse Estimation Leveraging Pseudo Labels and Self-Supervision. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 20531–20540.

Liu, M.; Tang, J.; Li, H.; Qi, J.; Li, S.; Wang, K.; Wang, Y.; and Chen, H. 2024. Spiking-PhysFormer: Camera-Based Remote Photoplethysmography with Parallel Spike-driven Transformer. arXiv:2402.04798.

Liu, X.; Fromm, J.; Patel, S.; and McDuff, D. 2020. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33: 19400–19411.

Liu, X.; Hill, B.; Jiang, Z.; Patel, S.; and McDuff, D. 2023a. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5008–5017.

Liu, X.; Narayanswamy, G.; Paruchuri, A.; Zhang, X.; Tang, J.; Zhang, Y.; Sengupta, S.; Patel, S.; Wang, Y.; and McDuff, D. 2023b. rPPG-Toolbox: Deep Remote PPG Toolbox. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Lu, H.; Han, H.; and Zhou, S. K. 2021. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12404–12413.

Lu, H.; Yu, Z.; Niu, X.; and Chen, Y.-C. 2023. Neuron Structure Modeling for Generalizable Remote Physiological Measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18589–18599.

McDuff, D. 2023. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55(9): 1–40.

Niu, X.; Han, H.; Shan, S.; and Chen, X. 2018. Synrhythm: Learning a deep heart rate estimator from general to specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 3580–3585. IEEE.

Niu, X.; Shan, S.; Han, H.; and Chen, X. 2019. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29: 2409–2423.

Niu, X.; Yu, Z.; Han, H.; Li, X.; Shan, S.; and Zhao, G. 2020. Video-based remote physiological measurement via cross-verified feature disentangling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 295–310. Springer.

Patro, B. N.; and Agneeswaran, V. S. 2024. SiMBA: Simplified Mamba-Based Architecture for Vision and Multivariate Time series. *arXiv preprint arXiv:2403.15360*.

Pilz, C. S.; Zaunseder, S.; Krajewski, J.; and Blazek, V. 2018. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1254–1262.

Poh, M.-Z.; McDuff, D. J.; and Picard, R. W. 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10): 10762–10774.

Shao, H.; Luo, L.; Qian, J.; Chen, S.; Hu, C.; and Yang, J. 2023. TranPhys: Spatiotemporal Masked Transformer Steered Remote Photoplethysmography Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Špetlík, R.; Franc, V.; and Matas, J. 2018. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference, Newcastle, UK*, 3–6.

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

Sun, Z.; and Li, X. 2024. Contrast-Phys+: Unsupervised and Weakly-Supervised Video-Based Remote Physiological Measurement via Spatiotemporal Contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5835–5851.

Tang, J.; Chen, K.; Wang, Y.; Shi, Y.; Patel, S.; McDuff, D.; and Liu, X. 2023. MMPD: Multi-Domain Mobile Video Physiology Dataset. In *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Verkruysse, W.; Svaasand, L. O.; and Nelson, J. S. 2008. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26): 21434–21445.

Wang, W.; Den Brinker, A. C.; Stuijk, S.; and De Haan, G. 2016. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7): 1479–1491.

Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; and Niu, Z. 2024. Frequency-domain MLPs are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36.

Yu, Z.; Li, X.; Niu, X.; Shi, J.; and Zhao, G. 2020. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, 27: 1245–1249.

Yu, Z.; Li, X.; and Zhao, G. 2019. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *30th British Machine Visison Conference: BMVC 2019. 9th-12th September 2019, Cardiff, UK*. The British Machine Vision Conference (BMVC).

Yu, Z.; Peng, W.; Li, X.; Hong, X.; and Zhao, G. 2019. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 151–160.

Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Cui, Y.; Zhang, J.; Torr, P.; and Zhao, G. 2023. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, 131(6): 1307–1330.

Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Torr, P. H.; and Zhao, G. 2022. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4186–4196.

Yue, Z.; Shi, M.; and Ding, S. 2023. Facial video-based remote physiological measurement via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, Y.; Zou, B.; Yang, F.; Lu, L.; Belkacem, A. N.; and Chen, C. 2021. Video-based physiological measurement using 3d central difference convolution attention network. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 1–6. IEEE.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *Forty-first International Conference on Machine Learning*.

Zou, B.; Guo, Z.; Chen, J.; and Ma, H. 2024. RhythmFormer: Extracting rPPG Signals Based on Hierarchical Temporal Periodic Transformer. *arXiv preprint arXiv:2402.12788*.