Journal of Advanced Digital Communications
https://www.sciltp.com/journals/jadc

Article

LightweightPhys: A Lightweight and Robust Network for Remote Photoplethysmography Signal Extraction

Yu Liu [1], Yinqiao Li [1], Yan He [1], Tao Wang [1,*] and Zhigao Zheng [2]

[1] Hubei Key Laboratory of Digital Education, Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
[2] School of Computer Science, Wuhan University, Wuhan 430072, China
∗ Correspondence: tmac@ccnu.edu.cn

How To Cite: Yu Liu; Li, Y.; He, Y.; et al. LightweightPhys: A Lightweight and Robust Network for Remote Photoplethysmography Signal Extraction. *Journal of Advanced Digital Communications* **2025**, *2*(1), 2. https://doi.org/10.53941/jadc.2025.100002

Received: 25 April 2025
Revised: 18 August 2025
Accepted: 29 August 2025
Published: 8 September 2025

Abstract: Remote Photoplethysmography (rPPG) has emerged as a promising technology for non-contact heart rate monitoring by analyzing subtle color variations in facial videos, which are caused by changes in blood volume. While traditional rPPG methods often struggle with environmental noise and motion artifacts, deep learning-based approaches have shown improved accuracy but typically come with high computational costs, limiting their applicability in real-time, resource-constrained scenarios. To address these challenges, this paper introduces LightweightPhys, a novel 3D-CNN-based network that combines Depthwise Separable Convolution (DSC) and a newly proposed Simulated Temporal Noise Suppression (Sim-TN) module. The DSC module significantly reduces the computational complexity of the model by decoupling spatial and channel-wise convolutions, making it more efficient for deployment on low-power devices. Meanwhile, the Sim-TN module enhances the robustness of feature extraction by effectively suppressing temporal noise, which is a common issue in rPPG signal processing due to environmental factors and subject movement. Extensive evaluations on two widely-used datasets, PURE and UBFC-rPPG, demonstrate that LightweightPhys achieves state-of-the-art performance in heart rate estimation while maintaining significantly lower computational overhead compared to existing deep learning models. This makes LightweightPhys particularly suitable for real-time health monitoring applications on resource-constrained devices, such as wearable gadgets and mobile health platforms.

Keywords: remote photoplethysmography; rPPG; LightweightPhys; depthwise separable convolution; Sim-TN; heart rate monitoring; deep learning

1. Introduction

In the field of healthcare, physiological signals such as average heart rate (HR), respiratory frequency (RF), and heart rate variability (HRV) play a crucial role in analyzing cardiovascular activity and have been widely adopted in diverse medical applications. Traditional methods for measuring HR, RF, and HRV primarily rely on electrocardiography (ECG) and contact-based photoplethysmography (cPPG) signals, which require specialized skin-contact devices for data collection. However, the use of such contact sensors in real-world healthcare scenarios often causes discomfort and inconvenience to subjects. For instance, prolonged wear of ECG electrodes may lead to skin irritation, while contact-based devices restrict mobility, interfering with daily activities.

With the rapid advancement of computer vision and non-contact sensing technologies, remote photoplethysmography (rPPG) has emerged as a promising technique that estimates physiological signals by capturing subtle facial skin color variations caused by blood volume pulsation in video streams. Compared to traditional contact-based methods like ECG, rPPG offers non-invasiveness, low cost, and compatibility with ubiquitous RGB cameras embedded in smartphones and surveillance systems. This enables remote measurement of HR, RF, and HRV, which

has seen rapid development.

The fundamental principle of rPPG lies in the physiological characteristics of the human body: cardiac contractions induce periodic changes in blood volume within the skin, leading to corresponding alterations in skin color. According to the Beer-Lambert law, the absorption and reflection of light by blood vary with these volumetric fluctuations. By analyzing such optical changes, critical physiological information like HR can be indirectly derived. Thus, rPPG-based physiological measurement depends on color variations caused by skin absorption and reflection of external light. Through modeling skin light reflection, hemoglobin concentration changes in capillaries can be inferred to estimate physiological parameters. However, practical applications of rPPG face significant challenges. Existing rPPG algorithms suffer from high computational costs and sensitivity to noise, such as motion artifacts and illumination variations. These issues result in insufficient robustness and accuracy in signal extraction, hindering effective utilization of rPPG signals in real-world scenarios.

To overcome these challenges, two main types of methods have been proposed in the field of rPPG-based physiological measurement. Traditional methods mainly include signal processing-based techniques. This method adopts chromaticity-based techniques and utilizes the skin reflection model to extract rPPG signals from facial videos. First, the region of interest (ROI) in each frame of the captured material is identified. Subsequently, spatial averaging is applied to the RGB channels to generate an initial signal waveform. Various signal processing techniques are then employed to refine these waveforms in order to reconstruct physiological signals. For example, Poh et al. utilized principal component analysis (PCA) and independent component analysis (ICA) to extract informative features for obtaining the remote photoplethysmographic (rPPG) signal [1]. However, such methods rely on handcrafted features, which limit their effectiveness under complex scenarios. In addition, researchers have explored nonlinear signal decomposition methods. Unlike traditional blind source separation techniques, these approaches assume that the color channel signals are nonlinear combinations of blood volume pulse (BVP) and noise signals. These methods often require prior assumptions regarding skin properties and light interaction, thereby restricting their applicability in real-world environments with varying illumination conditions and head movements.

On the other hand, neural network-based methods, such as DeepPhys, PhysNet, etc., utilize data-driven learning to extract rPPG signals and measure physiological parameters. These methods use convolutional neural networks (CNNs) to directly extract features from video sequences, eliminating the need for a separate signal processing stage. However, these neural network-based methods also have obvious drawbacks. They usually have high computational complexity, a large number of parameters, and high floating-point operations (FLOPs). For example, the number of parameters of some models can reach millions or even more, and the FLOPs are also at a high level, making it extremely difficult to train and deploy on edge devices with limited resources. The training process not only requires strong computational resource support but may also face problems such as slow running speed and high energy consumption when deployed on edge devices. In addition, they are still relatively sensitive to noise. In real-world scenarios, noise factors, such as lighting changes and human body movements, will interfere with signal extraction, resulting in a decrease in the accuracy of the extracted signals, and thus weakening their performance in practical applications.

To overcome these limitations, this study proposes a novel solution. We design the LightweightPhys network, which incorporates depthwise separable convolutions to reduce parameter counts and computational complexity, addressing the high computational overhead of traditional methods. Additionally, we introduce the Sim-TN feature enhancement module, integrating the SimAM attention mechanism with temporal normalization. This module effectively suppresses noise and improves signal extraction robustness, enabling accurate acquisition of rPPG signals even in complex environments.

Our contributions are summarized as follows:

- We introduce LightweightPhys, a novel 3D-CNN-based architecture that significantly improves the efficiency and robustness of rPPG signal extraction. By integrating Depthwise Separable Convolution (DSC) and a newly proposed Simulated Temporal Noise Suppression (Sim-TN) module, our method achieves state-of-the-art performance in heart rate estimation. The Sim-TN module effectively suppresses temporal noise caused by environmental factors and motion artifacts, while the DSC module reduces computational complexity without compromising accuracy. This combination results in lower mean absolute error (MAE) and root mean square error (RMSE), along with high Pearson correlation coefficients, across datasets with varying noise levels, demonstrating the model's robustness and generalization capability.
- LightweightPhys is specifically designed to minimize computational overhead, making it highly suitable for deployment in resource-constrained environments such as mobile devices and wearable gadgets. By reducing the number of parameters and FLOPs, our model significantly alleviates hardware strain while maintaining high responsiveness. This computational efficiency not only enables real-time rPPG signal extraction but also

facilitates broader adoption of non-contact heart rate monitoring in practical applications, such as remote health monitoring and fitness tracking.

- Unlike existing deep learning-based rPPG methods that often prioritize accuracy at the expense of computational cost, LightweightPhys achieves an optimal balance between accuracy and efficiency. The integration of DSC and Sim-TN modules ensures that the model remains lightweight while providing robust performance in noisy environments. This makes LightweightPhys a practical solution for real-world scenarios where both accuracy and computational efficiency are critical, such as in low-power devices or real-time health monitoring systems.

The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of related works. Section 3 presents a detailed description of our proposed LightweightPhys. Section 4 conducts extensive experiments to evaluate and analyze the performance of the proposed method. Finally, Section 5 concludes the paper by summarizing our contributions and discussing potential future directions.

2. Related Works

With the development of digital imaging, communication systems, and mobile devices, facial recognition technology has been widely applied in various fields. However, AI-driven facial forgery technologies such as Deepfakes pose a serious threat, making face anti-spoofing a research hotspot. Meanwhile, research on remote photoplethysmography (rPPG) in the fields of physiological signal detection and face anti-spoofing has also been evolving, which can be mainly divided into deep-learning-based methods and traditional methods.

2.1. Traditional Signal Processing Approaches for rPPG

Regarding traditional methods, they mainly involve rPPG signal extraction methods based on signal processing. These methods utilize the skin reflection model and adopt chrominance-based techniques to extract rPPG signals from facial videos. First, the region of interest (ROI) of each frame of the image is identified, then the RGB channels are spatially averaged to generate an initial signal waveform, and finally, the physiological waveform is reconstructed by optimizing the initial waveform through various signal processing techniques.

In image processing methods, Poh et al. used principal component analysis and independent component analysis to extract useful features for obtaining rPPG signals [1]. Wu et al. employed Eulerian video magniffcation to amplify and visualize subtle changes in facial videos caused by blood ffow, enabling better detection of heart rate information from non-contact video images [2]. Burzo et al. applied a multimodal approach to rPPG for affective computing applications, showing that remotely captured vital signs can improve the prediction of emotional valence when combined with facial expressions [3]. Hsu et al. used support vector regression to extract heart rate signals by inputting RGB channel signals and frequency domain features of ICA-decomposed signals [4]. Prakash and Tucker utilized bounded Kalman fflters and denoising algorithms to reduce motion effects and improve the extraction of HR and HRV features [5]. Liu et al. monitored emotions based on HRV signals extracted using rPPG, achieving accuracies of 97.59% and 96.90% for emotion-related HRV features LF and HF, respectively [6]. Sabour et al. extracted heart rate features from patient videos using rPPG and assessed patient emotions based on the relationship between the sympathetic nervous system and emotions [7]. Richard et al. optimized heart rate extraction from rPPG signals under multi-objective conditions using autocorrelation and independent component analysis [8]. Zhang et al. proposed a multimodal fusion method combining facial expressions, pupil, and rPPG signals, improving the accuracy of emotion recognition by 26.19% in an ideal environment dataset compared to single-modal methods [9]. However, traditional methods have poor robustness to noise and motion in complex environments and are easily affected by factors such as lighting changes and head movements.

2.2. Deep Learning Approaches for rPPG Estimation

In the aspect of rPPG research based on deep learning ,early researchers attempted to train heart rate estimation models using hand-crafted features.For instance, Hsu et al. utilized the frequency-domain features of the original RGB three-channel signals and the frequency-domain features of signals decomposed by Independent Component Analysis (ICA) as inputs, employing Support Vector Regression to estimate heart rate values [10]. Osman et al. used handcrafted features to classify the peak positions of Blood Volume Pulse (BVP) signals and subsequently measured heart rate based on these peaks [11].

Unlike traditional image processing methods, deep learning methods require higher computational complexity. Nardelli et al. demonstrated the possibility of emotion classiffcation based on HRV features extracted from ECG signals [12]. Harper and Southern proposed an end-to-end deep learning model for emotion assessment based on HRV, consisting of two concurrent streams with 8 convolutional layers [13]. Zhou et al. proposed a multimodal

(facial expressions + rPPG) valence-arousal emotion assessment model with a backbone of 19 convolutional layers, achieving superior performance compared to facial expression analysis [14]. Fan et al. applied a Transformer-based multimodal feature enhancement network using rPPG-extracted features, achieving good emotion recognition accuracy [15]. However, the self-attention mechanism and multi-head attention structure of Transformers require higher parameter processing and computational overhead.

Moreover, researchers have investigated the application of deep learning methods, which possess strong modeling capabilities, to the measurement of physiological parameters based on rPPG [16, 17]. Chen and McDuff [18] proposed DeepPhys, a novel end to end approach leveraging convolutional attention networks to extract physiological signals from video data, demonstrating signiffcant improvements in HRV features measurement accuracy. Yu et al. [19] explored the use of spatio-temporal networks to measure remote photoplethysmograph (rPPG) signals from facial videos, offering a robust solution for non-contact vital sign monitoring. Liu et al. [20] tackled on-device efffciency by proposing multi-task temporal shift attention networks, which facilitate real-time, accurate, and contactless vital sign measurement. Girish et al. introduced a multi-task dual-branch network model BigSmall, capable of recognizing expressions, respiration, and rPPG, reducing the number of feature parameters by cross-branch feature sharing with minimal impact on accuracy through feature fusion [21]. Yu et al. [22] introduced PhysFormer, a transformer-based model that utilizes temporal differences in facial videos to achieve state-of-the-art performance in physiological measurement. Joshi and Cho [23] contributed the iBVP dataset, an RGB-thermal rPPG dataset with high-resolution signal quality labels, providing a valuable resource for advancing research in remote physiological sensing. Luo et al. [24] proposed PhysMamba, an innovative framework that integrates SlowFast temporal difference Mamba for efffcient and accurate remote physiological measurement, pushing the boundaries of real-time monitoring.

2.3. Lightweight Architectures for rPPG Estimation

Deep learning has significantly advanced the development of remote photoplethysmography (rPPG) signal estimation. However, the high computational cost of existing models limits their deployment on mobile and embedded devices. To address this issue, various lightweight neural networks have been developed, aiming to reduce complexity while maintaining performance.

FastBVP-Net [25] is a typical lightweight rPPG network that employs a Multi-frequency Mode Signal Fusion (MMSF) mechanism. It characterizes different modes of the original signal in the decomposition module and reconstructs the Blood Volume Pulse (BVP) signal in complex noisy environments in the synthesis module. This design enables FastBVP-Net to efficiently estimate heart rate (HR) and heart rate variability (HRV) from 30-s or even 15-s facial videos with low computational burden, demonstrating excellent performance on benchmark datasets.

EfficientPhys [26] is based on the EfficientNet architecture and balances model accuracy and speed through a compound scaling strategy. It takes original video frames as input and incorporates custom normalization layers, self-attention modules, tensor shifting modules, and 2D convolution operations, enabling efficient and accurate spatiotemporal modeling. Nevertheless, it does not optimize for rPPG-specific temporal noise, resulting in limited robustness when facing interference such as motion artifacts.

As a typical ultra-light weight 3DCNN model, RTrPPG [27] greatly reduces computing overhead by simplifying network hierarchy and convolutional core design. Its core innovation is the adoption of a compact space-time feature extraction module to reduce redundant parameters while ensuring the frame rate.

As a key model compression technique, network pruning provides another important approach for the lightweight design of rPPG models. Its core idea is to remove redundant parameters and insignificant network structures, thereby reducing the computational cost of the model while minimizing the loss of accuracy. Addressing the issue of limited training samples in the rPPG field, Zhao et al. (2022) [28] proposed an rPPG network pruning method suitable for small-sample scenarios. By designing a sample-adaptive pruning criterion, this method avoids the problems of over-pruning or under-pruning caused by insufficient samples. To further improve pruning accuracy and model performance retention rate, Zhao et al. (2025) [29] put forward a weight-gradient joint pruning criterion. This criterion accurately identifies redundant parameters by simultaneously considering the importance of network weights and gradient information.

3. Methods

3.1. LightweightPhys Architecture

LightweightPhys is an optimized architecture derived from PhysNet [19], designed for efficient and robust Remote Photoplethysmography (rPPG) signal extraction. The network processes input tensors of size $4 \times 3 \times 300 \times 128 \times 128$ (batch, RGB channels, time, spatial resolution) and outputs a 1D rPPG signal, as illustrated in Figure 1

and Table 1. The architecture is tailored to reduce the computational complexity and noise sensitivity inherent in traditional rPPG methods by incorporating Depthwise Separable Convolution (DSC) and a novel Simulated Temporal Noise Suppression (Sim-TN) module.

The network begins with a 3D convolutional layer (Conv Layer 1,1) using a $3 \times 3 \times 3$ kernel to extract spatio-temporal features from the input video frames. This is followed by an Average Pooling (AvgPooling) layer, which downsamples the temporal dimension, reducing the computational load while preserving essential temporal information.

The architecture consists of several Conv Blocks (Blocks 2 to 9) that progressively process the extracted features. Each block is designed to enhance feature representation while suppressing noise. Blocks 2 to 5 and Blocks 7 to 9 employ 3D convolution combined with the Sim-TN module (TN+Conv+SimAM), followed by Batch Normalization (BatchNorm3D) and Max Pooling layers. The Sim-TN module plays a critical role in reducing noise artifacts caused by motion and illumination variations, ensuring robust feature extraction.

To further reduce computational complexity, Block 6 replaces the standard 3D convolution with Depthwise Separable Convolution (DSC) (TN+DepthwiseSeparableConv+SimAM). The DSC module decouples spatial and channel-wise convolutions, significantly reducing the number of parameters and FLOPs without compromising feature quality. This is followed by BatchNorm3D and Max Pooling layers, which help in maintaining the stability of the feature maps.

After the feature extraction and processing stages, a Transposed Convolution (ConvTranspose) layer is used to upsample the temporal dimension, recovering the temporal resolution necessary for accurate rPPG signal generation. Finally, a Conv Layer 10 with Average Pooling (AvgPooling) produces the 1D rPPG signal, which represents the heart rate information extracted from the input video frames.

The integration of DSC and Sim-TN modules results in a significant reduction in computational complexity. Specifically, the number of parameters is reduced by 54.5% (from 0.77M to 0.35M), and FLOPs are reduced by 27.5% (from 56.10G to 40.65G) compared to the original PhysNet architecture, as shown in Section 4. Despite this reduction in complexity, the Sim-TN module ensures robustness against noise, making the network suitable for real-time rPPG signal extraction on resource-constrained devices.

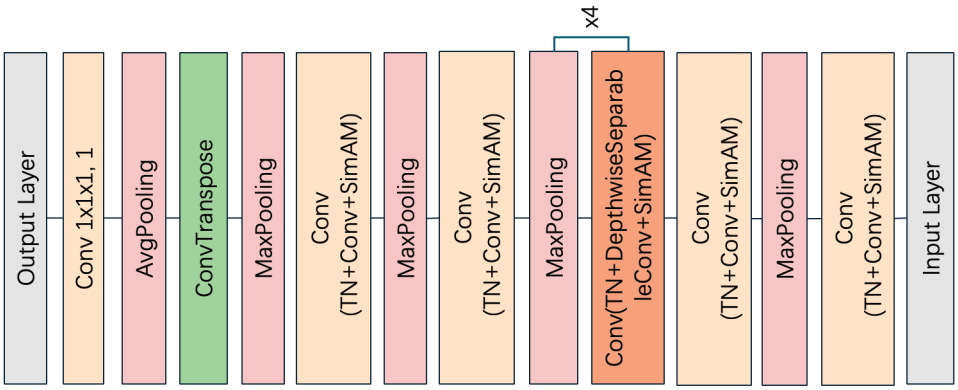


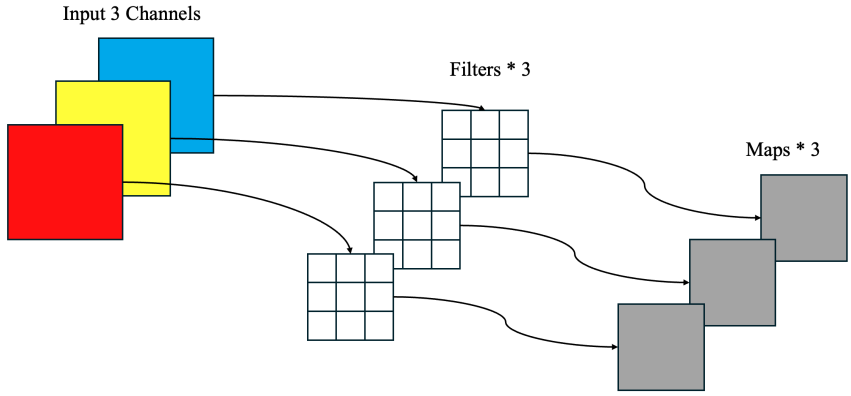
Figure 1. Architecture of LightweightPhys Network.

Table 1. Network Hyperparameter Overview.

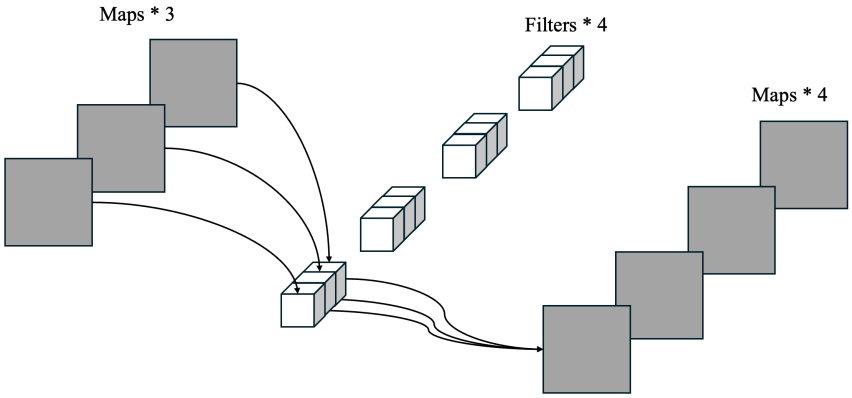
Layer/Module Type	Kernel Size	Stride	Padding
Initial 3D Convolutional Layer (Conv Layer 1,1)	$3 \times 3 \times 3$	(1,1,1)	(1,1,1)
Average Pooling Layer	$1 \times 2 \times 2$	(1,2,2)	(0,1,1)
Regular Convolutional Blocks (Blocks 2–5, 7–9)	$3 \times 3 \times 3$	(1,1,1)	(1,1,1)
Depthwise Separable Convolution Block (Block 6)	$3 \times 3 \times 1$ (Depthwise) + $1 \times 1 \times M$ (Pointwise)	(1,1,1)	(1,1,0)
Transposed Convolutional Layer	$3 \times 3 \times 3$	(2,1,1)	(1,0,0)

3.2. Depthwise Separable Convolution

Depthwise Separable Convolution divides standard 3D convolution into two efficient stages: depthwise convolution and pointwise convolution, reducing computational complexity while retaining feature representation.



(a) Depthwise Convolution



(b) Pointwise Convolution

Figure 2. Two Steps for Depthwise Separable Convolution.

In depthwise convolution, a single $D \times D \times 1$ filter is applied to each input channel (M) independently, extracting spatial features and producing intermediate feature maps of size $H' \times W' \times M$. This is followed by pointwise convolution, which uses $1 \times 1 \times M$ filters to combine these intermediate maps across channels, generating N output feature maps of size $H' \times W' \times N$, as shown in Figure 2.

The Depthwise Separable Convolution process is expressed as:

$$O'_{h',w',m} = \sum_{i=1}^{D} \sum_{j=1}^{D} D_{i,j,m} \cdot I_{h'+i-1,w'+j-1,m} \tag{1}$$

$$O_{h',w',n} = \sum_{m=1}^{M} P_{m,n} \cdot O'_{h',w',m} \tag{2}$$

where $D_{i,j,m}$ is the depthwise kernel of size $\mathbb{R}^{D \times D \times M}$, $P_{m,n}$ is the pointwise kernel of size $\mathbb{R}^{M \times N}$, $I_{h_i,w_j,m} \in \mathbb{R}^{H \times W \times M}$ are input feature maps, $O'_{h',w',m} \in \mathbb{R}^{H' \times W' \times M}$ are intermediate feature maps, and $O_{h',w',n} \in \mathbb{R}^{H' \times W' \times N}$ are output feature maps. Here, D is the kernel size, M is the number of input channels, N is the number of output channels, H and W are the height and width of input feature maps, and H' and W' are the height and width of output feature maps. The spatial indices assume a stride of 1 and appropriate padding to maintain output dimensions.

Theoretically, for 3x3 depthwise separable convolution, the computational cost is reduced by approximately 8 to 9 times with minimal accuracy loss when the number of output channels N is large, making the computational cost of Depthwise Separable Convolution approximately 1/9 of standard convolution [30]. Compared to standard convolution with $D^2 \times M \times N$ parameters, Depthwise Separable Convolution uses $D^2 \times M + M \times N$, achieving significant parameter and FLOPs reductions. In the proposed network, applied in ConvBlock 2–9, Depthwise Separable Convolution enables efficient deployment on edge devices while maintaining feature extraction (e.g., 54.5% parameter reduction from 0.77 M to 0.35 M and 27.5% FLOPs savings from 56.10 G to 40.65 G).

3.3. Sim-TN Module

The Sim-TN module enhances feature robustness for blood volume pulse (BVP) extraction by integrating SimAM attention and temporal normalization (TN), as shown in Figure 3.

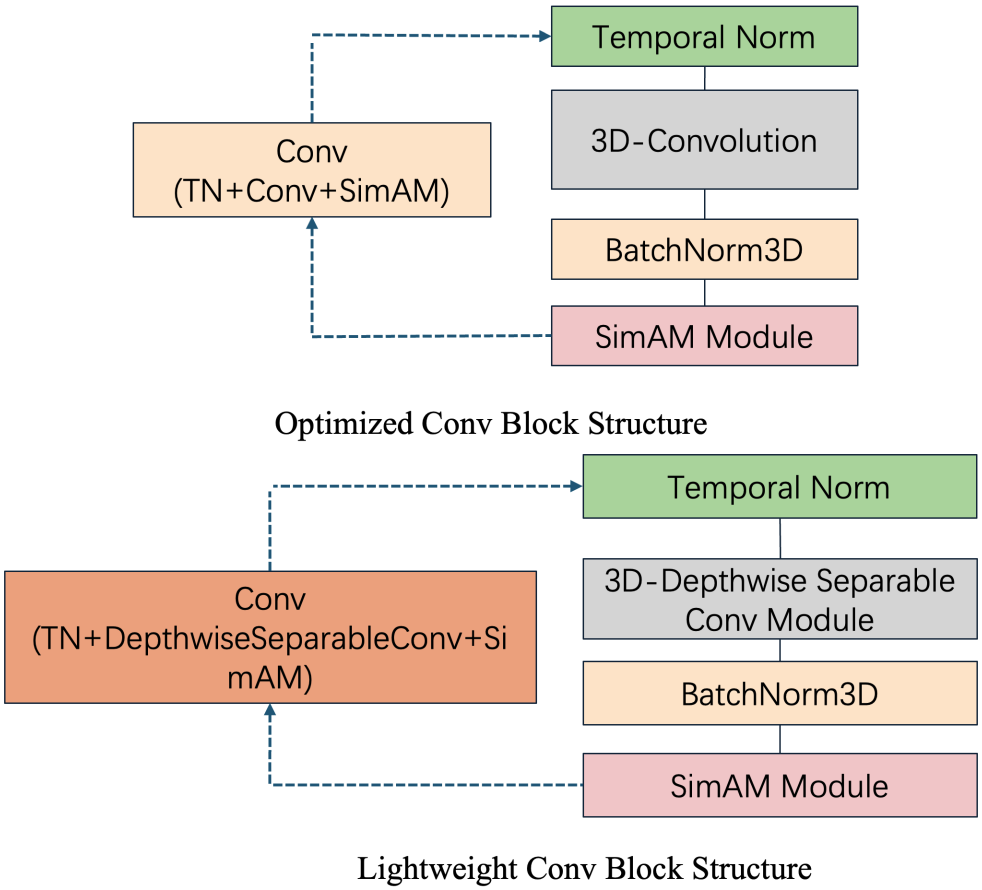


Figure 3. Feature Enhancement Modules in LightweightPhys.

SimAM, a parameter-free attention mechanism [31], generates 3D weights to prioritize BVP-related features over noise. It computes an energy function for each neuron in a feature map to measure its importance:

$$e_u = \frac{1}{H'W' - 1} \sum_{h',w'} \left(\frac{(x_{h',w'} - \mu_u)^2}{\sigma_u^2 + \lambda} + 2 \right) + \frac{(\mu_u - t)^2}{\sigma_u^2 + \lambda}, \tag{3}$$

where μ_u and σ_u^2 are the mean and variance of neurons in channel u, t is the target neuron, λ is a regularization term (typically 10^{-4}), and $H' \times W'$ is the spatial size of the feature map. Lower energy indicates higher importance, and weights are derived as:

$$\mathbf{F} = \frac{4(\sigma_u^2 + \lambda)}{e_u}. \tag{4}$$

Features are refined via:

$$\widetilde{\mathbf{X}} = \text{sigmoid}\left(\frac{1}{\mathbf{F}}\right) \odot \mathbf{X}, \tag{5}$$

where $\mathbf{X}$ is the input feature map, and $\odot$ denotes element-wise multiplication. This enhances BVP-relevant features while suppressing noise.

Temporal normalization (TN), inspired by the Shafer Reflectance Model [32], stabilizes temporal signals by removing long-term artifacts, such as illumination variations. The input signal $P_{i,j}(t)$, representing the signal at spatial position (i, j) and time t, is detrended using a moving average filter:

$$\tilde{P}_{i,j}(t) = P_{i,j}(t) - \frac{1}{W} \sum_{t'=t-W/2}^{t+W/2} P_{i,j}(t'), \tag{6}$$

where $\tilde{P}_{i,j}(t)$ is the detrended signal, W is the window size, and t' is the time index within the window. The detrended signal is then normalized to maintain consistent amplitude:

$$\hat{P}_{i,j}(t) = \frac{\tilde{P}_{i,j}(t)}{\sqrt{\frac{1}{T}\sum_{t=0}^{T}\tilde{P}_{i,j}(t)^2 + \epsilon}}, \tag{7}$$

where $\hat{P}_{i,j}(t)$ is the normalized signal, T is the signal length, and ϵ is a small constant (e.g., 10^{-8}) to avoid division by zero. This normalization enhances BVP signal stability by reducing temporal artifacts.

The Sim-TN module operates by first applying SimAM to refine spatial and channel-wise features, focusing on BVP-relevant patterns, followed by TN to normalize temporal signals. This sequential process ensures robust and accurate BVP extraction across diverse conditions, making it highly effective for real-world rPPG applications.

4. Experiments and Results

4.1. Datasets

To evaluate LightweightPhys, we used two standard datasets for remote photoplethysmography (rPPG): the PURE dataset [33] and the UBFC-rPPG dataset [34].

- PURE dataset: It includes 60 RGB videos from 10 subjects, recorded at 640 × 480 pixels and 60 FPS in a controlled laboratory setting with an ECO274CVGE camera. It features six scenarios (e.g., steady head positions, small rotations, and illumination changes),which can provide basic and controllable rPPG signal samples for model training, with ground-truth blood volume pulse (BVP) signals captured via a pulox CMS50E pulse oximeter,making it suitable for the model to learn basic physiological signal features. This dataset was used for training to optimize feature extraction.
- UBFC-rPPG dataset: It comprises 42 RGB videos from 42 subjects, recorded at 640 × 480 pixels and 30 FPS using a Logitech C920 webcam under unconstrained conditions (e.g., natural movements, ambient lighting),which can effectively evaluate the generalization ability of the model in complex environments. Ground-truth BVP signals were collected at 60 Hz with a CMS50E pulse oximeter and include richer types of noise,which can comprehensively verify the performance of the model under different noise levels. This dataset served as the test set to assess generalization in noisy scenarios.

Data preprocessing involved cropping facial regions with the MTCNN face detector [35], resizing to 128 × 128 pixels to match LightweightPhys's input ($N \times 3 \times T \times 128 \times 128$), and extracting 300-frame sequences (10 s at 30 FPS) using a sliding window (stride 0.5 s) to capture temporal context.

4.2. Experimental Setup

LightweightPhys was implemented in PyTorch 1.10.2 and trained on an NVIDIA RTX 3080 GPU (10 GB memory). It processes input tensors of size $N \times 3 \times T \times 128 \times 128$ ($N = 4$, $T = 300$, 3 RGB channels). The architecture includes nine convolutional blocks, with depthwise separable convolution (DSC) applied to blocks 2–9 ($3 \times 3 \times 1$ depthwise and $1 \times 1 \times$ M pointwise convolutions) for efficiency. The Sim-TN module combines SimAM attention ($\lambda = 10^{-4}$) and temporal normalization ($W = 60$, $\epsilon = 10^{-8}$) to enhance signal quality.

Training spanned 30 epochs using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), with an initial learning rate of 0.001 adjusted via OneCycleLR (peak 0.009). The loss function employed is the Negative Pearson Correlation Coefficient loss,which is designed to maximize the linear correlation between predicted rPPG signals and ground-truth BVP signals, defined as follows: For each sample i in a batch, the Pearson correlation coefficient r is calculated as:

$$r = \frac{N \cdot \sum(preds_i \cdot labels_i) - \sum preds_i \cdot \sum labels_i}{\sqrt{\left[N \cdot \sum preds_i^2 - \left(\sum preds_i\right)^2\right]}} \times \frac{1}{\sqrt{\left[N \cdot \sum labels_i^2 - \left(\sum labels_i\right)^2\right]}} \tag{8}$$

where $preds_i$ denotes the model-predicted rPPG signal, $labels_i$ represents the ground-truth BVP signal, and N is the length of the signal sequence (300 in this study).The total loss is computed by averaging (1 - r) over all samples in the batch:

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}(1 - r_i) \tag{9}$$

where B is the batch size (set to 4 in this study). This formulation directly optimizes the similarity between

the predicted and ground-truth signal waveforms, which is critical for accurate heart rate estimation in non-contact scenarios.

Data augmentation included random horizontal flipping (probability 0.5) and Gaussian noise ($\sigma = 0.01$). Post-processing used Fast Fourier Transform (FFT) and a 4th-order Butterworth bandpass filter (0.75–2.5 Hz, 45–150 bpm) to extract heart rates.

4.3. Evaluation Metrics

Model performance was evaluated using four metrics to compare predicted heart rates against ground-truth heart rates.

- Mean Absolute Error (MAE) : MAE repersents the average absolute difference between predicted and true heart rates (bpm).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|HR_{\text{pred},i} - HR_{\text{true},i}| \tag{10}$$

  where N is the number of samples, $HR_{\text{pred},i}$ is the predicted heart rate, and $HR_{\text{true},i}$ is the ground-truth heart rate.
- Root Mean Square Error (RMSE): It indicates the square root of average squared differences between predicted and true heart rates (bpm).

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(HR_{\text{pred},i} - HR_{\text{true},i})^2} \tag{11}$$

  where N is the number of samples, $HR_{\text{pred},i}$ is the predicted heart rate, and $HR_{\text{true},i}$ is the ground-truth heart rate.
- Mean Absolute Percentage Error (MAPE): MAPE represents the average absolute percentage difference between predicted and true heart rates (%).

$$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{HR_{\text{pred},i} - HR_{\text{true},i}}{HR_{\text{true},i}}\right| \tag{12}$$

  where N is the number of samples, $HR_{\text{pred},i}$ is the predicted heart rate, and $HR_{\text{true},i}$ is the ground-truth heart rate.
- Pearson Correlation Coefficient (R): It repersents the linear correlation between predicted and true heart rates, ranging from -1 to 1.

$$\text{R} = \frac{\sum_{i=1}^{N}(HR_{\text{pred},i} - \overline{HR}_{\text{pred}})(HR_{\text{true},i} - \overline{HR}_{\text{true}})}{\sqrt{\sum_{i=1}^{N}(HR_{\text{pred},i} - \overline{HR}_{\text{pred}})^2}} \cdot \frac{1}{\sqrt{\sum_{i=1}^{N}(HR_{\text{true},i} - \overline{HR}_{\text{true}})^2}} \tag{13}$$

  where N is the number of samples, $HR_{\text{pred},i}$ is the predicted heart rate, $HR_{\text{true},i}$ is the ground-truth heart rate, $\overline{HR}_{\text{pred}}$ is the mean predicted heart rate, and $\overline{HR}_{\text{true}}$ is the mean ground-truth heart rate.
- Computational Efficiency: It is quantified by the total number of trainable parameters (in millions, M) and floating-point operations per inference (in gigaflops, G).

4.4. Ablation Study

We conducted an ablation study on the UBFC-rPPG dataset to evaluate the individual and combined contributions of the DSC and Sim-TN modules. The results are summarized in Table 2, and we provide a detailed analysis of the impact of each module on model performance, complexity, and noise suppression capabilities.

The baseline model, which is a standard PhysNet without any DSC or Sim-TN modules, achieves an MAE of 3.253 bpm and RMSE of 8.745 bpm, with a correlation coefficient (R) of 0.880. However, this model requires 0.77 M parameters and 112.2G FLOPs, indicating a relatively high computational cost. While the baseline model performs reasonably well in terms of accuracy, its complexity makes it less suitable for resource-constrained environments.

When the DSC module is added to the baseline model, the computational complexity is significantly reduced, with the number of parameters dropping to 0.35 M and FLOPs decreasing to 40.65 G. This represents a 54.5% reduction in parameters and a 63.8% reduction in FLOPs. However, this reduction in complexity comes at the cost of accuracy, as the MAE increases to 4.220 bpm and RMSE to 9.224 bpm, with a correlation coefficient of 0.819. The degradation in performance can be attributed to the simplified feature extraction process introduced by the DSC

module, which may lose some critical information during the depthwise convolution operations.

In contrast, the addition of the Sim-TN module alone leads to a significant improvement in model accuracy. The MAE decreases to 2.483 bpm, RMSE to 5.482 bpm, and the correlation coefficient increases to 0.955. This demonstrates the effectiveness of the Sim-TN module in suppressing temporal noise, which is a common challenge in rPPG signal processing. The Sim-TN module enhances the model's ability to filter out noise artifacts, leading to more robust and accurate heart rate estimation. However, the computational complexity remains similar to the baseline model, as the Sim-TN module does not inherently reduce the number of parameters or FLOPs.

The full LightweightPhys model, which combines both the DSC and Sim-TN modules, achieves the best balance between accuracy and efficiency. With an MAE of 2.447 bpm, RMSE of 6.678 bpm, and a correlation coefficient of 0.969, it outperforms the baseline model in terms of accuracy while significantly reducing computational complexity. Specifically, the number of parameters is reduced by 54.5%, and FLOPs are reduced by 27.5% compared to the baseline. This combination allows the model to maintain high accuracy while being more suitable for deployment in resource-constrained environments.

The ablation study highlights the trade-offs between model complexity, accuracy, and noise suppression capabilities. The DSC module is effective in reducing computational complexity but at the cost of some accuracy, likely due to the loss of fine-grained features during the depthwise convolution process. On the other hand, the Sim-TN module excels in noise suppression, leading to significant improvements in accuracy, particularly in noisy environments. However, it does not inherently reduce computational complexity.

The combination of both modules in the LightweightPhys model demonstrates that it is possible to achieve a balance between efficiency and accuracy. By leveraging the strengths of both DSC and Sim-TN, the model is able to reduce complexity while maintaining high accuracy and robust noise suppression capabilities. This makes the LightweightPhys model particularly suitable for real-world applications where computational resources are limited, and accurate heart rate estimation is critical.

Table 2. Ablation Study of Sim-TN and DSC Modules on UBFC-rPPG.

Sim-TN	DSC Module	MAE (bpm)	RMSE (bpm)	MAPE (%)	Pearson (R)
✗	✗	3.253	8.745	8.064	0.880
✗	✓	4.220	9.224	8.199	0.819
✓	✗	2.483	**5.482**	4.162	0.955
✓	✓	**2.447**	6.678	**3.414**	**0.969**

Note: Bold values indicate the optimal performance among all ablation groups. ✓ denotes that the corresponding module (Sim-TN/DSC) is included in the model, while ✗ denotes that the module is excluded.

4.5. Comparative Results

To evaluate the performance of LightweightPhys, we compared it against both traditional and deep learning-based rPPG methods on the UBFC-rPPG dataset, with results summarized in Tables 3 and 4. Traditional methods, such as CHROM [36] and POS [37], rely on chrominance-based signal processing and achieve MAEs of 3.98 bpm and 4.10 bpm, respectively, with RMSEs of 8.72 bpm and 7.61 bpm. Their MAPE values range from 3.78% to 4.03%, and Pearson correlation coefficients (R) fall between 0.88 and 0.92. While these methods are computationally lightweight, their reliance on handcrafted features limits robustness in noisy environments, particularly under varying illumination or motion artifacts. In contrast, LightweightPhys achieves a significantly lower MAE of 2.45 bpm, RMSE of 6.68 bpm, MAPE of 3.41%, and an R of 0.97, demonstrating superior accuracy and robustness while maintaining computational efficiency.

Among deep learning approaches, PhysNet [19] serves as a baseline, with an MAE of 3.25 bpm, RMSE of 8.74 bpm, MAPE of 8.06%, and R of 0.88, but it incurs high computational costs (0.77M parameters, 112.2 G FLOPs). PhysFormer [22], a transformer-based model, requires 7.38 M parameters and 81.0G FLOPs, making it impractical for resource-constrained devices. PhysMamba [24] achieves the best accuracy with an MAE of 1.20 bpm, RMSE of 5.99 bpm, MAPE of 0.97%, and R of 0.95, using 0.56 M parameters and 94.6G FLOPs. LightweightPhys closely rivals PhysMamba's performance, with only a 1.25 bpm higher MAE and 0.69 bpm higher RMSE, while offering substantial computational savings: 37.5% fewer parameters (0.35 M) and 57% fewer FLOPs (40.65 G). This efficiency stems from the integration of depthwise separable convolutions (DSC), which reduce feature extraction complexity, and the Sim-TN module, which enhances robustness by suppressing noise through attention-based feature refinement and temporal normalization.

The balance of accuracy and efficiency in LightweightPhys makes it particularly well-suited for real-time applications on edge devices. On an NVIDIA RTX 3080, it processes 30 FPS video streams with a per-frame latency

of less than 33 ms, meeting real-time requirements. Compared to PhysNet, LightweightPhys not only reduces computational overhead by 54.5% in parameters and 63.8% in FLOPs but also improves accuracy across all metrics. While PhysMamba's slightly higher precision is notable, its higher computational demands limit its feasibility in low-power environments, such as mobile health monitors or embedded systems. LightweightPhys, by contrast, maintains near state-of-the-art performance with a fraction of the resources, enabling applications like remote health monitoring, where non-contact heart rate estimation is critical, and face anti-spoofing, where rPPG signals can verify liveness against Deepfake attacks.

Qualitatively, LightweightPhys excels in noisy, unconstrained scenarios, as evidenced by its high Pearson correlation (R = 0.97), which indicates strong linear agreement with ground-truth heart rates. However, a slight accuracy gap with PhysMamba suggests that DSC's simplified feature extraction may occasionally miss subtle signal details in highly complex scenarios. Future enhancements could explore hybrid architectures combining standard and separable convolutions to further close this gap while preserving efficiency. Overall, LightweightPhys represents a practical and robust solution for rPPG signal extraction, offering a compelling trade-off for real-world deployment.

Table 3. Performance Comparison on UBFC-rPPG.

Method	MAE (bpm)	RMSE (bpm)	MAPE (%)	R
CHROM [36]	3.98	8.72	3.78	0.88
POS [37]	4.10	7.61	4.03	0.92
PhysNet [19]	3.25	8.74	8.06	0.88
PhysMamba [24]	1.20	5.99	0.97	0.95
LightweightPhys	2.45	6.68	3.41	0.97

Table 4. Computational Complexity Comparison.

Method	Parameters (M)	FLOPs (G)
PhysNet [19]	0.77	112.2
PhysFormer [22]	7.38	81.0
PhysMamba [24]	0.56	94.6
LightweightPhys	0.35	40.65

LightweightPhys achieves competitive accuracy with a mean absolute error (MAE) of 2.45 bpm and a root mean square error (RMSE) of 6.68 bpm on the UBFC-rPPG dataset, while maintaining minimal computational cost. This performance outperforms the baseline PhysNet and is comparable to PhysMamba in noisy environments, demonstrating its robustness under challenging conditions.

The Simulated Temporal Noise Suppression (Sim-TN) module plays a critical role in enhancing the clarity of the Blood Volume Pulse (BVP) signal through its attention mechanism, which effectively suppresses noise and artifacts. This ensures the model's robustness against environmental disturbances such as motion and illumination variations. Additionally, the Depthwise Separable Convolution (DSC) module significantly reduces redundancy in the network, enabling real-time deployment on edge devices with limited computational resources.

The potential applications of LightweightPhys extend beyond health monitoring to include face anti-spoofing, where efficient rPPG signal extraction can be used to verify liveness and enhance security in biometric systems. However, a limitation of the model is a slight accuracy trade-off when compared to PhysMamba, particularly in scenarios with extremely high noise levels. This suggests that future work could explore hybrid architectures that combine the strengths of LightweightPhys and PhysMamba to further improve performance without compromising computational efficiency.

5. Conclusions

In this paper, we introduced LightweightPhys, an efficient and robust solution for Remote Photoplethysmography (rPPG) signal extraction, designed to address the challenges of computational complexity and noise sensitivity in traditional and deep learning-based methods. By integrating Depthwise Separable Convolution (DSC) and a novel Simulated Temporal Noise Suppression (Sim-TN) module, LightweightPhys significantly reduces computational overhead, achieving 0.35 million parameters and 40.65 GFLOPs, while maintaining high accuracy with a mean absolute error (MAE) of 2.45 bpm on the UBFC-rPPG dataset.

The Sim-TN module enhances the robustness of the model by effectively suppressing temporal noise and artifacts, ensuring reliable performance in noisy environments. Meanwhile, the DSC module reduces redundancy in the network, enabling real-time deployment on resource-constrained platforms such as mobile devices and wearable gadgets. LightweightPhys not only outperforms traditional rPPG methods but also competes with advanced deep learning models like PhysNet and PhysMamba, particularly in challenging conditions.

Future work should aim to optimize the trade-off between accuracy and computational cost and explore new use cases to maximize the potential of rPPG technology.

Author Contributions

Y.L. (Yu Liu): methodology, data curation, writing—original draft preparation; Y.L. (Yinqiao Li), Y.H.: visualization, investigation; T.W.: conceptualization, methodology, supervision; Z.Z.: writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This study was funded by National Natural Science Foundation of China under Grants 62477017, 62372333, 62277021 and 62277029, Fundamental Research Funds for the Central Universities, Central China Normal University, under Grants CCNU22JC011 and CCNU22QN013, Open Fund of National Engineering Research Center of Geographic Information System, China University of Geosciences, under Grant 2023KFJJ07, the Fundamental Research Funds for the Central Universities under Grant 2042023kf0135, the Key Research and Development Program of Hubei Province under Grant 2023BAB078, the Project funded by China Postdoctoral Science Foundation under Grant 2022M722459, and the Knowledge Innovation Program of Wuhan-Basic Research under Grant 2023010201010063.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data supporting the findings of this study are derived from third-party publicly available datasets. Specifically, the PURE dataset [33] can be accessed through academic repositories associated with its original publication, and the UBFC-rPPG dataset [34] is obtainable via its official distribution channels. The availability of the aforementioned datasets is subject to the relevant regulations of their original providers. The original contributions presented in this study are included in the article, and additional data related to the model training process (such as preprocessed video frames and training logs) can be obtained from the corresponding author (Tao Wang, email: tmac@ccnu.edu.cn) upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Poh, M.Z.; McDuff, D.J.; Picard, R.W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **2010**, *18*, 10762–10774.
2. Wu, H.Y.; Rubinstein, M.; Shih, E.; et al. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.* **2012**, *31*, 1–8.
3. Burzo, M.; McDuff, D.; Mihalcea, R.; et al. Towards sensing the influence of visual narratives on human affect. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012; pp. 153–160.
4. Hsu, G.S.; Ambikapathi, A.; Chen, M.S. Deep learning with time-frequency representation for pulse estimation from facial videos. In Proceedings of the 2017 IEEE international joint conference on biometrics (IJCB), Denver, CO, USA, 1–4 October, 2017; pp. 383–389.
5. Prakash, S.K.A.; Tucker, C.S. Bounded Kalman filter method for motion-robust, non-contact heart rate estimation. *Biomed. Opt. Express* **2018**, *9*, 873–897.

6. Zhu, J.; Ji, L.; Liu, C. Heart rate variability monitoring for emotion and disorders of emotion. *Physiol. Meas.* **2019**, *40*, 064004.
7. Sabour, R.M.; Benezeth, Y.; Marzani, F.; et al. Emotional state classification using pulse rate variability. In Proceedings of the 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 19–21 July 2019; pp. 86–90.
8. Macwan, R.; Benezeth, Y.; Mansouri, A. Heart rate estimation using remote photoplethysmography with multi-objective optimization. *Biomed. Signal Process. Control* **2019**, *49*, 24–33.
9. Zhang, J.; Zheng, K.; Mazhar, S.; et al. Trusted emotion recognition based on multiple signals captured from video. *Expert Syst. Appl.* **2023**, *233*, 120948.
10. Hsu, Y.; Lin, Y.L.; Hsu, W. Learning-based heart rate detection from remote photoplethysmography features. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4433–4437.
11. Osman, A.; Turcot, J.; El Kaliouby, R. Supervised learning approach to remote heart rate estimation from facial videos. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015.
12. Nardelli, M.; Valenza, G.; Greco, A.; et al. Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Trans. Affect. Comput.* **2015**, *6*, 385–394.
13. Harper, R.; Southern, J. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. *IEEE Trans. Affect. Comput.* **2020**, *13*, 985–991.
14. Zhou, K.; Schinle, M.; Stork, W. Dimensional emotion recognition from camera-based PRV features. *Methods* **2023**, *218*, 224–232.
15. Fan, H.; Zhang, X.; Xu, Y.; et al. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. *Inf. Fusion* **2024**, *104*, 102161.
16. Zhao, C.; Cao, P.; Hu, M.; et al. WTC3D: An Efficient Neural Network for Noncontact Pulse Acquisition in Internet of Medical Things. *IEEE Trans. Ind. Inform.* **2025**, *21*, 1547–1556.
17. Zhang, X.; Xia, Z.; Liu, L.; et al. Demodulation Based Transformer for rPPG Generation and Heart Rate Estimation. *IEEE Signal Process. Lett.* **2023**, *30*, 1042–1046.
18. Chen, W.; McDuff, D. Deepphys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 349–365.
19. Yu, Z.; Li, X.; Zhao, G. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv* **2019**, arXiv:1905.02419.
20. Liu, X.; Fromm, J.; Patel, S.; et al. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19400–19411.
21. Narayanswamy, G.; Liu, Y.; Yang, Y.; et al. Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2024; pp. 7914–7924.
22. Yu, Z.; Shen, Y.; Shi, J.; et al. Physformer: Facial video-based physiological measurement with temporal difference transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4186–4196.
23. Joshi, J.; Cho, Y. IBVP dataset: RGB-thermal RPPG dataset with high resolution signal quality labels. *Electronics* **2024**, *13*, 1334.
24. Luo, C.; Xie, Y.; Yu, Z. PhysMamba: Efficient Remote Physiological Measurement with SlowFast Temporal Difference Mamba. In Proceedings of the Chinese Conference on Biometric Recognition, Nanjing, China, 22–24 November 2024; pp. 248–259.
25. Zhuang, J.; Chen, Y.; Zhang, Y.; et al. FastBVP-Net: a lightweight pulse extraction network for measuring heart rhythm via facial videos. *arXiv* **2022**, arXiv:cs.CV/2206.12558.
26. Liu, X.; Hill, B.; Jiang, Z.; et al. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 5008–5017.
27. Botina-Monsalve, D.; Benezeth, Y.; Miteran, J. RTrPPG: An Ultra Light 3DCNN for Real-Time Remote Photoplethysmography. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 2145–2153.
28. Zhao, C.; Cao, P.; Xu, S.; et al. Pruning rPPG Networks: Toward Small Dense Network with Limited Number of Training Samples. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 2054–2063.
29. Zhao, C.; Zhang, S.; Cao, P.; et al. Pruning remote photoplethysmography networks using weight-gradient joint criterion. *Expert Syst. Appl.* **2025**, *282*, 127623.

30. Howard, A.G.; Zhu, M.; Chen, B.; et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
31. Yang, L.; Zhang, R.Y.; Li, L.; et al. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International conference on machine learning, PMLR, Virtual, 18–24 July 2021; pp. 11863–11874.
32. Wang, K.; Tang, J.; Wei, Y.; et al. A Plug-and-Play Temporal Normalization Module for Robust Remote Photoplethysmography. *arXiv* **2024**, arXiv:eess.IV/2411.15283.
33. Stricker, R.; Müller, S.; Gross, H.M. Non-contact video-based pulse rate measurement on a mobile service robot. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25–29 August 2014; pp. 1056–1062.
34. Bobbia, S.; Macwan, R.; Benezeth, Y.; et al. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.* **2019**, *124*, 82–90.
35. Zhang, K.; Zhang, Z.; Li, Z.; et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503.
36. De Haan, G.; Jeanne, V. Robust pulse rate from chrominance-based rPPG. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2878–2886.
37. Wang, W.; Den Brinker, A.C.; Stuijk, S.; et al. Algorithmic principles of remote PPG. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 1479–1491.