Physiological Information-Guided Network for Heart Rate Estimation from Near-Infrared Facial Video

Shuiping Gou*, *Senior Member, IEEE*, Kehong Liu*, *Student Member, IEEE*, Hantao Zhao, Nuo Tong *Member, IEEE*, Zhang Guo†, *Member, IEEE*, Wenbo Liu†, Qigong Sun†, Licheng Jiao, *Fellow, IEEE*

***Abstract*—Remote photoplethysmography (rPPG) is an video-based non-contact method for estimating heart rate (HR), which is crucial for reflecting physical health condition and preventing cardiovascular diseases. However, the majority of the existing non-contact HR estimation methods are constrained by the insufficient representation of heartbeat dynamics and temporal rhythm patterns, which restricts their abilities to capture the subtle variations and complexities of HR signals across different temporal scales. To address these issues, we propose a Physiological Information-Guided Neural Network (PhysioNN), which employs a beat-based attention module and a rhythm-based attention module to extract multi-scale temporal features from rPPG signals. Specifically, the Beat-Rhythm joint Attention Mechanism (BRAM) enables the model to focus on both edge features and sequential features of the heartbeat signal. The Physiological Feature Guided Regularization (PFGR) incorporates domain-specific physiological characteristics into the loss function during training, facilitating the model to maintain consistency with actual heartbeat signals. In addition, we build a well-organized database XDU-SenseTime, including 374 video data from 17 subjects. Our database includes various variations, establishing a less-constrained scenarios of HR estimation. Comprehensive experiments on public dataset MR-NIRP and private dataset XDU-SenseTime show that the proposed PhysioNN can achieve more accurate HR measurement.**

***Index Terms*—Remote Photoplethysmography (rPPG), Heart Rate Estimation, Physiological Information, Near-Infrared.**

I. Introduction

HEART rate (HR) is one of the most important signs of the human body, reflecting both physiological and mental conditions. HR measurement is crucial for various applications in continuous health monitoring and sleeping monitoring, especially during nighttime conditions. Traditional contact-based HR measurement requires direct contact with the body, difficult for long time monitoring or under certain circumstances (e.g., for newborns, burn patients, or drivers). Remote photoplethysmography (rPPG) is a video-based HR monitoring method that can detect pulsatile information caused by heart activity from subtle, invisible color changes in exposed skin, even from a distance [1], [2]. This method provides a non-contact and convenient way to estimate HR, attracting increasing attention from researchers [3]–[5].

However, the pulsatile signals are very weak and easily contaminated by environmental noise, typically caused by lighting variations and motion artifacts [6], [2]. To address these issues in real-world HR measurement, some methods have been proposed.

There are many non-contact HR measurement methods, one of these method is radar-based technologies, which includes frequency modulated continuous wave (FMCW) radar [7], [8] and pulse measurement radar [9], [10]. These researches extract HR signals by monitoring the small chest vibrations caused by heartbeat. However, they rely on expensive hardware and suffer from limitations in real-time performance and accuracy. Because of these limits, researchers have also started exploring the Red-Green-Blue (RGB) and NIR camera-based for HR estimation, as they provide a more accessible way to extract physiological signals.

RGB-based methods extract rPPG signals by analyzing subtle changes in facial skin color and achieve well performance under normal ambient lighting conditions [11]–[13]. However, RGB light primarily interacts with the reflected light from shallow skin layer, making it susceptible to skin tone difference, lighting changes, and motion artifacts, which leads to unstable rPPG signals. Moreover, RGB cameras are not suitable for night-time or low-light monitoring due to its low visibility in dark conditions.

In contrast, NIR light can integrate various wavelengths from the infrared spectrum and has greater penetration depth [14], allowing it to capture more stable physiological signals from subcutaneous tissues in low-light conditions, while exhibiting stronger robustness to lighting variations and skin tone differences [15], [16]. Therefore, many researchers start to apply NIR cameras as the equipment for rPPG signal acquisition.

In earlier studies, some signal-based HR measurement methods have been proposed. Zhang *et al.* use facial regions of interest (ROI) to extract rPPG signals and estimate HR through frequency analysis [17]. Nowara *et al.* and Cheng *et al.* employed signal decomposition and space transformation to extract rPPG signals [18], [19]. However, these methods are limited by assumptions like skin reflection model and linear combination of noise.

Recently, with the impressive performance of deep learning across various tasks, significant progress has been made in

This work was supported in part by the National Natural Science Foundation under Grant No. 62301395 and No. 62372358; in part by the Key Research and Development Program of Shaanxi under Grant No. 2024SF2-GJHX-35; and in part by the Shaanxi Province Postdoctoral Science Foundation under Grant No. 2023BSHEDZZ177. (*Corresponding authors: Zhang Guo, Wenbo Liu, and Qigong Sun; Co-first authors: Shuiping Gou, Kehong Liu*).

Shuiping Gou, Kehong Liu, Hantao Zhao, Nuo Tong, Zhang Guo and Licheng Jiao are with the Key Laboratory of Intelligent Perception and Image Understanding (IPIU), Xidian University, Xidian University, 710126, China (e-mail: shpgou@mail.xidian.edu.cn; kehongliu@stu.xidian.edu.cn; hantaozhao@stu.xidian.edu.cn; nuotong@xidian.edu.cn; guozhang@xidian.edu.cn; lchjiao@mail.xidian.edu.cn).

Wenbo Liu is with the WeiFang People's Hospital, Shandong Second Medical University, 261041, China (email: boboliucn@126.com).

Qigong Sun is with the Shanghai AI Lab and Applied Research Laboratory of SenseTime, 202172, China (email: sunqigong@sensetime.com).

HR estimation methods using deep models [20]–[22]. Earlier approaches use 2D convolution neural network (2D-CNN) to directly esimtate HR, which primarily focus on spatial learning without considering the temporal information [12]. Afterthat, Yu *et al.* [11] applied a spatial-temporal 3D-CNN to extract rPPG signals from both spatial and temporal dimensions, enabling more effective learning of the dynamic changes in facial features over time. Subsequently, to improve the representation learning, some approaches introduce complex strategies to separate real physiological signals from the noises, such as cross-verified feature disentangling [23] or dual-GAN [24]. Moreover, some transformer-based approaches have been employed to capture spatial-temporal features of rPPG signal and shown promising results for addressing the HR measurement task [25], [26].

While these approaches, including both signal-based and deep learning-based models, can estimate HR to some extent, the majority of the existing methods are constrained by the insufficient representation of heartbeat dynamics and temporal rhythm patterns, which restricts their abilities to capture the subtle variations and complexities of HR signals across different temporal scales.

Motivated by the discussions above, a robust HR measurement framework is proposed, called *PhysioNN* (Physiological Information-guided Neural Network), to effectively capture and analyze physiological signals from NIR facial videos. In this network, we design a *BRAM* (*B*eat-*R*hythm joint *A*ttention *M*echanism) module and *PFGR* (*P*hysiological *F*eature *G*uided *R*egularization) module to obtain a more stable and accurate HR estimation framework. Specifically, BRAM integrates *heartbeat-based* and *rhythm-based* attention modules, allowing the model to progressively focus on the edge features and sequential features of physiological signals; PFGR introduces the domain-specific physiological characteristics into loss functions during training to maintain consistency with actual heartbeat signals in terms of shape, trend, heartbeat count, intervals, and energy distribution. Furthermore, to better train our deep learning model, we additionally construct a well-organized database, called *XDU-SenseTime*, including 374 video data from 17 subjects. This database includes various noises, such as head movements and lighting changes, establishing a less-constrained scenarios of HR estimation.

The main contributions are summarized as follows:

- We propose a novel physiological information-guided network to improve HR estimation from NIR facial videos, which effectively addresses the limitations of insufficient physiological features across different temporal scales.
- The hierarchical integration of beat-based and rhythm-based attention modules enhances the model's robustness, enabling it to accurately capture heartbeat dynamics and temporal rhythm patterns of the heartbeat signal.
- Incorporating physiological knowledge into the training loss functions ensures alignment with real heartbeat signals in terms of shape, trend, heartbeat count, intervals, and energy distribution.
- Comprehensive experimental results demonstrate that PhysioNN exhibits superior performance and robustness over existing methods across various challenging scenarios.

The remainder of this paper is organized as follows. In Section II, we overview related work on vision-based physiological measurements. In Section III, we introduce the XDU-SenseTime database. Section IV describes the specific architectural details of the proposed PhysioNN framework. Section V analyzes the experimental results on two NIR datasets to demonstrate the superior performance of PhysioNN. Finally, we conclude this work in Section VI.

II. Related Work

This section provides a comprehensive review of the current methods for HR measurement using rPPG technology, as well as the existing public HR datasets. Moreover, the strengths and weaknesses of these methods are thoroughly analyzed and discussed.

A. Signal Processing-based Methods

Existing signal-based non-contact HR monitoring methods can be broadly categorized into RGB-based and NIR-based approaches. These methods utilize different types of optical signals to capture visible light reflections from the skin.

RGB cameras can detect subtle variations in skin tone caused by heartbeat by utilizing the rich color information. Recent studies have shown that these RGB camera-based methods have achieved good performance in non-contact HR monitoring task under well-lit environment [11], [12]. However, RGB cameras are sensitive to lighting changes and individual differences, making them unsuitable for night-time or low-light scenarios [17]. On the other hand, NIR cameras are robust to illumination variance and individual difference due to their abilities to penetrate deeper into the skin layers [27], thus offer a clear image even in poor lighting situations and full darkness.

Therefore, some non-contact HR estimation methods have used NIR cameras to capture physiological signals. In 2014, Jeanne *et al.* used NIR cameras to estimate HR under dynamic lighting conditions [28]. Van Gastel *et al.* demonstrated the feasibility of motion-robust HR estimation by simultaneously recording video images from multiple narrow band NIR wavelengths [16]. Magdalena *et al.* proposed a new denoising algorithm, named SparserPPG, based on robust principal component analysis (PCA) and sparse frequency spectrum estimation for HR estimation in driving environments [18].

According to current research, many HR estimation algorithms designed for RGB videos can also be applied to NIR videos. In 2008, Verkruysse *et al.* discovered that HR signals could be extracted from RGB facial videos captured by standard digital cameras under stable lighting conditions [29]. After that, Chen *et al.* recorded NIR facial videos using a webcam and successfully extracted HR by applying the similar methods [30]. Their studies found that the HR obtained from NIR videos were more accurate than those from RGB

videos under significant illumination variations, demonstrating the robustness of NIR cameras. Subsequently, Wu *et al.* applied Eulerian Video Magnification (EVM) technology on RGB videos to estimate HR by visualizing the invisible blood flow [31]. He *et al.* combined NIR cameras with EVM technology to measure HR and respiratory rate under dark conditions [32]. Moreover, Chen *et al.* successfully separated the environmental lighting noise by applying the Ensemble Empirical Mode Decomposition (EEMD) method to the rPPG signals of RGB videos. [33]. Zhang *et al.* used Empirical Mode Decomposition (EMD) techniques to extract rPPG signals from NIR videos under complex lighting conditions, and also conducted related experiments in in-car scenarios [34]. Cheng *et al.* proposed DCT-JBSS [19] to evaluate HR by combining Joint Blind Source Separation (JBSS) and Delay Coordinate Transformation (DCT) techniques.

The aforementioned signal processing-based methods demonstrate facial videos captured by NIR cameras can effectively extract HR information, and compared to standard RGB cameras, they offer superior stability and robustness to lighting variations.

B. Deep Learning-based Methods

In recent years, many researchers have used deep learning methods to address the challenges in remote HR estimation based on facial video.

Many RGB-based deep learning methods have been proposed. Chen *et al.* first combined CNN with attention mechanism to map video frames to physiological signals [35]. In the same year, Spetlik *et al.* introduced a two-step CNN method for HR estimation using facial images [12]. In 2019, Yu *et al.* developed a 3D-CNN model PhysNet leveraged spatiotemporal features to reconstruct rPPG signals from facial videos [11]. After that, Yu *et al.* also introduced a spatial-temporal video enhancement network and applied rPPGNet for HR estimation [36]. To suppress noise while preserving periodic physiological patterns, Niu *et al.* processed video sequences into MSTmaps and introduced a cross-verified feature disentangling strategy [23].

Beyond environmental and data noise, network structure significantly affects prediction accuracy. Liu *et al.* combined 2D and 3D-CNN with a spatial attention module to enhance temporal modeling [13]. They then introduced two efficient neural networks optimized for mobile and edge devices, balancing accuracy and computational cost [37]. Zou *et al.* explored a dynamic sparse attention mechanism to capture periodic features across multiple time scales [26].

Many deep learning approaches for NIR-based HR estimation have also been explored. Nowara *et al.* proposed an inverse attention network [38] to learn how to remove lighting noise and apply information from the facial regions in NIR videos. In order to denoise and enhance the extracted rPPG signal, Magdalena *et al.* used NIR light sources and sparse frequency spectrum estimation techniques to improve the accuracy of HR monitoring [18]. After that, Zhang *et al.* applied robust principal component analysis (RPCA) technique [17] to extract high-quality rPPG signals from facial regions in NIR videos by applying multi-region selection. Recently, transformer models have shown great potential in capturing global dependencies and modeling long-term temporal dynamics. Park *et al.* proposed a self-supervised RGB-NIR fusion transformer, called Fusion ViViT, to extract long-range local and global spatiotemporal features to enhance rPPG signal representation. Liu *et al.* proposed an information enhancement network to improve HR estimation accuracy and robustness by enhancing contextual, modality difference, and frequency information.

In summary, previous methods focused on improving spatiotemporal representation of facial videos but lack sufficient representation of heartbeat dynamics and temporal rhythm patterns, thus reducing the accuracy of HR estimation. In this work, we leverage features across different time scales to model rPPG signals, further enhancing the accuracy and robust of HR estimation.

C. Public Database for Remote HR Estimation

Due to the high cost of the equipment and the strict environmental requirements, building database for remote HR estimation is a challenging task. Existing publicly available NIR datasets are relatively limited. The MR-NIRP database [39] was first used in [18] for remote HR estimation using NIR videos. This database consists of 16 RGB videos and corresponding 16 NIR videos from 8 subjects. In 2018, Xiaobai *et al.* [40] introduced the OBF database, designed specifically for HRV analysis, which includes RGB and corresponding NIR videos and all scenarios in this database are well-controlled. Both of these datasets were recorded in well-lit lab conditions with subjects sit still in front of the camera. The VIPL-HR database [41] was introduced by Niu *et al.* in 2020, which has 6 different scenarios including head movements and lighting changes. However, the NIR videos in this dataset have low resolution and poor visibility.

Due to the limitations in dataset quality, MR-NIRP database is used in this study. We summarize the publicly available databases for remote HR estimation above, and all the existing databases are limited in either the recording scenarios or the video quality. Therefore, a well-organized database collected under less-constrained scenarios is needed for studying remote HR estimation methods in practice.

III. XDU-SenseTime Database

A. Device Setup and Data Collection

Our face video data is collected under different lighting scenarios, including low light and bright light to simulate day and night conditions with varying light levels in order to replicate real-world application scenarios.

The recording environmental setup is illustrated in Fig. 1. Videos are collected in both well-lit and dim indoor environments using RGB and NIR cameras SenseThunder mini and SenseN1 Monitor developed by Xi'an SenseTime Technology, with a resolution of 1280×720. In well-light conditions, the ceiling LED light is turned on to simulate daylight illumination. To simulate the low-light (dark) conditions, we turn

TABLE I
SPECIFICATIONS OF THE RECORDING DEVICES USED IN XDU-SENSETIME DATABASE.

Device	Specification	Setting	Output
Computer	Dell inspiration 15	Windows 10 OS	N/A
RGB Camera	SenseThunder mini	24fps 1280×720 color camera	RGB videos
NIR Camera	SenseN1 Monitor	24fps 1280×720 NIR camera	NIR videos
BVP recorder	CONTEC CMS50E	N/A	HR, SPO2 and BVP signals
lamp	N/A	50Hz	N/A

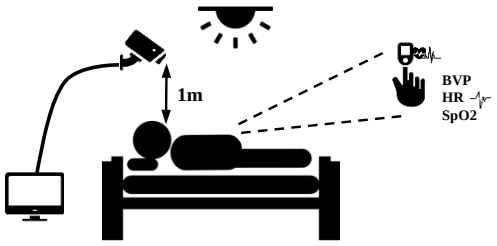


Fig. 1. Device and setup used in XDU-SenseTime database.

TABLE II
DETAILS OF THE FIVE RECORDING SCENARIOS IN THE XDU-SENSETIME DATABASE.

Scenario	Head Movement	Illumination	Exercise
1	S	B/D	No
2	SS	B/D	No
3	SR	B/D	No
4	MR	B/D	No
5	T	B/D	No
6	S	B/D	Yes

S = Still, SS = Signal Side, SR = Small Rotation, LR = Large Rotation, T = Talking, B = Bright Environment, D = Dim Environment

off the LED lamp and draw the curtains on to make sure the collect environment is dark enough.

In order to collect diverse facial angle variations of the subjects, we first ask the subjects lie flat on a simple bed with their face stay still, then encourage them to perform small and large facial rotations to simulate real-life scenarios. In order to cover a wider range of HR in our XDU-SenseTime dataset, we ask each subject to do some physical exercises before returning for dataset recording. The camera is positioned 1 meter directly above the subjects' head to record face videos. The dataset includes facial videos and captures physiological data such as HR, BVP curves, and SPO2, with labels synchronously recorded by CONTEC CMS50E fingertip pulse oximeter. The details of theses recording scenarios and device specifications can be found in Table III-A and Table I.

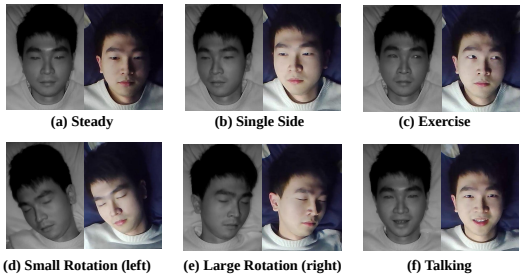


Fig. 2. Example video frames of XDU-SenseTime database, which has 6 different scenarios including NIR and the corresponding RGB facial videos.

B. Database Statistics

The SenseTime17ID dataset contains a total of 187 NIR videos and 187 RGB videos from 17 subjects, with 16 males and 1 female, all aged between 20 and 40 years old. Each video is recorded with a length of about 70s and a frame rate of 24 fps. The data collection environmental temperature is 24±3°C, with each video consisting of approximately 1680 frames. The example video frames of subject case7 captured by different conditions and devices are shown in Fig. 2.

To quantitatively demonstrate the lighting changes in XDU-SenseTime dataset, we have calculated the average gray-scale intensity of facial areas in all videos. The results are shown in Fig. 3, where we can see that the average gray-scale intensity varies from 42 to 207, covering complex lighting conditions including low light and bright light.

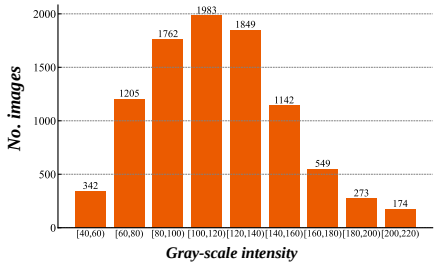


Fig. 3. The histogram of the average image intensity (gray-scale) for the videos recorded under the illumination-specific situations in XDU-SenseTime.

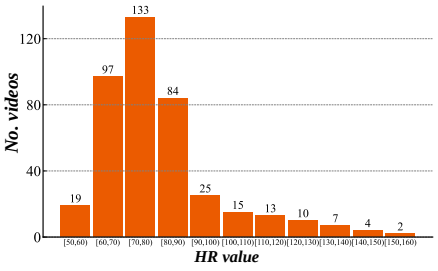


Fig. 4. The histogram of the ground-truth HR distribution in XDU-SenseTime.

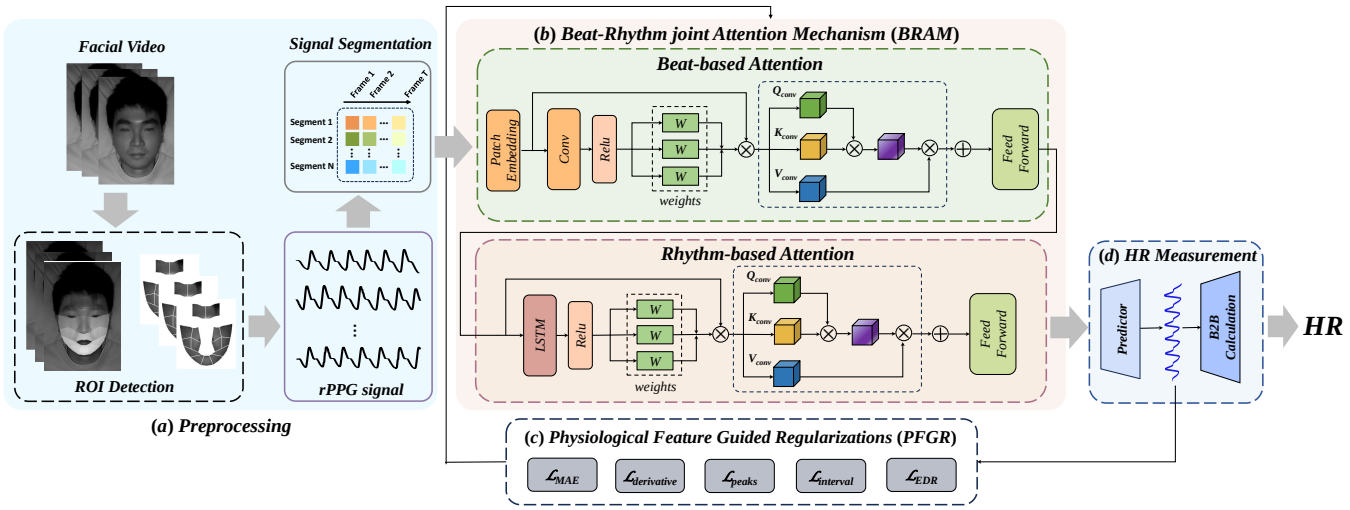


Fig. 5. Framework overview of the proposed PhysioNN for HR measurement. It contains four main components. (a) The *Preprocess* module to eliminate the irrelevant visual noises in the facial video. (b) The *BRAM* module allows the model to progressively focus on the edge features and sequential features of physiological signals by employing beat-based and rhythm-based attention modules. (c) the *PFGR* module introduces domain-specific physiological characteristics into loss functions during training to maintain consistency with the real heartbeat signal. (d) The *HR Measurement* module outputs the target rPPG signal and calculates the corresponding HR value by using B2B calculation.

Fig. 4 shows the histogram of ground truth HR, from which we can see that the real HR in XDU-SenseTime dataset distributed from 45 to 157, covering a broad spectrum of HR values. The wide range of HR in XDU-SenseTime mitigates the gap between the HR range in dataset and the HR distribution appears in daily-life scenarios. Thus can better simulate the diversity of HR in populations. In addition, the wide range of HR distribution allows for more comprehensive training, enabling models to capture a wider range of HR patterns.

IV. Method

This section describes the details of the proposed PhysioNN framework, as shown in Fig. 5. It consists of four modules: the *Preprocessing* module, the *Beat-Rhythm joint Attention Mechanism* (BRAM) module, the *Physiological Feature Guided Regularization* (PFGR) module, and the *HR Measurement* module.

Before being fed into the network, the original data from the NIR facial videos must be processed by the preprocessing module (in Section IV-A), which detects and extracts ROIs, and segments the facial signals. Then the signal segments are fed into the BRAM module to accurately capture the short-term and long-term features of the heartbeat signals (in Sections IV-B1 and IV-B2). Next, the output signals from the BRAM module are further constrained by multiple regularization terms to maintain the consistency with actual heartbeat signals (in Section IV-C). Finally, we apply the Beat-to-Beat (B2B) calculation method to evaluate the precise HR values (in Section IV-D).

A. Preprocessing

*1) **Detect ROIs**:* Our method achieves HR estimation by detecting and analyzing subtle and periodic changes in facial blood flow caused by the cardiac cycle [42]. Typically, these changes appear most obvious on specific facial areas, called ROIs, such as the forehead, two cheeks, and chin areas, where blood vessels are highly concentrated and relatively less affected by head movements [42]–[44]. We thus detect these ROIs to capture and analyze facial signals. The following **Alg**. 1 presents the main flow of ROI detection.

Note that, even though the subject is typically instructed to stay still during remote physiological measurements, small movements of human head are still inevitable due to natural behaviors such as breathing and blinking. Therefore, we need to separate movement noises from the images to improve signal quality.

Given a facial video $V \in \mathbb{R}^{T\times H\times W}$, where T, H and W are frame length, height and width, respectively. We first obtain facial key points from every image, then calculate the optical flow equation based on image gradient, and get the new positions of every facial key points despite the head movements.

*2) **Extract and Segment Pulse Signals**:* In remote HR measurement, preprocessing the color intensity signal derived from facial ROIs (detrending, normalization, filtering, and denoising) improves signal quality and reduces the impact of noise. Given the input signal $s(t)$, The specific process is as follows:

- Detrend the input signal: $s_d(t) = s(t) - p(t)$, where, $s_d(t)$ is the detrended signal, $p(t)$ is the fitted polynomial, which can be formulated as $p(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_d t^d$. This can remove baseline drift as well as some linear or nonlinear trends within the signal.
- Normalize the signal using Z-score: $s_z(t) = \frac{s_d(t)-\mu}{\sigma}$, where, μ is the mean of the $s_d(t)$, and σ is the standard deviation of the $s_d(t)$. This transforms the data into a standard normal distribution, thus remove the effects

Algorithm 1 Optical Flow-based ROI detection

Input: video $V = \{f_1, f_2, \cdots, f_T\}$ with T frames
Output: ROI region F

f_1 is selected as the initial frame
for $t = 1$ to T **do**
  $p_{t-1} = CD(f_{t-1})$ // obtain key points using corner detection
  $u, v = f_t$ // u and v are the horizontal and vertical components of the optical flow
  $I_x \times u + I_y \times v + I_t = 0$ // I is image gradient, optical flow equation
  $p_t(x', y') = p_{t-1}(x, y) + (u, v)$ // calculate key points new positions
end for
$F_t = Crop(f_t, p_t)$ // crop ROI from the whole frame
$F = \sum_{t=1}^{T} F_t$
return F

caused by factors such as lighting variations, noise, and motion artifacts.

- Filter the signal using Band-Pass filter: $s_f(t) = Filter(s_z(t), f_{low}, f_{high})$, where, $s_f(t)$ is the filtered signal, $Filter()$ represents the Band-Pass Filter, f_{low} and f_{high} represents the corresponding lower and upper cutoff frequencies of the filter. In this paper, we use a frequency range of 0.75 Hz to 4 Hz, corresponding to a HR of 45 to 240 bpm.
- Denoise the signal using wavelet transform: $s_n(t) = \mathcal{W}^{-1}(\hat{c}_i(t))$, $s_n(t)$ is the denoised signal, $\mathcal{W}^{-1}$ is the inverse wavelet transform operation, and $\hat{c}_i$ represents the denoised multiscale wavelet coefficients. The wavelet transform is employed to denoise the signal, producing a clean, noise-free rPPG signal.
- Feature Preparation: Given the input signal $X \in \mathbb{R}^L$, where L is the total length of the rPPG signal, we corp the signal into M segments, which can be represented as $x \in \mathbb{R}^{M \times T}$, where T is the length of each segment.

B. Beat-Rhythm joint Attention Mechanism (BRAM)

1) ***Beat-based Convolutional Attention Module****:* The structure of the Beat-based Convolutional Attention module is shown in Fig. 5 (b). We begin by applying the differentiation $diff(\cdot)$ and pad operation $pad(\cdot)$ to the input signal x to capture the dynamic characteristics and trends of the input signals. We then apply one-dimensional convolution to the differentiated signal to extract the relevant local feature $C = W_c * x + b_c$, where W_c is the convolution kernel weight matrix, b_c is the bias. This one-dimensional convolution can capturing short-term variations within the signal, such as the rise and fall of each heartbeat, which are important features of heartbeat rhythm and variability.

After that, we apply three separate convolutional layers to generate different sets of weights: query (W_Q^b), key (W_K^b), and value (V_Q^b) weights, which are crucial for capturing the attention weights that relate to the varying time scales in HR signal. These weights are then normalized to the range $[0, 1]$ to prevent issues with vanishing or exploding gradients. The above process can be formulated as:

$$\begin{aligned} W_Q^b &= norm(Conv_Q(pad(diff(x))))^T, \\ W_K^b &= norm(Conv_K(pad(diff(x))))^T, \\ W_V^b &= norm(Conv_V(pad(diff(x))))^T, \end{aligned} \tag{1}$$

where $norm(\cdot)$ represents normalization and $Conv_Q$, $Conv_K$, and $Conv_V$ are the convolution kernels designed to capture different aspects of the heartbeat signal.

Finally, we apply matrix multiplication to the normalized convolutions across time steps and feature channels to compute three new features: Q_b, K_b, and V_b:

$$\begin{aligned} Q_b &= x \cdot W_Q^b, \\ K_b &= x \cdot W_K^b, \\ V_b &= x \cdot W_V^b. \end{aligned} \tag{2}$$

This process captures the relationships between signal features, such as the timing of heartbeats, their intervals, and rhythmic patterns. To summarize, the beat-level self-attention calculation can be expressed as:

$$Attention_b = Softmax(Q_b K_b^T / \sqrt{d}) V_b, \tag{3}$$

where d is query feature dimension and the output signal x_b can be expressed as:

$$x_b = Conv(Attention_b(x)). \tag{4}$$

This attention mechanism focuses on the model's attention on key heartbeat-related features, such as peak locations and RR intervals that are important for HR estimation. By filtering out irrelevant components and noise, the Beat-based Convolutional Attention module improves the model's ability to extract meaningful physiological features, enhancing its performance in accurately modeling HR dynamics, even in the presence of subtle lighting variations.

2) ***Rhythm-based Recurrent Attention Module****:* To capture the periodic trends and long-term dependencies of heartbeat signals, we design a Rhythm-based Attention module. The output $x_b \in \mathbb{R}^{N \times T}$ of beat-based attention mechanism is fed into the rhythm-based attention mechanism, where N is the number of batch size and T is the length of each segment. We use an LSTM network to extract temporal features from x_b, as LSTM is effective at learning dependencies over time and is well-suited for handling the sequential nature of heartbeat signals. The corresponding temporal features can be represented as:

$$\text{LSTM}_Q(x_b), \text{LSTM}_K(x_b), \text{LSTM}_V(x_b). \tag{5}$$

Subsequently, we apply three separate linear layers to map the LSTM outputs LSTM_Q, LSTM_K, and LSTM_V into three sets of weights:

$$\begin{aligned} W_Q^r &= \frac{\text{Linear}_Q(\text{LSTM}_Q) - min(\text{LSTM}_Q)}{max(\text{LSTM}_Q) - min(\text{LSTM}_Q)}, \\ W_K^r &= \frac{\text{Linear}_K(\text{LSTM}_K) - min(\text{LSTM}_K)}{max(\text{LSTM}_K) - min(\text{LSTM}_K)}, \\ W_V^r &= \frac{\text{Linear}_V(\text{LSTM}_V) - min(\text{LSTM}_V)}{max(\text{LSTM}_V) - min(\text{LSTM}_V)}, \end{aligned} \tag{6}$$

where $Linear_Q$, $Linear_K$ and $Linear_V$ are three linear layers, $min(\cdot)$ and $max(\cdot)$ are the maximum and minimum value of the signal, which are used to normalize the signal into the range $[0, 1]$. These weights represent the importance of different parts of the signal at various time scales.

Finally, the signal x_b is multiplied with the normalized weights W_Q, W_K, and W_V through matrix operations to generate the Query(Q_r), Key(K_r), and Value(V_r), therefore the rhythm-based attention mechanism can be formulated as:

$$Attention_r = Softmax(Q_r K_r^T / \sqrt{d}) V_r. \tag{7}$$

The output of the rhythm-based attention mechanism can be formulated as:

$$x_r = LSTM(Attention_r(x_b)). \tag{8}$$

In summary, the rhythm-based attention module enhances the model's ability to focus on long-term rhythmic patterns, such as sustained heartbeat trends and periodical variations. By capturing these global trend features, it improves the model's capacity to accurately estimate HR dynamics over time.

C. Physiological Feature Guided Regularization (PFGR)

The structure of the Physiological Feature Guided Regularization is shown in Fig. 5 (c). Incorporating domain-specific physiological characteristics into loss function can help the model maintain consistency with real heartbeat signals, thus enhancing the accuracy and physiological plausibility of the output rPPG signals. Studies show that the cardiac cycle-related features can be categorized into the following three types [45].

1) Heartbeat Dynamics: Heartbeat dynamics refer to the variation pattern of the heartbeat signal. These dynamics are captured through features such as the signal's amplitude (overall waveform) and instant fluctuations, which are addressed by the MAE loss and derivative MAE loss functions, respectively.

MAE loss $\mathcal{L}_{\text{MAE}}$ measures the overall difference between the predicted and actual BVP signals, reflecting the accuracy of the overall signal. It can capture the global heartbeat dynamics by focusing on the general shape of the output rPPG signals. The derivative MAE loss $\mathcal{L}_{\text{derivative}}$ measures the difference within heartbeat fluctuations over time. It captures the dynamic variations of the rPPG signal, emphasizing temporal changes in the heartbeat dynamics.

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{9}$$

$$\mathcal{L}_{\text{derivative}} = \frac{1}{N-1} \sum_{i=1}^{N-1} |\Delta y_i - \Delta \hat{y}_i|, \tag{10}$$

where N denotes the number of samples. y_i and $\hat{y}_i$ are the ground truth and predicted HR values of the i th subject and Δy_i and $\Delta \hat{y}_i$ are the ground truth and predicted changing rate of the i th subject, respectively.

2) Pulse Wave Morphology: Pulse Wave Morphology includes the shape and characteristics of pulse wave, such as the peak positions, time intervals, and the rise and fall rates of the waveform.

$\mathcal{L}_{\text{peaks}}$ and $\mathcal{L}_{\text{interval}}$ represents heartbeat count regularization and hearbeat interval regularization. These two regularizations ensure that the heartbeat positions of the predicted signal align with the real BVP signal, as well as to help the model to capture subtle temporal dynamic changes and adapt to individual diversity.

$$\mathcal{L}_{\text{peaks}} = |N_{\text{pred}} - N_{\text{target}}|, \tag{11}$$

$$\mathcal{L}_{\text{interval}} = \text{mean}\left(|\text{diff}(N_{\text{pred}}) - \text{diff}(N_{\text{target}})|\right), \tag{12}$$

where N denotes the number of samples. N_{pred} and N_{target} are HR peaks of the predicted and ground truth. diff$(\cdot)$ represents differences between consecutive heartbeats and mean$(\cdot)$ is the average function.

3) Temporal Rhythm Patterns: Temporal rhythm patterns refer to the rhythmic characteristics of the rPPG signal in the time domain, such as energy distribution in a fixed time interval, which reflects the dynamic change and rhythmic variations of the signal.

The energy distribution regularization $\mathcal{L}_{\text{EDR}}$ ensures the output signal closely matches the energy distribution of the ground truth BVP signal. By applying $\mathcal{L}_{\text{EDR}}$, the model can foucus on the periodicity and energy fluctuations, while minimizing the impact of noise on energy distribution.

$$E_k = \sum_{t=k}^{k+T} x_t^2, \tag{13}$$

$$\mathcal{L}_{\text{EDR}} = \text{mean}\left(|\Delta E_k|\right) \tag{14}$$

where k represents the starting position of the current time window, T denotes the sampling length, x_t represents the amplitude of the signal at time t, and E_k represents the signal energy within the k th window.

Our proposed model is trained using an loss function, which can be formulated as follows:

$$\mathcal{L} = \alpha(\mathcal{L}_{\text{MAE}} + \mathcal{L}_{\text{derivative}}) + \beta(\mathcal{L}_{\text{peaks}} + \mathcal{L}_{\text{interval}} + \mathcal{L}_{\text{EDR}}), \tag{15}$$

where α and β are used as coefficients to adjust the relative importance of each loss component. In this paper, we set α and β as 0.6 and 0.1. Each of these loss components is responsible for capturing the desired properties.

D. HR Measurement

The structure of the *HR Measurement* module is shown in Fig. 6. The output of the *BRAM* module is first fed into the *Predictor* to obtain the one-dimension rPPG signal. Then apply *B2B calculation* to calculate the corresponding HR value.

The *Predictor* maps the high-dimensional features to the final one-dimensional output. We use a *Conv1D* layer to capture local dependencies and a *Linear* layer to adjust the

output dimension, ensuring accurate and structured signal reconstruction.

In order to perform real-time and accurate HR calculation, we design a Beat-to-Beat (B2B) HR calculation method, which directly detects the locations of two consecutive heartbeats locations (R-R Interval, RRI).

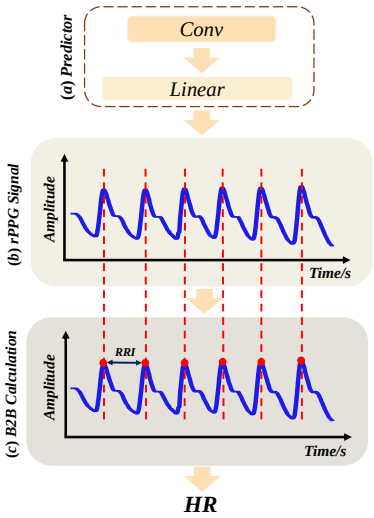


Fig. 6. The illustration of the *HR Measurement* module.

By obtaining HR information from each individual heartbeat interval, the B2B method can capture real-time HR changes in less-constrained scenarios. The formula for B2B HR calculation is as follows:

$$HR = \frac{T}{mean(p^{(m)^{(2)}} - p^{(m)^{(1)}}, \cdots, p^{(m)^{(K)}} - p^{(m)^{(K-1)}})}, \tag{16}$$

where, T is the window length, $p^{(m)^{(i)}}$ is the position of i th heartbeat.

V. Experimental Evaluation

A. Datasets and experimental setups

In our experiments, We utilizes two NIR datasets MR-NIRP and XDU-SenseTime to evaluate the performance of our proposed method PhysioNN.

1) Datasets: We provide evaluations on one public database MR-NIRP [39] and one private XDU-SenseTime database. The details about these database can be found in Section II and III.

2) Experimental setup: Our proposed method is performed under the Pytorch framework with an NVIDIA GeForce RTX 4090 GPU. The preprocessing settings include a low-pass filtering threshold = 0.7 Hz, a high-pass filtering threshold = 2 Hz, a data sampling rate = 30 Hz, a label sampling rate = 60 Hz, and a sliding window length is 30 seconds for signal segmentation. The model is trained with Adam optimizer and the learning rate and Dropout are 0.001 and 0.11, respectively. We perform a 5-fold cross-validation for both MR-NIRP and XDU-SenseTime datasets and use 3 folds for training, 1 fold for validation and 1 fold for testing.

3) Evaluation Metrics: To perform a quantitative analysis of the HR estimation model proposed in this paper, four quantitative evaluation metrics are employed: Mean Absolute Error (MAE), Root Mean Square Error ($RMSE$), Pearson Correlation Coefficient (r). Here, T represents the total number of frames in the video sequence, x_t represents the predicted HR value at frame t, and y_t represents the ground truth HR value corresponding to x_t. The evaluation metrics above are calculated based on HR_{error} between the predicted HR value and the true HR value:

$$HR_{error} = HR_{predict} - HR_{gt}. \tag{17}$$

1) *Mean Absolute Error*: It is used to assess the average absolute error (HR_{error}) between the predicted HR values ($HR_{predict}$) and ground truth HR values (HR_{gt}). A smaller MAE indicating a more accurate predictive capability of the model:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|x_t - y_t|. \tag{18}$$

2) *Root Mean Absolute Error*: It measures the average magnitudes of differences between predicted HR values ($HR_{predict}$) and ground truth HR values (HR_{gt}). A smaller $RMSE$ indicates a more accurate predictive capability of the model:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(HR_{error})^2}. \tag{19}$$

3) *Pearson Correlation Coefficient*: It measures the degree of linear relationship between the predicted HR ($HR_{predict}$) and the ground truth HR (HR_{gt}) and has values between -1 and 1. It is calculated as the covariance between the two variables divided by the product of their respective standard deviation.

$$r = \frac{\sum_{t=1}^{T}(x_t - \overline{x_t})(y_t - \overline{y_t})}{\sqrt{\sum_{t=1}^{T} x_t - \overline{x_t}^2}\sqrt{\sum_{t=1}^{T} y_t - \overline{y_t}^2}}. \tag{20}$$

Moreover, we apply three time-domain indicators to measure HR variability: the Mean value MeanNN, the standard deviation SDNN, and the square root of the mean square of the difference RMSSD. Here, n represents the total number of RR intervals (the time interval between two consecutive R-waves), RR_i is the i th RR interval.

1) *MeanNN*: The average value of all *NN* intervals, reflecting the overall rhythm of heartbeat:

$$MeanNN = \frac{1}{n}\sum_{i=1}^{n} RR_i. \tag{21}$$

2) *SDNN*: The standard deviation of *NN* intervals, reflecting the overall variability of the HR:

$$SDNN = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(RR_i - MeanNN)^2}. \tag{22}$$

3) *RMSSD*: The root mean square of the successive differences between *NN* intervals, reflecting the short-term variability of the HR:

$$RMSSD = \sqrt{\frac{1}{n-1}\sum_{i=2}^{n}(RR_i - RR_i - 1)^2}. \quad (23)$$

B. Main Comparsion

1) Intra-Dataset Testing: First, we conduct intra-dataset testing on all the datasets used in this paper. The proposed PhysioNN is compared with a lot non-contact methods, including signal-based methods (GREEN [29], POS [46], ICA [47], SCF [30], and SCICA [48]) and learning-based methods (PhysNet [11], DeepPhys [12], EfficientPhys [37], DCT-JBSS [19], Nowara2021 [38], Contrast-Phys [49], TS-CAN [13] and RhythmFormer [26]).

Performance on MR-NIRP Dataset. Table III shows the comparison results of different models on the MR-NIRP dataset. It is clear from Table III that traditional methods such as SCF, and SCICA show poor performance. This is because these methods rely on manually designed features, which are insufficient to capture all the information in the HR signal. In contrast, deep learning-based methods outperform traditional ones due to their excellent feature extraction capabilities. However, DeepPhys and PhysNet do not achieve satisfactory performance because they are end-to-end models that estimate the HR by using the entire facial area, which introduces additional noise from background, motion, or lighting. Nowara2021 uses inverse attention network to denoise interference from background regions. DCT-JBSS increases the dimensionality of facial single-channel signals, enabling effective extraction of signal information from multiple facial ROIs. Contrast-Phys generates spatiotemporal blocks from facial videos through preprocessing. As a result, they achieved better performance than DeepPhys and PhysNet, as sufficient physiological features and information are crucial for HR estimation.

However, the proposed method still outperforms these approaches because the two-level attention mechanism enables the model to progressively focus on the edges and sequential features of the physiological signals, reducing the impact of ambient light and noise. Additionally, by incorporating domain-specific physiological characteristics into the optimization strategy, the output signal can maintain consistency with the actual heartbeat signal. Among the deep learning methods, the proposed PhysioNN achieved very competitive results, with 0.98 bpm in MAE, 1.10 bpm in RMSE, and 0.95 in Pearson Correlation Coefficient.

Performance on XDU-SenseTime. Table IV shows the comparsion results on XDU-SenseTime dataset, the results indicate that traditional methods perform poorly due to their limited noise modeling capabilities. End-to-end deep learning methods such as DeepPhys, PhysNet, and EfficientPhys fail to capture the temporal and periodic information of the rPPG signal, which leads the models to overlook some significant signal features. RhythmFormer uses a dynamic sparse attention mechanism to extract periodic features at multiple time scales, while TS-CAN applies an attention module to lead the model to focus on image pixels that containing physiological features, reducing noise amplification. As a result, these two methods perform better than others.

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED METHOD ON THE MR-NIRP DATASET. THE BEST RESULTS ARE IN BOLD.

Method	$MAE \downarrow$ (bpm)	$RMSE \downarrow$ (bpm)	$r \uparrow$
Signal-based Method			
SCF [30]	2.29	4.48	0.88
SCICA [48]	2.71	5.86	0.81
Learning-based Method			
DCT-JBSS [19]	1.75	2.56	0.95
DeepPhys [12]	7.78	16.8	-0.03
PhysNet [11]	3.07	7.55	0.66
Nowara2021 [38]	2.34	4.46	0.85
Contrast-Phys [49]	2.68	4.77	0.85
Ours	**0.90**	**1.45**	**0.97**

The proposed PhysioNN method outperforms the others models on all metrics, demonstrating that our approach can better extract physiological features across different temporal scales.

Additionally, the test results on this dataset are slightly worse than those on the publicly available MR-NIRP dataset. This can be attributed to the fact that the proposed XDU-SenseTime dataset is collected in less-constrained environments, whereas the MR-NIRP dataset consists of still scenes.

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED METHOD ON THE XDU-SENSETIME DATASET. THE BEST RESULTS ARE IN BOLD.

Methods	$MAE \downarrow$ (bpm)	$RMSE \downarrow$ (bpm)	$r \uparrow$
Signal-based Method			
GREEN [29]	5.74	7.28	0.77
POS [46]	4.15	5.91	0.82
ICA [47]	4.77	5.86	0.86
Learning-based Method			
DeepPhys [12]	6.08	2.71	0.51
EfficientPhys [37]	4.09	5.98	0.83
PhysNet [11]	4.21	6.36	0.88
TS-CAN [13]	2.40	5.19	0.95
RhythmFormer [26]	3.56	4.78	0.82
Ours	**1.86**	**2.47**	**0.98**

2) Cross-Dataset Testing: In addition to intra-dataset testing, we also perform cross-dataset testing to demonstrate the generalization ability of the proposed method. Specifically, the model trained on one dataset is directly validated on another dataset. For example, MR-NIRP $\rightarrow$ XDU-SenseTime indicates that the model is trained on MR-NIRP and tested on XDU-SenseTime. The parameter settings for the cross-dataset experiments are the same as those used in the original intra-dataset experiments. Table V presents the comparison results with other existing methods. It is clear that our proposed method significantly outperforms others in cross-dataset testing, demonstrating its strong generalization ability in various scenarios.

TABLE V
EXPERIMENTAL RESULTS OF OUR METHOD ON CROSS-DATASET TESTING. THE BEST RESULTS ARE IN BOLD.

Methods	MR-NIRP → XDU-SenseTime			XDU-SenseTime → MR-NIRP		
	MAE	RMSE	r	MAE	RMSE	r
Signal-based Method						
GREEN [29]	5.15	6.72	0.85	5.42	6.59	0.87
POS [46]	4.28	5.81	0.82	6.38	7.46	0.85
ICA [47]	4.86	5.94	0.84	5.69	6.82	0.87
Learning-based Method						
DeepPhys [12]	3.42	4.58	0.92	3.57	4.42	0.95
EfficientPhys [37]	4.78	6.13	0.87	7.48	9.72	0.86
PhysNet [11]	2.70	3.58	0.82	4.68	5.32	0.82
TS-CAN [13]	5.82	6.82	0.85	6.78	8.51	0.86
RhythmFormer [26]	6.78	8.41	0.72	9.44	11.32	0.74
Ours	**1.80**	**1.853**	**0.97**	**2.89**	**2.489**	**0.98**

C. HRV Estimation

To better evaluate the performance of the proposed method, Table VI shows the HRV estimation results across different scenarios in XDU-SenseTime database. It can be clearly observed from the Table VI that the MeanNN is the lowest in *Dim Environment*, indicating that the proposed method can provide stable HR monitoring under low-light conditions and deliver accurate HR estimates. However, the values of SDNN and rMSSD are higher in *Small Rotation* and *Large Rotation*, suggesting that the interference caused by head movement can affect the estimation accuracy. Notice that in the *Talking* scenarios, all metrics show higher error values, which is due to the interference from facial muscle movement during speech. Overall, the proposed method PhysioNN performs well in all scenarios across all metrics, demonstrating that it can effectively perform HR detection in various scenes, with stronger robustness to noise in low-light conditions.

TABLE VI
HRV ESTIMATION RESULTS ON XDU-SENSETIME DATASET.

Scenarios	MeanNN	SDNN	rMSSD
Still	0.753	2.076	3.427
Single Side	0.751	3.439	5.476
Small Rotation	1.489	3.739	5.635
Large Rotation	1.035	3.546	5.682
Talking	1.380	3.984	6.416
Bright Environment	0.559	2.535	3.851
Dim Environment	0.353	1.962	3.541

D. Ablation Studies

This section introduces the ablation study results of the proposed method on the MR-NIRP and XDU-SenseTime datasets to analyze the impact of each component on the experimental performance.

1) BRAM: To validate the effectiveness of BRAM module, the network is trained using four models: the baseline transformer, baseline+beat-based attention module, baseline+rhythm-based attention module and the baseline+BRAM. The comparsion results are displayed in Fig.7.

In the frequency domain, as shown in Fig.7 (a), we present comparison results of the pulse waveform between the predicted rPPG signals and the ground truth rPPG signal. It can be observed that the rPPG signal predicted by the original baseline method hard to match the real signal in terms of peak, valley and the overall signal trend. After incorporating the rhythm-based attention module, the predicted rPPG signal better aligns with the overall trend of the ground truth signal. By applying the BRAM module, the baseline+BRAM model not only can better fit the overall trend, but can also accurately locate the peaks and troughs of the ground truth signal. This validate the effectiveness of beat-base attention module in capturing short term physiological features like peaks and troughs.

In the time domain, as shown in Fig.7 (b), we present comparison of the power spectral density (PSD) between the predicted rPPG signals and the ground truth rPPG signal. The PSD can illustrate the distribution of signal power across different frequencies. It is evident that the baseline method exhibits a poor fit in the frequency domain and fail to capture the signal's main frequency components. The baseline+beat-based attention module model shows slight improvement, with the frequency characteristics of the signal enhanced to some extent. The baseline+BRAM method performs the best, with the highest degree of fit in the PSD, accurately capturing both the frequency components and noise distribution of the signal. This further proves that rhythm-based attention module can capture the overall trend and frequency distribution of the PPG signal in the entire temporal sequence.

Moreover, the quantitative comparsion results are also proposed in Table VII. It can be observed that the baseline method shows a 6.36 bpm higher MAE compared to the proposed baseline+BRAM model. This primarily because BRAM can extract multi-scale temporal features from rPPG signals, resulting in better fit for the real signal, thereby improving HR estimation performance.

TABLE VII
ABLATION EXPERIMENTAL RESULTS OF VARIOUS SCENARIOS ON XDU-SENSETIME DATASETS.

baseline	BRAM	PFGR	MAE	RMSE
✓			8.32	9.47
✓	✓		1.96	2.78
✓		✓	7.96	9.13
✓	✓	✓	1.86	2.59

2) PFGR: To validate the effectiveness of PFGR module, the proposed model is compared with the baseline model, and the results are presented in Table VII. After applying PFGR, the HR MAE decreased from 8.32 bpm to 7.96 bpm. This improvement is primarily due to PFGR's heartbeat count and heartbeat interval regularization, which further constrain the model from overfitting to abnormal signals, ensuring that the estimation results align with actual physiological patterns.

3) B2B: To validate the effectiveness of the B2B HR calculation method, the FFT and B2B method are compared using the proposed model, as shown in Fig. 8. On the MR-NIRP public dataset, the MAE decreased from 0.977 bpm to 0.905 bpm, representing a 7% reduction. On the XDU-SenseTime dataset, the MAE decreased to 1.864 bpm, reflecting a 53%

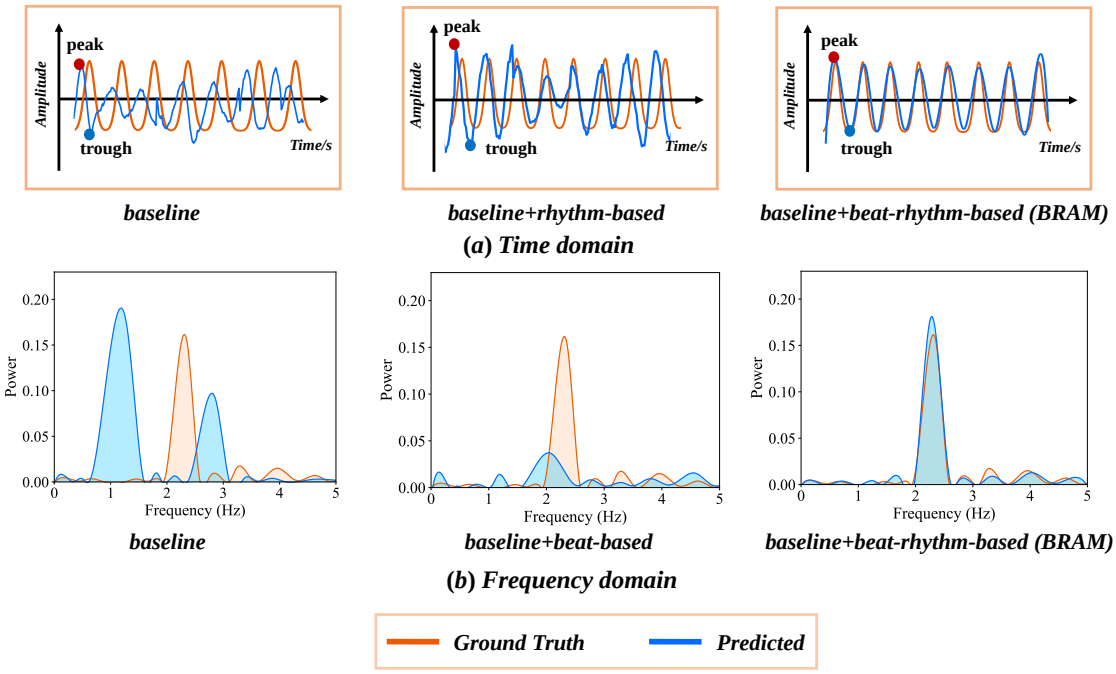


Fig. 7. Ablation experimental results of HR estimation on XDU-SenseTime datasets. The blue and orange line represent the predicted and ground-truth rPPG signals, respectively. (a) the frequency domain comparsion. (b) the Time domain comparsion.

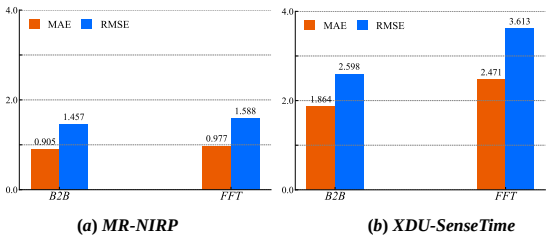


Fig. 8. The impacts of different HR calculations methods on MR-NIRP and XDU-SenseTime datasets.

TABLE VIII
ABLATION EXPERIMENTAL RESULTS OF DIFFERENT COMBINATION STRATEGIES OF VARIOUS LOSSES.

Setting	XDU-SenseTime			MR-NIRP		
	MAE	RMSE	r	MAE	RMSE	r
w/o $\mathcal{L}_{\text{MAE}}$, $\mathcal{L}_{\text{derivative}}$	2.88	4.17	0.92	4.38	5.99	0.92
w/o $\mathcal{L}_{\text{peaks}}$,$\mathcal{L}_{\text{interval}}$	2.47	3.61	0.94	4.17	5.48	0.95
w/o $\mathcal{L}_{\text{EDR}}$	1.96	2.67	0.96	0.97	1.38	0.95
all	0.90	1.45	0.97	1.86	2.47	0.98

reduction. This improvement is attributed to the B2B method's direct reliance on the intervals between individual heartbeats, which demonstrates significant advantages in capturing HR signals.

E. Parameter Analysis

1) Analysis of Multiple Losses: We trained our model on XDU-SenseTime and MR-NIRP datasets, removing one type of losses at a time to assess the effectiveness of each type of losses. The results are shown in Table VIII. As demonstrated by the results, when $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{derivative}}$, $\mathcal{L}_{\text{peaks}}$ and $\mathcal{L}_{\text{interval}}$, and $\mathcal{L}_{\text{EDR}}$ are removed separately, the results all showed some performance degradation. However, the best performance is achieved when all losses are incorporated into the model. This indicates that the combination of losses is crucial for improving HR estimation performance. Therefore, it is evident that the combined loss function significantly enhances the estimation accuracy.

2) Hyperparameters α and β: To assess the impact of the hyperparameters α and β, we provide a quantitative analysis and present the findings in Fig. 10. It can be observed that optimal results are obtained with α at 0.6 and β at 0.1. Thus, we adopt these parameter settings in all experiments.

3) Robust HR Estimation: 12 challenging scenarios are selected from the MR-NIRP and XDU-SenseTime datasets to evaluate the robustness of the proposed PhysioNN method, as shown in Fig. 9. Among these scenarios, the signals in the "Still" scenario exhibited the best fit. In the "Bright" and "Talk" scenarios, the rPPG signals are not entirely consistent with the ground truth, while their peaks, troughs, and periodic patterns are fully aligned. Since our proposed method uses the B2B approach to estimate HR, the predicted rPPG signals only need to match the trends of the ground truth signals.

Notably, the proposed method demonstrated strong robustness even in the presence of significant noise, as seen in scenarios like "Single Side" and "Large Rotation." This indicates that our method can effectively suppresses noise interference, accurately captures physiological information within the signals, and maintain stable HR estimation even in noisy

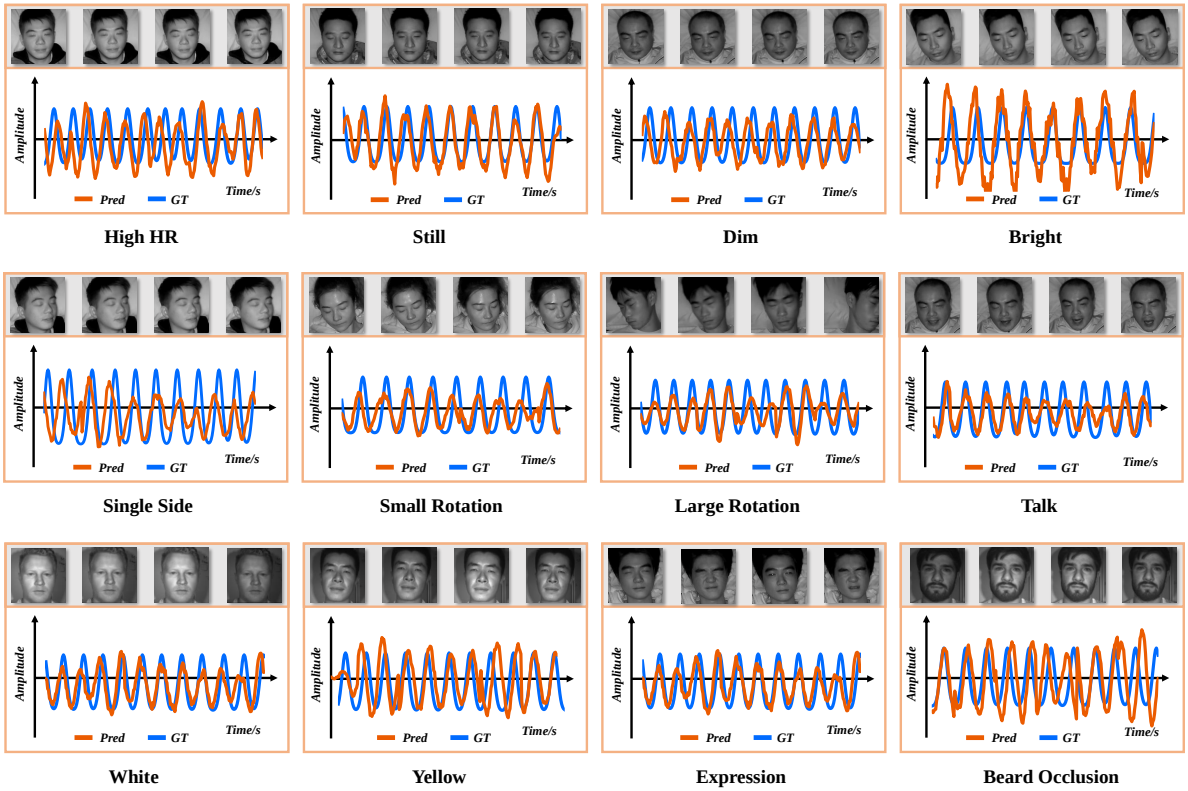


Fig. 9. Visualization of quantitative results in twelve challenging scenarios according to two benchmark physiological datasets (MR-NIRP and XDU-SenseTime). We show the predicted rPPG signals (Pred) and the corresponding ground-truth PPG signals (GT) in orange and blue curves, respectively. Some representative video frames are placed on the top of the signal diagram.

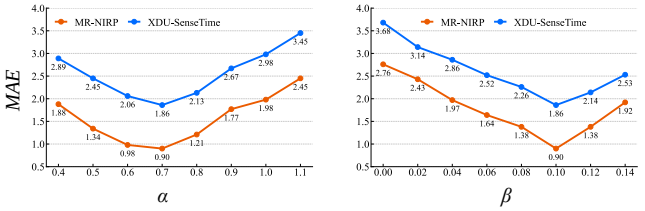


Fig. 10. The impacts of different α and β values on the MR-NIRP and XDU-SenseTime datasets.

scenarios.

VI. Conclusions

In this work, a novel physiological information-guided network is proposed to effectively extract multi-scale temporal features from rPPG signals, addressing the limitations in HR estimation caused by the lack of physiological information across different temporal scales. Our findings contribute to the development of non-contact health monitoring systems by demonstrating the feasibility of accurate HR detection from facial videos under low-light conditions. This research lays the foundation for the wider adoption of rPPG technology in continuous vital sign monitoring, particularly in settings where comfort, convenience, and long-term tracking are essential.

Despite the encouraging performance, the proposed method still has some limitations. The proposed method only utilizes NIR facial videos without considering RGB-NIR multi-modal fusion. Moreover, the proposed method has strong dependence on annotated data and requires supervised training. The model performance will degrade without sufficient high-quality data. In future work, we plan to explore the effectiveness of the proposed approach for other rPPG-based physiological measurement tasks such as blood pressure and investigate the multi-modal source of rPPG signals to enhance information extraction ability.

References

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

[2] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, "Video-based heart rate measurement: Recent advances and future prospects," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 10, pp. 3600–3615, 2019.

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

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

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

[6] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, "Remote heart rate measurement from face videos under realistic situations," in *Procedings of the IEEE/CVF Internatinal Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4264–4271.

[7] P. Wang and Q. He, "Heart rate estimation and performance analysis using mimo radar with dispersed antennas," in *Procedings of the IEEE/CVF International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[8] Z. Yang and Z. Bao, "Short-time heart rate estimation based on 60-ghz fmcw radar," in *Procedings of the IEEE MTT-S International Wireless Symposium (IWS)*, 2023, pp. 1–3.

[9] J. H. Park and S. C. Park, "Complex range resolution model of point scatterers in lfm chirp pulse radar," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[10] Y. Sha, J. Faber, S. Gou, B. M. Liu, W. Li, S. Schramm, H. Stoecker, T. Steckenreiter, D. Vnucec, N. Wetzstein, A. Widl, and K. Zhou, "A multi-task learning for cavitation detection and cavitation intensity recognition of valve acoustic signals," *arXiv preprint arXiv: 2203.01118*, 2022.

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

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

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

[14] M. Hu, G. Zhai, D. Li, Y. Fan, H. Duan, W. Zhu, and X. Yang, "Combination of near-infrared and thermal imaging techniques for the remote and simultaneous measurements of breathing and heart rates under sleep situation," *PLoS ONE*, vol. 13, 2018.

[15] N. Martinez, M. Bertran, G. Sapiro, and H.-T. Wu, "Non-contact photoplethysmogram and instantaneous heart rate estimation from infrared face video," in *Proceedings of the IEEE International Conference on Image Processing*, 2019, pp. 2020–2024.

[16] M. van Gastel, S. Stuijk, and G. de Haan, "Motion robust remote-ppg in infrared," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 5, pp. 1425–1433, 2015.

[17] Z. Zhang, C.-h. Fu, L. Zhang, and H. Hong, "Near infrared video heart rate detection based on multi-region selection and robust principal component analysis," in *Proceedings of the International Conference on Image and Graphics*, 2023, pp. 37–47.

[18] E. Magdalena Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1272–1281.

[19] J. Cheng, P. Wang, R. Song, Y. Liu, C. Li, Y. Liu, and X. Chen, "Remote heart rate measurement from near-infrared videos based on joint blind source separation with delay-coordinate transformation," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2020.

[20] Y. Ma, X. Jiang, Z. Xia, M. Gabbouj, and X. Feng, "Casqnet: Intrinsic image decomposition based on cascaded quotient network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2661–2674, 2021.

[21] X. Zhang, F. Zhang, and C. Xu, "Joint expression synthesis and representation learning for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1681–1695, 2022.

[22] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3141–3149.

[23] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, "Video-based remote physiological measurement via cross-verified feature disentangling." in *Proceedings of the European Conference on Computer Vision*, 2020.

[24] H. Lu, H. Han, and S. K. Zhou, "Dual-gan: Joint bvp and noise modeling for remote physiological measurement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 399–12 408.

[25] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. Torr, and G. Zhao, "Physformer: Facial video-based physiological measurement with temporal difference transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4176–4186.

[26] B. Zou, Z. Guo, J. Chen, and H. Ma, "Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer," *arXiv preprint arXiv: 2402.12788*, 2024.

[27] S. Joardar, A. Chatterjee, and A. Rakshit, "A real-time palm dorsa subcutaneous vein pattern recognition system using collaborative representation-based classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 4, pp. 959–966, 2015.

[28] V. Jeanne, M. Asselman, B. den Brinker, and M. Bulut, "Camera-based heart rate monitoring in highly dynamic light conditions," in *Proceedings of the International Conference on Connected Vehicles and Expo*, 2013, pp. 798–799.

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

[30] J. Chen, Z. Chang, Q. Qiu, X. Li, G. Sapiro, A. Bronstein, and M. Pietikäinen, "Illumination invariant heart rate estimation from videos," in *Proceedings of the International Conference on Image Processing Theory, Tools and Applications*, 2016, pp. 1–6.

[31] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, 2012.

[32] X. He, R. Goubran, and F. Knoefel, "Ir night vision video-based estimation of heart and respiration rates," in *Proceedings of the IEEE Sensors Applications Symposium*, 2017, pp. 1–5.

[33] L. K. C. H. W. H. C. Y. Chen D., Wang J. and L. S, "Ir night vision video-based estimation of heart and respiration rates," vol. 15, no. 1, pp. 618–627, 2015.

[34] Q. Zhang, Y. Zhou, S. Song, G. Liang, and H. Ni, "Heart rate extraction based on near-infrared camera: Towards driver state monitoring," *IEEE Access*, vol. 6, pp. 33 076–33 087, 2018.

[35] R. Spetlik, V. Franc, J. Cech, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proceedings of the British Machine Vision Conference*, 2018.

[36] Z. Yu, W. Peng*, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement," in *Proceedings of the International Conference on Computer Vision*, 2019.

[37] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4997–5006.

[38] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "The benefit of distraction: Denoising camera-based physiological measurements using inverse attention," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4955–4964.

[39] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Near-infrared imaging photoplethysmography during driving," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2020.

[40] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, and G. Zhao, "The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection," in *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition*, 2018, pp. 242–249.

[41] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2020.

[42] S. Kwon, J. Kim, D. Lee, and K. Park, "Roi analysis for remote photoplethysmography on facial video," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015, pp. 4938–4941.

[43] M. A. Hassan, A. S. Malik, D. Fofi, N. Saad, B. Karasfi, Y. S. Ali, and F. Meriaudeau, "Heart rate estimation using facial video: A review," *Biomedical Signal Processing and Control*, vol. 38, pp. 346–360, 2017.

[44] S. Huynh, R. K. Balan, J. Ko, and Y. Lee, "Vitamon: measuring heart rate variability using smartphone front camera," in *Proceedings of the International Conference on Embedded Networked Sensor Systems*, 2019, pp. 1–14.

[45] X. Dong and W. Si, "Heartbeat dynamics: A novel efficient interpretable feature for arrhythmias classification," *IEEE Access*, vol. 11, pp. 87 071–87 086, 2023.

[46] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2017.

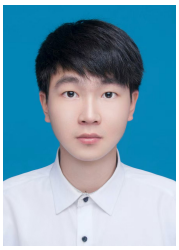
[47] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2010.

[48] F. Zhao, M. Li, Y. Qian, and J. Z. Tsien, "Remote measurements of heart and respiration rates for telemedicine," *PLoS ONE*, vol. 8, no. 10, p. e71384, 2013.
[49] Z. Sun and X. Li, "Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 492–510.

Shuiping Gou (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science and technology from Xidian University, Xi'an, China, in 2000 and 2003, respectively, and the Ph.D. degree in pattern reconition and intelligent system from Xidian University, in 2008. She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. Her research interests include machine learning, data mining, and remote sensing image analysis.

Kehong Liu (Student Member, IEEE) received the B.E degree in computer science and technology from Xi'an University of Science and Technology, Xi'an, China, in 2023. She is currently pursuing a M.S. degree with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China, Xidian University, Xi'an.

Her research interests include remote heart rate estimation, signal processing, medical image registration and artificial intelligence.



Hantao Zhao received the B.S. degree from China Jiliang University, Hangzhou, China, in 2021. He is currently pursuing his M.S. degree with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China, Xidian University, Xi'an.

His research interests include remote heart rate estimation, deep learning, and signal processing.

Nuo Tong (Member, IEEE) received her B.S. and Ph.D. degrees in intelligent information processing from Xidian University, in 2015 and 2020. She is now a lecturer in the Academy of Advanced Interdisciplinary Research, Xidian University.

Her research interest includes multi-organ medical image segmentation, medical image reconstruction, medical image detection and grading.

Zhang Guo (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2013, 2016 and 2021, respectively. He is currently an Assistant Professor at the School of Artificial Intelligence, Xidian University.

His main research interests include intelligent monitoring of physiological time series, neuromorphic computing and artificial intelligence.

Wenbo Liu received M.D. degree in Binzhou Medical University, Binzhou, China, in 1998, and the Ph.D. degree from the Fourth Military Medical University, Xi'an, China, in 2012. He is currently an associate professor at Shandong Second Medical University.

His research interests include brain protection mechanisms, biological monitoring technology.

Qigong Sun received the B.S. degree in intelligence science and technology and the Ph.D. degree in circuit and system from Xidian University, Xi'an, China, in 2015 and 2021, respectively.

He is currently a Principal Investigator with Shanghai AI Lab and the Director of Applied Research Laboratory of SenseTime. His research interests include machine learning and image processing

Licheng Jiao (Fellow, IEEE) received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, Xi'an, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China, and the International Research Center of Intelligent Perception and Computation. His research interests include intelligent information processing, image processing, machine learning, and pattern recognition.

Dr. Jiao is a member of the IEEE Xi'an Section Execution Committee, the President of the Computational Intelligence Chapter, the IEEE Xi'an Section, and the IET Xi'an Network, the Chairperson of the Awards and Recognition Committee, the Vice-Board Chairperson of the Chinese Association of Artificial Intelligence, a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, and an Expert of the Academic Degrees Committee of the State Council.