Multimedia Tools and Applications (2024) 83:44699–44728
https://doi.org/10.1007/s11042-023-16794-9

Review on remote heart rate measurements using photoplethysmography

Ru Jing Lee[1] · **Saaveethya Sivakumar**[1] · **King Hann Lim**[1]

Received: 19 November 2022 / Revised: 28 July 2023 / Accepted: 31 August 2023 /
Published online: 19 October 2023

Abstract
Remote photoplethysmography (rPPG) gains recent great interest due to its potential in contactless heart rate measurement using consumer-level cameras. This paper presents a detailed review of rPPG measurement using computer vision and deep learning techniques for heart rate estimation. Several common gaps and difficulties of rPPG development are highlighted for the feasibility study in real-world applications. Numerous computer vision and deep learning methods are reviewed to mitigate crucial issues such as motion artifact and illumination variation. In comparison, deep learning approaches are proven more accurate than conventional computer vision methods due to their adaptive pattern learning and generalization characteristics. An increasing trend of applying deep learning techniques in rPPG can improve effective heart rate estimation and artifact removal. To consider more realistic disturbances into account, additional vital signs and large training datasets are crucial to improve the accuracy of heart rate estimations. By taking the benefit of contactless and accurate estimation, the application of rPPG can be greatly adopted in real-world activities, especially in precision sports.

Keywords Remote photoplethysmography · Heart rate measurement · Signal processing · Computer vision · Deep learning

1 Introduction

Heart rate (HR) is one of the most important physiological signs, especially for the early detection of cardiovascular catastrophes and cardiac arrhythmia [1]. Generally, the methods for HR monitoring can be classified as contact and non-contact type measurement. For

✉ Ru Jing Lee
19948665@student.curtin.edu.au

Saaveethya Sivakumar
saaveethya.s@curtin.edu.my

King Hann Lim
glkhann@curtin.edu.my

[1] Department of Electrical and Computer Engineering, Curtin University Malaysia, CDT 250, Miri 98009, Sarawak, Malaysia

contact type measurement, the electrocardiogram (ECG)-based HR detection [2] can provide reproducible and reliable detection in most clinical diagnoses. Ten adhesive ECG electrodes [3] are required to attach to the specific body locations to measure bio-potential output. Hence, it restricts the moving flexibility of patient in the usage scenario of an ECG machine. On the other hand, photoplethysmography (PPG) [4] is a widely used alternative solution to overcome the limitations while offering a reliable measurement. This non-invasive optical technique utilizes a light source and photo-detector to sense the blood volume pulse (BVP) in the blood vessels beneath the skin during blood circulation. According to Beer-Lambert's law [5], for the attenuation of light under a specific wavelength, the light reflected through the skin can be considered through the process of diffusion and scattering. The light absorbed by the skin is proportional to the penetration of light into the skin and the concentration of hemoglobin in the blood. The hemoglobin concentration variation in a cardiac cycle causes a small variation between the reflected and transmitted light intensity that is recorded as a PPG signal by a photodetector for HR measurement. This approach is commonly integrated into wearable devices such as pulse oximeters, smartwatches, and smartphones.

Remote health monitoring has recently gained public interest, because of its entirely non-contact properties, especially during the pandemic. The widespread physical social distancing and stay-at-home orders during the COVID-19 pandemic have caused obstacles in clinical physical examinations and created a pressing need for remote healthcare and telemedicine predominantly [6, 7]. Thus, it is engaged to find an innovative or efficient adaptation of existing technology to support remote monitoring methods. Remote photoplethysmography (rPPG) could be a vital HR screening and monitoring method for the infected and quarantined individuals in terms of telemedicine during the overwhelming global pandemic to avoid any risk of healthcare-associated infections on healthcare professions [8]. As having the same operation as the conventional contact PPG, rPPG uses a digital camera that functions as the photodetector to record the subtle changes on the skin, particularly the face of the subject from a few meters away, while ambient light acts as the light source. The frontal face video of a subject is captured with a camera and passed through a set of algorithms and models to perform HR estimation.

This technique was first proposed by Verkruysse et al. [9] used a consumer-level camera with ambient light to estimate HR successfully in the year 2008. After that, several conventional methods were proposed to reduce the impacts of motion artifact and illumination variation, targeted to improve the HR estimation accuracy. However, these methods typically made some hypothetical linear assumptions on subject head movement and skin reflections, which may not hold in real cases. Due to the advancement in Deep Learning (DL) recently, the application of DL in the digital signal processing field has beneficial and empowered object detection, e.g., people detection in [10], and several image processing, such as image super-resolution in [11], and image denoising in [12]. Since this field relies heavily on object detection and image processing, rPPG research follows this DL trend by using end-to-end methods or creating hybrid ways combining deep learning and traditional methods. Both methods have outperformed conventional computer vision methods. This paper systematically reviews both conventional and DL approaches for rPPG HR measurement.

The authors endeavored to address several significant research questions, as outlined in Table 1. The rest of the paper is organized as follows. Section 3.1 presents and analyzes various conventional methods. Section 3.5 categorizes the DL-based methods into end-to-end and hybrid methods. Section 3.8 discusses the current challenges and gaps in the field. Section 3.9 explores the potential real-world applications of the methods. Finally, Section 4 concludes the paper and suggests some future directions for research.

Table 1 Questions on reviewing rPPG

• What are the available types of rPPG techniques in the current research?
• How do different types of rPPG techniques work and what are the latest findings and improvements in this field?
• What are the main sources of error or variability in rPPG measurements and how can these be minimized or controlled?
• What are the limitations and key challenges of rPPG technology and how researchers attempted to address them?
• What is the current development using machine learning and deep learning in rPPG?
• What are the potential applications of rPPG technology, and how have these applications been tested in real-world settings?
• What are the privacy considerations surround rPPG technology, particularly in sensitive contexts such as healthcare or surveillance?

2 Review methodology

A comprehensive understanding of the various types of studies in this domain is necessary to review the current rPPG methods systematically. Therefore, an abundance of information from various sources was gathered to provide a comprehensive overview of the current state of the art. The research strategy sought to identify relevant studies published from 2008 to the beginning of 2023.

The terms and phrases used in different studies are inconsistent in this field. Therefore, the search and screening were performed using a combination of the following keywords: 'remote photoplethysmography', 'remote', 'photoplethysmography', 'rPPG', 'imaging', 'non-contact', 'contactless', 'contact-free', 'camera', 'webcam', 'video', 'facial video', 'camera-based', 'video-based', 'blood volume pulse', 'pulse', 'pulse estimation', 'heart rate', 'heart rate estimation', 'heart rate measurement', 'heart rate monitoring', 'physiological measurement', ' physiological signals', 'deep learning', 'end-to-end', 'convolutional neural network', 'attention network', 'spatial-temporal', 'spatio-temporal network', 'generative adversarial networks', 'super resolution', 'transformer', along with their abbreviations sometimes.

The information sources included conferences, symposium proceedings, periodicals, research articles, and books. In addition, a Google Scholar search using the listed keywords is conducted to obtain a large number of relevant published studies.

- IEEE eXplore (ieeexplore.ieee.org)
- Elsevier (elsevier.com)
- Springer (link.springer.com)
- ACM Digital Library (dl.acm.org)
- Other reputed research journals
- Other reputed conference/symposium proceedings

Figure 1 presents the PRISMA flow diagram for this systematic literature review (SLR). This review focused on rPPG experiments that employed facial video as their input modality and sought to infer heart rates from the video frames. We concentrated on research that used consumer- and commercial-level recording equipment, such as webcams and consumer cam-

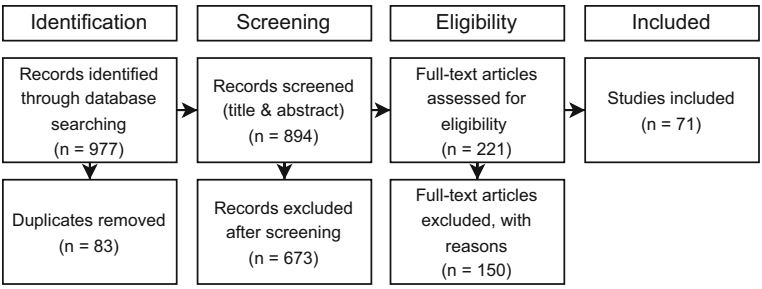

Fig. 1 PRISMA flow diagram for this systematic literature review (SLR)

eras. In our studies, the used specialized tools, such as high-speed and infrared cameras were excluded in the comparison because they were not representative in this field, where RGB cameras are the most accessible and popular tool for the public. But to make comparisons, we briefly addressed a few of these research. This information sourcing did not include the techniques, algorithms, and research that were not specifically about rPPG but were utilized in some of the chosen studies. 71 papers that fit the criteria and related directly to rPPG were included.

3 Result

This section analyses all the selected studies classified into two categories: conventional computer vision methods and deep learning methods. The studies on conventional computer vision methods are arranged according to the critical stages in the conventional workflow and evaluated based on their novelty and performance, as presented by the original studies. Meanwhile, the studies on deep learning methods are divided into two sections: hybrid and end-to-end. In the hybrid section, the studies are also grouped based on their involvement in the significant stages of the conventional workflow and discussed based on their novelty

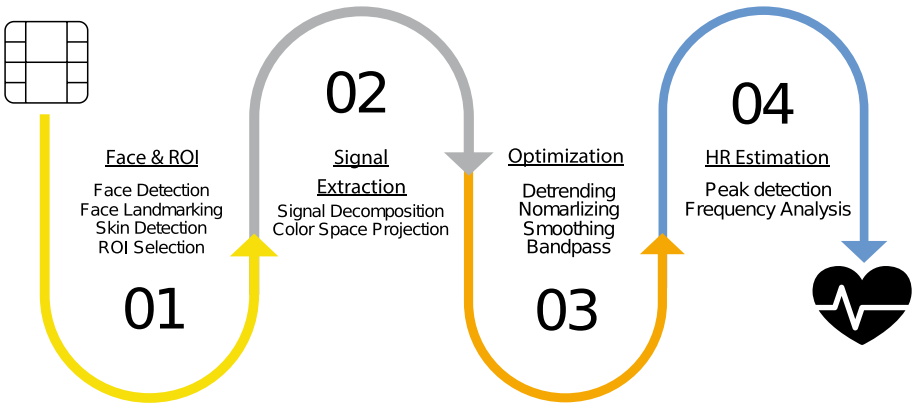

Fig. 2 Basic workflow for conventional computer vision methods

and performance. In the end-to-end section, the studies are classified according to their deep learning architecture and evaluated based on their novelty and performance.

3.1 Conventional computer vision methods

Most remote HR estimation methods share the general workflow as shown in Fig. 2. The workflow begins with a recorded or live video of a subject through a camera. Next, a pre-processing step is to crop out the human face region in each video frame and define a region of interest (ROI) within the bounded region (Fig. 2). The raw rPPG signal is then extracted from the pixels within ROI by various methods. After passing through some signal optimizations, for e.g. filtering out unwanted frequencies, the HR estimation process is proceeded by mathematical analysis. In some literature studies, an optimization algorithm is applied before signal extraction. Table 2 summarises this review's discussed conventional methods. On the other hand, Fig. 3 presents the breakdown of the included studies by classifying their features.

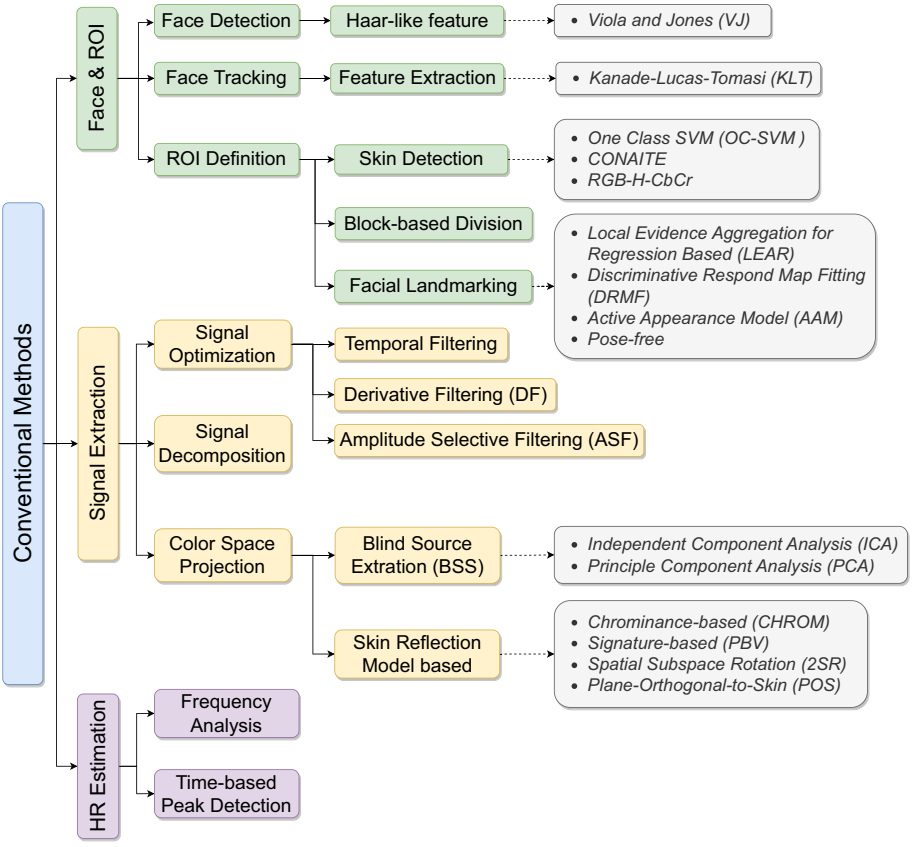

Fig. 3 An overview on discussed conventional computer vision methods

Table 2 Summary of down-selected conventional signal processing methods

Publication	Face detect & track	ROI	Signal extraction	HR estimation*	Database	Result
Verkrusse et al. [9]	Manually fixed	Entire	Green channel, butterworth	F	Private	-
Poh et al. [13]	VJ	60% width	ICA, bandpass	P	Private	-
Poh et al. [14]	VJ	60% width	ICA, bandpass, detrending	P	Private	-
Lewandowska et al. [15]	Manually fixed	Forehead	PCA, bandpass	F	Private	-
Wu et al. [16]	Manually fixed	Face	EVM	-	-	
Haan et al. [17]	VJ	Selected skin pixels, face	CHROM, FIR bandpass	F	Private	-
McDuff et al. [18]	VJ, LEAR	Face, ex. eyes region	ICA, bandpass. detrending	P	Private	-
Lam et al. [19]	VJ, landmark	Cheeks	FastICA, temporal filter	F	MAHNOB-HCI	RSME=8.9
Wang et al. [20]	CSK, OC-SVM	Manually	Pixel-level CHROM, spatial pruning, temporal filter	F	Private	-
Feng et al. [21]	VJ, KLT	Cheeks	GRD, bandpass	F	Private	-
Li et al. [22]	VJ, DRMF, KLT	Selected skin pixels, Face (below eyes)	DRLSE,NLMS, temporal filter	F	MAHNOB-HCI	RSME= 7.62, r =0.81
Haan et al. [23]	-	-	PBV	P	Private	-
Wang et al. [24]	OC-SVM	Selected skin pixelson face	2SR	P	Private	-
Wang et al. [25]	CSK	Selected skin pixels, face	POS		Private	-
Fouad et al. [26]	VJ, KLT, skin seg	Forehead, cheeks (skin pixels)	BSS based	F	Private	-
Macwan et al. [27]	VJ, KLT, Skin seg	Selected skin pixels, Face	MAICA	F	UBFC-RPPG, MMSE-HR	MMSE-HR, MAE=3.91, r=0.86
Gudi et al. [28]	AAM	Selected skin pixels, face	POS, butterworth	F	VicarPPG, PURE, MAHNOB-HCI	MAE=2.64; MAE=0.3; MAE=12.8

*F indicates Fast Fourier Transform (FFT). P indicates Peak Detection

3.2 Face detection and ROI

Face detection is the first step in the entire rPPG, and sometimes, it is associated with a face-tracking algorithm to reduce motion artifacts that require additional processing. As the detected face border still includes a greater scale of non-facial skin regions, most researchers choose specific parts as ROI on the detected face for additional HR estimates rather than the full detected face.

3.2.1 Face detection

The HR estimation is entirely based on the estimated PPG signal resulting from the subtle skin color variation caused by the cardiac pulse cycle. Thus, it is essential to undergo the preprocessing step to detect the face in each frame correctly, prior to choosing the favorable ROI within the bounding for subsequent raw signal extraction. In early studies [9, 15, 16], face detection was absent with directly selected fixed region as ROI.

The Viola and Jones (VJ) algorithm is the state-of-the-art object detection using a boosted cascade of simple features. It was frequently applied in various studies [13, 14, 17, 22, 29]. It utilizes the Haar feature and Integral image to detect and form a rectangular frame that bounds the subject's facial image. It becomes an iconic face detection algorithm among a variety of conventional and DL-based rPPG methods because of its automated face detection, high detection rate, and availability in the computer vision library of OpenCV and MATLAB. The benefit of using a face detection algorithm, like the VJ algorithm, is to provide a simple removal for undesirable and uninterested regions at the beginning of the entire workflow.

3.2.2 Face tracking

The implementation of face detection in every frame consumes large processing power and hence, it makes low feasibility in carrying out the real-time application. To tackle this matter, tracking the face over time in a smaller bounding box reduces the computational cost in the process. Lucas et al. [30] proposed the Kanade-Lucas-Tomasi (KLT) face tracker by inserting the facial bounding box from VJ face detection; the track-able points of the subject's face are then generated and tracked over time. It makes the re-detecting process serves only when the movements are severe. The combination of the VJ face detector algorithm with the KLT face tracker is commonly involved in several studies emphasizing their real-time performance, such as [21, 22, 26].

3.2.3 ROI definition

In early fundamental studies [9, 15, 16], the subject is required to stay still to enable manual ROI selection to form a fixed spatial region on the subject's entire face or forehead. Some studies [13, 17, 25] equipped with VJ face detection and use the VJ's result as their ROI without any selection. Assuming a sufficient and stable light condition, these solutions might still cause some motion artifacts because of head and eye movements, resulting in inaccuracy signal extraction and HR estimation [31].

Choosing on suitable ROI for raw rPPG signal extraction is critical to improve the accuracy of HR estimation. Using VJ algorithm to bound a "face rectangle" may result in imprecise HR measurement due to the unavoidable non-facial pixels at the corners of the rectangle. The result can be worse when considering head movement [22]. Compared to the entire face, the

cheeks and forehead are usually selected as excellent ROIs because of the high blood vessel distribution [31]. However, more studies applied either skin detection or facial landmark algorithm to obtain favorable and adaptive ROI.

Skin selection is a process that removes all the non-skin pixels, such as the background, eyes, eyebrows, hair, and so on while maintaining the skin pixels. Besides the simple skin detection algorithm used in Han et al. [17], Wang et al. [20] used an OC-SVM classifier to distinguish skin and non-skin pixels. Besides that, Fouad et al. [26] proposed a novel approach to skin detection and segmentation. The purpose is to identify skin pixels and remove non-skin pixels from the VJ's bounding box of the face prior to ROI selection. The comparison between CONAITE [32] and RGB-H-CbCr [33] was made to select the best skin detection method. As a result, the CONAITE with the lower RMSE (6.06 bpm lower at the most) is more robust in differentiating between skin and non-skin areas. With the skin segmentation process, Fouad et al. [26] proposed their own ROI definition with three small rectangular regions located at the forehead and cheeks.

A facial landmark detection algorithm is an alternative or optional step to identify the locations of key landmark points of the unique facial components on facial images or videos [34]. It is excellent in generating unevenly shaped ROI or various ROIs simultaneously. There are plenty of facial landmark detectors, and were applied to different studies. In McDuff et al. [18], the Local Evidence Aggregation for Regression-Based (LEAR) facial landmark detector [35] was employed to generate the selected ROI, excluding a region around the eyes by locating the x-y coordinates of one hundred twenty-five thousand pixels at the mean. LEAR is a regression-based approach combined with a probabilistic graphical model-based face shape model.

Li et al. [22] proposed the Discriminative Respond Map Fitting (DRMF) facial detector [36] to define a face mask-like shaped ROI with sixty-six facial landmarks. On the other hand, Gudi et al. [28] proposed the Active Appearance Model (AAM) [37] to define ROI containing only the upper region of the face excluding the eyes. AAM identifies the landmark location by inserting the learned global facial shape patterns into the testing images [34]. Lam et al. [19] proposed the Pose-free facial landmark [38] that was used to sixty-six facial landmarks for all frames in the video.

Po et al. [39] applied the detection of face using VJ, KLT, and Speeded up robust features (SURF) [40] before the ROI selection. Its novelty is presented as the whole facial region in each face image is divided into non-overlapped blocks. Later, the SNR maps were calculated with mean-shift clustering to reflect the rPPG signal quality distribution on the face. An adaptive thresholds in SNR maps was obtained to generate the block-wise adaptive ROI

3.3 rPPG signal extraction

Instead of using an ideal light source in the contact PPG, rPPG uses the uncontrolled ambient light from the surroundings as the light source to estimate HR. Since the light source is environment-dependent, it significantly causes unwanted noises due to motion and illumination. Hence, it requires more signal processing to be applied during raw signal extraction. After the ROI was selected and tracked, it was split into the three RGB channels. To reduce the camera quantization error, the three channels averaged spatially among all the pixels within the selected ROI; thus, the raw RGB traces can be obtained. The traditional approaches to optimize and extract useful rPPG signals can be classified into signal decom-

position and color space projection ways as a result of improving the signal-to-noise ratio (SNR).

3.3.1 Signal decomposition

A bandpass filtering process is performed on the spatial means to remove the non-physiological information, mainly the undesired high and low-frequency noise. The frequency band (0.75 to 4 Hz) is commonly picked, corresponding to the HR in the normal range of 42 to 240 bpm [41]. This process is carried out by a common bandpass filter, for e.g. Butterworth, Hamming window FIR, and is involved in every conventional rPPG method. These are fundamental steps to extract the useful signal, which was first included in [9].

3.3.2 Signal optimization

To mitigate the influences of the motion artifact and illumination variation, temporal filtering is commonly used in several studies instead of bandpass filtering only. The robust temporal filtering includes detrending, normalization, smoothing, and bandpass filtering. The concept was first proposed by [14] and implemented in several studies [18, 19, 22, 26]. These studies all underwent temporal filtering using a similar process. Firstly, a detrending filter using a procedure based on a smoothness priors' approach, proposed by Tarvainen et al. [42] is used to minimize low-frequency, slow, and non -stationary trend of the signal. The smoothing parameter λ was set differently, varying from 10 to 2000 in different studies. Secondly, normalization is carried out, as periodicity is the main interest. Thirdly, a moving-average filter smooths out the random noise. Lastly, the bandpass filter, as mentioned previously.

Besides bandpass-based filtering, amplitude selective filtering (ASF) was proposed in Wang et al. [43] to reduce the effect of head motion. The pulsatile amplitude is exploited as a bio-metric signature by distinguishing the frequency components of motion noise and BVP signal because the spectral amplitude of motion is much larger than the spectral amplitude of the pulse. Yang et al. [44] proposed a derivative filter (DF) was proposed to provide a motion-tolerant signal extraction in the next step. The three-order derivative of the Gaussian filter becomes low during smooth changes but high during steep changes. Thus, it was utilized to select subtle color variation under large motions as it depicts smoother trajectories at a temporal scale compared to the motion interference.

3.3.3 Color space projection

Compared to signal decomposition methods, the color space projection methods emphasize the application of different algorithms to extract useful rPPG signals from the raw traces. Usually, the more accurate and robust rPPG signal can be extracted with the use of the dimension reduction concept to tolerate small motion artifacts and illumination variations. Blind source separation (BSS) was implemented in rPPG studies to distinguish the desired RGB signal from noise and motion artifacts. This method assumes that the pulse signal shows the strongest periodicity and ignores the distortion caused by periodic motion. The representative BSS techniques that are widely applied are independent component analysis (ICA) [13, 14] and principle component analysis (PCA) [15].

The ICA algorithm for rPPG was explained in Poh et al. [13, 14]. It chose green channel trace as reported by Verkruysse et al. [9] and utilized the joint approximate diagonalization

of eigenmatrices (JADE) algorithm developed by Cardoso et al. [45] to perform motion-artifact removal by separating the fluctuations caused by BVP from the raw RGB signals. Several studies included this method in their works, such as [18, 19, 46]. After that, the PCA algorithm was proposed in Lewandowska et al. [15] and applied in other works such as Wang et al. [20]. The author claimed that it is better performance and higher computational efficiency and resulting a similar HR estimation accuracy with ICA. The work also showed sufficient accuracy in estimating pulse rate based on the small rectangular region on the forehead compared to the whole face. However, both ICA and PCA methods showed lower accuracy in excessive movement.

Chrominance-based (CHROM) algorithm [17] was first proposed to address the motion intolerance weakness of the BSS techniques. It considered the pulse as a linear combination of RGB channels under a standardized skin-tone assumption. The influence of motion artifacts could be eliminated through the track using color difference signals. Experimental results showed higher motion-robustness compared to BSS methods in the presence of periodic motion. With one hundred seventeen subjects in vigorous motion, and the HR estimated through a simple peak detector, the correct pulse rate at 48% of the time, while ICA and PCA scored 4% and 11%, respectively. In Wang et al. [20], the CHROM was used as a baseline to construct a pixel-based rPPG sensor integrated with motion-compensated pixel-to-pixel pulse extraction, optimized spatial redundancy and temporal filtering, resulting in a more motion tolerance algorithm.

Shortly, the CHROM's author proposed the BVP signature-based PBV [23] indicating better motion robustness than BSS methods and CHROM. It used the signature of blood volume changes in different wavelengths of RGB to gain the pulse-induced color changes from motion noise. With 117 subjects on different gym equipment, the improved result could be obtained from a similar experiment setup in CHROM, the correct pulse rate increased to 68% from 60% (CHROM), and SNR increased from -5dB (CHROM) to -4dB.

The Spatial Subspace Rotation (2SR) algorithm [24] was proposed to claim its higher performance for subjects with dark skin tone or body motions in complex illuminance conditions. Its core concept is to exploit the skin-pixel distribution and estimate the temporal rotation of skin pixels subspace in RGB for deriving the pulse rate. The author claimed that the 2SR method outperformed ICA, CHROM, and PBV. Based on their private benchmark dataset containing 54 videos, the SNR of 2SR was obtained as 6.55dB compared to ICA, CHROM, and PBV were all lower than 5dB, and the highest PCC at 0.94 compared to others ranged at 0.64-0.78.

An alternative concept, the plane-orthogonal-to-skin (POS) [25] was proposed. It defined a pulsatile plane orthogonal to the skin tone in temporally normalized RGB space for pulse extraction. It also compared GREEN, PCA, ICA, CHROM, PBV, 2SR, and POS based on their private benchmark dataset containing 60 videos. In fitness challenges, the model-based methods (CHROM, PBV, and POS) surpassed the non-model-based methods (GREEN, PCA, and ICA) in motion robustness.

3.4 HR estimation

Conventionally, the extracted rPPG signal relies on mathematical steps using frequency analysis or peak detection methods to perform HR estimation. In GREEN [9], it proved that the green channel contains the most sufficient and significant PPG signal, compared to red and blue channels. Usually, after the rPPG signal extraction, the filtered green channel is taken out and normalized.

3.4.1 Peak detection

The peak detection method is a relatively simpler method that detects the peaks on the temporal rPPG signal to calculate the heart rate without converting to the frequency domain. The recovered rPPG signal, however, frequently exhibits noise and has a lower temporal precision than the contact type due to its non-contact nature. For example, Mcduff et al. [47] proposed an automated method for the detection of the systolic and diastolic peaks and capturing the peak-to-peak time (SD-PPT) from the rPPG signal in the time domain. The estimated HR is the inverse of SD-PPT.

3.4.2 Frequency analysis

The extracted rPPG signal is converted from the spatial domain to the frequency domain by FFT or DCT. The power spectral density (PSD) distribution is estimated as the function of frequency using Welch's method. The frequency with the maximal power response is recognized as the rPPG signal. The HR over a particular period is estimated as (1). This method is more commonly adopted in studies compared to the peak detection approach.

$$HR_{estimated} = 60 * f_{rPPG} \tag{1}$$

3.5 Deep learning method

The introduction of DL methods has recently achieved better generalization and higher accuracy because it can learn the pattern of spatial and temporal features simultaneously in the training phase as compared to the signal processing methods. As shown in Table 2, most rPPG methods in the early development used their own recorded datasets, which were not publicly accessible. This issue has made difficulties in the evaluation and generalization of different rPPG methods. The situation has turned over as several common datasets were introduced and used by various DL-based methods. Table 3 summarizes the publicly available datasets commonly used in training and testing at the state-of-the-art. Learning-based approaches can often be divided into two categories, i.e., hybrid and end-to-end methods. Hybrid methods require a portion of conventional methods while serving as compensation or enhancement throughout the entire pipeline, from video input to numerically estimated HR. In contrast, the end-to-end method intakes video sources and outputs with numerical estimated BVP or HR without human intervention. Figure 4 presents an overview of all discussed approaches in this review, and Table 4 summarizes all the discussed deep learning methods in this review.

3.6 Hybrid methods

The DL approaches can be applied to enhance and compensate for the major steps, including face detection, signal extraction/optimization, and HR estimation in the conventional pipeline. Multiple enhancements in a single approach are possible. A few selected approaches in each major step are discussed below.

3.6.1 Face detection

Hsu et al. [48] proposed a customized face and landmark detection network to address inconsistent facial regions and landmarks detected across frames caused by the variation of the facial regions given by the VJ method. It combined a Single Shot MultiBox Detector (SSD) face detector [49] and a double dropout regression network for landmark localization. The

Table 3 Common used dataset in rPPG research

Dataset	Subject	Video	Physiological signal	Description
PURE	10	RGB 640*480@30fps	HR, SpO2, PPG	6 different head motions
COHFACE	40	RGB MPEG4 640*480@20fps	BVP, RR	2 illumination settings
MAHNOB-HCI	30	RGB H.264 780*580@51fps	ECG, RR	Recorded while watching video 6 different views
UBFC-RPPG	42	RGB 8-bit 640*480@30fps	HR, PPG	Recorded while play time-intensive game
MMSE-HR	40	RGB 1040*1392@25fps	HR, RR, BP	Subject with different skin tones
VIPL-HR	107	HR, SpO2, BVP	Recorded by different devices	

SSD face detector is a feed-forward-based convolutional network optimized for real-time object detection. By the SSD-detected face, the CNN-based landmark detector renders the 68 coordinates of landmarks on the face. This work was evaluated on the 300W benchmark database [50], and it resulted in a normalized error of 5.02, compared to RLBF's 5.16 and SDM's 5.70. For subsequent steps, another CNN trained by a set of Time-Frequency Representation images converted from a collection of facial videos with gold-standard pulses was used to estimate HR.

Tran et al. [51] proposed an adaptive pulsatile plane approach to eliminate the light reflection changes induced by motions. The state-of-the-art semantic segmentation Deeplabv3+ was implemented to segment the skin pixels from the facial region pre-detected by any face detection and tracking methods, where the commonly used VJ was implemented in the approach. The celebA-HQ dataset was used to train the model for skin pixel segmentation. The subsequent conventional methods were included to extract the rPPG signal and estimate HR. The author also demonstrated this practical application in a hospital.

3.6.2 Signal extraction

Perepelkina et al. [52] proposed HeartTrack, a 3D spatio-temporal attention neural network, to solve the lack of real-life biomedical data as the deep learning methods typically require large amounts of training data. The face detection is preprocessed by the RetinaNet network independently. The key feature in this work, HeartTrack, is a 3D spatio-temporal attention neural network used to do three major works simultaneously: choose the best-fit ROI from pre-processed face detection for pulse detection in each frame, the optimal nonlinear function of color channels, and to complete signal filtering using temporal information simultaneously. The output is then fed to a feed-forward 1D CNN for HR estimation. The model was trained with private datasets (MoLi-ppg-1 and MoLi-ppg-2) and tested on UBFC-RPPG.

In Song et al. [53], the rough rPPG pulse signal extracted through CHROM is inputted to the proposed denoising method, PulseGAN. This generative adversarial network (GAN) is capable of outputting with a denoised pulse waveform, thereby the significant improvement in the accuracy of HR, interbeat interval (IBI), and heart rate variability (HRV). It was trained and tested with various public datasets, including UBFC-RPPG, PURE, VIPL-HR, MAHNOB-HCI, and in-house BSIPL-RPPG databases in terms of intra-dataset and cross-dataset.

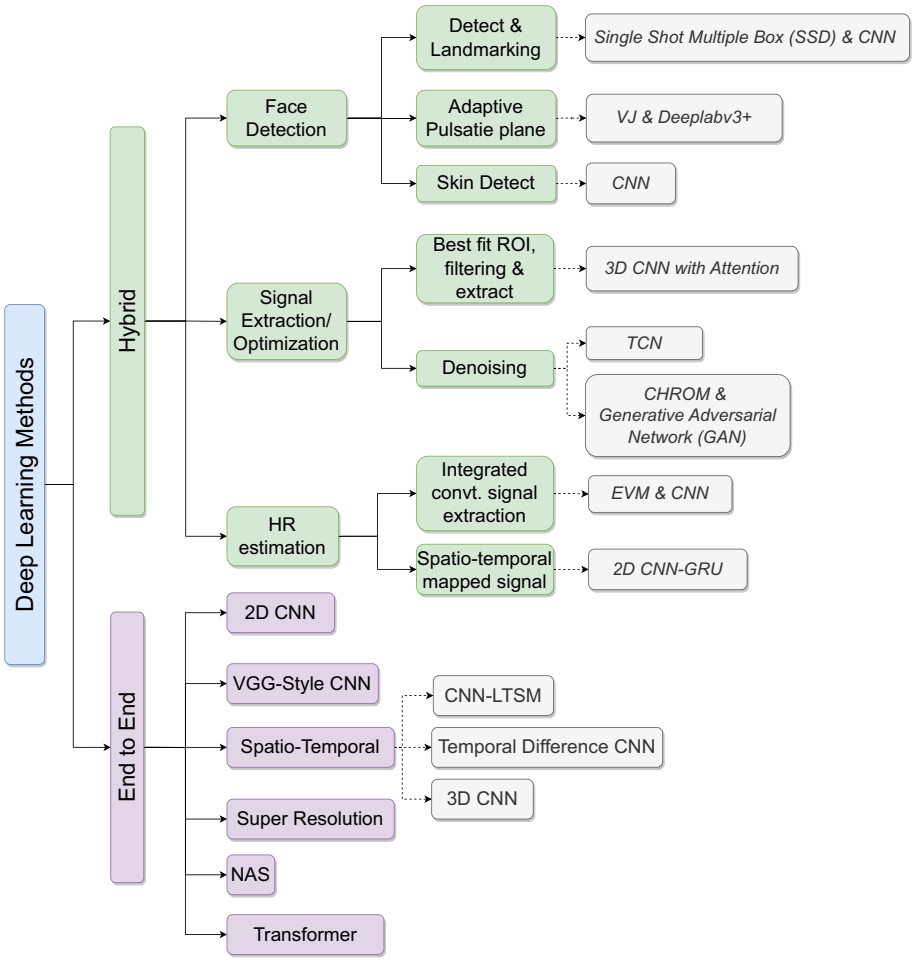

Fig. 4 An overview of discussed DL methods

Similarly, Lu et al. [54] proposed Dual-GAN that involves two GANs, BVP-GAN and Noise-GAN, trained jointly to model the BVP predictor and noise distribution. The input video is first preprocessed to form a spatial-temporal map (STMap) as a raw representation of the noisy BVP signal. The first GAN model, BVP-GAN, learns the mapping from STMap to BVP, and the second GAN model, Noise-GAN, learns the noise distribution. The two GANs work together to enhance and promote each other's capabilities via adversarial learning. It exhibited superior performance relative to conventional methods and outperformed PulseGAN [53] in terms of lowest SD, MAE, RMSE, and greatest correlation coefficient across all intra-datasets evaluation, including UBFC-rPPG, PURE, and VIPL-HR. Dual-GAN was also trained on PURE and tested on UBFC-rPPG for cross-dataset testing, resulting in the lowest MAE of 0.74, RMSE of 1.02, and highest correlation coefficient 0.997.

Wu et al. [55] proposed a novel error compensation recurrent neural network by incorporating the raw face images and the raw rPPG pulse signal extracted via CHROM to compute the corresponding correction term to enhance performance. It was trained using two in-house

Table 4 Summary of down-selected DL methods

Publication / model	Type**	Network used	Training dataset	Testing dataset	Result			
					SD[a]	MAE[b]	RSME[c]	r[d]
HR-CNN [60]	EE	CNN	1) COHFACE	1) COHFACE	-	8.10	10.78	0.29
			2) PURE	2) PURE		1.84	2.37	0.98
			3) MAHNOB-HCL	3) MAHNOB-HCL		7.25	9.24	0.51
DeepPhys [61]	EE	VGG style CNN	RGB Video I*	MAHNOB-HCI	-	4.57	-	-
EVM-CNN [58]	H	CNN	MMSE-HR	MMSE-HR	6.85	-	6.95	0.98
STVEN-rPPGnet [62]	EE	3D CNN	1) OBF	MAHNOB-HCI	5.57	4.03	5.93	0.88
			2) MAHNOB-HCI					
PhysNet [63]	EE	3D CNN	OBF	MAHNOB-HCI	7.84	5.96	7.88	0.76
AutoHR [64]	EE	NAS	1) VIPL-HR	1)VIPL-HR	8.48	5.68	8.68	0.72
			2) MAHNOB-HCL	2) MAHNOB-HCI	4.73	3.78	5.10	0.86
			3) VIPL-HR	3) MMSE-HR	5.71	-	5.87	0.89
HeartTrack [52]	H	3D CNN	MoLi-ppg-1*	UBFC-RPPG	-	-	3.368	0.983
		1D CNN	MoLi-ppg-2*					
RhythmNet [59]	EE	CNN-RNN	VIPL-HR	1) VIPL-HR	8.11	-	8.14	0.76
				2) MAHNOB-HCI	3.97		3.99	0.87
				3) MMSE-HR	5.45		5.49	0.84
Meta-rPPG [65]	EE	CNN-LSTM	1)MAHNOB-HCI	1)MAHNOB-HCI	4.90	3.01	3.68	0.85
			2)UBFC-rPPG	2)UBFC-rPPG	7.12	5.97	7.42	0.53
PulseGAN [53]	H	GAN	VIPL-HR	MAHNOB-HCI	-	-	6.53	0.71
Dual-GAN [54]	H	GAN	1) UBFC-rPPG	1) UBFC-rPPG	-	0.44	0.67	0.99
			2) PURE	2) PURE	-	0.82	1.31	0.99
			3) VIPL-HR	3) VIPL-HR	7.63	4.93	7.68	0.81
			4) PURE	4) UBFC-rPPG	-	0.74	1.02	0.997
Yue et al. [66]	EE	RBPN,	1) MMVS*	1) MMVS*	-	3.32	4.87	0.93
		ResNet-10	2) DEAP	2) DEAP		4.23	5.47	0.89

Table 4 continued

Publication / model	Type**	Network used	Training dataset	Testing dataset	Result SD[a]	MAE[b]	RSME[c]	r[d]
AND-rPPG [56]	H	TCN	1) UBFC-rPPG	1) UBFC-rPPG	4.01	3.15	4.75	0.92
			2) COHFACE	2) COHFACE	0.92	7.83	8.06	0.63
Wu et al. [55]	H	1D CNN	1) Color Temp*	1) UBFC-rPPG		3.28	3.95	
			2) Passenger*	2) PURE		2.13	2.83	
				3) Driver*		5.92	8.04	
TS-CAN [67]	EE	2D CNN	AFRL	MMSE-HR	-	3.41	7.82	0.84
RTrPPG [68]	EE	3D CNN	VIPL-HR	VIPL-HR	-	3.99	-	0.73
RADIANT [69]	EE	Transformer	Synthetic data	1) UBFC-rPPG	3.45	2.91	4.52	
				2) COHFACE	7.41	8.01	10.12	
Physformer [70]	EE	Transformer	1) VIPL-HR	1) VIPL-HR	7.74	4.97	7.79	0.78
			2) MAHNOB-HCI	2) MAHNOB-HCI	3.87	3.25	3.97	0.87
			3) OBF	3) OBF	-	-	0.804	0.998
			4) VIPL-HR	4) MMSE-HR	5.22	2.84	5.36	0.92
Physformer++ [71]	EE	Transformer	1) VIPL-HR	1) VIPL-HR	7.65	4.88	7.62	0.80
			2) MAHNOB-HCI	2) MAHNOB-HCI	3.90	3.23	3.88	0.87
			3) OBF	3) OBF	-	-	0.765	0.998
			4) VIPL-HR	4) MMSE-HR	5.09	2.71	5.15	0.93

[a]Standard deviation in BPM
[b]Mean absolute error in BPM
[c]Root square mean error in BPM
[d]Pearson correlation coefficient
*Own private dataset
**H indicates hybrid, EE indicates end to end

datasets. These datasets included recordings of the subject in various color temperatures and as a passenger in a moving vehicle on sunny and inclement days. Additionally, the model was evaluated on PURE, UBFC, and an in-house dataset containing recordings of the subject driving a moving vehicle. Lower RMSE and MAE in all datasets showed the model's better performance. Due to their well-controlled conditions, PURE and UBFC datasets improved less. The model performed better in the noisier in-house dataset, including a real driving scenario with head motion and illumination variations. The model reduced MAE by 66.7% in this dataset. The study's results indicate that the proposed compensation network with error mapping can significantly enhance the robustness of rPPG in noise-heavy conditions.

Lokendra et al. [56] proposed AND-rPPG, a denoising rPPG method capable of effectively mitigating the facial expression-based noise from the temporal signal. It incorporated action units (AUs) for analyzing facial expressions and denoising temporal signals. Several non-causal Temporal Convolutional Networks (TCN) were used to denoise different facial regions. For the pre- and post-processing, the CLNF Openface was employed to generate ROI, while the BSS-based Multi-kurtosis optimization and FFT were responsible for signal extraction and HR estimation. The approach was trained and tested on UBFC-rPPG and COHFACE. The author also proved the easy integration of this method into other state-of-the-art methods to improve their performance.

The emergence of novel coronavirus pneumonia, COVID-19, has garnered worldwide attention. To address the low-performance problem in conventional methods caused by the lack of facial information, particularly when wearing masks, Zheng et al. [57] proposed a non-end-to-end CNN-based residual network model. The proposed model utilizes the location of human eyeballs to locate the frontal ROI, generates spatio-temporal feature images, and determines their authenticity. Only the signal that passes this test undergoes FFT processing on the chrominance signal using CHROM [17] to predict the HR. The RMSE after correction is improved to 4.65 bpm, which is 0.42 bpm higher than without correction. Moreover, the correlation is improved from 0.85 to 0.95 on their private dataset, in which their subjects wore masks.

3.6.3 HR estimation

An improvement in HR estimation can be achieved via DL approaches for better regression modeling. Qiu et al. [58] applied the conventional face detection followed by EVM-CNN to regress the HR from the facial features. The Eulerian Video Magnification (EVM) [16] is used to extract face color changes that correspond to the heart rate information within a time interval. The CNN for HR estimation was mainly constructed by depthwise separable convolutions that combined depthwise convolutions and pointwise convolutions to reduce the computational burden and model size. EVM-CNN was intra-dataset trained and tested on MMSE-HR.

RhythmNet [59] utilizes the spatio-temporal map for HR estimation. After landmarking by the SeetaFace face detector on each frame, the HR signals from multiple ROI volumes are encoded to a spatio-temporal map. The CNN-RNN deep network consisting of 2D CNN and Gated Recurrent Unit (GRU) was trained on VIPL-HR to predict the HR from the spatial-temporal maps. It was tested on MAHNOB-HCI and MMSE-HR. The result showed that the diverse image acquisition conditions in VIPL-HR trained the RhythmNet to be a more robust HR estimator. The challenging factors, i.e., video compression, illumination, and head movement, were further tested.

3.7 End-to-end DL methods

A DL-based method is typically recognized as an end-to-end method if it takes video frames as input and outputs an HR estimation without supervised human intervention as the intermediate steps. The following methods are classified and discussed based on the types of networks used. However, some methods require prepossessing facial landmarking or outputting the rPPG (estimated BVP) information instead of the numerical HR estimation. Those methods are still classified as end-to-end frameworks because they were trained in an end-to-end manner.

3.7.1 Convolutional-based networks (CNN)

Various CNN-based architectures, such as 2D CNN, convolutional attention network (CAN), recurrent convolutional network (RNN), temporal shift CNN, and 3D CNN, have been widely adopted in rPPG-based HR estimation. This is because these models enable feature subspace mapping with minimal need for domain-specific knowledge. Spetlik et al. [60] first proposed a concept of end-to-end methods, HR-CNN. It is a two-step 2D CNN consisting of a signal extractor and HR estimator CNN. The extractor was trained to output the rPPG signal to form a sequence of the subject's facial images. The output rPPG signal was fed into the estimator for single scalar HR estimation. Evaluated with COHFACE, MAHNOB-HCI, and PURE, it outperformed the conventional methods (CHROM and 2SR) in terms of RMSE, MAE, and PCC.

Chen et al. [61] proposed DeepPhys, a convolution attention network (CAN) based on VGG-style 2D CNN, that consisted of a jointly trained motion model and appearance model to provide an end-to-end network that accurately estimates HR and breathing rate from RGB or infrared videos. Based on the skin reflection model, the motion model takes the normalized frame difference as input motion representation and outputs the estimated HR. The appearance model takes the same input but acquires the attention mechanism like human eyes to learn the difference between frames, thus assisting the motion learning by generating a spatial mask responding to stronger signals to skin areas. DeepPhys was evaluated on MAHNOB-HCI, resulting in -8.98dB in terms of SNR.

3.7.2 Spatio-temporal network

The mainstreams of spatio-temporal models are 3D CNN, RNN, or temporal difference based. A 3D CNN takes spatial and temporal information in the videos into account and typically performs better. In contrast, a 2D CNN lacks the ability to learn the temporal context and utilizes the spatial information of video only. In RNN based module, 2D CNN is first deployed to extract spatial features, and the RNN is used to propagate the spatial feature in the temporal domain. On the other hand, the temporal difference is a relatively straightforward approach by applying Temporal Shift Module (TSM) [72]. TSM is intended to shift a subset of channels along the temporal axis prior to 2D convolution, thereby enhancing the information exchange between adjacent frames and facilitating spatio-temporal modeling with negligible computational overhead and parameter cost.

Lee et al. [65] proposed Meta-rPPG, an end-to-end supervised learning method. It introduced the application of meta-learning and transductive inference in rPPG estimation. Meta-rPPG consists of a convolutional encoder, rPPG estimator, and synthetic gradient generator modeled by a CNN, a LSTM, and a shallow Hourglass network. The meta-learner

performs well when training data is abundant, with only a slight deviation of distribution from the testing set. To cope with out-of-distribution samples, e.g., a new video sequence, the modeled synthetic gradient generator involves performing transductive inference and a prototypical distance minimizer, resulting in a fast adaptation in a self-supervised fashion.

Liu et al. proposed TS-CAN [67], a two-branch CAN that utilizes one branch for motion modeling and the other for extracting spatial features. This architecture is comparable to DeepPhys [61] but with the additional capability of incorporating TSM to conduct temporal shifts on the tensor before it is fed into each convolutional layer of the motion branch. This mimics the effects of 3D CNN and enables spatio-temporal modeling. The model was trained on the AFRL dataset, cross-tested with MMSE-HR, and outperformed conventional methods and DeepPhys with a lower MAE of 3.41, RMSE of 7.82, and higher correlation coefficient 0.84, SNR of 2.92.

In STVEN-rPPGnet [62], two spatio-temporal networks were combined to form a two-stage end-to-end STN. The STVEN model is used to overcome the issue of highly compressed input video frames by outputting the spatially enhanced video. The rPPGnet model consisting of the skin-based attention module and partition constraint module, can be used individually to accurately estimate the HR and HRV. The author claimed better performance could be achieved if the rPPGnet and STVEN models were trained jointly.

Yu et al. [63] developed PhysNet based on the three different architectures, i.e. 3D CNN, 3D CNN with temporal encoder-decoder (ED) structure, and 2D CNN combined with RNN for further comparison. All variants were trained and validated under the OBF dataset. The result showed that 3DCNN-ED achieved the best performance in HR estimation compared to the worst case, RNN's 3.139 (using LSTM as RNN). The 3DCNN-ED PhysNet was then cross-tested with MAHNOB-HCI, and it outperformed HR-CNN and other 2D CNN-based models. Besides HR and HRV estimation, emotion recognition is also supported in [63].

3D CNN network with deep architecture requires high computational expense due to the simultaneous processing of spatial and temporal dimensions. Training a 3D CNN network using large dataset sis computationally expensive and time-consuming, which may make it unsuitable for real-time applications. To address these issues, Botina et al. [68] proposed RTrPPG, a lightweight 3D CNN architecture. The proposed 3D CNN architecture balances heart rate measurement precision and inference time by decreasing the input size for fast inference. The model employing YUV as the color space for skin segmentation provides the optimal balance between real-time, signal quality, and heart rate measurement performance. Consequently, the MAE of 3.99 bpm is comparable to PhysNet's 3.87 bpm in the VIPL-HR dataset, while the GPU and CPU inference procedure improved by approximately 88%, from 51.77 ms to 2.32 ms in GPU and from 245.57 ms to 28.03 ms in CPU.

3.7.3 Super-resolution network

Yue et al. [66] proposed rPPGRNet and THRNet to achieve end-to-end framework for the ease of training. In rPPGRNet, their specialized rPPGPM module was integrated with a recurrent back projection network (RBPN) to generate SR-upscaled facial images with enhanced and recovered rPPG signals from low-resolution facial videos. The author claimed that the integration of rPPGPM is prioritized to enhance and recover rPPG information in a facial image during SR processing, which was absent in the similar state-of-the-art RBPN approaches in Song et al. [73] and McDuff et al. [74]. The THRNet is designed based on ResNet-10 with

a modification of including temporal-wise attention mechanism to keep the model focusing on the pulsating information of blood flow. The model was trained and evaluated on DEAP and its own database, MMVS, with preprocessed facial landmarks.

3.7.4 Neural architecture search

Yu et al. [64] proposed AutoHR, an end-to-end baseline consisting of a powerful searched backbone with the designed Temporal Difference Convolution (TDC) to discover a well-suited backbone for remote HR measurement. The TDC inputted original RGB frames to discover the temporally normalized frame difference. The neural architecture search (NAS) was used to find out the best-suited 3D CNN backbone for HR measurement, enforced by a hybrid loss function with time and frequency constraints. The spatio-temporal data augmentation was involved for better representation learning. The model was trained on VIPL-HR and MAHNOB-HCI and conducted intra-dataset testing, respectively. For cross-dataset validation, the VIPL-HR trained model was tested on MMSE-HR.

3.7.5 Transformer network

The Transformer architecture has been widely used in natural language processing tasks such as BERT [75], BART [76], GPT-2 [77], and GPT-3 [78] and computer vision classification applications such as ViT [79] and DE:TR [80]. The objective of rPPG-based HR estimation is to estimate pulse signals from multiple temporal signals, containing mainly the pulse signal corrupted by noise. Consequently, the pulse signal generates a robust correlation between the temporal signals. The Transformer architecture can efficiently learn this due to its exceptional multi-head attention mechanism.

Yu et al. proposed TransPPG [81], a method for automatic rPPG feature representation on a 3D mask face that employs the ViT architecture to estimate the pulse signal from temporal signals extracted from frame differences and detect the liveness representation. TranRPPG obtains superior or state-of-the-art performance in 3D mask face presentation attack detection (PAD) by being more lightweight and efficient.

Gupta et al. proposed RADIANT [69] that combines Multilayer perceptron (MLP) and Transformer architecture. The MLP linear projection transforms temporal signals into signal embeddings in a novel manner to improve the rPPG feature representation and facilitate the learning of pertinent rPPG feature representations. The Transformer architecture is utilized for denoising and estimation of the cardiovascular pulse. Pre-training using synthetic temporal signals [82], large-scale dataset ImageNet and data augmentation were employed to address the problem of limited training data and under-fitting. On UBFC-rPPG and COHFACE, the model outperformed DeepPhys [61], HR CNN [60], META-rPPG [65] and other CNN-based architectures with lower RMSE at 4.52 and 10.12, respectively.

On the other hand, Yu et al. [70] proposed Physformer consisting of cascaded temporal difference transformer blocks to adaptively accumulate local and global spatio-temporal features to improve rPPG representation. The key highlight of this model is that it can be readily trained from scratch on public rPPG datasets, in contrast to the majority of transformer networks that require pretraining on large-scale datasets such as ImageNet. In addition, the authors propose a label distribution learning approach and a dynamic constraint in the frequency domain to provide informative supervision for PhysFormer and mitigate the problem of overfitting. PhysFormer achieved state-of-the-art performance for the intra-dataset testing on VIPL-HR, MAHNOB-HCI, and OBF without requiring heavy preprocessing procedures.

Furthermore, the model trained with VIPL-HR was cross-tested with MMSE-HR, PhysFormer outperformed traditional, non-end-to-end learning, and end-to-end learning-based methods by achieving the lowest SD of 5.22, MAE of 2.84, RMSE of 5.36, and the highest correlation coefficient at 0.92.

Yu et al. also introduced Physformer++ [71] utilizing a SlowFast temporal difference transformer with two pathways, and incorporates periodic- and cross-attention mechanisms. In contrast to PhysFormer, which only employs the slow pathway, Physformer++ uses both pathways to leverage temporal contextual and periodic rPPG clues from facial videos more effectively. In cross-dataset testing, PhysFormer++ has surpassed its predecessor and other methods that its predecessor surpassed. It achieved this with a lower standard deviation of 5.09, a mean absolute error of 2.71, a root mean square error of 5.15, and a higher correlation coefficient of 0.93.

3.8 Gaps & influencing challenges

Remote HR measurement has attracted considerable research interest since the introduction of rPPG. The COVID-19 pandemic has further stimulated the development of remote physiological monitoring techniques. Nevertheless, several challenges and limitations remain in this field. This section discusses the gaps and influencing challenges affecting the performance and applicability of remote HR measurement methods. Table 5 summarizes the identified gaps from state-of-the-art research for future research directions suggestion.

Firstly, the public dataset should be more comprehensive and diverse. Currently, different public datasets mainly focus on either motion artifact or illumination variation as the main challenge for rPPG. For instance, PURE is designed to evaluate the effect of head motion artifacts on rPPG, while COHFACE is used to assess the impact of illumination variation. Additionally, demographic diversity is an important factor. The majority of existing camera-based physiological measurement datasets have been collected in Europe, the United States, or China. Images of participants with paler skin tones predominate. The typical age of the subject is also youthful, with a narrow age range. In addition, other factors such as multiple object detection, camera-to-subject distance, environmental conditions, and camera specifications have not been emphasized and surmounted in the majority of studies. In supervised methods, this may result in a trained model with poor performance in real-world applications, as its performance is extremely dependent on the training dataset. To enhance the robustness of a model, it is necessary to propose a comprehensive, fair, well-balanced, and diverse dataset.

Table 5 Summary of identified gaps in state-of-the-art rPPG researches

Identified gaps	Description
Inadequate in terms of completeness and variety, public rPPG dataset	Limitations in trained model's performance due to insufficient variations reflect real-world scenarios
Limited ROI options and absence of research besides face region	Over-reliance on facial images in state-of-the-art methods
Lack of research related hardware and software configuration in this field	Dependence of performance on the choice of camera sensor and video compression technique
Insufficient research on rPPG data security and protection	Potential for misuse of rPPG data by unauthorized or malicious parties

The face is typically chosen as the ROI for rPPG HR estimation approaches due to its larger planar skin area and more visible blood capillaries, particularly in the cheeks and forehead regions. However, this technique may have limitations due to its reliance on facial features on the entire face region. In order to increase the number of possible application scenarios when only a portion of the face is accessible, such as when a face mask is worn, a method must be able to manage diverse input skin areas. For example, Zheng et al. [57] proposed a non-end-to-end CNN-based residual network model tailored to the case of face mask use. In DeeprPPG [83], a hybrid DL method with spatio-temporal convolutional networks was trained to accept video clips of various input skin regions as input for HR estimation.

From the hardware side, the contact devices for collecting ground truth label data can exhibit biases [84]. Dasari et al. [85] examined the estimation biases of rPPG methods across various demographics and found similar biases with contact-based devices and environmental conditions. In supervised methods, this may result in a trained model with poor performance in real-world applications, as its performance is extremely dependent on the training dataset.

Besides the bias in ground truth sensors, the cameras also reveal their limitations and biases as they vary widely in specifications. Cameras are designed with certain operating criteria and often optimized for lighter skin types. Sensitivity is usually highest in the middle of the camera's frequency range, and dark or very light skin types may saturate the pixels and lose physiological variations. Image sensor qualities and parameters, such as sensor type, color filter, frequency bands, bit depth, and pixels, affect the sensitivity of measurements. Other hardware factors, such as lens, aperture, shutter speed, and ISO, can also influence the image content. Camera internal software properties or controls, such as resolution, frame rate, and auto controls, impact image quality and may vary depending on the hardware and bandwidth. However, it is hard to characterize the impact of each parameter due to the many combinations.

Heart rate (HR) is an important physiological indicator that reflects an individual's health condition. Consequently, the majority of studies in this field concentrate on HR estimation. However, other vital signs are also important for detecting and monitoring a variety of diseases. For example, blood pressure (BP) can reveal cardiovascular diseases such as hypertension, whereas SpO2 can indicate cardio-respiratory health by revealing whether or not a person has sufficient oxygen. In fact, there are fewer studies on estimating BP and SpO2 remotely than HR. Therefore, there are many research opportunities to investigate the feasibility of estimating multiple vital signs at the same time. HR estimation alone is insufficient for clinical applications because estimating multiple vital signs simultaneously is more practical [86]. With the growing need for telemedicine and remote healthcare technologies [87], the ability to estimate and track multiple vital signs remotely can offer healthcare professionals a comprehensive understanding of their patient's health conditions, leading to better diagnosis, treatment, and disease management.

In addition to the previously mentioned gaps, data privacy is a primary concern for camera-based solutions. This is particularly true for rPPG studies, as they frequently use consumer-grade cameras for data collection and monitoring. These camera-based methods pose a risk to biometric data security, necessitating stringent privacy protocols to safeguard individual privacy. However, fewer studies are addressing this; there are lots of research opportunities to investigate the feasibility of security algorithms in rPPG. For example, in PulseEdit [88], a novel security algorithm capable of editing physiological signals to conceal a person's cardiac activity and physiological status without altering or distorting the original visual appearance was proposed to protect the user's physiological signal from being disclosed. Similarly, Sun et al. [89] proposed PrivacyPhys, a model that modifies rPPG in facial videos captured from

online video meetings or video-sharing platforms in order to safeguard against malicious capture and ensure privacy.

Back to the current development in rPPG, due to the nature of camera-based systems, the most critical and urgent issues in this field, motion artifact and illumination variation, are reduced to a degree by each proposed rPPG solution, but there are still a number of problems and challenges that could affect performance.

Video compression algorithms aim to reduce a video's bit rate while maintaining its fidelity, facilitating its transmission bandwidth and storage. Due to limited storage capacity, the majority of consumer-level cameras compress the recorded videos, as demonstrated by the datasets in Table 3. This compression process inevitably degrades video quality and introduces errors in extracting rPPG signals, particularly for conventional methods that heavily rely on the color information encoded in the original videos. Compression algorithms frequently disregard the subtle temporal variations in pixel values that reflect the PPG signal to save bits. For instance, Woyczyk et al. [90] reported that the efficacy of their model was severely impacted by video compression, specifically the MPEG-4 format. Mcduff et al. [91] found a linear correlation between the PPG SNR and constant rate compression factors. In contrast, Rapczynski et al. [92] observed that HR estimation was more resistant to variations in video resolution and color subsampling. In other words, the future dataset should contain multiple videos with varying compression levels to conduct robust analysis or train a robust model.

Consistent performance in different HR ranges has to be ensured. In Li et al. [93], with the first open challenge on remote physiological signal sensing, RePSS 2020, the top three teams all performed the best within 77 to 90 bpm, and the MAE values are significantly higher when tested on either higher or lower HR rate ranges. The inconsistencies showed that the state-of-the-art methods still require improvement to adapt to the broader bpm range. Therefore, it is still debatable whether good statistical indicators, including MAE, RMSE, SNR, and PCC, accurately reflect the model's high promise across the human HR range.

Additionally, data privacy could pose a challenge for rPPG research. Researchers must carefully consider the potential hazards of such testing and ensure that their studies are conducted with the subjects' privacy and ethical concerns in mind. This includes respecting the autonomy and dignity of research participants, obtaining informed consent before collecting rPPG data, safeguarding the confidentiality and security of the data, and ensuring that participants have the right to withdraw or access their data. However, conducting real-world tests with rPPG may present additional ethical challenges, especially when subjects are not informed consent or when testing is conducted in public or uncontrolled settings.

3.9 Potential applications

As rPPG prevents wearing discomfort, and skin allergies, integration into infant monitoring is possible and appropriate. Physiological monitoring, especially HR monitoring for a newborn in the neonatal intensive care unit (NICU), is a crucial practice to avoid bradycardia and apnea. To this end, Huang et al. [94] first proposed a newborn-optimized spatio-temporal neural network and dataset for HR monitoring to overcome newborns' side-face posture and limited valid ROI compared to adults. Villanroel et al. [86] conducted a clinical study to develop a multi-task algorithm that segments skin areas and estimates vital signals only when the infant is present in the view of the camera without clinical interventions. Neonates are generally less mobile than adults, and the hospital environment provides controlled illumination, such as when they are resting in a small incubator. This makes rPPG a prospective neonatal monitoring application. Nevertheless, the integration of multiple sensors and signal fusion

can overcome the limitations of camera-based measurements, such as when the face or body is partially covered.

Tran et al. [51] began a practical demonstration by deploying their rPPG system for monitoring the patient's HR at a hospital, which is a first step towards telemedicine and telehealth. In addition, the COVID-19 pandemic has brought telemedicine and telehealth to the forefront, particularly in situations where remote monitoring and diagnosis are required. Annis et al. [87] studied the feasibility and effectiveness of a remote patient monitoring (RPM) program for COVID-19 patients. The program included features such as remote monitoring of vital signs, symptom tracking, and two-way video conferencing. The study revealed a reduction in hospitalizations and visits to the emergency department, as well as an increase in patient satisfaction. Camera-based tools are considered to be the most convenient and user-friendly remote monitoring tools available to the general public [8]. However, one significant challenge in using these tools is the greater variability in illumination conditions, motion, and frame quality, compared to data capture performed in hospital settings with gold-standard sensors.

In addition to clinical applications, numerous studies, including [95–97], found that HR monitoring has benefits for sports activities in terms of preventing the emergence of a state of fatigue and improving cardiac regulation. This is because HR is one of the significant psycho-physiological variables that may determine sports performance. The idea of heart rate monitoring was first mooted in archery sport officially in the World Archery Cup back to the year 2012. A contact-type heart rate sensor that would be attached just below the archer's knee was developed and implemented in top archery tournaments [98]. After a decade, the non-contact camera-based HR monitoring solution was introduced publicly and used to broadcast the archers' heart rates in the top tournament, Tokyo Olympics 2020, on TV. The system consists of four sets of cameras installed 12 meters away from athletes and pattern-recognition software for tracking the face color changes and calculating the HR.

For entertainment purposes, the obtained HR changes during various phases of the competitions were broadcast concurrently so that the television audience could experience the stress and body's adrenaline rush [99]. More importantly, it is beneficial for the optimization of long-term sports performance. In Clemente et al. [100], it was proved any mismatch of physiological and psychophysical factors could deteriorate archers' accuracy and precision. It also concluded that experienced archers exhibit better accuracy and, at the same time, a lower heart rate compared to inexperienced archers because of the better arousal control. With this solution, the archer's physiological signals with associated performance under real stress can be precisely analyzed to outcome with targeted training for the archer's strength, endurance, and mental training.

Currently, wearable devices, such as smartwatches and smart bands, are very attractive for sports, fitness, training, and wellness. These contact PPG devices allow users to keep tracking their physiological conditions as prevention and avoid excessive training. However, the water and sweat deposited in the straps usually cause an allergic reaction in terms of rash and irritation on the skin. Thus, it makes it possible for the non-contact rPPG technique to take place and provide similar HR monitoring. In Wang et al. [101], the algorithm was proposed targeted for reducing motion artifacts in fitness training.

Using in-car cameras to monitor a driver's physiological state is a method to prevent and reduce traffic accidents brought on by human factors such as fatigue and drowsiness. The monitoring may detect abnormal vital signs and alert the driver immediately or use this information to stop a vehicle safely, thus helping to avoid tragedy and accidents. Several driver monitoring-focused approaches are proposed to detect fatigue and drowsiness. For instance, in Nowara et al. [102], the application of a NIR camera to deal with the abrupt

changes in the driver's face's illumination during the day and night. Du et al. [103] proposed a multimodal fusion technique integrating both eyelid features and rPPG extracted from RGB video for driving fatigue detection.

The pervasive use of smartphones with video recording capabilities enables the incorporation of rPPG techniques into non-invasive vital physiological signs assessment. For instance, MobilePhys [104] utilizes both front and rear cameras on a smartphone to generate high-quality self-supervised labels for training personalized contactless camera-based PPG models. This is because current state-of-the-art neural models are typically trained using high-quality videos with gold-standard physiological labels, which are not applicable to complex real-world contexts for mobile systems. However, the limited computational resource in end devices poses a challenge for rPPG methods. Casalino et al. [105] proposed a prototype of a client-server architecture mobile app that performs the signal and video processing on the server while the smartphone handles the video acquisition and the interaction with the user.

As a basic biometric authentication method, a conventional face recognition system is vulnerable to paper-attack and video replay attacks. The rPPG technique can serve as a face spoofing detector to improve the face recognition system. For example, Yao et al. [106] created a weighted spatial-temporal map by utilizing multiple smaller ROIs with varying weights. This approach was employed to simplify the decision-making process for a customized EfficientNet [107] model. Yu et al. [81] proposed TransrPPG to extract pulse signals from frame difference and render them as a 3D face mask for face liveness detection.

4 Conclusion

A comprehensive review of conventional computer vision and DL-based approaches for contactless heart rate estimations, its possible applications, and current research gaps are discussed in this paper. The conventional computer vision methods require pixel-based processing and filtering to retrieve ROIs in the face. Subsequently, these ROIs are processed with signal extraction sequentially to estimate HR. On the other hand, DL approaches are proven to direct estimate HR from images/video. It can efficiently reduce the noise caused by motion and illumination, thus improving the performance of HR measurement. Future research should prioritize addressing influencing challenges, such as illumination variations and motion artifacts, while considering other factors, such as consistent performance across the human heart rate (HR) range and the impact of video compression. The strong reliance on facial ROI indicates the lack of ROI selections in this field. In addition, the influence of hardware on efficacy has received little consideration. In addition to estimating HR, future studies should consider other vital signs, such as breathing rate variations.

To achieve this, realistic and diverse datasets comprising detailed information should be made available to enable the proper benchmarking of increasingly advanced deep learning (DL) techniques. However, research efforts should not be limited to the evaluation of dataset performance alone. Additional research should investigate potential clinical applications, including neonatal monitoring, telemedicine, and telehealth. Non-clinical applications, such as sports or fitness activity monitoring, in-car fatigue monitoring, face spoofing detection, and mobile-based applications, may be viewed as preliminary stages before more complex and strictly regulated clinical uses are implemented. However, it is essential to note that user privacy is always a concern in camera-based applications and research. Researchers should take precautions and conduct research to prevent the leakage of physiological information.

These recommendations can guide future research efforts to advance the field of non-contact HR estimation via camera-based analysis.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Gonzalez E, Peã R, Avila A, Munoz D (2018) Applications to improve the assistance of first aiders in outdoor scenarios, pp 175–196. https://doi.org/10.1016/B978-0-12-812130-6.00010-X
2. Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H (2018) A review on wearable photoplethysmography sensors and their potential future applications in health care. Int J Biosens Bioelectron 4(4):195–202. https://doi.org/10.15406/ijbsbe.2018.04.00125
3. Ashley EA, Niebauer J (2004) Cardiology explained. Remedica, http://www.ncbi.nlm.nih.gov/books/nbk2204
4. Johnston W, Mendelson Y (2005) Extracting heart rate variability from a wearable reflectance pulse oximeter. In: Proceedings of the IEEE 31st annual northeast bioengineering conference, 2005, pp 157–158. https://doi.org/10.1109/NEBC.2005.1431971
5. Swinehart DF (1962) The beer-lambert law. J Chem Educ 39(7):333. https://doi.org/10.1021/ED039P333
6. Watson AR, Wah R, Thamman R (2020) The value of remote monitoring for the covid-19 pandemic. Telemed e-Health 26(9):1110–1112. https://doi.org/10.1089/tmj.2020.0134
7. Tang Z, Hu H, Xu C, Zhao K (2021) Exploring an efficient remote biomedical signal monitoring framework for personal health in the covid-19 pandemic. Int J Environ Res Public Health 18(17):9037. https://doi.org/10.3390/ijerph18179037
8. Rohmetra H, Raghunath N, Narang P, Chamola V, Guizani M, Lakkaniga NR (2021) Ai-enabled remote monitoring of vital signs for covid-19: methods, prospects and challenges. Computing:1–27. https://doi.org/10.1007/s00607-021-00937-7
9. Verkruysse W, Svaasand LO, Nelson JS (2008) Remote plethysmographic imaging using ambient light. Opt Express 16(26):21434–21445. https://doi.org/10.1364/OE.16.021434
10. Fuertes D, del-Blanco CR, Carballeira P, Jaureguizar F, García N (2022) People detection with omnidirectional cameras using a spatial grid of deep learning foveatic classifiers. Digit Signal Process 126. https://doi.org/10.1016/j.dsp.2022.103473
11. Wu W, Wang T, Wang Z, Cheng L, Wu H (2022) Meta transfer learning-based super-resolution infrared imaging. Digit Signal Process:103730. https://doi.org/10.1016/j.dsp.2022.103730
12. Sheng J, Lv G, Wang Z, Feng Q (2022) Srnet: sparse representation-based network for image denoising. Digit Signal Process 130:103702. https://doi.org/10.1016/j.dsp.2022.103702
13. Poh M-Z, McDuff DJ, Picard RW (2010) Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt Express 18(10):10762–10774. https://doi.org/10.1364/OE.18.010762

14. Poh M-Z, McDuff DJ, Picard RW (2011) Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans Biomed Eng 58(1):7–11. https://doi.org/10.1109/TBME.2010.2086456
15. Lewandowska M, Rumiński J, Kocejko T, Nowak J (2011) Measuring pulse rate with a webcam — a non-contact method for evaluating cardiac activity. In: 2011 Federated conference on computer science and information systems (FedCSIS), pp 405–410
16. Wu H-Y, Rubinstein M, Shih E, Guttag J, Durand F, Freeman W (2012) Eulerian video magnification for revealing subtle changes in the world. ACM Trans Graph 31(4). https://doi.org/10.1145/2185520.2185561
17. Haan G, Jeanne V (2013) Robust pulse rate from chrominance-based rppg. IEEE Trans Biomed Eng 60(10):2878–2886. https://doi.org/10.1109/TBME.2013.2266196
18. McDuff D, Gontarek S, Picard RW (2014) Improvements in remote cardiopulmonary measurement using a five band digital camera. IEEE Trans Biomed Eng 61(10):2593–2601. https://doi.org/10.1109/TBME.2014.2323695
19. Lam A, Kuno Y (2015) Robust heart rate measurement from video using select random patches. In: 2015 IEEE international conference on computer vision (ICCV), pp 3640–3648. https://doi.org/10.1109/ICCV.2015.415
20. Wang W, Stuijk S, Haan G (2015) Exploiting spatial redundancy of image sensor for motion robust rppg. IEEE Trans Biomed Eng 62(2):415–425. https://doi.org/10.1109/TBME.2014.2356291
21. Feng L, Po L-M, Xu X, Li Y, Ma R (2015) Motion-resistant remote imaging photoplethysmography based on the optical properties of skin. IEEE Trans Circuits Syst Video Technol 25(5):879–891. https://doi.org/10.1109/TCSVT.2014.2364415
22. Li X, Chen J, Zhao G, Pietikäinen M (2014) Remote heart rate measurement from face videos under realistic situations. In: 2014 IEEE conference on computer vision and pattern recognition, pp 4264–4271. https://doi.org/10.1109/CVPR.2014.543
23. De Haan G, Van Leest A (2014) Improved motion robustness of remote-ppg by using the blood volume pulse signature. Physiol Meas 35(9):1913. https://doi.org/10.1088/0967-3334/35/9/1913
24. Wang W, Stuijk S, Haan G (2016) A novel algorithm for remote photoplethysmography: spatial subspace rotation. IEEE Trans Biomed Eng 63(9):1974–1984. https://doi.org/10.1109/TBME.2015.2508602
25. Wang W, Brinker AC, Stuijk S, Haan G (2017) Algorithmic principles of remote ppg. IEEE Trans Biomed Eng 64(7):1479–1491. https://doi.org/10.1109/TBME.2016.2609282
26. Fouad RM, Omer OA, Aly MH (2019) Optimizing remote photoplethysmography using adaptive skin segmentation for real-time heart rate monitoring. IEEE Access 7:76513–76528. https://doi.org/10.1109/ACCESS.2019.2922304
27. Macwan R, Benezeth Y, Mansouri A (2019) Heart rate estimation using remote photoplethysmography with multi-objective optimization. Biomed Signal Process Control 49:24–33. https://doi.org/10.1016/j.bspc.2018.10.012
28. Gudi A, Bittner M, Lochmans R, Gemert J (2019) Efficient real-time camera based estimation of heart rate and its variability. In: 2019 IEEE/CVF International conference on computer vision workshop (ICCVW), pp 1570–1579. https://doi.org/10.1109/ICCVW.2019.00196
29. Adelabu MA, Imoize AL, Adesoji KE (2022) Enhancement of a camera-based continuous heart rate measurement algorithm. SN Comput Sci 3(4):1–16. https://doi.org/10.1007/s42979-022-01179-w
30. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: International joint conference on artificial intelligence. http://andrewd.ces.clemson.edu/courses/cpsc482/papers/LK81_stereoRegistration.pdf
31. Kwon S, Kim J, Lee D, Park K (2015) Roi analysis for remote photoplethysmography on facial video. In: 2015 37th Annual international conference of the ieee engineering in medicine and biology society (EMBC), pp 4938–4941. https://doi.org/10.1109/EMBC.2015.7319499
32. Conaire CO, O'Connor NE, Smeaton AF (2007) Detector adaptation by maximising agreement between independent data sources. In: 2007 IEEE conference on computer vision and pattern recognition, pp 1–6. https://doi.org/10.1109/CVPR.2007.383448
33. Abdul Rahman NA, Wei KC, See J (2007) Rgb-h-cbcr skin colour model for human face detection. Faculty of Information Technology, Multimedia University 4
34. Wu Y, Ji Q (2019) Facial landmark detection: A literature survey. Int J Comput Vision 127(2):115–142. https://doi.org/10.1007/s11263-018-1097-z
35. Martinez B, Valstar MF, Binefa X, Pantic M (2013) Local evidence aggregation for regression-based facial point detection. IEEE Trans Pattern Anal Mach Intell 35(5):1149–1163. https://doi.org/10.1109/TPAMI.2012.205

36. Asthana A, Zafeiriou S, Cheng S, Pantic M (2013) Robust discriminative response map fitting with constrained local models. In: 2013 IEEE conference on computer vision and pattern recognition, pp 3444–3451. https://doi.org/10.1109/CVPR.2013.442
37. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(6):681–685. https://doi.org/10.1109/34.927467
38. Yu X, Huang J, Zhang S, Yan W, Metaxas DN (2013) Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: 2013 IEEE international conference on computer vision, pp 1944–1951. https://doi.org/10.1109/ICCV.2013.244
39. Po L-M, Feng L, Li Y, Xu X, Cheung TC-H, Cheung K-W (2018) Block-based adaptive roi for remote photoplethysmography. Multimed Tool Appl 77(6):6503–6529. https://doi.org/10.1007/s11042-017-4563-7
40. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (surf). Comput Vis Image Underst 110(3):346–359. https://doi.org/10.1016/j.cviu.2007.09.014
41. Bousefsaf F, Maaoui C, Pruski A (2013) Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. Biomed Signal Process Control 8(6):568–574. https://doi.org/10.1016/j.bspc.2013.05.010
42. Tarvainen MP, Ranta-aho PO, Karjalainen PA (2002) An advanced detrending method with application to hrv analysis. IEEE Trans Biomed Eng 49(2):172–175. https://doi.org/10.1109/10.979357
43. Wang W, Den Brinker AC, Stuijk S, De Haan G (2017) Amplitude-selective filtering for remote-ppg. Biomed Opt Express 8(3):1965–1980. https://doi.org/10.1364/BOE.8.001965
44. Yang Z, Yang X, Wu X (2019) Motion-tolerant heart rate estimation from face videos using derivative filter. Multimed Tool Appl 78(18):26747–26757. https://doi.org/10.1007/s11042-019-07849-x
45. Cardoso J-F (1999) High-order contrasts for independent component analysis. Neural Comput 11(1):157–192. https://doi.org/10.1162/089976699300016863
46. Hsu Y, Lin Y-L, Hsu W (2014) Learning-based heart rate detection from remote photoplethysmography features. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4433–4437. https://doi.org/10.1109/ICASSP.2014.6854440
47. McDuff D, Gontarek S, Picard RW (2014) Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera. IEEE Trans Biomed Eng 61(12):2948–2954. https://doi.org/10.1109/TBME.2014.2340991
48. Hsu G-S, Ambikapathi A, Chen M-S (2017) Deep learning with time-frequency representation for pulse estimation from facial videos. In: 2017 IEEE international joint conference on biometrics (IJCB), pp 383–389. https://doi.org/10.1109/BTAS.2017.8272721
49. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: single shot multibox detector. In: European conference on computer vision, Springer, pp 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
50. Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M (2013) 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: 2013 IEEE international conference on computer vision workshops, pp 397–403. https://doi.org/10.1109/ICCVW.2013.59
51. Tran Q-V, Su S-F, Sun W, Tran M-Q (2021) Adaptive pulsatile plane for robust noncontact heart rate monitoring. IEEE Trans Syst Man Cybern Syst 51(9):5587–5599. https://doi.org/10.1109/TSMC.2019.2957159
52. Perepelkina O, Artemyev M, Churikova M, Grinenko M (2020) Hearttrack: convolutional neural network for remote video-based heart rate monitoring. In: 2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 1163–1171. https://doi.org/10.1109/CVPRW50498.2020.00152
53. Song R, Chen H, Cheng J, Li C, Liu Y, Chen X (2021) Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. IEEE J Biomed Health Inform 25(5):1373–1384. https://doi.org/10.1109/JBHI.2021.3051176
54. Lu H, Han H, Zhou SK (2021) Dual-gan: joint bvp and noise modeling for remote physiological measurement. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12399–12408. https://doi.org/10.1109/CVPR46437.2021.01222
55. Wu B-F, Wu Y-C, Chou Y-W (2022) A compensation network with error mapping for robust remote photoplethysmography in noise-heavy conditions. IEEE Trans Instrum Meas 71:1–11. https://doi.org/10.1109/TIM.2022.3141149
56. Birla L, Puneet G (2021) And-rppg: A novel denoising-rppg network for improving remote heart rate estimation. Comput Biol Med 141:105146. https://doi.org/10.1016/j.compbiomed.2021.105146
57. Zheng K, Ci K, Li H, Shao L, Sun G, Liu J, Cui J (2022) Heart rate prediction from facial video with masks using eye location and corrected by convolutional neural networks. Biomed Signal Process Control 75:103609. https://doi.org/10.1016/j.bspc.2022.103609

58. Qiu Y, Liu Y, Arteaga-Falconi J, Dong H, Saddik AE (2019) Evm-cnn: Real-time contactless heart rate estimation from facial video. IEEE Transactions on Multimedia. 21(7):1778–1787. https://doi.org/10.1109/TMM.2018.2883866
59. Niu X, Shan S, Han H, Chen X (2020) Rhythmnet: end-to-end heart rate estimation from face via spatial-temporal representation. IEEE Trans Image Process 29:2409–2423. https://doi.org/10.1109/TIP.2019.2947204
60. Špetlík R, Franc V, Matas J (2018) Visual heart rate estimation with convolutional neural network. In: Proceedings of the british machine vision conference, Newcastle, UK, pp 3–6. http://bmvc2018.org/contents/papers/0271.pdf
61. Chen W, McDuff D (2018) Deepphys: video-based physiological measurement using convolutional attention networks. In: European conference on computer vision, Springer, pp 356–373. https://doi.org/10.1007/978-3-030-01216-8_22
62. Yu Z, Peng W, Li X, Hong X, Zhao G (2019) Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In: International conference on computer vision (ICCV). https://doi.org/10.1109/ICCV.2019.00024
63. Yu Z, Li X, Zhao G (2019) Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: 30th British machine vision conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019, p 277. BMVA Press. https://bmvc2019.org/wp-content/uploads/papers/0186-paper.pdf
64. Yu Z, Li X, Niu X, Shi J, Zhao G (2020) Autohr: a strong end-to-end baseline for remote heart rate measurement with neural searching. IEEE Signal Process Lett 27:1245–1249. https://doi.org/10.1109/LSP.2020.3007086
65. Lee E, Chen E, Lee C-Y (2020) Meta-rPPG: remote Heart Rate Estimation Using a Transductive Meta-learner, pp 392–409. https://doi.org/10.1007/978-3-030-58583-9_24
66. Yue Z, Ding S, Yang S, Yang H, Li Z, Zhang Y, Li Y (2021) Deep super-resolution network for rppg information recovery and noncontact heart rate estimation. IEEE Trans Instrum Meas 70:1–11. https://doi.org/10.1109/TIM.2021.3109398
67. Liu X, Fromm J, Patel S, McDuff D (2020) Multi-task temporal shift attention networks for on-device contactless vitals measurement. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33, pp 19400–19411. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2020/file/e1228be46de6a0234ac22ded31417bc7-Paper.pdf
68. Botina-Monsalve D, Benezeth Y, Miteran J (2022) Rtrppg: an ultra light 3dcnn for real-time remote photoplethysmography. In: 2022 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 2145–2153. https://doi.org/10.1109/CVPRW56347.2022.00233
69. Gupta, AK, Kumar R, Birla L, Gupta P (2023) Radiant: better rppg estimation using signal embeddings and transformer. In: 2023 IEEE/CVF winter conference on applications of computer vision (WACV), pp 4965–4975. https://doi.org/10.1109/WACV56688.2023.00495
70. Yu Z, Shen Y, Shi J, Zhao H, Torr P, Zhao G (2022) Physformer: facial video-based physiological measurement with temporal difference transformer. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4176–4186. https://doi.org/10.1109/CVPR52688.2022.00415
71. Yu Z, Shen Y, Shi J, Zhao H, Cui Y, Zhang J, Torr P, Zhao G (2023) Physformer++: facial video-based physiological measurement with slowfast temporal difference transformer. International Journal of Computer Vision. 131:1–24. https://doi.org/10.1007/s11263-023-01758-1
72. Lin J, Gan C, Han S (2019) Tsm: temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7083–7093. https://doi.org/10.1109/ICCV.2019.00718
73. Song R, Zhang S, Cheng J, Li C, Chen X (2019) New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2019.103535
74. McDuff D (2018) Deep super resolution for recovering physiological information from videos. In: 2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 1448–14487. https://doi.org/10.1109/CVPRW.2018.00185
75. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers), pp 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423
76. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461

77. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
78. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
79. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: International conference on learning representations. https://openreview.net/forum?id=YicbFdNTTy
80. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, pp 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
81. Yu Z, Li X, Wang P, Zhao G (2021) Transrppg: remote photoplethysmography transformer for 3d mask face presentation attack detection. IEEE Signal Process Lett 28:1290–1294. https://doi.org/10.1109/LSP.2021.3089908
82. Niu X, Han H, Shan S, Chen X (2018) Synrhythm: learning a deep heart rate estimator from general to specific. In: 2018 24th international conference on pattern recognition (ICPR), pp 3580–3585. https://doi.org/10.1109/ICPR.2018.8546321
83. Liu S-Q, Yuen PC (2020) A general remote photoplethysmography estimator with spatiotemporal convolutional network. In: 2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020), pp 481–488. https://doi.org/10.1109/FG47880.2020.00109
84. Bickler P, Feiner J, Severinghaus J (2005) Effects of skin pigmentation on pulse oximeter accuracy at low saturation. Anesthesiology 102:715–719. https://doi.org/10.1097/00000542-200504000-00004
85. Dasari A, Arul Prakash SK, Jeni L, Tucker C (2021) Evaluation of biases in remote photoplethysmography methods. npj Digit Med 4. https://doi.org/10.1038/s41746-021-00462-z
86. Villarroel M, Chaichulee S, Jorge J, Davis S, Green G, Arteta C, Zisserman A, McCormick K, Watkinson P, Tarassenko L (2019) Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit. npj Digit Med 2:128. https://doi.org/10.1038/s41746-019-0199-5
87. Annis T, Pleasants S, Hultman G, Lindemann E, Thompson J, Billecke S, Badlani S, Melton G (2020) Rapid implementation of a covid-19 remote patient monitoring program. Journal of the American Medical Informatics Association: JAMIA 27. https://doi.org/10.1093/jamia/ocaa097
88. Chen M, Liao X, Wu M (2022) Pulseedit: editing physiological signals in facial videos for privacy protection. IEEE Trans Inf Forensics Secur 17:457–471. https://doi.org/10.1109/TIFS.2022.3142993
89. Sun Z, Li X (2022) Privacy-phys: facial video-based physiological modification for privacy protection. IEEE Signal Process Lett 29:1507–1511. https://doi.org/10.1109/LSP.2022.3185964
90. Woyczyk A, Fleischhauer V, Zaunseder S (2021) Adaptive gaussian mixture model driven level set segmentation for remote pulse rate detection. IEEE J Biomed Health Inform 25(5):1361–1372. https://doi.org/10.1109/JBHI.2021.3054779
91. McDuff DJ, Blackford EB, Estepp JR (2017) The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), pp 63–70. https://doi.org/10.1109/FG.2017.17
92. Rapczynski M, Werner P, Al-Hamadi A (2019) Effects of video encoding on camera-based heart rate estimation. IEEE Trans Biomed Eng 66(12):3360–3370. https://doi.org/10.1109/TBME.2019.2904326
93. Li X, Han H, Lu H, Niu X, Yu Z, Dantcheva A, Zhao G, Shan S (2020) The 1st challenge on remote physiological signal sensing (repss). In: 2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 1274–1281. https://doi.org/10.1109/CVPRW50498.2020.00165
94. Huang B, Chen W, Lin C-L, Juang C-F, Xing Y, Wang Y, Wang J (2021) A neonatal dataset and benchmark for non-contact neonatal heart rate monitoring based on spatio-temporal neural networks. Eng Appl Artif Intell 106:104447. https://doi.org/10.1016/j.engappai.2021.104447
95. Jiménez-Morgan S, Molina Mora J (2017) Effect of heart rate variability biofeedback on sport performance, a systematic review. Appl Psychophysiology Biofeedback 42:1–11. https://doi.org/10.1007/s10484-017-9364-2
96. Bricout V-A, DeChenaud S, Favre-Juvin A (2010) Analyses of heart rate variability in young soccer players: the effects of sport activity. Auton Neurosci 154(1–2):112–116. https://doi.org/10.1016/j.autneu.2009.12.001
97. Mahmood, NH, Uyop N, Zulkarnain N, Che Harun FK, Kamarudin MF, Linoby A (2011) Led indicator for heart rate monitoring system in sport application. In: 2011 IEEE 7th international colloquium on signal processing and its applications, pp 64–66. https://doi.org/10.1109/CSPA.2011.5759843
98. Archery W (2012) Revealing the secrets of an archer's body for the world to see. https://worldarchery.sport/news/93733/revealing-secrets-archers-body-world-see, Accessed 28 Sept 2022

99. W C (2021) Archery debuts heart-rate graphics on broadcast of the Olympic Games. https://worldarchery.sport/news/200395/archery-debuts-heart-rate-graphics-broadcast-olympic-games, Accessed 28 Sept 2022
100. Clemente F, Couceiro M, Rocha R, Mendes R (2011) Study of the heart rate and accuracy performance of archers, vol 11, pp 434–437. http://efsupit.ro/images/stories/imgs/JPES/2011/12/10Art_66.pdf
101. Wang W, Brinker A, Stuijk S, Haan G (2017) Robust heart rate from fitness videos. Physiological Meas 38:1023–1044. https://doi.org/10.1088/1361-6579/aa6d02
102. Nowara EM, Marks TK, Mansour H, Veeraraghavan A (2018) Sparseppg: towards driver monitoring using camera-based vital signs estimation in near-infrared. In: 2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 1353–135309. https://doi.org/10.1109/CVPRW.2018.00174
103. Du G, Zhang L, Su K, Wang X, Teng S, Liu PX (2022) A multimodal fusion fatigue driving detection method based on heart rate and perclos. IEEE Trans Intell Transp Syst 23(11):21810–21820. https://doi.org/10.1109/TITS.2022.3176973
104. Liu X, Wang Y, Xie S, Zhang X, Ma Z, McDuff D, Patel S (2022) Mobilephys: personalized mobile camera-based contactless physiological sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 6(1) https://doi.org/10.1145/3517225
105. Casalino G, Castellano G, Nisio A, Pasquadibisceglie V, Zaza G (2022) A mobile app for contactless measurement of vital signs through remote photoplethysmography. In: 2022 IEEE international conference on systems, man, and cybernetics (SMC), pp 2675–2680. https://doi.org/10.1109/SMC53654.2022.9945406
106. Yao C, Wang S, Zhang J, He W, Du H, Ren J, Bai R, Liu J (2021) rppg-based spoofing detection for face mask attack using efficientnet on weighted spatial-temporal representation. In: 2021 IEEE international conference on image processing (ICIP), pp 3872–3876. https://doi.org/10.1109/ICIP42928.2021.9506276
107. Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning. Proceedings of Machine Learning Research, vol 97, pp 6105–6114. PMLR. https://proceedings.mlr.press/v97/tan19a.html