Progress in
Medical Devices

Zentime

REVIEW ARTICLE

Transformer-based methods for remote photoplethysmography in non-contact physiological measurement: A systemic review

Miaomiao Peng[1], Rongguo Yan[1], Xudong Guo[1,2]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.
[2]State Key Laboratory of Cardiovascular Diseases and Medical Innovation Center, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai 200093, China.

Corresponding author: Xudong Guo.

Address correspondence to: Xudong Guo, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu District, Shanghai 200093, China.
E-mail: guoxd@usst.edu.cn.

Received January 8, 2026; Accepted February 25, 2026; Published September 18, 2026

DOI: 10.61189/773422coislv

Abstract

Traditional contact-based physiological monitoring techniques are limited by their comfort and convenience, making them unsuitable for continuous and unobtrusive monitoring in medical environments and daily healthcare applications. Remote photoplethysmography (rPPG) is a non-contact physiological sensing technique that estimates vital physiological signals, such as heart rate and respiration rate, from subtle skin color variations captured in facial videos. However, the practical application of rPPG remains challenging because the physiological signals extracted from facial regions of interest are weak and can be easily overwhelmed by non-stationary noise, such as changes in ambient illumination and subject head motion, resulting in a low signal-to-noise ratio. Recently, Transformer architectures have demonstrated strong capabilities in global dependency modeling through self-attention mechanism, enabling effective extraction of long-range temporal physiological features while suppressing global interference. These advantages provide a new paradigm to overcome the performance bottlenecks of conventional rPPG methods. This review provides a comprehensive overview of recent progress on Transformer-based rPPG research. Specifically, it covers multiple research directions, including pure transformer end-to-end model, CNN-Transformer hybrid model, multi-stream information fusion frameworks, lightweight models for edge deployment, self-supervised learning frameworks, and extended applications in physiological monitoring. Additionally, this review discusses current limitations and challenges of Transformer-based rPPG systems and highlights future research opportunities for developing new non-contact physiological monitoring technologies.

Keywords: Transformer, Remote photoplethysmography, Physiological measurement

Highlights

- This review systematically summarizes recent advances in Transformer-based methods for remote photoplethysmography (rPPG), covering pure Transformer architectures, CNN-Transformer hybrid frameworks, multi-stream fusion strategies, lightweight models, and self-supervised learning frameworks.
- This review highlights the advantages of the Transformer self-attention mechanism over CNN-based approaches in modeling long-range, quasi-periodic physiological signals, thereby enhancing measurement robustness in complex real-world scenarios.
- This review further discusses current challenges in rPPG research, including limited data availability, high computational cost, and the restricted range of measurable physiological parameters, and outlines future directions such as data-efficient learning, model compression for edge deployment, and multi-task learning for comprehensive physiological assessment.

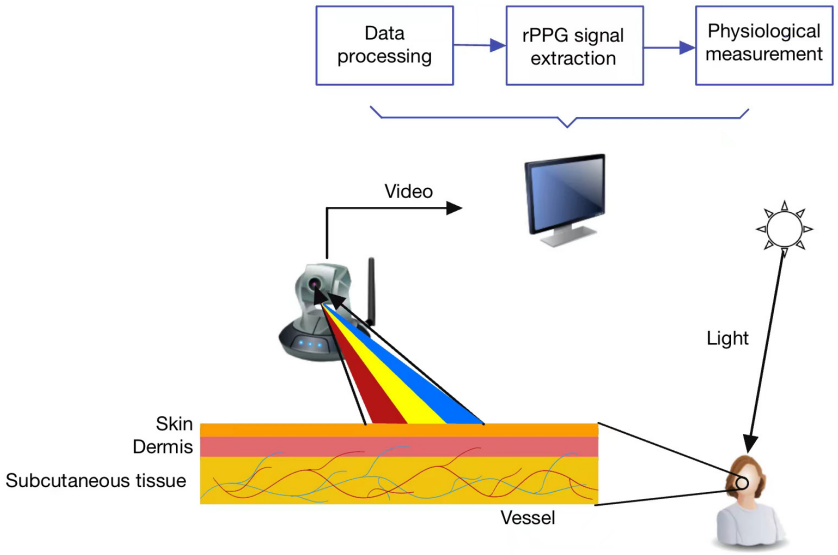

Figure 1. The schematic diagram of rPPG principle. Figure illustrates the basic principle of rPPG based on dichromatic reflection model, revealing how subtle periodic skin color changes caused by blood perfusion are separated from illumination noise to extract physiological pulse signals.

1 INTRODUCTION

With the rapid advancement of artificial intelligence (AI) and computer vision technologies, non-contact physiological monitoring using wearable devices has become a research hotspot [1, 2]. Remote photoplethysmography (rPPG) utilizes conventional Red-Green-Blue (RGB) cameras to capture facial videos and extracts blood volume pulse (BVP) signals from subtle skin color variations. rPPG has shown broad application prospects in telemedicine, affective computing, human-computer interaction, and intelligent security systems due to its non-invasive nature, ease of use, and low cost [3]. However, in real life, rPPG signals are very weak and can be overwhelmed by environmental noise, such as light variations, head motion, sensor noise, and facial expressions. Hence, the primary technical challenge in rPPG research is the reliable extraction of clean physiological signals from complex and noisy backgrounds.

In the deep learning era, convolutional neural networks (CNNs) have achieved remarkable success in rPPG tasks. CNN- based approaches can automatically learn mappings from raw video frames to heart rate, without the need for manual feature extraction. Nevertheless, conventional CNN architectures exhibit limitations in modeling long-range temporal dependencies. Specifically, the receptive field of CNNs is limited by the size and depth of convolutional kernels, which may hinder the effective modeling of rhythmic relationships between temporally distant signal components in rPPG tasks characterized by long-range periodicity. In addition, the local connectivity mechanism of CNNs limits the mechanism of global spatiotemporal information, resulting in limited robustness against complex and nonstationary spatiotemporal noises [4].

Recently, the Transformer framework, originally developed for natural language processing (NLP), has also revolutionized the field of computer vision (CV) due to its powerful self-attention mechanism. Representative models such as the Vision Transformer (ViT) divide images into patches and employ self-attention to model relationships among all patches globally. This design enables the construction of models with large receptive fields covering the entire image and provides strong global context modeling capability, making Transformers particularly suitable for tasks involving complex spatiotemporal dependencies [5]. In rPPG, physiological rhythms such as heart rate exhibit global spatiotemporal characteristics. Signals extracted from different facial regions of interest (ROIs) are intrinsically correlated, while noise sources may also demonstrate spatial correlations. The global modeling ability of the Transformers is therefore well suited for capturing these correlations, facilitating more effective separation of physiological rhythms from multi-ROI temporal signals and reduce global noise [6].

1.1 Remote imaging photoplethysmography

Photoplethysmography (PPG) is an economical and noninvasive optical technique. The fundamental principle of PPG is based on Beer-Lambert law [7, 8]. When light illuminates human skin, part of the light is absorbed, while the remaining portion is reflected or scattered by blood, bone, or tissue. Because hemoglobin exhibits strong wavelength-dependent absorption characteristics, and the blood volume in the microvascular bed changes cyclically during the cardiac cycle due to vasodilation and vasoconstriction, the intensity of the reflected or transmitted light also varies periodically in synchrony in time with the heartbeat [9, 10]. Traditional contact-based PPG sensors are widely applied in clinical settings; however, their reliance on direct skin contact limits their applicability in certain scenarios. For instance, in burn patient monitoring, neonatal monitoring, driver state monitoring, and public health screening during epidemics, contact-based measurements may cause discomfort, increase the risk of infection, and interfere long-term monitoring.

To address these issues, rPPG utilizes conventional commercial cameras or dedicated optical imaging devices together with ambient light sources to capture video recording of the face or other exposed skin area. Physiological signals are then estimated from subtle temporal color changes in diffusely reflected light from the skin surface, as illustrated in **Figure 1**. Dichromatic Reflection Model states that light reelected from skin can be decomposed into two components: specular reflec-

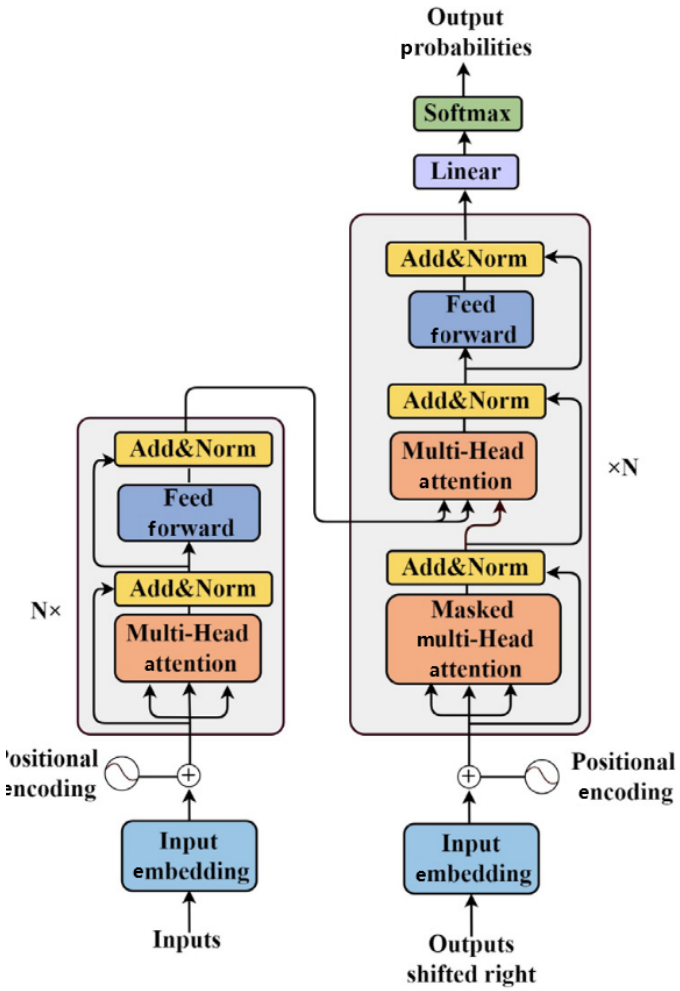

Figure 2. The transformer-model architecture. Figure displays the core structure of standard Transformer, focusing on multi-head self-attention mechanism which realizes global feature modeling.

tion and diffuse reflection. Specular reflection is mainly determined by skin surface and illumination direction and is considered a major source of noise. On the contrary, diffuse reflection originates from subsurface tissue interactions and contains critical physiological pulse information associated with blood perfusion [11]. This model can be formulated as follows:

$$I_c(t) = \alpha_c(t) \cdot C_c + \beta_c(t) \cdot S_c(t)$$

where $I_c(t)$ denotes the pixel intensity of a color channel (R, G, or B) at time t; C_c represents the baseline skin color; $\alpha_c(t)$ describes variations in illumination intensity; $S_c(t)$ denotes heartbeat-induced skin color changes; and $\beta_c(t)$ represents the modulation coefficient of cardiac pulsation on the observed illumination signal. The core objective of rPPG is to accurately recover $S_c(t)$ from the weak signal $\beta_c(t) \cdot S_c(t)$, which is contaminated by the dominant noise component $\alpha_c(t)$.

Early rPPG signal extraction methods primarily relied on traditional signal processing techniques, statistical models, and physical optics models. These approaches, through sophisticated algorithms, attempted to separate weak physiological pulsations from raw pixel signals contaminated by various noise sources. Representative methods include blind source separation (BSS)-based methods such as independent component analysis (ICA), principal component analysis (PCA), spatial component independent component analysis (SCICA), and joint blind source separation (JBSS), as well as model-based approaches, including CHROM, normalized least mean squares (NLMS), and plane-orthogonal-to-skin (POS) methods [12-18]. Although these classical approaches can achieve promising results under constrained experimental conditions, they mostly rely on manually extracted features and strong prior assumptions. Consequently, their performance deteriorates substantially in real-world scenarios involving complex illumination variations, significant head motion, and other non-stationary disturbances.

1.2 Transformer

The transformer architecture was originally proposed for the sequence-to-sequence tasks such as machine translation and is built upon an encoder-decoder framework with self-attention mechanism. Its core computation component is the multi-head self-attention (MHSA) mechanism, as shown in **Figure 2**. The attention operation can be formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where Q (Query), K (Key), and V (Value) are linear projections of the input vector X, and d_k denotes the dimensionality of the key vectors used for scaling to prevent excessively small gradients during optimization. In the multi-head mechanism, attention computations are performed in multiple subspaces simultaneously, enabling the model to capture information from different representation subspaces and thereby enhancing feature representation capability [5].

With the introduction of Transformers into the field of computer vision, representative architectures (e.g., Vision Transformer [ViT]) have demonstrated strong performance in various visual tasks. In ViT, an input image is divided into a series of overlapping or non-overlapping patches, which are then flattened into vectors and combined with positional embeddings before being fed into the Transformer encoder as a sequential input. By stacking multiple layers of self-attention and feed-forward networks, ViT can model global spatial dependencies within an image and has achieved competitive performance compared with CNNs of similar scale on several computer vision benchmarks [19, 20]. The strong global context modeling capability of Transformers is particularly advantageous for rPPG tasks, which require the extraction of long-range physiological rhythms from video sequences. First, physiological

Table 1. Analysis of rPPG methods based on Transformer

Type	Model name	Year	Method	Description
Pure Transformer	PhysFormer [24]	2022	Transformer + TDT	An end-to-end video Transformer that introduces temporal difference (TDT) modules to enhance quasi-periodic rPPG temporal features.
	RADIANT [25]	2024	Transformer + Signal embeddings	Employs specially designed signal embeddings and multi-head attention to model pulse-related multi-correlations in the subspace of color variations.
CNN-Transformer hybrid	CMRPPGFormer [26]	2024	Transformer + 3D CNN	Uses a 3D-CNN and modulation Transformer to effectively fuse local and global spatiotemporal features for accurate reconstruction of rPPG signals.
	Remote HR estimation via CNNs with Transformers [27]	2023	Transformer + LFFM	Enhances Transformers with a local feed-forward module, relative and absolute positional encodings, and temporal multi-scale modules to remedy limited local modeling capacity.
	ACTNet [28]	2024	Transformer + CNN	Adopts a dual-branch architecture with an attention-based CNN and a Transformer, where the former captures subtle color changes and the latter models long-range temporal dependencies, and their features are fused via a feature coupling unit.
	Non-contact heart rate estimation from photoplethysmography using EEMD and convolution-Transformer network [29]	2024	Transformer + EEMD + CNN	First applies EEMD to decompose rPPG into heart-rate–informative IMFs and then feeds them into parallel CNN–Transformer branches to separately learn local and global representations.
Multi-stream methods	Dual-path TokenLearner (dual-TL) [30]	2024	Transformer + dual-TL	Based on a dual-path TokenLearner, where a spatial TL aggregates correlations across multiple facial ROIs and a temporal TL learns quasi-periodic heartbeat patterns.
	MaskFusionNet [31]	2024	Transformer + MFB	Leverages mask-reconstruction pretraining and a dual-stream multi-scale fusion block to jointly model multiple facial regions and both long and short temporal segments.
	Dual-Path Periodic Transformer (DPPT) [32]	2025	Transformer + TPT + CPT	Constructs parallel temporal-periodic and channel-periodic Transformer paths to jointly exploit periodicity in the temporal and channel domains.
	PhysFormer++ [33]	2023	Transformer + SlowFast	Extends PhysFormer by introducing a SlowFast dual-pathway and periodic/cross-attention modules to better capture temporal context and periodic cues.
Lightweight models	ViT-rPPG [34]	2023	Transformer	Uses a ViT backbone to first segment facial skin regions and build spatiotemporal feature maps, then applies a Transformer to extract rPPG-related physiological features.
	EfficientPhys [35]	2023	Transformer + Normalization	Takes raw video frames as input and removes face detection and other preprocessing steps, simplifying the pipeline while preserving estimation accuracy.
	Spiking-PhysFormer [36]	2025	Transformer + SNNs	Introduces SNNs into Transformer blocks via a parallel spiking Transformer and simplified spiking self-attention to substantially reduce computational and energy cost.
	FacePhys [37]	2025	Transformer + TSD	Builds a memory-efficient rPPG model grounded in temporal–spatial state-space duality to jointly satisfy model scalability, cross-dataset generalization, and real-time operation.
Self-supervised learning frameworks	TranPhys [38]	2023	Transformer	A contrastive-learning-based self-supervised framework that, without labels, uses data augmentation and a spatiotemporal Transformer to predict facial pulse signals.
	Self-supervised RGB-NIR fusion video vision transformer framework for rPPG estimation [39]	2022	Transformer + RGB-NIR	Employs an RGB–NIR fusion ViViT architecture with contrastive self-supervised learning to robustly estimate heart rate under challenging illumination and scene conditions.
	PhySU-Net [40]	2025	Transformer + Traditional methods	Generates physiologically relevant pseudo-labels from traditional rPPG algorithms and image masking to self-supervise the pretraining of a long-context Transformer.

Extended applications	BPNet [41]	2024	Transformer	Builds a unified signal, feature, and user-information branch network that fuses rPPG signals, their derivatives, and demographic features for calibration-free non-contact blood pressure estimation.
	PulseNet [42]	2024	Transformer + STDA	Targets non-contact pulse-condition diagnosis by extracting spatiotemporal features with a Transformer, performing hierarchical multi-scale fusion and multi-task learning, and aggregating local spatiotemporal differences via an STDA module.
	Uni-rPPGNet [43]	2024	Transformer + 3D CNN	Combines shallow 3D CNNs and an hourglass-shaped Transformer, with 3D ShuffleAttention to enhance inter-channel communication for accurate HRV-oriented rPPG extraction.
	SHINE [44]	2026	Transformer	The first Transformer-inspired system for non-contact SpO_2 estimation that uses supervised contrastive learning to fuse multiple RGB RoR combinations and improve fairness and robustness across diverse skin tones.

signals such as heart rate exhibit quasi-periodic temporal characteristics over relatively long durations. Transformer's self-attention can directly model interactions between arbitrary temporal frames within a sequence, enabling effective capture of global periodic patterns across entire video clips. This capability is essential for recovering stable cardiac rhythm from noise [21]. Unlike CNNs, which typically require deep architectures to enlarge the receptive field for long-range dependency modeling, Transformers can establish global temporal relationships more directly and efficiently. Second, noise in rPPG signals, especially illumination-induced interference, often exhibits global spatial and temporal characteristics that affect large regions or even the entire video frame. In calculating attention, Transformer can integrate information across the full spatiotemporal domain, thereby improving the identification and suppression of noise patterns. In contrast, the local receptive fields of CNNs may limit their ability to handle complex non-local disturbances. Transformers are generally less sensitive to variations in noise distribution and environmental conditions, resulting in improved generalization performance. As a result, Transformer-based architectures have emerged as a powerful tool for addressing key challenges in rPPG research, particularly in global feature modeling, long-range dependency learning, and flexible multimodal representation learning [22, 23].

1.3 Data preprocessing

Data preprocessing is a critical component of rPPG pipelines, directly influencing the quality of physiological signal extraction and the performance of subsequent deep learning models. A typical preprocessing workflow begins with face detection and alignment to ensure that the ROI is consistently located across video frames. Facial regions such as the forehead and cheeks are commonly selected as ROIs due to their relatively rich blood perfusion and reduced susceptibility to motion artifacts. In addition, advanced segmentation techniques can be employed to further refine ROI section by excluding occluded regions or non-skin areas. After ROI selection, color normalization and illumination correction are often performed to mitigate the impact of lighting variations. Subsequently, spatiotemporal representations are constructed, such as stacking pixel values over time to generate spatial-temporal maps or by extracting temporal signals from each ROIs. These representations are then used as inputs for Transformer-based models, enabling effective modeling of long-range temporal dependencies and global contextual information. In recent studies, data augmentation techniques—such as random cropping, brightness jittering, and temporal perturbation—have also been incorporated to enhance model robustness and generalization. Overall, a well-designed preprocessing pipeline is essential for enhancing the signal-to-noise ratio and ensuring reliable physiological signal measurement under real-world conditions.

2 RESEARCH PROGRESS

With the widespread application of Transformers architectures in CV, increasing attention has been directed toward their strong of spatiotemporal modeling capabilities for rPPG-based physiological signal measurement. According to differences in network architecture and design philosophy, existing Transformer-based rPPG approaches can be categorized into several major groups, including pure transformer end-to-end models, CNN-Transformer hybrid architectures, multi-stream information fusion models, lightweight models for edge deployment, self-supervised learning frameworks, and extended applications in physiological measurement. This chapter provides a comprehensive overview of these categories of methods. Combined with the representative studies summarized in **Table 1**, the discussion mainly focuses on network design principles, performance characteristics, and application scenarios. As illustrated in **Figure 3**, the general workflow of Transformer-based rPPG methods typically consists of several key stages.

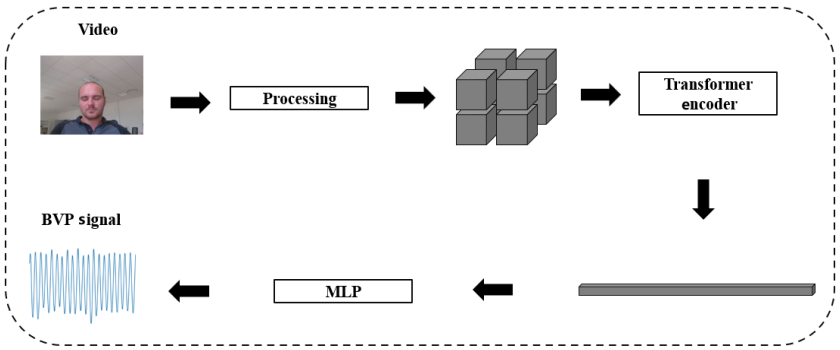

Figure 3. Transformer-based architecture. Figure depicts the overall workflow of Transformer-based rPPG algorithm, including video preprocessing, Transformer feature encoding and final physiological BVP signal regression. Collectively, the three figures systematically interpret the physical principle of rPPG, core Transformer architecture and the complete technical pipeline of Transformer-driven non-contact physiological measurement.

2.1 End-to-end models based on pure transformer

Primary studies in this field mainly focused on directly applying standard ViT architectures or their variants to rPPG tasks by constructing end-to-end deep learning frameworks for mapping video sequences to physiological signals [45]. The core of these methods is to employ Transformers as the primary backbone network and exploit their strong global dependency modeling capability to directly learn spatiotemporal features from serialized videos. The models subsequently regress either physiological parameters, such as heart rate or complete rPPG waveform. It generally starts by converting the video data into spatiotemporal maps or patch-based representations, which are then fed into Transformer encoders where multiple layers of self-attention are utilized to capture long-range spatiotemporal dependencies.

A key advantage of Transformer architectures in rPPG is their exceptional ability to model long-range temporal dependencies, which is particularly crucial for physiological signal extraction. Unlike CNNs, whose receptive fields are limited by kernel size and network depth, Transformers employ self-attention mechanisms to directly capture relationships between arbitrary frames within a video sequence, regardless of their temporal distance. This global attention mechanism enables the model to effectively distinguish subtle, periodic physiological signals (e.g., cardiac pulsations) from non-periodic or globally distributed noise sources, including sudden illumination changes, head motion, and facial expression changes. In rPPG applications, the target physiological signal is often weak and temporally distributed, while noise may exhibit both local and global spatiotemporal characteristics. By modeling pairwise interactions across the entire sequence, Transformers can better isolate the quasi-periodic physiological rhythms while suppressing non-stationary disturbances. Consequently, Transformer-based approaches not only improve the accuracy of physiological parameter estimation but also enhance the robustness and generalization of rPPG models in real-world unconstrained environments.

PhysFormer is one of the earliest studies to introduce a pure Transformer architecture into the rPPG field [24]. A key innovation of PhysFormer is the proposed Temporal difference transformer, which performs global attention by modeling feature differences between adjacent frames. This design enhances sensitivity to the quasi-periodicity of the rPPG signal while suppressing non-periodic motion-related noise. In addition, PhysFormer incorporates Label Distribution Learning and frequency-domain dynamic constraints, which function similarly to curriculum learning by providing more specific signals during training and reducing the risk of overfitting. Unlike many vision Transformer models that rely on large-scale pretraining datasets, PhysFormer can be trained directly from scratch using relatively small rPPG data while still achieving strong cross-dataset generalization performance. Consequently, it has become an important baseline framework for subsequent Transformer-based rPPG research.

RADIANT views the problem as noise reduction and feature representation enhancement [25]. The framework introduces a new signal embedding strategy in which the face is divided into multiple ROIs and temporal signals are extracted independently from each region. Then, these one-dimensional signals are mapped to a high-dimensional embedding space using multilayer perceptrons (MLPs), thereby improving feature representation capability while simultaneously reducing local noise. The resulting signal embedding are then fed into a Transformer encoder, where the global attention mechanism integrates information across different ROIs. The multi-head attention mechanism further facilitates feature learning in different representation subspaces, enabling more effective suppression of region-specific noise. In addition, RADIANT improves model generalization capability through the incorporation of synthetic data generation and data augmentation strategies.

End-to-end rPPG models based on pure Transformer architectures fully exploit the global modeling capability of self-attention mechanisms and have demonstrated substantial potential for capturing long-range periodic physiological patterns. Nevertheless, these approaches also have common problems. First, the computational complexity of standard self-attention increases quadratically with sequence length, leading to substantial computational cost for long video sequences. Second, model performance is closely related to data quality, and the risk of overfitting exists when training data are limited [8].

2.2 CNN-transformer hybrid architectures

To simultaneously exploit the advantages of both CNNs in extracting local features and Transformers in learning long-range dependencies, several CNN-Transformer hybrid structures have been proposed [28]. These approaches integrate CNNs' local perception capabilities with Transformers' global modeling power, aiming to achieve optimal performance while maintaining computational efficiency. This hybrid design philosophy has become a main research direction in rPPG. Depending on the way CNN and Transformer components are combined, hybrid models can be categorized into two types: sequential and parallel architectures.

2.2.1 Sequential hybrid architectures

Sequential hybrid architectures usually adopt a CNN-encoder and Transformer-decoder model. In this setup, the CNN module serves as a feature extraction backbone, efficiently capturing low-level spatiotemporal patterns in the video, such as edges, textures, and subtle color changes. Then, the extracted feature maps are tokenized and fed into a Transformer encoder, which uses self-attention to model long-range dependencies and aggregate global contextual information for downstream rPPG signal or heart rate prediction.

CMRPGFormer is a typical sequential hybrid architecture [26]. Its initial stage is a 3D-CNN module that extracts local spatiotemporal features from video data, capturing the dynamics of motion and appearance within small temporal windows. These local features are subsequently processed BY a ViViT module, where multi-head self-attention models complex global dependencies across both temporal and spatial dimensions, enabling effective learning of complex spatiotemporal patterns. By progressively integrating local to global information, the model outperforms standalone CNN or Transformer architectures while remaining computationally efficient. Guo et al. proposed another sequential hybrid structure with refinements in the Transformer module [27]. Specifically, a local feed-forward network is appended after the MHSA to compensate for the Transformer's shortcoming in capturing neighboring feature interactions. Moreover, the use of both relative and absolute positional encodings allows the model to better preserve temporal sequence information, which is very important for accurately modeling the quasi-periodic nature of rPPG signals. This design has demonstrated robustness across multiple public datasets.

2.2.2 Parallel hybrid architectures

Parallel hybrid architectures are commonly designed as two or more network branches, with each branch designed to handle different types or scales of information. Features from these branches are subsequently integrated through a dedicated fusion module to generate the final representation.

ACTNet is a dual-branch hybrid network for respiratory rate (RR) estimation [28]. One branch is an attention-based CNN, which captures subtle color variations between frames and highlights local spatiotemporal features. The other branch is a Transformer network, which models long-range temporal dependencies across the entire video sequence, thereby encoding global information. Features from both branches are fused using a feature coupling unit, enabling the network to leverage both local detail and global context simultaneously.

Liu et al. proposed a hybrid approach combining traditional signal processing with deep learning [29]. This method uses ensemble empirical mode decomposition (EEMD) to decompose the raw rPPG signal into a series of intrinsic mode functions (IMFs). IMFs containing effective heart rate information are selected and subsequently processed by a parallel CNN-Transformer network. In this step, the CNN branch extracts local features from those selected IMFs, while the Transformer branch captures global patterns across the sequence. This combination leverages the signal denoising and separation advantages of EEMD while incorporating deep networks for feature learning, resulting in an efficient and lightweight framework for detecting heart rates.

2.3 Multi-stream information fusion methods

In complex real-world scenarios, rPPG signals are often contaminated by various interferences, such as illumination variations, head motion, occlusions, and video compression artifacts. Single-stream models often struggle to distinguish spatiotemporal regions containing reliable physiological information from those dominated by noise [6, 46]. Consequently, recent works have increasingly explored Transformer-based architecture from the perspective of multi-stream and dual-stream information fusion [30-33]. By constructing parallel branches across different dimensions—such as spatial, temporal, channel, or scale domains—and subsequently performing interaction and fusion among these branches, these approaches aim to enhance the robustness and stability of rPPG measurements.

Dual-path TokenLearner represents a typical spatial-temporal dual-path modeling framework [30]. The core idea is to employ learnable tokens for adaptive aggregation of informative contextual features through separate spatial and temporal paths. Specifically, the Spatial TokenLearner (S-TL) identifies combinations of facial ROIs that contain stronger pulse signal while automatically suppressing regions affected by facial expression, occlusions, or localized illumination artifacts. In parallel, the Temporal TokenLearner (T-TL) mainly emphasizes the quasi-periodic structure of heartbeat by selecting and preserving the most informative temporal segments associated with physiological periodicity, thereby reducing non-periodic interference caused by head movement. Finally, spatial and temporal tokens are jointly modeled in the prediction head to achieve

synergistic fusion of global spatiotemporal information. Compared with methods that rely on fixed ROIs or manually selected temporal windows, Dual-Path TokenLearner demonstrates improved cross-dataset generalization and robustness across diverse scenarios owing to its adaptive dual-stream information aggregation strategy. MaskFusionNet performed multi-stream fusion based on a multi-region, multi-scale approach [31]. During the pre-training phase, a mask reconstruction task based on tube masking is introduced, forcing the model to recover rPPG time series from randomly masked facial regions. This strategy encourages the model to learn more distributed and robust physiological representations. In fine tune, a Multi-Scale Fusion Block (MFB) is used to integrate features extracted from dual-stream networks across different temporal and spatial scales. This design allows the model to capture subtle but important short-term physiological variations while simultaneously leveraging long-term contextual information to suppress occasional noise. Consequently, the framework demonstrates enhanced robustness against occlusions and local motion artifacts.

Dual-Path Periodic Transformer (DPPT) extends multi-stream information fusion from temporal domain to channel domain [32]. DPPT employs two parallel Transformer branches: the Temporal Periodic Transformer (TPT) that models rhythmic temporal dependencies, and the Channel Periodic Transformer (CPT) that captures periodic variations across different color channels such as RGB. By explicitly decoupling and jointly modeling temporal and channel domain periodicity, followed by feature fusion at later stages, DPPT can better adapt to channel response variations caused by differences in skin tone and illumination conditions, greatly improving heart rate estimation accuracy and enhancing cross-database robustness on public datasets such as PURE and UBFC-rPPG. PhysFormer++ further extends the original PhysFormer architecture by adopting a dual-path SlowFast-style video Transformer framework [33]. The model consists of a slow pathway and a fast pathway that respectively capture low-frequency long-term trends and high-frequency short-term change. Temporal difference attention and cross-attention mechanisms are employed to facilitate information exchange between the two paths. The slow pathway focuses on stable periodic physiological rhythms, whereas the fast pathway is designed to better model rapid motion and respiratory-related interference. Fusing information from both paths generates more robust rPPG representations for complex dynamic scenarios.

2.4 Lightweight models for edge deployment

Regarding practical applications on resource-constrained platforms like mobile devices, wearable devices and in-vehicle terminals, Transformer-based rPPG methods must address the challenge of reducing model parameters, computational complexity, memory consumption, and power usage, while maintaining satisfactory physiological measurement accuracy [35]. Recent studies have increasingly focused on the development of lightweight and computationally efficient architectures [34, 36, 37]. These approaches seek to balance performance and resource overhead through simplified backbone networks, optimized spatiotemporal modeling methods, and energy-efficient neural computing approaches.

ViT-rPPG represents one of the earliest attempts to directly apply the standard ViT architecture to rPPG task [34]. The framework first segments facial skin regions and constructs compact spatiotemporal feature maps through temporal stacking. These maps are then divided into patches and fed into a ViT backbone for end-to-end regression from pseudo-image representations to rPPG signal. By adapting the patch size and network scale to the rPPG features, ViT-rPPG showed the feasibility of directly learning spatiotemporal representations using Transformers on mixed datasets. At the same time, the framework greatly reduced computational cost compared with the original ViT, providing an important structural reference for subsequent light-weight Vision Transformers in rPPG applications.

EfficientPhys approaches lightweight optimization from the perspective of preprocessing simplification and unified end-to-end learning [35]. Conventional deep learning-based rPPG pipelines usually need a bunch of preprocessing steps, including face detection, face alignment, ROI extraction, normalization, and color space transformation, all of which introduce additional computational overhead. EfficientPhys, in contrast, directly processes raw facial video frames and employs an efficient convolutional network or lightweight Transformer backbone to automatically learn rich physiological representations, thus eliminating the need for aforementioned preprocessing steps. Experimental demonstrated that compared to the original model, the framework greatly improves inference speed and hardware efficiency on mobile and edge devices even at a cost of moderate reduction in overall accuracy (<33%).

Spiking-PhysFormer further advances lightweight rPPG modeling by integrating event-driven spiking neural networks (SNNs) with Transformer architectures [36]. Because SNNs perform computations only when neuron spikes occur, they exhibit extremely low power consumption on neuromorphic hardware. Spiking-PhysFormer introduces parallel spike-driven Transformer blocks and a lightweight spiking self-attention mechanism, achieving a 10.1% reduction in power consumption while maintaining performance comparable to that of PhysFormer. The power consumed by the transformer block itself has been decreased by a factor of 12.2, demonstrating great potential for mobile and edge computing applications.

FacePhys addresses the trade-off among model scale, cross-database generalization, and real-time performance from a system-level perspective [37]. The framework proposes a memory-friendly rPPG approach according to dual temporal-spatial

state-space modeling. Specifically, FacePhys introduces a transferable heart-state representation and recursively updates physiological states along the temporal dimension in a manner similar to state-space models, instead of performing global self-attention across all video frames. This can greatly reduce memory usage and inference latency. FacePhys reduces both cross-dataset estimation error while maintaining a memory footprint of approximately 3.6 MB and a single-frame inference latency of about 10 ms, thereby enabling robust real-time heart rate estimation in resource-constrained environments involving compressed transmission and edge-side computation.

While lightweight Transformer-based rPPG models offer significant advantages in computational efficiency, memory footprint, and deployment flexibility on mobile or edge devices, these benefits often come at the cost of reduced accuracy. Compared to full-scale Transformer models, lightweight variants typically exhibit moderately higher heart rate estimation errors and lower signal-to-noise ratio (SNR) performance on benchmark datasets. Despite these trade-offs, many lightweight models still achieve acceptable performance for practical applications, especially when combined with robust preprocessing and data augmentation strategies. Therefore, balancing between efficiency and estimation accuracy remains a key consideration in the design of rPPG models for real-world deployment.

2.5 Self-supervised learning frameworks

Traditional supervised learning paradigm for rPPG tasks is highly dependent on large-scale, high-quality, and diverse labeled datasets. However, the synchronized acquisition of facial video recordings and high-precision physiological signals in real-world scenarios is both expensive and subject to substantial privacy and ethical constraints. Consequently, most existing rPPG datasets remain limited in scale and diversity, which restricts the effective training and cross-scenario generalization of high-capacity Transformer models. Nevertheless, self-supervised learning (SSL), which leverages pre-training tasks that do not require manual annotations, has demonstrated significant success in speech and computer vision applications. Accordingly, SSL has been gradually introduced into Transformer-based rPPG research to alleviate the issue of data scarcity [8].

TranPhys is a representative framework that combines contrastive learning with a pure Transformer architecture [38]. Its core idea is to construct positive pairs from the same source and negative pairs from different sources through a series of spatiotemporal data augmentation strategies, including brightness jittering, random cropping, and temporal perturbations. The network structure consists of a stem module for extracting coarse-grained local features and a spatiotemporal Transformer responsible for aggregating global contextual information and explicitly modeling long-term physiological rhythms. TranPhys performs pre-training entirely without true heart rate labels and achieves performance superior to several supervised baseline methods after fine-tuning with only a small amount of labeled data. These results indicate that contrastive self-supervised representations are of great significance for improving the generalization capability of rPPG models in real-world scenarios.

Park et al. further extended self-supervised learning into a multimodal setting by proposing Fusion ViViT, the first Transformer-based RGB-NIR fusion self-supervised framework for rPPG [39]. The method leverages the complementary characteristics of RGB and near-infrared (NIR) modalities to construct an end-to-end Fusion ViViT network capable of jointly extracting long-range spatiotemporal representations from multimodal video sequences. By introducing contrastive learning objectives across both modalities and augmented views, the framework encourages the learning of modality-invariant physiological representations that remain robust under severe illumination variations, head motions, and environmental disturbances. Simultaneously, the Transformer's self-attention mechanism automatically weights the contributions of RGB and NIR information during feature fusion, enabling dynamic multimodal integration. Experimental results showed that the framework achieved short-term heart rate estimation performance comparable to mainstream methods using 30-second averaging windows, while maintaining robust transferability in challenging scenarios such as driving environments.

PhySU-Net represents another important direction that combines pseudo-labeling and reconstruction-based self-supervised strategies in long-context Transformer rPPG networks [40]. On the one hand, traditional rPPG algorithms are employed to generate noisy pseudo-labels from unlabeled videos, which serve as weak supervisory signals. On the other hand, the framework introduces a masked feature reconstruction task that forces the model to focus on dynamic spatiotemporal pattern related to physiological rhythm during the recovery of masked regions. By performing these two types of self-supervised pre-training strategies on large-scale unlabeled datasets, PhySU-Net learns feature representations with enhanced long-term contextual awareness and noise robustness. After supervised fine-tuning on public datasets such as OBF, VIPL-HR, and MMSE-HR, the framework outperformed many existing methods, with the benefits of self-supervised pre-training being particularly significant.

2.6 Extended applications in physiological measurement

Early rPPG studies primarily focused on the reliable estimation of average heart rate (HR). However, with the continuous advancement of model representation capability and the increasing demand for multi-task physiological analysis, recent research has gradually expanded toward more comprehensive physiological measurement applications, including blood pres-

sure (BP), heart rate variability (HRV), blood oxygen saturation (SpO_2), and even pulse diagnosis in Traditional Chinese Medicine (TCM). Such kind of tasks require more detailed characterization of rPPG waveforms across temporal, spectral, and channel domains, thereby placing higher demands on long-range dependency modeling and multi-source information fusing capabilities provided by Transformer architectures.

BPNet is a typical Transformer-based architecture for non-contact blood pressure estimation (NC-BPE) [41]. The network consists of three parts: a Signal Branch, a Feature Branch, and a Predictor branch. The Signal Branch processes the rPPG waveform together with with its temporal derivatives, thereby preserving subtle changes in the original physiological signal. The Feature Branch extracts high-level statistical and morphological features from various representation spaces, including temporal and frequency-domain representations. On this basis, the Predictor Branch integrates user-specific information to joint jointly estimate systolic and diastolic blood pressure values. Through multi-source information fusion at the architecture level, BPNet achieves calibration-free BP estimation across different individuals and a wide range of blood pressure conditions. In addition, the framework demonstrates favorable computational efficiency and model complexity characteristics, indicating substantial potential for clinical applications.

PulseNet extends Transformer to the higher-level semantic task of pulse diagnosis in TCM [42]. To the multi-scale spatiotemporal feature fusion strategy and employs specialized attention module to discriminate fine-grained pulse characteristics from facial video sequences. This approach provides a novel technical pathway for non-contact intelligent TCM diagnosis and further demonstrates the transferability of rPPG features to more complex health evaluation tasks.

Uni-rPPGNet further expands Transformer-based physiological measurement toward HRV estimation, which imposes more stringent requirements on temporal modeling [43]. The network uses a hybrid 3D-CNN and Transformer structure. Shallow 3D-CNN layers are used to extract local spatiotemporal features, while a channel interaction module enhances information exchange among different color channels. In deeper layers, an hourglass-shaped Transformer architecture captures global spatiotemporal dependencies across different scales, enabling more precise reconstruction of rPPG waveform in both temporal and frequency domains. Experimental results demonstrated that Uni-rPPGNet achieves a balance between estimation accuracy and inference efficiency for various time- and frequency-domain HRV indices on the UBFC-rPPG and PURE datasets, providing a feasible approach for remote autonomic nervous system (ANS) evaluation using rPPG.

SHINE is a Transformer-based framework specifically designed for SpO_2 estimation from rPPG signals [44]. Its core idea is to integrate Transformer's global modeling capability with rich representations based on multi-channel ratios-of-ratios (RoRs) to capture subtle optical changes associated with blood oxygen changes. SHINE constructs multiple RGB-channel RoR combinations within the feature space, and each combination is modeled through dedicated parallel branches to learn corresponding temporal patterns. Furthermore, a quality-weighted fusion strategy is employed to suppress noisy predictions while assigning greater importance to higher-quality channel combinations. Furthermore, a quality-weighted fusion strategy is employed to suppress noisy predictions while assigning greater importance to higher-quality channel combinations. The framework also incorporates supervised contrastive learning to improve the learning of discriminative SpO_2-sensitive representations under various individual characteristics and skin tones. Further supported by a large-scale multi-skin-tone rPPG-SpO_2 dataset, SHINE outperformed existing systems on both public and self-collected datasets, particularly improving estimation accuracy for darker skin tones, which represents an important issue related to fairness and model generalizability.

3 DATASETS AND EVALUATION METRICS

Public benchmark datasets and unified evaluation standards provide the foundation for performance comparison, generalization evaluation, and practical deployment of rPPG algorithms. Differ from conventional computer vision tasks, rPPG datasets differ not only in video quality and scene diversity, but also in the accuracy of physiological signal acquisition, sensor types, and synchronization strategies for annotation. These factors directly influence the comparability of model training and evaluation results. In this section, a systematic review of existing mainstream rPPG datasets and evaluation metrics is presented.

3.1 Datasets

Existing rPPG datasets can be generally divided into two categories: datasets collected under controlled laboratory conditions and datasets acquired in near-naturalistic settings. The former is typically recorded under stable illumination conditions with limited subject motion, making them suitable for evaluating the upper-bound performance of algorithms. In contrast, the latter better reflects real-life applications in terms of illumination variation, head movement, posture diversity, and recording equipment, thereby providing a more suitable benchmark for evaluating model robustness and cross-domain generalization capability.

Moreover, different rPPG datasets vary considerably in participant numbers, skin-tone distribution, and acquisition devices. These variations can significantly affect model training, evaluation consistency, and cross-dataset transferability. **Table 2** summarizes several representative datasets commonly used in rPPG field.

Table 2. Summary of public datasets

Dataset	Subjects	Videos	Label	Description
UBFC-rPPG [47]	42	640×480; 30 fps	PPG, HR	Recorded under stress-inducing conditions with pronounced head motion and illumination changes, and widely used as a benchmark for evaluating algorithm robustness.
PURE [48]	160	640×480; 30 fps	PPG, SpO_2	Consists of ten independent sequences of approximately ten minutes each, with synchronized acquisition of multiple physiological signals, providing a relatively large data volume and diverse scenarios.
COHFACE [49]	40	640×480; 20 fps	PPG	Recorded in social-interaction scenarios, containing rich facial expressions and natural interactions, and is particularly sensitive to motion artifacts.
VIPL-HR [50]	100	960×720, 1920×1080, 640×480; 60 fps, 30 fps	PPG, HR, SpO_2	Recorded with heterogeneous devices in highly challenging environments, characterized by unstable illumination, large head movements, and variable frame rates.
MAHNOB-HCI [51]	34	1040×1392; 24 fps	ECG	Contains videos from multiple emotion-elicitation experiments with synchronously recorded ECG and other physiological signals.
OBF [52]	100	920×1080; 60 fps	PPG, ECG, RR	Recorded in a strictly controlled laboratory environment with minimal illumination and motion disturbances, yielding high-quality data with low noise.
MMPD [53]	33	1280×720; 30 fps	PPG, HR	Comprises facial videos collected under varying illumination and motion conditions.

3.2 Evaluation metrics

To quantitatively evaluate the performance of rPPG algorithms in estimating heart rate and other physiological parameters, a variety of error-based and correlation-based metrics are commonly used for comprehensive assessment. The commonly used evaluation metrics include the following.

(1) Mean Absolute Error, MAE

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

(2) Root Mean Square Error, RMSE

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

(3) Mean Absolute Percentage Error, MAPE

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

(4) Pearson's Correlation Coefficient, ρ

$$\rho = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

(5) Standard Deviation of Absolute Error, SDAE

$$e_i = |\hat{y}_i - y_i|, \quad \bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i$$

$$\mathrm{SDAE} = \sqrt{\frac{1}{N\text{-}1}\sum_{i=1}^{N}(e_i - \bar{e})^2}$$

(6) Signal-to-Noise Ratio, SNR

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right)$$

(7) Concordance Correlation Coefficient, CCC

$$\mathrm{CCC} = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2}$$

4 CONCLUSION

This paper provides a systematic review of the recent research developments in Transformer-based approaches for rPPG, covering several aspects: pure Transformer end-to-end models, CNN-Transformer hybrid models, multi-stream information fusion, lightweight models for edge deployment, self-supervised and multi-modal learning, and extended multi-task applications. We compared the network design philosophies, strengths, and limitations of representative methods. The powerful global dependency modeling capability and self-attention mechanism of the Transformer enable effective capture of long-term quasi-periodic physiological rhythms, suppression of global illumination and motion noise, and superior multi-source heterogeneous information fusion compared with CNN-based methods. These developments have enhanced the accuracy and robustness of non-contact physiological parameter estimation in complex, naturalistic scenarios. Additionally, we summarized commonly used rPPG datasets and evaluation metrics, highlighting that evaluation protocols and data sets

characteristics significantly influence experimental conclusions and providing reference baselines for future research.

Despite these achievements, Transformer-based rPPG still faces some challenges. First, existing datasets are mostly collected in controlled environments or limited population groups, resulting in significant distribution bias. It not only restricts the comprehensive exploitation of large-scale Transformer representational power but also limits systematic fairness and reliability assessment. Second, although some studies have explored model lightweighting and spiking neural networks, Transformer models generally involve large parameter counts and high computational demands. Deploying high-precision rPPG models on resource-constrained mobile devices, wearable terminals, or edge platforms while achieving stable real-time (or close-to-real-time) remains challenging. Third, most studies focus primarily on average heart rate estimation, but simultaneous, high-accuracy measurement of other physiological parameters, such as HRV, breathing frequency, SpO_2, and BP, is still at a nascent stage. Due to the complex interdependencies among these signals, we do not yet have a unified, physically interpretable paradigm for joint multi-task modeling.

Future directions may involve self-supervised, pseudo-labeling, and semi-supervised learning to exploit large-scale unlabeled video resources, thereby enhancing model generalization and data efficiency while minimizing annotation costs. Architectural hybrid designs such as sparse attention, low-rank decomposition, and hybrid state-space model combined with convolutional attention, could further improve computational efficiency and scalability. Hardware-oriented solutions, including Spiking Neural Networks and Neuromorphic Chips, can be integrated to reduce latency and energy consumption while maintaining high accuracy for long-term and continuous monitoring. Finally, the establishment of unified evaluation protocols, reporting standards, and open benchmarks that aligned with clinical standards is also essential to facilitate the translation of Transformer-based rPPG technology from the laboratory to the clinic and industrial application.

DECLARATIONS

Author contributions

Miaomiao Peng performed the data analyses and wrote the manuscript; Miaomiao Peng, Rongguo Yan and Xudong Guo contributed significantly to analysis and manuscript preparation.

Funding

This research received no external funding.

Data availability

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that there are no competing interests associated with this work.

Acknowledgements

Not applicable.

REFERENCES

[1] Pirzada P, Wilde A, Harris-Birtill D. Remote photoplethysmography for heart rate and blood oxygenation measurement: A review. IEEE Sens J. 2024;24(15):23436-23453. https://doi.org/10.1109/JSEN.2024.3405414

[2] Hassan MA, Malik AS, Fofi D, Saad N, Karasfi B, Ali YS, et al. Heart rate estimation using facial video: A review. Biomed Signal Process Control. 2017 Sep 1;38:346-360. https://doi.org/10.1016/j.bspc.2017.07.004

[3] Zhao Y, Wang X, Che T, Bao G, Li S. Multi-task deep learning for medical image computing and analysis: A review. Comput Biol Med. 2023 Feb;153:106496. https://doi.org/10.1016/j.compbiomed.2022.106496

[4] Špetlík RFV, Čech J, Matas J. Visual heart rate estimation with convolutional neural network. In: British Machine Vision Conference 2018; 2018 Sep 3-6; Newcastle upon Tyne, UK. Durham (UK): BMVA Press; 2018. p. 271.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. 31st Annual Conference on Neural Information Processing Systems (NIPS 2017); Long Beach, CA. Red Hook (NY): Curran Associates; 2017. p. 5998-6008.

[6] Zou B, Guo Z, Chen J, Zhuo J, Huang W, Ma H. Rhythmformer: Extracting patterned rppg signals based on periodic sparse attention. Pattern Recognit. 2025 Aug;164:111511. https://doi.org/10.1016/j.patcog.2025.111511

[7] Chen W, Yi Z, Lim LJR, Lim RQR, Zhang A, Qian Z, et al. Deep learning and remote photoplethysmography powered advancements in contactless physiological measurement. Front Bioeng Biotechnol. 2024 Jul 17;12:1420100. https://doi.org/10.3389/fbioe.2024.1420100

[8] Xiao H, Liu T, Sun Y, Li Y, Zhao S, Avolio A. Remote photoplethysmography for heart rate measurement: A review. Biomed Signal Process Control. 2024 Feb;88 Part B:105608. https://doi.org/10.1016/j.bspc.2023.105608

[9] Jaiswal KB, Meenpal T. rPPG-FuseNet: Non-contact heart rate estimation from facial video via RGB/MSR signal fusion. Biomed Signal Process Control. 2022 Sep;78:104002. https://doi.org/10.1016/j.bspc.2022.104002

[10] Usman M, Sobotka M, Ruminski J. Stream-net: Spatio-temporal feature fusion network for robust rPPG signal measurement in remote health monitoring. Knowl-Based Syst. 2025 Sep 27;326:114080. https://doi.org/10.1016/j.knosys.2025.114080

[11] Chen W, McDuff D. Deepphys: Video-based physiological measurement using convolutional attention networks. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. ECCV 2018; Springer, Cham. Springer International Publishing; 2018. p. 356-373. https://doi.org/10.1007/978-3-030-01216-8_22

[12] Poh MZ, McDuff DJ, Picard RW. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt Express. 2010 May 10;18(10):10762-10774. https://doi.org/10.1364/OE.18.010762

[13] Lewandowska M, Rumiński J, Kocejko T, Nowak J. Measuring pulse rate with a webcam — a non-contact method for evaluating cardiac activity. In: Proc 2011 Federated Conf Comput Sci Inf Syst; Sep 18-21; Szczecin, Poland. Institute of Electrical and Electronics Engineers; 2011. p. 405-410.

[14] Fan Q, Li K. Noncontact imaging plethysmography for accurate estimation of physiological parameters. J Med Biol Eng. 2017 Jun 20;37(5):675-685. https://doi.org/10.1007/s40846-017-0272-y

[15] Sun Y, Sijung H, Vicente AP, Jonathon AC, Yisheng Z, Stephen EG. Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise. J Biomed Opt. 2011 Jul 1;16(7):077010. https://doi.org/10.1117/1.3602852

[16] de Haan G, Jeanne V. Robust pulse rate from chrominance-based rPPG. IEEE Trans Biomed Eng. 2013;60(10):2878-2886. https://doi.org/10.1109/TBME.2013.2266196

[17] Li X, Chen J, Zhao G, Pietikäinen M. Remote heart rate measurement from face videos under realistic situations. In: Proc IEEE CVPR. 2014. p. 4264-4271. https://doi.org/10.1109/CVPR.2014.543

[18] Wang W, den Brinker AC, Stuijk S, de Haan G. Algorithmic principles of remote PPG. IEEE Trans Biomed Eng. 2017;64(7):1479-1491. https://doi.org/10.1109/TBME.2016.2609282

[19] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations 2021; 2021 May 4-8; Virtual Conference. Montreal (QC): ICLR; c2021. p. 1-18.

[20] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. ACM Comput Surv. 2022;54(10s):Article 200:1-41. https://doi.org/10.1145/3505244

[21] Xiao H, Li L, Liu Q, Zhu X, Zhang Q. Transformers in medical image segmentation: A review. Biomed Signal Process Control. 2023 Jul;84:104791. https://doi.org/10.1016/j.bspc.2023.104791

[22] Liu X, Narayanswamy G, Paruchuri A, Zhang X, Tang J, Zhang Y, et al. rPPG-toolbox: deep remote PPG toolbox. In: Globerson A, Hardt M, Levine S, Saenko K, editors. 37th Annual Conference on Neural Information Processing Systems (NeurIPS 2023); 2023 Dec 10-16; New Orleans, LA. Red Hook (NY): Curran Associates; c2023. p. 68485-68510.

[23] Nerella S, Bandyopadhyay S, Zhang J, Contreras M, Siegel S, Bumin A, et al. Transformers and large language models in healthcare: A review. Artif Intell Med. 2024 Aug;154:102900. https://doi.org/10.1016/j.artmed.2024.102900

[24] Yu Z, Shen Y, Shi J, Zhao H, Torr P, Zhao G. PhysFormer: Facial video-based physiological measurement with temporal difference transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), LA. 2022. p. 4176-4186. https://doi.org/10.1109/CVPR52688.2022.00415

[25] Gupta AK, Kumar R, Birla L, Gupta P. RADIANT: Better rPPG estimation using signal embeddings and transformer. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), HI, USA. 2023. p. 4965-4975. https://doi.org/10.1109/WACV56688.2023.00495

[26] Ma X, Wang Z, Liu X, Kuang H. CMRPPGFormer: 3-D spatio-temporal convolutional modulation transformer network for remote heart rate estimation. IEEE Sens J. 2024;24(19):30275-30286. https://doi.org/10.1109/JSEN.2024.3407816

[27] Guo Z, Chen H, Lin L, Zhou W, Yang M, Ying N, et al. Remote heart rate estimation via convolutional neural networks with transformers. J Franklin Inst. 2023;360(17):13149-13165. https://doi.org/10.1016/j.jfranklin.2023.10.013

[28] Chen H, Zhang X, Guo Z, Ying N, Yang M, Guo C. ACTNet: Attention based cnn and transformer network for respiratory rate estimation. Biomed Signal Process Control. 2024;96:106497. https://doi.org/10.1016/j.bspc.2024.106497

[29] Liu K, Wu S, Li T, Gou S, Wang X, Guo Z. Non-contact heart rate estimation from photoplethysmography using eemd and convolution-transformer network. In: 2024 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA); 2024 Jun 14-16; Xi'an, China. IEEE; 2024. p. 1-6. https://doi.org/10.1109/CIVEMSA58715.2024.10586459

[30] Qian W, Guo D, Li K, Zhang X, Tian X, Yang X, et al. Dual-path tokenlearner for remote photoplethysmography-based physiological measurement with facial videos. IEEE Trans Comput Soc Syst. 2024;11(3):4465-4477. https://doi.org/10.1109/TCSS.2024.3356713

[31] Zhang Y, Shi J, Wang J, Zong Y, Zheng W, Zhao G. MaskFusionNet: A dual-stream fusion model with masked pre-training mechanism for rppg measurement. EEE Trans Circuits Syst Video Technol. 2024;34(11):11521-11534. https://doi.org/10.1109/TCSVT.2024.3422849

[32] Son J, Choi YS. Dual-path periodic transformer for enhancing rPPG estimation. In: 2025 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC); 2025 Jul 7-10; Seoul. IEEE; 2025. p. 1-5. https://doi.org/10.1109/ITC-CSCC66376.2025.11137683

[33] Yu Z, Shen Y, Shi J, Zhao H, Cui Y, Zhang J, et al. PhysFormer++: Facial video-based physiological measurement with slowfast temporal difference transformer. Int J Comput Vis. 2023;131(6):1307-1330. https://doi.org/10.1007/s11263-023-01758-1

[34] Sun W, Sun Q, Sun H, Sun Q, Jia R. ViT-rPPG: A vision transformer-based network for remote heart rate estimation. J Electron

Imaging. 2023;32(2):023024. https://doi.org/10.1117/1.JEI.32.2.023024

[35] Liu X, Hill B, Jiang Z, Patel S, McDuff D. EfficientPhys: Enabling simple, fast and accurate camera-based cardiac measurement. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2-7; HI, USA. IEEE; 2023. p. 4997-5006. https://doi.org/10.1109/WACV56688.2023.00498

[36] Liu M, Tang J, Chen Y, Li H, Qi J, Li S, et al. Spiking-PhysFormer: Camera-based remote photoplethysmography with parallel spike-driven transformer. Neural Netw. 2025 May;185:107128. https://doi.org/10.1016/j.neunet.2025.107128

[37] Wang K, Tang J, Wang Y, Liu X, Fan Y, Ji J, et al. FacePhys: State of the Heart Learning. arXiv [Preprint]. 2025 [cited 2026 Jul 1]. Available from: https://arxiv.org/abs/2512.06275

[38] Wang R, Sun H, Hao R, Pan A, Jia R. TransPhys: Transformer-based unsupervised contrastive learning for remote heart rate measurement. Biomed Signal Process Control. 2023;86:105058. https://doi.org/10.1016/j.bspc.2023.105058

[39] Park S, Kim BK, Dong SY. Self-supervised RGB-NIR fusion video vision transformer framework for rPPG estimation. IEEE Trans Instrum Meas. 2022;71:1-10. https://doi.org/10.1109/TIM.2022.3217867

[40] Savic M, Zhao G. PhySU-Net: Long temporal context transformer for rppg with self-supervised pre-training. In: Antonacopoulos A, Chaudhuri S, Chellappa R, Liu CL, Bhattacharya S, Pal U, editors. Pattern Recognition. ICPR 2024. Cham: Springer Nature; 2025. p. 228-243. https://doi.org/10.1007/978-3-031-78341-8_15

[41] Manullang MCT, Lin YH, Chou NK. A transformer-based network for estimating blood pressure using facial videos. IEEE Sens J. 2025;25(1):1969-1977. https://doi.org/10.1109/JSEN.2024.3496115

[42] Zhao Z, Zhou Y, Zhang S, Chen S, Wen C, Xu Q, et al. PulseNet: Multi-task learning-based non-contact pulse condition diagnosis using multi-scale fusion and transformer. Knowl Based Syst. 2024;302:112333. https://doi.org/10.1016/j.knosys.2024.112333

[43] Liu X, Zuo J, Ma X, Kuang H. Uni-rPPGNet: Efficient and lightweight remote heart rate variability measurement. In: 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC); 2024 Sep 20-22; Chongqing, China. IEEE; 2024. p. 19-23. https://doi.org/10.1109/ITNEC60942.2024.107332500

[44] Agarwal V, Saikia T, Kumar Gupta A, Gupta P. SHINE: Synergizing transformers with contrastive learning for thriving rPPG-based SpO2 estimation. Expert Syst Appl. 2026;296:129190. https://doi.org/10.1016/j.eswa.2025.129190

[45] Chowdhury MH, Reaz MBI, Ali SHM, Khan MS, Chowdhury MEH. ROSE-Net: Leveraging remote photoplethysmography to estimate oxygen saturation using deep learning. Biomed Signal Process Control. 2025;100:107105. https://doi.org/10.1016/j.bspc.2024.107105

[46] Yan Z, Zhong Y, Xu H, Zhang W, Yi S, Shu L, et al. PhysMamba: State space duality model for remote physiological measurement. arXiv [Preprint]. 2024 [cited 2026 Jul 1]. Available from: https://arxiv.org/abs/2408.01077

[47] Bobbia S, Macwan R, Benezeth Y, Mansouri A, Dubois J. Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognit Lett. 2019;124:82-90. https://doi.org/10.1016/j.patrec.2017.10.017

[48] Stricker R, Müller S, Gross HM. Non-contact video-based pulse rate measurement on a mobile service robot. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication; 2014 Aug 25-29; Edinburgh, UK. IEEE; 2014. p. 1056-1062. https://doi.org/10.1109/ROMAN.2014.6926392

[49] Heusch G, Anjos A, Marcel S. A reproducible study on remote heart rate measurement. arXiv [Preprint]. 2017 [cited 2026 Jul 1]. Available from: https://arxiv.org/abs/1709.00962

[50] Niu X, Han H, Shan S, Chen X. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In: 14th Asian Conference on Computer Vision; 2018 Dec 2-6; Perth, Australia. Cham (Switzerland): Springer; c2019. p. 562-576. https://doi.org/https://doi.org/10.1007/978-3-030-20873-8_36

[51] Wang H, Ahn E, Kim J. Self-supervised representation learning framework for remote physiological measurement using spatio-temporal augmentation loss. In: 36th AAAI Conference on Artificial Intelligence; 2022 Feb 22–Mar 1; Virtual Conference. Palo Alto (CA): AAAI Press; c2022. p. 2431-2439. https://doi.org/10.1609/aaai.v36i2.20143

[52] Li X, Alikhani I, Shi J, Seppanen T, Junttila J, Majamaa-Voltti K. The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018); 2018 May 15-19; Xi'an, China. IEEE; 2018. p. 242-249. https://doi.org/10.1109/FG.2018.00043

[53] Tang J, Chen K, Wang Y, Shi Y, Patel S, McDuff D. MMPD: Multi-domain mobile video physiology dataset. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2023 Jul 24-27; Sydney, Australia. IEEE; 2023. p. 1-5. https://doi.org/10.1109/EMBC40787.2023.10340857