كلية الهندسة بشبرا
FACULTY OF ENGINEERING AT SHOUBRA
ERJ
PRINT ISSN 3009-6049
ONLINE ISSN 3009-6022

ENGINEERING RESEARCH JOURNAL (ERJ)
Volume (55),Issue (3)
July 2026, pp:102-116
https://erjsh.journals.ekb.eg

Lightweight Multi-scale TCN for Heart Rate Estimation

Ahmed Mehrez [1,*], Abdelwahab Al-sammak [1], Shady Y. El-mashad [1]

[1]*Electrical Engineering Department, Faculty of Engineering at Shoubra, Benha University, Cairo, Egypt*
**Corresponding author*

E-mail address: ahmed.mehrez@feng.bu.edu.eg, asammak@feng.bu.edu.eg , shady.elmashad@feng.bu.edu.eg

Abstract: Remote photoplethysmography (rPPG) provides a non-contact heart rate (HR) estimation from facial videos. However, real-world deployment is distorted by motion artifacts, illumination variations, and the high computational cost of deep learning and transformer models. In this paper, we propose a lightweight multi-scale temporal convolutional network (TCN) designed for CPU-based edge devices. The framework dynamically extracts green-channel signals from three facial regions (forehead, left cheek, and right cheek) and a background reference using MediaPipe landmarks. These 1-D temporal traces are processed by a four-block TCN with parallel multi-scale kernels, exponential dilation, and squeeze-and-excitation attention, enabling the model to capture both fast systolic peaks and slower diastolic trends. Regularization via ROI dropout further improves generalization. The entire system uses only 0.1 million parameters and requires 5.5 MFLOPs. Evaluated on two public datasets under a subject-disjoint protocol, our method achieves a mean absolute error (MAE) of 0.86 BPM on UBFC-rPPG and 1.95 BPM on COHFACE, outperforming traditional and many deep learning baselines while being an order of magnitude more efficient than transformer-based models. The results demonstrate that compact, landmark-guided temporal networks can deliver near-state-of-the-art accuracy with minimal computational resources, making them suitable for real-time, on-device physiological monitoring.

Keywords: rPPG, TCN, Heart rate.

1. Introduction

Recently, Interest in vital signs has increased. Heart disease and general health issues have become very common with technological and industrial advancements, which have led to a decrease in physical activity. This, coupled with diets based on processed and ready-made foods, has significantly increased the risk of cardiovascular disease. Therefore, monitoring general cardiovascular health indicators becomes essential for most people. This monitoring must be accessible through various methods to facilitate tracking anywhere, especially for individuals at high risk of heart disease. Accurate monitoring of vital signs such as heart rate (HR) is essential for early detection, prevention, and management of health conditions [1], [2].

Traditional monitoring methods include electrocardiography (ECG) [3] and pulse oximetry [4]. These methods offer high accuracy and provide reliable monitoring of general health. For stable operation, the devices must be in direct contact with the body to ensure the sensors function correctly. Any movement that affects the sensor's contact with the body will completely decrease the device's operation and accuracy. Therefore, the patient must remain stable throughout the monitoring period. This may not be applicable in all cases or acceptable to all patients. If a person wants to live in a normal way, tracking with bulky devices is not possible. With significant technological advancements, wearable devices have become available, some with relatively acceptable accuracy. Furthermore, the stability of the wrist is essential for the proper functioning of these wearable devices. Wearable devices with the required precision may be expensive for some individuals, and some patients may not be able to use them [5].

The limitations of these contact-based methods can be solved by using remote monitoring techniques. Remote techniques are those that do not require contact to measure physiological signals. Remote photoplethysmography (rPPG) is a promising technique for measuring the blood volume pulse signal (BVP) using standard cameras[6]. It works by detecting subtle color changes in the captured skin areas. These variations are invisible to the human eye but can be extracted from facial videos using signal processing. rPPG was introduced in the early 2000s. This technique relies on the absorption and reflection of light by blood flow. Accordingly, this operation allows heart rate and related indicators to be obtained from video without the need to attach any sensors to the body.

Most rPPG systems use standard RGB cameras because they are widely available and low-cost. However, thermal cameras can detect heat changes from blood flow (useful in low light), and near-infrared (NIR) cameras penetrate deeper into tissue, offering better robustness to skin tone variations. Each sensor type has trade-offs in cost, privacy, and environmental sensitivity. From an algorithmic perspective, rPPG methods can generally be categorized into three groups. Traditional signal-processing methods explicitly extract physiological signals using color-space analysis and handcrafted models. Deep-learning approaches learn physiological representations directly from facial video data through convolutional neural networks. More recently, transformer-based architectures have been introduced to capture long-range temporal dependencies using self-

attention mechanisms. Each category offers different trade-offs between computational efficiency, robustness, and estimation accuracy.

Beyond heart rate (HR), rPPG has been shown to estimate respiratory rate, blood pressure, oxygen saturation (SpO_2), heart rate variability (HRV), and even stress or emotional states. This makes rPPG a promising tool for telemedicine, driver monitoring, fitness tracking, and clinical diagnostics. However, real-world deployment remains challenging due to motion artifacts, illumination changes, and computational constraints.

The main advantage of the RGB-based rPPG is its ease of integration with commonly used devices such as smartphones. Since rPPG only requires a camera, it offers a low-cost and widely accessible approach for physiological monitoring. It enables many applications in telemedicine, fitness tracking, driver monitoring, human-computer interaction, and clinical environments[7].

The rPPG technology faces significant challenges in real-life applications. Changes in illumination (such as ambient light or shadows) increase the noise that hides the weak pulse signal[8]. Furthermore, the movements of the subject, including head rotation, movement, and facial expressions, cause motion distortions that degrade the signal extracted from the area of interest (ROI) and signal stability. Facial geometry, skin tone variations, and facial features further complicate signal extraction due to differences in light reflection and absorption. Other problems include video compression distortions and general camera noise, all of which reduce signal quality and accuracy [8]. Overcoming these challenges enables the development of robust and widely applicable rPPG models that are suitable for real environments.

Methods for extracting signals for rPPG typically involve three stages. The first stage is the preprocessing, where the face is detected, and Regions of Interest (ROIs) within the face are determined. Typically, the skin is segmented to avoid the eye regions and areas where hair usually grows, such as the eyebrows and mustache. The second stage involves using the model to identify the relationship and extract physiological information from the raw signal. The final stage is the post-processing, which extracts the final heart rate using a bandpass filter to filter the generated signal to get only frequencies associated with a normal heart rate.

Despite significant progress, existing methods face a trade-off between estimation accuracy and computational efficiency. Traditional approaches are lightweight but less robust under challenging conditions, whereas modern deep-learning and transformer-based methods achieve higher accuracy at the expense of substantially increased computational complexity. Addressing this trade-off remains a key challenge for practical edge-device deployment. Recent works in rPPG have been driven by deep learning and transformer architectures. While these models achieve state-of-the-art accuracy, they typically process raw video frames directly. This end-to-end spatial processing requires massive computational resources, making it highly impractical for real-time deployment on edge devices. Conversely, traditional signal-processing methods are computationally lightweight because they work on pre-extracted 1D signals, but they often lack the robustness to handle complex environmental noise and motion artifacts.

To bridge this gap, our work proposes a hybrid framework that shifts the deep learning workload to the temporal domain. Rather than forcing a heavy neural network to process raw spatial pixels, we perform explicit signal extraction as a preprocessing step. By feeding these 1D traces into a constrained Multi-scale Temporal Convolutional Network (TCN), our model achieves competitive accuracy with a fraction of the computational cost. Specifically, the main contributions of this work are summarized as follows:

1. A Hybrid Signal-to-Target Learning Paradigm: Rather than processing spatial pixel data, our framework performs traditional explicit signal extraction as a preprocessing step. The resulting 1D temporal signals are fed directly into the network. By dedicating the deep learning architecture to mapping the relationship between pre-extracted physiological signals and the ground-truth heart rate, we achieve reduced computational costs.
2. Explicit Environmental Noise Modeling: Our preprocessing step extracts 1D green channel signals from facial landmarks beside a dedicated background signal. The background signal acts as an environmental reference. This provides the network with an explicit reference for illumination noise. This reference allows the model to suppress environmental noise rather than implicitly learning it from raw pixels.

2. Related work

Several methods have been developed to extract signals from facial videos using rPPG. Initially, traditional signal analysis was used for signal extraction and processing. This evolved to use convolutional neural networks (CNNs) [9]and deep learning to achieve acceptable accuracy. More recently, with the development of Transformers and Attention mechanisms, new methods have appeared to achieve superior accuracy in controlled environments. In this section, these methods are discussed along with their advantages and limitations.

2.1 Traditional methods

Early rPPG methods mainly depended on traditional signal processing techniques designed to estimate the physiological information from RGB facial videos. One of the earliest approaches was the Green channel method, which proposed that the green channel contains stronger physiological information. This strong relation is due to the hemoglobin light absorption characteristics. Although

computationally efficient, the Green method is highly sensitive to illumination variation and motion artifacts [10].

To improve robustness, de Haan and Jeanne proposed the CHROM method. They proposed chrominance-based color transformations to suppress motion-caused distortions and enhance pulse-related component extraction[6]. Later, Wang et al. proposed the Plane-Orthogonal-to-Skin (POS) algorithm. POS uses projections of RGB signals onto a plane orthogonal to the skin tone direction to improve the extraction of the pulse signal under varying lighting conditions [11]. These traditional methods achieved promising performance in controlled environments. However, they still suffer from significant limitations under large head movements, facial expressions, illumination changes, and video compression artifacts.

Most traditional models employ specific ROIs in the physiological signal estimation. The selection of ROIs has a significant impact on signal extraction. The model analyzes facial motion patterns to suppress regions likely to contain motion-related artifacts. While the forehead and cheeks are typically the most preferred areas, other regions have their own importance. For example, the glabella has the strongest correlation with pulse due to the strength and closeness of blood vessels to the skin's surface[12]. However, it can be affected by facial movements depending on the individual's face geometry. Instead of treating the entire face regions with the same weight, as in other methods, weighting should be applied to specific areas.

2.2 Deep learning methods

To address the limitations of handcrafted features, and with the advancement of deep learning, deep learning-based methods have gained traction. CNN-based methods have become increasingly popular for rPPG estimation. DeepPhys proposed a convolutional attention mechanism to learn motion representations related to physiological signals directly from video frames [13]. Although DeepPhys demonstrated the effectiveness of end-to-end learning for physiological estimation, its reliance on frame-level spatial processing results in relatively high computational requirements. Furthermore, the learned representations can be sensitive to illumination variations and dataset-specific appearance characteristics.

PhysNet further improved end-to-end learning by employing spatiotemporal convolutions to estimate continuous rPPG waveforms [14]. PhysNet achieved higher accuracy than earlier CNN-based methods by directly reconstructing the BVP waveform. However, its 3D convolutional architecture presents computational overhead and memory consumption. This makes deployment on resource-constrained edge devices challenging.

The TS-CAN method estimates physiological signals from facial videos by combining spatial and temporal modeling in an efficient deep network. It uses temporal shift modules to capture motion detected across frames without heavy 3D convolutions, while an attention mechanism highlights informative facial regions related to blood flow. The TS-CAN produces a robust rPPG signal even under variations in lighting and subject movement by jointly learning appearance and motion features[15]. Despite its improved efficiency compared to 3D-CNN approaches, TS-CAN still processes raw facial frames and therefore requires learning both spatial and temporal representations simultaneously. This increases computational complexity and may reduce generalization when appearance characteristics differ significantly across datasets.

Deep learning methods outperform traditional techniques because they can learn robust spatial and temporal features from data. They outperform traditional methods in controlled settings and show better handling of some artifacts through data-driven feature learning. However, some challenges still remain. Challenges include over-fitting to specific datasets, poor generalization to unseen lighting conditions, skin tones, and motion patterns. It may achieve limited generalization when evaluated across different datasets or environmental conditions [16].

2.3 Transformer-based methods

More recently, transformer-based architectures have been explored for rPPG due to their ability to model long-range temporal dependencies via self-attention mechanisms. PhysFormer[17] used the temporal difference transformers to model global spatiotemporal interactions and enhance quasi-periodic rPPG patterns. While PhysFormer significantly improved the modeling of long-range temporal dependencies, the self-attention mechanism introduces considerable computational and memory costs. Consequently, inference often requires dedicated GPU resources for practical operation. The modified version, PhysFormer++[18], further improves feature representation using SlowFast temporal difference transformers for physiological measurement. PhysFormer++ enhanced temporal representation through a SlowFast architecture. However, the improved accuracy comes at the expense of increased model complexity, parameter count, and computational requirements, making edge-device deployment more difficult.

More recently, RhythmFormer[19] proposed periodic sparse attention mechanisms that reduce irrelevant attention computations while focusing on periodic rPPG patterns. Transformer-based methods have demonstrated state-of-the-art performance in several benchmark datasets due to their ability to capture long-range dependencies and complex temporal dynamics. RhythmFormer introduced periodic sparse attention to reduce redundant attention computations while focusing on periodic physiological patterns. Although more efficient than conventional transformer architectures, it still relies on computationally demanding frame-level feature extraction and remains substantially more resource-intensive than lightweight temporal models.

However, transformer-based models also have several limitations. Standard self-attention mechanisms can cause a quadratic computational complexity with respect to sequence length. The computational complexity results in a high memory consumption and computational cost. This challenge becomes more present when processing high-resolution videos or long temporal sequences required for accurate physiological estimation. Additionally, transformer models often require large training datasets and extensive computational resources, which limit their deployment on edge devices and real-time healthcare systems. Despite recent advances, achieving robust, lightweight, and generalized transformer-based rPPG systems remains an active research direction.

Despite differences in network architectures, many recent deep-learning and transformer-based rPPG methods rely on a preprocessing stage that detects and tracks the face before extracting a facial ROI for analysis. For example, PhysFormer and PhysFormer++ employ MTCNN-based face detection[20], while RhythmFormer relies on the rPPG-Toolbox[21] preprocessing pipeline, which can be configured to provide a static or dynamically enlarged face bounding box at fixed frame intervals. Although these approaches provide a standardized input for end-to-end learning, the resulting ROI pixels contain both physiological information and non-physiological spatial content, requiring the network to simultaneously learn physiological signal extraction and temporal modeling.

A fundamental trade-off in current rPPG research is highlighted in this paper. Traditional signal-processing methods are computationally efficient, but their performance is sensitive to motion and illumination variations. Deep learning and transformer-based approaches achieve robust accuracy by learning directly from raw video data. However, they suffer from increased computational complexity and memory requirements. Consequently, there is a need for a framework that preserves the computational cost efficiency of explicit traditional signal extraction while utilizing the adaptability of modern learning-based methods. This observation motivates the proposed lightweight signal-to-target learning framework.

3. Methodology

Remote photoplethysmography (rPPG) aims to recover subtle cardiovascular-induced skin reflectance variations from standard RGB videos. The physiological signal embedded within facial videos is extremely weak and is often distorted by motion artifacts, illumination variance, compression noise, and subject-dependent appearance variations. These challenges become more severe in lightweight systems where computational constraints limit the use of large spatiotemporal architectures.

To address these limitations, we propose a lightweight ROI-guided temporal framework for rPPG estimation that combines adaptive physiological region extraction with efficient multi-scale temporal modeling.

The proposed framework is specifically designed to operate under constrained computational settings while preserving sensitivity to periodic rPPG dynamics. Unlike conventional approaches that process the entire facial image using computationally expensive 2D or 3D feature extractors, the proposed method directly models temporally evolving physiological traces extracted from carefully selected facial regions. This strategy significantly reduces spatial redundancy while maintaining the information most relevant to blood-volume pulse (BVP) estimation. As shown in Fig. 1, the overall pipeline consists of four major stages:

- Dynamic physiological ROI extraction
- Multi-channel temporal signal construction
- Lightweight multi-scale temporal modeling
- Continuous pulse reconstruction and heart-rate estimation

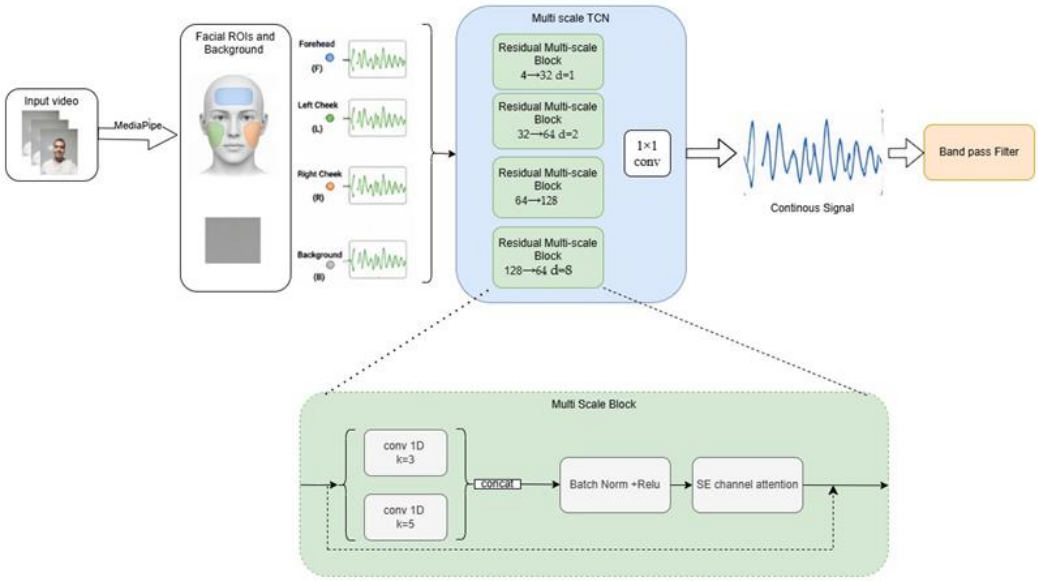

Figure 1 Overview of the proposed rPPG framework.

3.1 Dynamic Physiological ROI Extraction

The first stage of the proposed framework focuses on extracting stable physiological measurements from facial videos. Since rPPG signals originate from subtle skin reflectance changes caused by blood perfusion, selecting reliable skin regions is critical for robust estimation. Most models employ an outer face detection or large facial bounding-box ROIs during preprocessing [22]. Such static ROIs may become suboptimal under facial motion and pose variation, as they rely primarily on the model to learn how to distinguish movements or changes. However, the proposed framework utilizes the MediaPipe Face Landmarker [23] to dynamically localize dense facial landmarks for every frame. This enables continuous adaptation of the ROI geometry according to facial movement throughout the sequence.

Three physiologically informative facial regions are selected: the forehead (Rf), the left cheek (Rl), and the right cheek (Rr). These regions are preferred in the rPPG literature due to their relatively high skin visibility, high capillary density, and lower sensibility to non-rigid facial expressions compared to regions surrounding the mouth or eyes.

To operationalize the extraction, polygonal masks are constructed from predefined landmark subsets using convex hull[24] generation. Let L denote the set of coordinate points returned by the landmarker mesh. The precise landmark indices mapped to each anatomical region are defined as follows:

- **Forehead (R_f):** Indices {10, 109, 67, 107, 70, 63, 105, 66}
- **Left Cheek (R_l):** Indices {234, 93, 94, 95, 96, 97, 132, 133, 143, 144, 145, 146, 147, 150, 134}
- **Right Cheek (R_r):** Indices {361, 362, 363, 364, 365, 366, 367, 373, 374, 375, 376, 377, 378, 379, 380}

For a given frame $I_t \in R^{H\times W\times 3}$ at time t, where H and W are the height and width of the frame with 3 color channels RGB, the convex hull polygon H_k for region $k \in \{f, l, r\}$ is generated from its corresponding landmark coordinates. The binary spatial mask $M_{k,t} \in \{0, 1\}^{H \times W}$ is formulated via polygon filling. Compared with fixed bounding-box approaches, this landmark-guided strategy provides several key advantages. This adaptive strategy improves spatial consistency under rigid and non-rigid facial motion, and reduces contamination from non-skin pixels (e.g., background elements, hair, or eyes). Moreover, it improves structural stability against out-of-plane head rotations. The adaptive masks ensure that the extracted physiological traces remain strictly aligned with the underlying vascular regions even during continuous head movements.

3.2 Physiological Signal Representation

After ROI localization, the framework extracts temporal physiological traces from the green channel of the video sequence. The green channel (G) is prioritized because hemoglobin shows stronger optical absorption within green wavelengths (520–580 nm) compared to the surrounding tissue [25]. Consequently, pulsatile blood-volume variations introduce the most observable modulation in this spectral range. While the red channel penetrates deeper, it is dominated by static tissue reflectance, whereas the blue channel suffers from shallow penetration and high scattering. Thus, the green channel delivers the highest physiological signal-to-noise ratio under standard visible-light imaging conditions. Specifically, recent comparative studies[26] have demonstrated that the green channel yields higher correlation coefficients and lower error rates in heart rate estimation across varied recording conditions and subject activities when compared to the red and blue channels. For each facial mask $M_{k,t}$, the spatial average of the green channel intensity is computed at each frame, yielding a raw 1D temporal signal:

$$y_{k(t)} = \frac{\left(\Sigma\, G_{t(x,y)} \bullet M_{k,t}(x,y)\right)}{\left(\Sigma\, M_{k,t}(x,y)\right)} \quad (1)$$

where $G_{t(x,y)}$ represents the green channel pixel value, and $M_{k,t}(x, y)$ is the mask for ROI number k at coordinates (x, y) at frame t. Unlike end-to-end image-based networks that implicitly compress raw frames through heavy spatial convolutions, our approach explicitly constructs compact physiological descriptors before temporal modeling. This domain-specific preprocessing minimizes spatial redundancy and bypasses the vast computational overhead of processing full-frame video maps.

However, facial intensity variations are not exclusively driven by cardiovascular perfusion. Ambient illumination drifts, camera auto-exposure adaptations, and motion-induced shadowing can produce macro-level intensity fluctuations that mask or artificially mimic physiological patterns. To decouple these global disturbances from the true blood-volume pulse, an environmental background reference trace $y_{bg(t)}$ is extracted from a static 20 × 20 pixels non-facial region located at the top-left corner of each frame.

Because this background region contains no physiological information, its signal variation reflects purely scene-level lighting variations. By jointly modeling the facial traces alongside this background trace, the deep temporal network is given explicit cues to isolate globally shared illumination noise from true localized physiological periodicity.

The comprehensive representation for an entire video session is assembled into a multi-channel temporal feature matrix $S \in R^{N\times 4}$:

$$S = \begin{bmatrix} y_{f(1)} & y_{l(1)} & y_{r(1)} & y_{bg(1)} \\ y_{f(2)} & y_{l(2)} & y_{r(2)} & y_{bg(2)} \\ y_{f(3)} & y_{l(3)} & y_{r(3)} & y_{bg(3)} \\ \cdots & \cdots & \cdots & \cdots \\ y_{f(N)} & y_{l(N)} & y_{r(N)} & y_{bg(N)} \end{bmatrix} \quad (2)$$

where N is the total frame count of the video session. This compact matrix formulation compresses millions of video pixels into a dense four-channel coordinate

representation while preserving the multi-spatial physiological signals necessary for robust modeling.

3.3 Temporal Window Construction

To facilitate localized feature extraction and allow real-time operational capability, the full session matrix S is segmented into overlapping temporal windows before neural network processing. Each extracted window segment is defined as a tensor $X_0 \in R^{T \times 4}$, where the temporal window duration is fixed at T = 100 frames. At a standard video acquisition rate of f_s = 20-30 Hz, each processing window captures approximately 3-5 seconds of continuous physiological information. The window stride is set to 50 frames, yielding an exactly 50% overlap between adjacent segments. This windowing configuration balances two trade-offs:

1. Providing enough periods of cardiac cycles (approximately 3 to 6 beats per window, assuming a heart rate range of 60–120 BPM) to ensure stable spectral and temporal correlation profiling.
2. Restricting the sequence length to minimize the computational and operational memory complexities during forward passes on edge CPUs.

The overlapping sliding window mechanism serves as a reliable temporal data augmentation step during training, expanding the number of effective training instances without requiring additional data collection. Before being channeled into the network architecture, each individual window undergoes independent Z-score normalization along its temporal axis. For each channel c∈ {1, 2, 3, and 4}, the normalized window sequence $\hat{X}_{(t,c)}$ is computed via:

$$\hat{X}_{(t,c)} = \frac{(X_{(t,c)} - \mu_c)}{(\sigma_c + \varepsilon)} \quad (3)$$

where $\varepsilon = 10^{-6}$ is a stabilization parameter to prevent zero-division, and μ_c, σ_c are the mean and standard deviation for channel c. This normalization isolates the alternating current (AC) component, which carries the vascular pulsation information, by removing the large direct current (DC) baseline offset caused by skin pigmentation, static tissue thickness, and average ambient brightness levels.

3.4 Lightweight Multi-Scale Temporal Modeling

The core challenge in remote physiological assessment lies in distinguishing the weak, quasi-periodic cardiovascular pulse wave from the highly correlated artifacts introduced by motion and environmental lighting shifts. To solve this challenge under stringent edge-computing limitations, we introduce a customized lightweight multi-scale Temporal Convolutional Network [27].

Unlike recurrent networks (LSTMs or GRUs), which suffer from sequential processing latencies and gradient degradation over long sequences, 1D Temporal Convolutional Networks allow fully parallelized computations, exhibit stable gradient behaviors through residual loops, and provide explicit control over receptive field sizes. Furthermore, they require fewer parameters and lower computational overhead compared to transformer architectures that utilize expensive O (T^2) self-attention matrix operations, making them more suitable for deployment on standard edge CPUs.

The proposed Lite-Robust TCN is structurally organized into four stacked residual temporal blocks characterized by a progressively increasing dilation sequence:

$$d_i = 2^{i-1} \text{ for block } i \in \{1, 2, 3, 4\} \quad (4)$$

The incorporation of dilated convolutions enables an exponential expansion of the model's receptive field across successive layers without requiring deep architectural configurations or oversized kernel dimensions. Let k denote the kernel size; the receptive field R_F of a stacked network layer can be computed as:

$$R_F = 1 + \Sigma (k_i - 1) \bullet d_i \quad (5)$$

By applying dilations d ∈ {1, 2, 4, 8}, the model's temporal observation window expands to structurally cover localized systolic transitions, complex multi-site pulse wave morphologies, and long-range periodic cardiac patterns within a highly compressed block sequence.

3.5 Multi-Scale Temporal Feature Extraction

The design of each temporal block in the Lite-Robust TCN is a specialized dual-branch multi-scale feature extraction framework. The physiological foundation behind this design is the non-uniform temporal profiles intrinsic to cardiac waveforms. A typical blood-volume pulse wave is composed of asymmetric phases: a fast, high-frequency systolic upstroke, followed by a slower, lower-frequency diastolic decay. Modeling these asymmetric temporal dynamics with a uniform kernel size often causes the network to over-smooth peak locations or miss broader rhythmic trends. To address this, each block splits its incoming feature channels $H_i \in R^{Cin \times T}$ into two parallel 1D convolutional paths:

$$H_{path1} = Conv1D(H_{i-1}, k = 3, dilation = d_i)$$
$$H_{path2} = Conv1D(H_{i-1}, k = 5, dilation = d_i) \quad (6)$$

The first path uses a tight kernel size k=3 to extract high-frequency local variations and prevent the temporal blurring of sharp peaks. The second path employs a wider kernel size k=5 to capture low-frequency macro-rhythms and broader waveform contextual trends. Both paths use identical dilation factors d_i tailored to that specific block layer. Crucially, the total target output width C_{out} of the block is split equally between these two branches, such that each branch outputs exactly C_{out} /2 feature channels. This split-channel architecture restricts the computational FLOP footprint while maximizing multi-scale feature extraction capabilities. The feature representations from both branches are fused via concatenation along the channel dimension, where H_{fused} is the concatenation result of H_{path1} and H_{path2}:

$$H_{fused} = Concat(H_{path1}, H_{path2}) \quad (7)$$

The combined feature map is then routed through a Batch Normalization layer, a Rectified Linear Unit (ReLU) non-linearity, and a Squeeze-and-Excitation channel attention block before being combined with the layer's identity shortcut. The network tracks dimensions explicitly through the following feed-forward cascade:

- **Input State:** $X \in R^{4\times 100}$
- **Residual Block 1 (d=1):** Transforms 4 - 32 feature maps via parallel kernel allocations $C_{out}/2 = 16$.
- **Residual Block 2 (d=2):** Transforms 32 - 64 feature maps $C_{out}/2 = 32$.
- **Residual Block 3 (d=4):** Transforms 64 - 128 feature maps $C_{out}/2=64$.
- **Residual Block 4 (d=8):** Compresses 128 - 64 feature maps $C_{out}/2 =32$ to prepare for output projection.

The concatenated multi-scale feature map, denoted H_{fused}, is passed through Batch Normalization and a ReLU activation. Let this resulting fused feature tensor be defined as $U \in R^{C\times T}$, where C is the total number of channels and the temporal window T = 100 frames.

3.6 Motion-Aware Channel Attention

A core challenge in practical rPPG setups is that head movements or expression changes affect different facial regions unevenly. For instance, speaking or jaw movements introduce substantial non-rigid motion noise across the cheek regions while leaving the forehead relatively stable. Conversely, head tilts or lighting changes may corrupt the forehead region while the cheeks remain clear. To dynamically handle these spatial reliability changes without expanding the model's footprint, a lightweight Squeeze-and-Excitation (SE) attention module is embedded within every temporal block [28].

The SE module performs channel-wise feature recalibration. This module adaptively prioritizes cleaner physiological streams over corrupted ones. Given a multi-scale feature tensor U, the block operates via a two-stage process:

1. **Squeeze Function:** The temporal dimension T is squeezed using Global Average Pooling to generate a global channel descriptor vector $z \in R^{C\times 1}$:

$$z_c = \left(\frac{1}{T}\right) \bullet \Sigma\, U_c(t) \quad (8)$$

Where T is the temporal window size and $U_c(t)$ is the fused feature for channel c at time t. This operation condenses the full temporal profile of each channel into a single scalar value representing its global activation energy across the window.

2. **Excitation Function:** To capture non-linear inter-channel dependencies, a bottleneck structure with a reduction ratio r=4 is implemented using two fully connected layers:

$$s = \sigma\big(W_2 \bullet ReLU(W_1 \bullet z)\big) \quad (9)$$

where $W_1 \in R^{\frac{C}{4}\times C}$ compresses the descriptor vector, $W_2 \in R^{C\times \frac{C}{4}}$ expands it back to the original channel size, and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the Sigmoid activation. The output vector $s \in R^{C\times 1}$ represents the calculated attention weights for each feature map.

Finally, the element-wise multiplication of the attention weights and the original feature map is scaled as:

$$\tilde{U}_c = s_c \bullet U_c \quad (10)$$

Through this mechanism, if a specific facial region encounters heavy motion artifacts, the network assigns lower weights to its corresponding feature maps. This allows cleaner physiological streams (such as the forehead) to guide the prediction, enhancing motion robustness without requiring expensive spatial tracking updates.

3.7 Continuous Pulse Reconstruction

The final feature map $H_4 \in R^{64\times 100}$ resulting from the fourth residual block contains highly dense, motion-corrected temporal representations. This feature representation is projected into a single-channel raw rPPG pulse estimate $\hat{y}_m \in R^{1\times 100}$ using a point-wise 1×1 convolutional layer:

$$\hat{y}_m = Conv1D(H_4, k = 1) \quad (11)$$

Because inference is performed on overlapping sliding windows, individual segment predictions must be combined to result in a unified, continuous pulse waveform for the entire video session. We implement an overlap-averaging reconstruction strategy. Let M denote the total number of window segments extracted from a video sequence. An accumulation vector $v_{pred} \in R^N$ and a weight tracker vector $v_{weight} \in R^N$ are initialized to zero. For each window index $m \in \{0, \ldots, M-1\}$, the segment prediction is mapped to its original position in the full sequence. The final continuous raw reconstructed rPPG waveform is obtained by element-wise division.

The mixing of this window decreases the boundary discontinuities at the segment transitions, smooths out the localized estimation errors, and keeps the temporal continuity in long video. To eliminate remaining out-of-band noise, such as high-frequency camera jitter or low-frequency baseline drifts from global lighting shifts, the reconstructed signal is filtered using a zero-phase 4th-order Butterworth bandpass filter. The cutoff frequencies are strictly bounded between 0.7 Hz and 2.5 Hz, which limits the heart rate range to physiologically realistic values (42–150 BPM).

3.8 Hybrid Supervision Strategy

Optimizing rPPG models purely through Mean Squared Error often proves insufficient. The MSE Loss forces the network to minimize point-to-point amplitude errors. However, it can cause the model to learn the general shape of the waveform with shifting in phase or frequency under noisy conditions. Therefore, we utilize a hybrid loss

optimization that combines MSE, Pearson correlation, and spectral losses:

$$L_{total} = \alpha L_{MSE} + \beta L_{Pearson} + \gamma L_{Spectral} \quad (12)$$

The loss hyper-parameters are set to a balanced ratio of α = 1.0, β = 5.0, and γ = 1.0 for stable multi-domain optimization.

1. Temporal Domain Supervision (L_{MSE}): The mean squared error forces the model to align the overall amplitude and value distribution of the predicted waveform ŷ with the ground-truth BVP signal y at time (t):

$$L_{MSE} = \left(\frac{1}{T}\right) \bullet \Sigma \left(\hat{y}(t) - y(t)\right)^2 \quad (13)$$

2. Phase Domain Supervision ($L_{Pearson}$): To enforce morphological synchronization and avoid phase alignment shift, we utilize the negative Pearson correlation coefficient. This loss term focuses on the synchronization of systolic peaks independent of amplitude scale:

$$L_{Pearson} = 1 - \frac{[\Sigma(\hat{y}-\mu_{\hat{y}})(y-\mu_y)]}{sqrt\left(\Sigma(\hat{y}-\mu_{\hat{y}})^2 \bullet \Sigma(y-\mu_y)^2\right)} \quad (14)$$

Where ŷ and y denote the respective temporal means within the current target window.

3. Frequency Domain Supervision ($L_{Spectral}$): Spectral loss is used to ensure frequency and spectral consistency estimation under motion. This loss converts both the predicted and ground-truth signals to the frequency domain using a one-dimensional Fast Fourier Transform (FFT). The Power Spectral Densities (P) are computed and normalized into probability distributions. The spectral loss is formulated as the Kullback-Leibler [29] divergence between the two frequency distributions:

$$L_{Spectral} = \Sigma P_{y(f)} \bullet \log\left(\frac{\left(P_{y(f)}+\delta\right)}{\left(P_{\hat{y}(f)}+\delta\right)}\right) \quad (15)$$

where $\delta = 10^{-8}$ is an anti-log stabilization parameter set according to [29]. This hybrid optimization strategy guides the model across multiple domains, allowing it to predict clear, stable, and frequency-accurate heart rate waveforms even under computational constraints.

3.9 Regularization via ROI Dropout

To further protect the network from overfitting to a specific facial area and to ensure robust generalization during localized motion corruption, we implement an online data augmentation technique termed ROI Dropout within the data loader [30].

During the training phase, for each forward step, a randomized masking check is executed. With a probability of $P_{drop} = 0.20$, the network selects one of the three primary facial input channels at random (either Forehead-Green, LeftCheek-Green, or RightCheek-Green) and sets its entire temporal window sequence to zero. This regularization strategy prevents the network from relying exclusively on a single spatial region for pulse tracking. Instead, it forces the network to learn how to extract physiological patterns from any combination of available clean facial areas. This mechanism directly supports the model's ability to maintain accurate predictions when a subject's facial region is obscured or heavily distorted by motion artifacts.

4. Experimental results

Performance of the proposed model is evaluated against traditional, deep learning models, and Transformer-based models. To evaluate the efficiency and generalization capabilities of the proposed lightweight Multi-scale TCN architecture, systematic validation experiments were conducted on two public datasets.

4.1 Datasets details

To evaluate the proposed multi-scale lite TCN, the performance is benchmarked on two datasets, UBFC-rPPG and COHFACE. The datasets details are listed in Table 1; they represent different forms of challenges.

UBFC-rPPG [31]: The dataset includes 42 participants playing a stress-inducing math game. The logitech C920 HD Pro webcam is used to record videos at 640×480 resolution and a frame rate of 30 fps. The capturing environment was characterized by good illumination with a varying amount of sunlight. Videos are saved at raw resolution with no compression (bitrate: 220Mbps) to capture subtle variations in face colors. The dataset contains one video per participant, with an average length of one minute (a total of 42 videos).

COHFACE[32]: The dataset includes 40 subjects. Four videos are recorded for each subject under different environmental conditions: clean (suitable lab illumination) and natural (lab lights are off, only sunlight from windows). Therefore, the dataset includes 160 video samples with an average length of one minute. The logitech HD C525 webcam is used to record videos at 640×480 resolution and a frame rate of 20 fps. All videos are saved with high compression (bitrate: 250kbps). High compression leads to compression distortions that interfere with the subtle skin color variations.

Table 1 Datasets details and differences

Dataset	Videos	Subjects	Main Challenge	Illumination Source	Compression	Ground Truth
UBFC-rPPG	42	42	Spontaneous Reaction	Indoor	No compression(data rate~220Mbps)	CMS50E (PPG)
COHFACE	160	40	Illumination Compression rate	Lab/Natural	High(data rate ~ 250 kbps)	BVP

4.2 Implementation details

To maximize training throughput and optimize CPU execution time, the dynamic ROI tracking and four-channel temporal signal construction are executed entirely as an offline pipeline prior to model training and evaluation. For each video sequence in the database, the raw spatial average intensities for the four ROIs are tracked for each frame and saved directly to non-volatile storage. During the initialization of both the training and testing data loaders, these 4 channel signals are streamed directly into memory.

The framework is evaluated using a 5-fold subject-disjoint cross-validation protocol. In each fold, approximately 80% of the subjects are used for training, and the remaining 20% are reserved for testing. Subjects appearing in the testing fold are never included in the corresponding training set. No separate validation set is employed, and the model is trained for all epochs. The reported performance metrics correspond to the average results across all five folds. Prior to temporal segmentation, all video sequences were standardized to a common frame rate. The UBFC-rPPG videos were down sampled from their original frame rate to 20 fps to ensure a consistent temporal resolution across datasets and ensure uniform model training and evaluation.

Using a temporal window length of 100 frames and a stride of 50 frames, each video sequence is segmented into overlapping temporal samples. For the UBFC-rPPG dataset, each fold contains 9 testing subjects and 33 training subjects (each subject generating approximately 23 temporal windows in average). Similarly, for the COHFACE dataset, each fold contains 8 testing subjects and 32 training subjects (each subject generating approximately 92–93 temporal windows). Across the entire dataset, this preprocessing generates 972 samples for UBFC-rPPG and 3714 samples for COHFACE. Due to small variations in video duration, the number of generated windows differs slightly across subjects and recordings. On average, each fold contains approximately 759 training samples and 207 testing samples for UBFC-rPPG, and 2970 training samples and 740 testing samples for COHFACE.

The network is trained for a maximum duration of E = 50 epochs. The batch size is configured to B = 64 samples to balance between memory caching and gradient vectorization on edge processors. A dynamic Cosine Annealing learning rate scheduling strategy [33] is used in network optimization. The execution is initialized with a starting warm up learning rate of max = 10^{-3}, which decays dynamically through the training path following a cosine curve down to reach a minimum threshold of min = 0. Network parameters are updated using the Adam optimizer with a decoupled weight decay regularizer set to lambda = 10^{-4} to overcome over fitting across the multi-scale kernels.

This dynamic schedule provides critical structural benefits during optimization. The initial learning rate of 10^{-3} ensures a warm-up gradient descent during initial epochs, which allows the model to quickly learn spatial weights. As training progresses into advanced epochs, the learning operation slows down due to the smooth decay of the cosine function. This gradual deceleration allows the model to fine-tune phase relationships and the precise peak alignments. This strategy prevents gradient oscillations around narrow local minima.

To prove the lightweight, mobile-deployable nature of the proposed rPPG framework, the entire system was trained, deployed, and tested on a low power processor, an Intel Core i5-1135G7 CPU (11th Generation, 2.40 GHz). No external graphical acceleration (GPU) was used. This hardware profile highlights the edge viability of our system. While state-of-the-art spatio-temporal transformer models require high-end GPUs (like NVIDIA RTX series) to tolerate real-time inference matrices, our model runs seamlessly within standard CPU allocations. This hardware shows that the framework can be deployed directly as an on-device utility inside mobile applications without thermal issues or high power consumption.

4.3 Results and Discussion

Performance is evaluated using three standard statistical metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Pearson correlation coefficient (r). Since the proposed framework combines explicit physiological signal extraction with learning-based temporal modeling, its performance is compared against representative traditional signal-processing methods as well as deep-learning and transformer-based rPPG approaches. The comparative results are presented in Table 2.

As shown in Table 2, results demonstrate that the proposed Multi-scale TCN bridges the performance gap between heavy spatial deep networks and lightweight operational constraints. On the UBFC-RPPG dataset, the proposed model achieves an MAE of 0.86 bpm, an RMSE of 1.64 bpm, and a high linear correlation coefficient of r = 0.91. This represents a substantial improvement over traditional mathematical source-separation techniques. Specifically, the model drops the MAE by over 3.20 bpm compared to CHROM and POS, while improving the correlation coefficient. When compared with end-to-end convolutional models such as DeepPhys and PhysNet, which implicitly extract representations from full spatial video arrays, the proposed method achieves lower estimation errors. It outperforms DeepPhys MAE = 3.25 bpm and PhysNet MAE = 2.95bpm on UBFC-RPPG. This indicates that compressing spatial video matrices into compact, landmark-guided multi-site physiological signals prior to training can preserve sufficient physiological information for accurate heart-rate estimation. Instead, by reducing spatial noise outside the designated regions of interest, the network can allocate more of its capacity to modeling temporal physiological variations.

On the highly challenging COHFACE dataset, which includes ambient illumination fluctuations and motion degradation, the proposed architecture maintains strong stability, yielding an MAE of 1.95 bpm, an RMSE of 2.40 bpm, and a correlation of r = 0.72. On this partition, spatially driven baselines degrade noticeably. For example, PhysNet's error escalates to 5.38 bpm, and DeepPhys drops to an r of 0.34. Our model's flexibility under these conditions may be attributed to the combination of dual-branch temporal convolutions and the SE attention mechanism. The parallel kernel branches (sizes 3 and 5) capture temporal information at different resolutions, while the temporal Squeeze-and-Excitation layer adaptively reweights feature channels according to their relative importance.

Although the computationally intensive transformer-based architectures achieve the highest accuracy due to their vast parameter capacity, our proposed lightweight Multi-scale TCN delivers competitive performance while narrowing the accuracy-computation gap. When evaluated against standard traditional pipelines and representative deep convolutional neural networks, our framework achieves competitive accuracy, establishing a favorable trade-off between minimized computational complexity and physiological estimation accuracy.

Table 2 Performance comparison of the proposed Multi-scale TCN against representative traditional, deep-learning, and transformer-based models on the UBFC-RPPG and COHFACE datasets.

Method		**UBFC**			**COHFACE**		
		MAE (bpm)↓	RMSE (bpm)↓	r ↑	MAE (bpm) ↓	RMSE (bpm)↓	r ↑
Traditional	GREEN[25]	6.01	7.87	0.29	-	-	-
	CHROM[12]	4.06	8.83	0.27	7.82	10.8	0.67
	POS[11]	4.08	7.62	0.68	8.14	10.57	0.49
Deep-learning	Ts-CAN[15]	1.70	2.72	0.92	-	-	-
	DeepPhys[13]	3.25	6.81	0.65	4.56	7.84	0.34
	PhysNet[34]	2.95	3.67	0.97	5.38	10.76	0.48
Transformer	PhysFormer[17]	0.52	0.71	0.98	-	-	-
	PhysFormer++[18]	0.51	0.69	0.98	-	-	-
	RhythmFormer[19]	0.5	0.87	0.99	1.17	3.36	0.97
Light CNN	Multi scale TCN(proposed)	0.86	1.64	0.91	1.95	3.87	0.72

It should be noted that the compared methods belong to different methodological categories, including traditional signal-processing approaches, CNN-based models, transformer-based architectures, and the proposed hybrid signal-to-target learning framework. To maximize consistency, results were obtained either from the corresponding original publications or by evaluating the publicly available official implementations on the same datasets and evaluation protocol whenever possible. Nevertheless, differences in preprocessing pipelines (e.g., face detection, ROI selection), input representation (raw ROI pixels vs. pre-extracted 1D signals), and hardware (GPU vs. CPU) are inherent to these different methodological families. Therefore, the comparison is intended to provide a representative benchmark of accuracy–efficiency trade-offs across different rPPG paradigms, while preserving the original characteristics of each method. In addition to accuracy, computational complexity and runtime characteristics are reported separately to provide a more comprehensive evaluation of practical deployment suitability.

4.4 Computational analysis

The primary contribution of this work lies in achieving a favorable trade-off between minimized hardware complexity and physiological estimation accuracy. Unlike most deep-learning and transformer-based rPPG methods that process facial ROI pixels directly, the proposed framework operates on compact physiological signals extracted during preprocessing. This design shifts the learning task from joint spatial-temporal representation learning to temporal modeling only. This design modification results in reducing the computational burden of the neural network while introducing only a limited preprocessing cost.

When evaluated against both transformer and representative deep learning models, our framework strongly reduces computational requirements while maintaining competitive error boundaries. To validate this architectural efficiency, Table 3 compares the total parameter volume, floating-point operations (FLOPs) per frame, and root mean squared error (RMSE) of the proposed network against prominent baseline methods from our comparative evaluation.

As demonstrated by the empirical resource profiles in Table 3, the proposed model achieves a substantial reduction of hardware requirements. Our network requires approximately 0.1M parameters, which represents an approximate 39x parameter reduction over deep 2D spatial models (DeepPhys, TS-CAN) and nearly a 100x architectural reduction when contrasted with state-of-the-art transformer baselines like PhysFormer++.

More crucially, the computational complexity of our system drops to a minimal footprint of just 5.5M FLOPs. Standard end-to-end networks are severely constrained on edge hardware because they must continuously propagate massive 3D or 2D spatial pixel tensors through deep convolutional or self-attention blocks, creating unsustainable memory caching overhead and high floating-point dependencies.

Our pipeline addresses this constraint by completely isolating spatial image-processing redundancies prior to network propagation using landmark-guided region-of-interest (ROI) extraction. Since the neural layers are specialized strictly to process pre-extracted, 1D temporal multi-site trajectories, the model achieves an exceptionally low complexity of 5.5M FLOPs per frame. This computational profile allows the system to achieve an RMSE of 1.64 bpm, establishing a favorable balance between estimation accuracy and computational efficiency that runs seamlessly in real-time within standard single-thread CPU bounds on low-power edge mobile processors.

4.5 Error Distribution and Agreement Analysis

To evaluate the error distributions of our proposed model, scatter plots, error distribution histograms and Bland-Altman analyses are presented for COHFACE (left) and UBFC-rPPG (right) datasets. Fig. 2 shows the scatter plot of predicted versus ground-truth heart rates (top), the error distribution histogram (center), and the Bland-Altman plot (down). The scatter points cluster closely around the identity line in the two datasets, indicating strong agreement. The error histogram is approximately centered at zero with a slight positive bias, and 95% of the errors lie within ±4 bpm for COHFACE and ±2 BPM for UBFC-rPPG. The Bland-Altman plots show a negligible mean bias for results of COHFACE with 95% limits of agreement ranging from -4.51 to 4.04 for the COHFACE and -2.87 to 2.98 for the UBFC-rPPG. These results demonstrate that our lightweight multi-scale TCN achieves accurate and unbiased heart rate estimation, making it suitable for real-time, CPU-based rPPG applications.

Results for two videos are shown in Fig. 3, one from the COHFACE dataset (top) and one from the UBFC-rPPG dataset (bottom), with representative predicted BVP signals (blue dashed) overlaid on the ground-truth BVP (red solid). Both examples show the first 10 seconds. The COHFACE example (top panel) exhibits a slightly noisier predicted signal, especially during the first 2-3 seconds, but the overall periodic structure remains correct. The model occasionally underestimates the peak amplitude, yet the heart rate is still accurately estimated (error 1 BPM). For the UBFC example (bottom panel), the predicted waveform follows the ground truth in both shape and phase. The systolic peaks are accurately localized, and the amplitude modulation is well captured. This high similarity is reflected in the Pearson correlation coefficient of 0.94 for this video. The higher signal-to-noise ratio in UBFC-rPPG is expected because this dataset includes more controlled motion and lighting variations.

Despite this, the model is able to recover the cardiac frequency in both datasets, which proves the model is accurate on datasets with different conditions. The slightly lower results on COHFACE indicate that more data augmentation may help to improve the performance on more challenging real-world videos.

Table 3 Computational and Performance Metrics

Method	Type	Parameters (M)	MACs	RMSE
DeepPhys	2D-CNN	3.91	744	6.81
TS-CAN	2D+shift	3.91	744	2.72
PhysNet	3D-CNN	0.78	429	3.67
PhysFormer	Transformer	7.38	293	0.71
PhysFormer++	Transformer	9.79	311	0.69
RhythmFormer	Transformer	3.25	240	0.87
Multi scale TCN(proposed)	Multi scale TCN	0.1	5.5	1.64

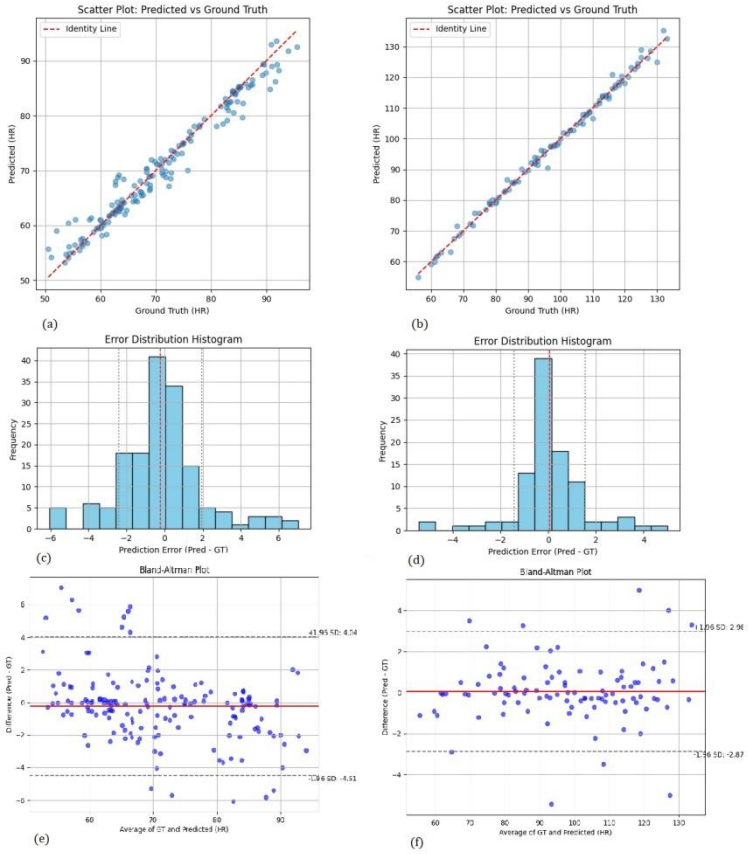

Figure 2 Quantitative evaluation on the COHFACE (left) and UBFC-rPPG (right) datasets. Top row: Scatter plots of predicted vs. ground-truth heart rates (red circles) with the identity line (dashed black). Middle row: Histograms of prediction error (predicted – ground truth). Bottom row: Bland-Altman plots showing the difference against the average heart rate

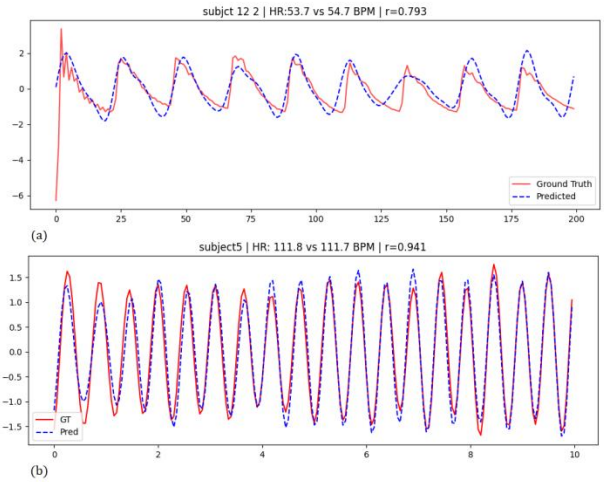

Figure 3 Representative BVP waveform predictions. Ground-truth (red) and predicted (blue) signals for a 10-second segment from COHFACE (top) and UBFC-rPPG (bottom), demonstrating close phase and shape alignment.

4.6 Runtime Analysis

To evaluate the practical deployment capability of the proposed framework, runtime measurements were performed on an Intel Core i5-1135G7 CPU without GPU acceleration. Table 4 reports the average runtime per frame/ per window for each pipeline stage. The results show that the proposed temporal learning model introduces only a small computational overhead, requiring 3.14 ms for TCN inference per temporal window. The majority of the computational cost originates from the facial landmark detection stage (2.057 ms per frame), while ROI extraction and signal preprocessing require less than 1 ms per frame. The complete processing pipeline requires approximately 4.05 s to process a full video sequence of length 60 seconds at 20 fps.

Table 4 Runtime analysis of the proposed framework measured on an Intel Core i5-1135G7 CPU. Processing times are reported for individual stages of the pipeline.

Stage	Time(ms)
Background extraction(per frame)	0.068
Face landmark detection(per frame)	2.057
ROI extraction (3 regions) (per frame)	0.263
Preprocessing (normalization, per window)	0.05
TCN inference (per window)	3.14
Overlap-add reconstruction (per window)	0.012
Total processing time (average for 60-second videos)	4053

Since the model operates exclusively on low-dimensional temporal signals rather than spatial video frames, its computational cost remains suitable for deployment on resource-constrained edge devices. The total processing time of 4.05 seconds for a 60-second video corresponds to a speedup factor of ≈15× relative to real-time. The cost per frame, including all stages, is approximately 3.4 ms, enabling a throughput of >290 FPS.

4.7 Ablation study

To validate the contribution of each proposed component, we conducted an incremental ablation study on the COHFACE dataset. Starting from a minimal baseline, a single kernel (k=3) 1D TCN with three facial ROI inputs, no SE attention, no ROI Dropout, and standard MSE loss, we sequentially added: (1) hybrid loss, (2) background reference channel, (3) multi scale kernels (parallel k=3 and k=5), (4) SE channel attention, and (5) ROI Dropout.

As a reference, we also evaluated a traditional configuration consisting of the proposed dynamic ROI extraction followed by conventional Green-channel heart-rate estimation without any learning component. As shown in Table 5, this configuration achieved an MAE of 6.43 BPM. Replacing the fixed signal-processing estimation stage with the proposed learning-based baseline reduced the MAE to 3.12 BPM, demonstrating the benefit of learning the relationship between the extracted signals and the ground-truth heart-rate signal. The baseline model tends to bias its predictions toward the dominant heart-rate range of the training distribution. For subjects with heart rates significantly below or above the population mean, the baseline produces predictions regressed toward the mean, resulting in large absolute errors. The Pearson correlation between predicted and ground-truth BVP waveforms is also low, indicating poor temporal alignment.

The addition of the hybrid loss improved waveform similarity and temporal consistency between the predicted and reference signals. As a result, the model became more responsive to heart-rate variations rather than producing predictions concentrated around the dominant range of the training data. However, estimation errors remain relatively high, indicating that improving waveform correlation alone is insufficient to fully capture all physiological variations. Adding the explicit environmental noise reference (background intensity) provides the network with information related to global illumination variations. This reduces errors in videos with strong lighting variations.

Table 5 Incremental ablation study on the COHFACE dataset

Component	MAE
GREEN+ Dynamic ROI (no learning)	6.425
Base Model	3.121
Hybrid loss	2.897
Background	2.608
Multi-scale	2.237
SE attention	2.039
ROI Dropout	1.958

Introducing parallel k=3 and k=5 convolutions with exponential dilation provides the largest single improvement. The multi-scale design provides temporal information at different resolutions, which appears to improve the modeling of both short-term and long-term signal variations. The multi-scale design appears particularly beneficial for lower-HR samples, suggesting that information at different temporal resolutions helps represent slower physiological variations more effectively. However, errors remain higher in recordings containing substantial motion artifacts, suggesting that temporal modeling alone is insufficient when some facial regions become unreliable.

Adding Squeeze-and-Excitation channel attention enables the model to dynamically reweight the three facial ROI channels based on their signal quality. The SE module allows the network to place greater emphasis on more informative temporal representations. This directly reduces errors in high-motion samples, as the network can selectively attend to reliable physiological streams. The improvement (−0.20 BPM) is consistent across all motion severities.

Finally, ROI Dropout randomly masks entire facial ROI channels during training. This method prevents over-fitting to any particular face region. The improvement (−0.08 BPM) indicates that the model generalizes better to test samples where one ROI is temporarily blocked or corrupted. The full model achieves a 37% relative MAE reduction compared to baseline (3.12 to 1.96 BPM), indicating that each component contributes to addressing different limitations of the baseline configuration.

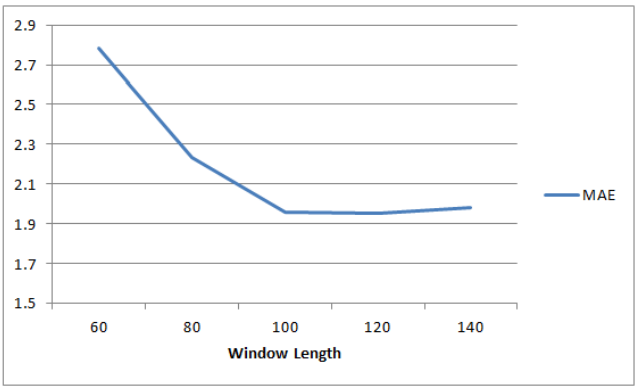

Figure 4 Sensitivity analysis of the window Length on the COHFACE dataset

Effect of window length: The impact of the input window length is also investigated on model performance. Using the full proposed architecture, we trained and evaluated models on COHFACE with window lengths of 60, 80, 100, 120, and 140 frames (stride fixed at half the window length to maintain 50% overlap). As shown in Fig. 4, the MAE decreases from 2.78 BPM (60 frames) to 1.95 BPM (100 frames). Shorter windows contain fewer cardiac cycles and do not provide enough temporal context to estimate HR. Therefore, it results in lower accuracy. Increasing the window length to 140, results in a slight decrease in accuracy to 1.98 BPM. Therefore, a window length of 100 frames provides long temporal context, which is sufficient for accurate heart rate estimation. Longer windows do not improve accuracy but increase the computational cost. Therefore, a window length of 100 frames is adopted for all experiments in this paper.

Effect of the loss function: The effect of the proposed loss components was evaluated using four configurations: MSE only, MSE + Pearson, MSE + Spectral, and the complete hybrid loss. The results are presented in Table 6. The MSE-only configuration produced the highest error (2.55 BPM). Adding the Pearson term reduced the MAE to 2.14 BPM, whereas the spectral term alone reduced it to 2.32 BPM. The improvement obtained with the Pearson term was therefore larger than that obtained using spectral supervision alone.

Combining the three loss components resulted in the lowest error (1.95 BPM). These results show that each loss term contributes different information during training. The MSE loss constrains the overall prediction error. Furthermore, the Pearson loss promotes the similarity of predicted and ground-truth waveforms, and the spectral loss helps to preserve the frequency characteristics of the signal. The combination of the three objectives resulted in the most accurate estimates on the COHFACE dataset.

Table 6 Ablation of the hybrid loss COHFACE dataset

Loss Type	MAE
Mse	2.55
Mse+ Pearson	2.14
Mse+Spectral	2.32
Hybrid	1.95

5. Conclusion

In this work, a lightweight multi-scale temporal convolutional network is presented for remote heart rate estimation. By replacing full-frame spatiotemporal processing with adaptive landmark-guided ROI extraction, we reduced the input space to a compact four-channel temporal representation. The proposed architecture uses parallel convolutions with increasing dilation to model multi-scale cardiac dynamics, while squeeze-and-excitation attention and ROI dropout provide robustness to motion and overfitting. Our hybrid loss ensures that the network learns not only waveform amplitude but also phase alignment and spectral content.

The experimental results on two challenging datasets confirm the effectiveness of the approach. On UBFC-rPPG, the model achieved an MAE of 0.86 BPM and an RMSE of 1.64 BPM; on COHFACE, it attained an MAE of 1.95 BPM and an RMSE of 3.87 BPM. With only 0.1 M parameters and 5.5 MFLOPs per frame, the model operates in real time on a standard CPU, without any GPU acceleration. Compared to transformer-based methods that require tens of millions of parameters, our system offers a favourable trade-off between accuracy and computational cost.

Future work will explore lightweight temporal attention mechanisms and adaptive fusion of multiple physiological channels to further improve robustness. Additionally, we plan to extend the framework to estimate other cardiovascular parameters, such as blood pressure and oxygen saturation, using the same low-complexity design.

REFERENCES

[1] R. Macwan, Y. Benezeth, and A. Mansouri, "Heart rate estimation using remote photoplethysmography with multi-objective optimization," *Biomed. Signal Process. Control*, vol. 49, pp. 24–33, Mar. 2019, doi: 10.1016/j.bspc.2018.10.012.

[2] R. Karthick, M. S. Dawood, and P. Meenalochini, "Analysis of vital signs using remote photoplethysmography (RPPG)," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 12, pp. 16729–16736, Dec. 2023, doi: 10.1007/s12652-023-04683-w.

[3] L. Sörnmo and P. Laguna, "Electrocardiogram (ECG) Signal Processing," in *Wiley Encyclopedia of Biomedical Engineering*, 1st ed., M. Akay, Ed., Wiley, 2006. doi: 10.1002/9780471740360.ebs1482.

[4] "Accuracy of pulse oximeters in estimating heart rate at rest and during exercise. | British Journal of Sports Medicine." Accessed: May 12, 2026. [Online]. Available: https://bjsm.bmj.com/content/25/3/162.short

[5] H. W. Loh *et al.*, "Application of photoplethysmography signals for healthcare systems: An in-depth review," *Comput. Methods*

Programs Biomed., vol. 216, p. 106677, Apr. 2022, doi: 10.1016/j.cmpb.2022.106677.

[6] G. de Haan and V. Jeanne, "Robust Pulse Rate From Chrominance-Based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013, doi: 10.1109/TBME.2013.2266196.

[7] "Intelligent Remote Photoplethysmography-Based Methods for Heart Rate Estimation from Face Videos: A Survey." Accessed: May 12, 2026. [Online]. Available: https://www.mdpi.com/2227-9709/9/3/57

[8] N. Nguyen, L. Nguyen, H. Li, M. Bordallo López, and C. Álvarez Casado, "Evaluation of video-based rPPG in challenging environments: Artifact mitigation and network resilience," *Comput. Biol. Med.*, vol. 179, p. 108873, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108873.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[10] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Opt. Express*, vol. 18, no. 10, pp. 10762–10774, May 2010, doi: 10.1364/OE.18.010762.

[11] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic Principles of Remote PPG," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017, doi: 10.1109/TBME.2016.2609282.

[12] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013, doi: 10.1109/TBME.2013.2266196.

[13] W. Chen and D. McDuff, "DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks," in *Computer Vision – ECCV 2018*, vol. 11206, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11206. , Cham: Springer International Publishing, 2018, pp. 356–373. doi: 10.1007/978-3-030-01216-8_22.

[14] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," jultika.oulu.fi. Accessed: Mar. 15, 2026. [Online]. Available: https://oulurepo.oulu.fi/handle/10024/28818

[15] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 19400–19411. Accessed: Apr. 20, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/e1228be46de6a0234ac22ded31417bc7-Abstract.html

[16] Z. Yang, H. Wang, and F. Lu, "Assessment of Deep Learning-Based Heart Rate Estimation Using Remote Photoplethysmography Under Different Illuminations," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 6, pp. 1236–1246, Dec. 2022, doi: 10.1109/THMS.2022.3207755.

[17] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. Torr, and G. Zhao, "PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 4176–4186. doi: 10.1109/CVPR52688.2022.00415.

[18] Z. Yu *et al.*, "PhysFormer++: Facial Video-Based Physiological Measurement with SlowFast Temporal Difference Transformer," *Int. J. Comput. Vis.*, vol. 131, no. 6, pp. 1307–1330, Jun. 2023, doi: 10.1007/s11263-023-01758-1.

[19] B. Zou, Z. Guo, J. Chen, J. Zhuo, W. Huang, and H. Ma, "RhythmFormer: Extracting patterned rPPG signals based on periodic sparse attention," *Pattern Recognit.*, vol. 164, p. 111511, Aug. 2025, doi: 10.1016/j.patcog.2025.111511.

[20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[21] X. Liu *et al.*, "rPPG-Toolbox: Deep Remote PPG Toolbox," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 68485–68510, Dec. 2023.

[22] J. A. S. Y. Jayasinghe, S. Katsigiannis, and L. Malasinghe, "Comparative Study of Face Tracking Algorithms for Remote Photoplethysmography," in *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Nov. 2023, pp. 1–7. doi: 10.1109/ICECET58911.2023.10389182.

[23] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," Jun. 01, 2019, *arXiv*. doi: 10.48550/arXiv.1906.08172.

[24] G. R.L., "An efficient algorith for determining the convex hull of a finite planar set," *Inf. Process. Lett.*, vol. 1, no. 4, pp. 132–133, Jun. 1972, doi: 10.1016/0020-0190(72)90045-2.

[25] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light.," *Opt. Express*, vol. 16, no. 26, pp. 21434–21445, Dec. 2008, doi: 10.1364/OE.16.021434.

[26] R. C. Ontiveros, M. Elgendi, G. Missale, and C. Menon, "Evaluating RGB channels in remote photoplethysmography: a comparative study with contact-based PPG," *Front. Physiol.*, vol. 14, Dec. 2023, doi: 10.3389/fphys.2023.1296277.

[27] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," Apr. 19, 2018, *arXiv*: arXiv:1803.01271. doi: 10.48550/arXiv.1803.01271.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141. Accessed: May 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html

[29] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep Label Distribution Learning With Label Ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017, doi: 10.1109/TIP.2017.2689998.

[30] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient Object Localization Using Convolutional Networks," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 648–656. Accessed: May 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Tompson_Efficient_Object_Localization_2015_CVPR_paper.html

[31] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognit. Lett.*, vol. 124, pp. 82–90, Jun. 2019, doi: 10.1016/j.patrec.2017.10.017.

[32] G. Heusch, A. Anjos, and S. Marcel, "A Reproducible Study on Remote Heart Rate Measurement," Sep. 04, 2017, *arXiv*: arXiv:1709.00962. doi: 10.48550/arXiv.1709.00962.

[33] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," arXiv.org. Accessed: May 21, 2026. [Online]. Available: https://arxiv.org/abs/1608.03983v5

[34] Z. Yu, X. Li, and G. Zhao, "Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks," Jul. 31, 2019, *arXiv*: arXiv:1905.02419. doi: 10.48550/arXiv.1905.02419.