Deep learning based remote-photoplethysmography measurement from short-time facial video

Bin Li[1], Wei Jiang[1], Jinye Peng[1]*, Xiaobai Li[2].

[1] School of Information Science and Technology, Northwest University, Xi'an, China

[2] Center for Machine Vision and Signal Analysis, University of Oulu, Oulu

E-mail: `pjy@nwu.edu.cn`

Keywords:Remote photoplethysmography, heart rate estimation,spatial-temporal convolution, fusion attention mechanism, short-time monitoring

Abstract. Objective: Efficient non-contact heart rate (HR) measurement from facial video has received much attention in health monitoring. Past methods relied on prior knowledge and unproven hypothesis to extract rPPG signals, e.g., manually designed regions of interest (ROIs) and skin reflection model. Approach: This paper presents a short-time end-to-end HR estimation framework based on facial features and temporal relationships of video frames. In the proposed method, a deep 3D multi-scale network with cross-layer residual structure is designed to construct an autoencoder and extract robust remote photoplethysmography (rPPG) features. Then, a spatial-temporal fusion mechanism is proposed to help the network focus on features related to rPPG signals. Both shallow and fused 3D spatial-temporal features are distilled to suppress redundant information in the complex environment. Finally, a data augmentation strategy is presented to solve the problem of uneven distribution of HR in existing datasets. Main results: The experimental results on four face-rPPG datasets show that our method overperforms the state-of-the-art methods and requires fewer video frames. Compared with the previous best results, the proposed method improves the RMSE by 5.9% , 3.4% 21.4% on the OBF dataset (intra-test), COHFACE dataset (intra-test) and UBFC dataset (cross-test), respectively. Significance: Our method achieves good results on diverse datasets (i.e., highly compressed video, low-resolution and illumination variation), demonstrating that our method can extract stable rPPG signals in short time.

1. Introduction

Heart rate (HR) is a vital physiological parameter for humans. Accurate and fast HR measurement method conduces to efficient health surveillance and disease diagnosis, which has become an important research topic in many applications, including health monitoring [1], emotion recognition [2], 3D mask face attack detection [3], physiological signal privacy protection [4].

The HR monitoring method in medical treatment is based on contact devices, including skin-contact ECG sensors and pulse oximetry PPG optical device. PPG

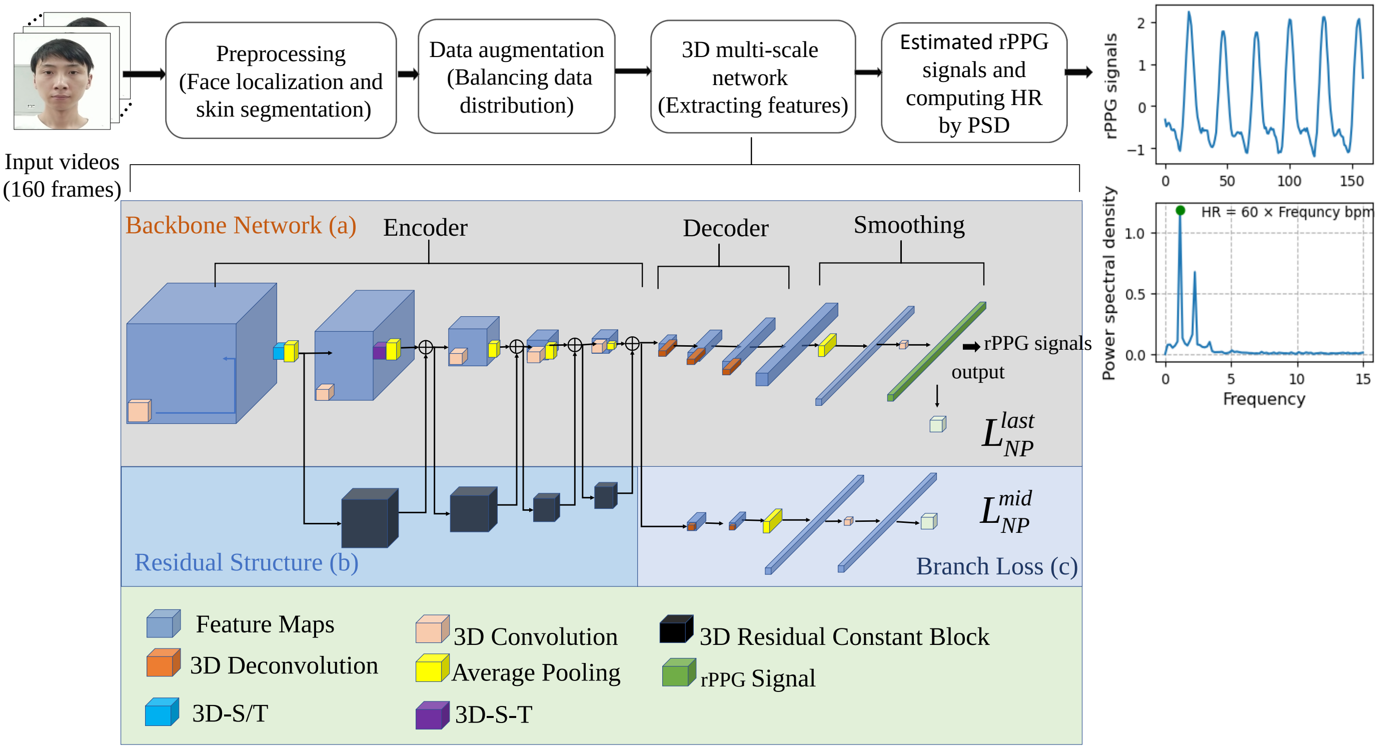


Figure 1: The overall process of heart rate measurement. Among them,3D multi-scale network is used to extract features related rPPG signals, including (a) (b) (c) three parts, (a) is considered as backbone network for feature extraction, (b) (c) as auxiliary structure, (b) is used to transfer residual features and (c) is a branch after the encoder of (a) for supervision.

is a method based on light absorption changes caused by blood flow during cardiac cycles [5]. These changes affect the scattering and reflection of the skin to light, which ultimately leads to subtle changes in skin color [6]. Combining the imaging and PPG techniques and apply it to face images, remote photoplethysmography (rPPG) is another promising method that remotely extracts pulse wave from the periodic fluctuations of blood volume [7]. However, rPPG signals are easily disturbed by external noise(e.g., low video resolution, head movement and illuminance change), so it is necessary to improve signal processing algorithms.

Recently, many significant solutions for non-contact HR measurement based on facial videos have been proposed [8–11], which extract features associated with heart rate signals based on facial videos to achieve non-contact heart rate monitoring. Compared with the traditional photoplethysmography (PPG) and electrocardiography (ECG) measurements which need to be in contact with the human body for a long time, the non-contact physiological measurement methods based on facial video are more convenient and user-friendly. However, the efficiency and accuracy of those measurement algorithms are not satisfactory, due to low video resolution, head movement, illuminance change, facial expression and insufficient diversity of existing datasets.

Existing methods for non-contact HR measurement from facial videos can be divided into the following methods, including the method based on blind source signal separation(BBS), knowledge-based approaches and deep neural network (DNN). The traditional methods use the color space transformation and signal decomposition to calculate the HR value, which rely on hand-designed features and specific environment settings (e.g., bright light source and non-rigid facial motion.) [12–17]. knowledge-based approaches conduct signal decomposition or linear combination to recover the rPPG signal by pre-assumptions and prior knowledge [18–20]. The DNN based algorithms are dedicated to establish a network to extract features from face region and calculate HR [21–26]. These deep learning approaches do not require complex pre-processing steps, and with the expansion of datasets and improvements, which can drive non-contact healthcare applications and low-cost easy-to-use diagnostics [27].

In this paper, we focus on the method based on DNN. To achieve an accurate and practical HR measurement method, the following issues should be considered. First, in order to make non-contact HR measurement more practical, the algorithm should make use of shorter video frame sequence. Existing non-contact heart rate measurement methods usually take longer periods of time, e.g., 10 s [28–31] or 30 s [15, 22–24, 32–34], they need to capture more video sequence for computing and the subject should maintain in longer time. In addition, the deep learning approach that requires capturing longer video frames has high time delay if applied to real-time heart rate monitoring. Second, when the duration of the captured face video becomes shorter, fewer complete signal cycles can be extracted and corresponding temporal relationships that can be used become weaker. Third, the input video sequence contains foreground (face region) and background, how to build an HR measurement model without complex pre-processing, e.g., face feature points localization, ROI segmentation, etc., is also an issue to be considered. In addition, although masks or ROIs set on the cheeks or forehead can be effective in making the model focus on areas with more relevant information, it can be misleading in scenes like a masked forehead or a single-sided light source on face.

Regarding the issues above, most existing DNN models are difficult to measure the HR accurately and effectively from a short-time facial video. Therefore, a deep 3D spatial-temporal convolution network with multi-scale features fusion is presented, which can accurately and efficiently measure HR in short-time facial video without additional information (e.g., facial mask, ROI detection, etc.). First, an encoder-decoder network structure is employed to obtain the multi-scale facial features in the video. Second, a spatial-temporal fusion attention mechanism is presented to guide the network focus on those features that are highly related to physiological signals and strengthens the spatial-temporal association of each frame. Third, a 3D residual constant block is designed to transmit the physiological characteristics during the network training. Finally, data augmentation strategy is proposed to deal with the problem of uneven HR distribution in the existing datasets.

The proposed end-to-end network structure is shown in Figure 1. First, the facial region of each frame is generated based on the face position of the first frame. Then, a skin segmentation algorithm is used to preserve the skin region. Next, those regions are fed into the proposed network through a data augmentation strategy and generate rPPG signals. Finally, the mean heart rate is calculated from the recovered rPPG signals by the power spectral density (PSD) method.

The main contributions of our study are as follows:

1. An efficient end-to-end spatial-temporal multiscale network is proposed, which is able to reconstruct rPPG signals and measure HR from only 5s facial video.

2. A novel 3D fusion attention mechanism is proposed, which aggregates both spatial-temporal and multiscale features, and helps the network focus on the features of interest in each frame and entire video sequence.

3. A new dataset PHY-100 with measured Blood Volume Pulse (BVP) and synchronized facial video is presented, which contains 100 videos from 100 subjects. The experiments are conducted on four datasets to verify the effectiveness of the proposed network.

The rest of this paper is organized as follows. Section 2 reviews video-based remote physiological measurement algorithms in recent years. In Section 3, the proposed method is described in detail. The experimental results and discussion are presented in section Section 4 and Section 5, respectively. Finally, Section 6 summarizes the study.

2. Related Work

In this section, we briefly introduce the methods used in the research on HR measurement and illustrate the feasibility of some data augmentation strategies in the rPPG measurement task.

2.1. Traditional methods on rPPG measurement

Verkruysse et al. [12] discussed the PPG signals can be remotely measured using an RGB camera under natural light for the first time. This conclusion inspired subsequent studies in the field of rPPG. Previous traditional methods have focused on analyzing small color transformations of RIOs for rPPG measurements. Blind source separation (BSS) techniques were used to separate features in early research on non-contact cardiac pulse measurement. Furthermore, Lewandowska et al. [35] proposed the principal component analysis to selects the most likely component to identify the underlying PPG signal, which can effectively improve the signal-to-noise rate(SNR) of the PPG signal obtained from facial videos. Ming-Zher et al. [13] utilized independent component analysis (ICA) to remove the noise unrelated to the rPPG information. However, these methods do not consider the inevitable impact of illumination variation and head movement. De Haan et al. [18] first proposed a skin optical model for a motion scene, and linearly combined RGB channels were used to calculate chrominance signals to reduce motion noise. Subsequently, many improved methods were proposed. The pixel-wise calculation of chrominance features improves the motion robustness of the rPPG [36]. Tulyakov et al. [16] created a matrix compensation framework to dynamically select suitable ROIs to recover a more robust rPPG signal. 2SR [37] used spatial subspace rotation to reduce the impact of head movement. Wang et al. [19] designed a projection plane orthogonal to the skin-tone for rPPG measurements. In addition, Balakrishnan et al. [20] proposed a novel motion-based method that tracks multiple landmark points to obtain pulse signals from the head movement caused by cardiovascular circulation. However, the subject needs to remain stationary with no additional movement.

Traditional methods focus on handicrafted features and customized ROIs of skin with intense blood flow, but lose the physiological information in other facial skin that are equally important. Furthermore, most of the methods rely on prior knowledge, unproven assumptions, and fixed scene settings, these methods tend to target a specific class of problems, so their robustness is not high when applied in real-world scenarios.

2.2. Deep learning methods on rPPG measurement

With the expansion of the scale and diversity of datasets, more data-driven methods are used to explore the effectiveness of remote HR monitoring.

Hsu et al. [9] first used a deep learning framework for real-time HR prediction. The generated time-domain map was fed into the VGG-16 model to evaluate multiple HRs, and the final prediction value was obtained through majority voting. Spetlk et al. [21] proposed a two-stage network based on convolution: the extractor was used to extract potential rPPG components, and then the predictor was used to predict the HR (HR-CNN). A novel spatial-temporal representation map built by the ROIs to map pattern change was proposed by Niu et al. [31], they constructed a general-to-specific model by using large-scale synthetic rhythm signals to enhance generality [24]. Yu et al. [25] discussed the ability of various spatial-temporal convolution structures to recover timing-related rPPG signals, including RNN-LSTM and 3D convolution (physnet). Recently, some programs have been proposed for different issues. Yu et al. [34] proposed a video enhancement network that can effectively improve the information loss from data compression (rPPGnet). Lee et al. [33] proposed a transductive metalearner with adaptive weight adjustment to apply real deployment. Hu et al. [30] proposed an effective time-domain attention network to reduce the long-term noise (ETA-rPPGNet). Lokendra et al. [38] initiate the utilization of Action Units (AUs) for denoising temporal signals by facial expression.

These methods intend to design a reasonable end-to-end network structure and recover physiological parameters in real scenarios by strengthening the temporal correlation and reducing redundant information.

2.3. Data augmentation strategies attempted on rPPG tasks

In addition, data-driven strategies rely on diverse and large-scale datasets; it is necessary to construct a suitable data augmentation method for the current HR prediction task.

Image-based data augmentation strategies are summarized in [39], including geometric transformation, random cropping, color space conversion, and random mask blocks.

These methods are applied to some tasks (e.g., image classification, medical image analysis) to improve generality, but they focus on detailed texture information instead of the shallow information of skin color changes in the current task. Heart rate prediction differs from other logistic regression video tasks (e.g., pose classification), the prediction results obtained by simply selecting facial video subsequences may be inconsistent with the original sequence. Niu et al. [28] designed a data augmentation strategy to change the distribution of HR on the original dataset based on a spatial-temporal map, but did not provide a detailed algorithm and verify its feasibility.

3. Method

3.1. Data augmentation strategy

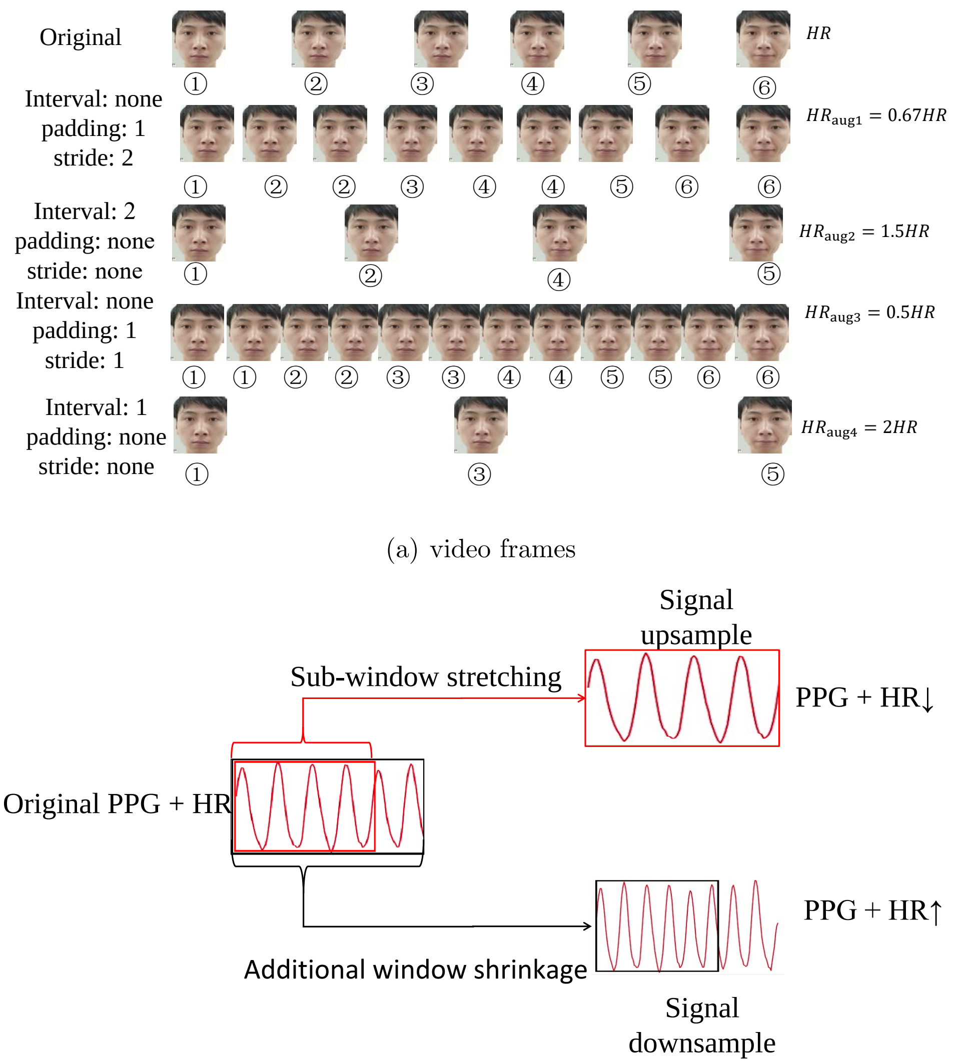
In the data-driven method, the data distribution affects the performance of the network[40, 41]. Especially in the practical application of HR measurement, the HR of most people distribute in a narrow range which leads to an uneven distribution in existing datasets and make the network produce boundary effects (i.e., a certain frequency domain range has a small amount of data in the training set, which will cause a large amount of deviation when testing in this range data). To solve this problem, a data augmentation strategy based on sampling operation is proposed.

Video downsampling strategy is used to expand the upper limit of HR in the datasets. For example, the 1.5T(T represents a length of time) frames sequence correspond to an average HR of 60 bpm, which is downsampled to 1T frames when the interval is set to 2. Therefore, the sequence length is reduced and the corresponding HR is 1.5× the original ($60 \cdot 1.5 = 90$ bpm). Upsampling strategy is similarly used to extend the low limit of the HR, and the video sequences are augmented by adjacent frames interpolation. Such as the 0.5T frames image sequence correspond to an average HR of 120 bpm are upsampled to 1T frames, the adjacent images are selected to extend the sequence by setting the stride and padding (e.g., the number of stride and padding are both set to 1). Hence the HR is reduced to 0.5× the original HR (60 bpm).

The HR transformation can be expressed by the following formula.

$$\begin{aligned} &HR_{aug} = HR \times \frac{interval + 1}{interval} \times \frac{padding}{padding + stride} \\ &interval, padding, stride \in N^{+}. \end{aligned} \tag{1}$$

where HR_{aug} represents the heart rate which is generated by data augmentation strategy from the original HR. *interval* is the successive interval size of a video sequence. *stride* and *padding* are the number of steps and padding in upsampling. A detailed example is presented in Figure 2(a). In the training, the PPG signal corresponding to each image is changed equally. A simple signal sampling strategy can be seen in



(a) video frames

(b) PPG signals

Figure 2: The detailed processing of data augmentation strategy on video frames and PPG signals. In (a), the HR value and video frames are adjusted by different parameters. The corresponding signal sampling process is shown in (b). In the process of upsampling, the sub-window of the original signal is selected to stretch to the original length and the heart rate decrease. While downsampling, the adjacent additional window is selected to shrink to the original length and the heart rate increase.

Figure 2(b), where the changed PPG signal maintains the same length as the original signal after the sampling strategy, and the corresponding waveform period and HR is changed.

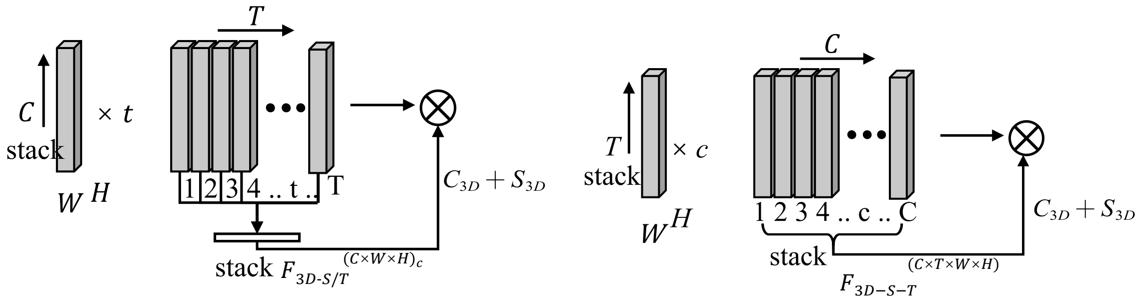
Table 1: The structure of the backbone network. ConvB_i represents a convolution block composed of multiple 3D convolution filters and an average pooling layer, where ConvB_1 and ConvB_2-5 contain one and two 3D convolutions, respectively. DConv_x denotes the 3D deconvolution filter. SGAP means a spatial global average pooling and then the spatial dimension is removed by squeeze operation, and Smooth1d signifies a convolution block composed of multiple 1D convolution filters. Each convolutional layer has the BatchNorm layer (instance normalization). ReLU, ELU is used as the activation function of ConvB_i and DConv_x individually. The size of feature maps was reduced one by one after ConvB_1,2,3,4,5, which is regarded as an encoder, while in DConv_1,2,3, temporal dimension is restored as a decoder end.

	Layername	*Outputsize*	*Kernelsize*
Encoder	ConvB_1	$16 \times T \times \frac{W}{2} \times \frac{H}{2}$	1×5×5
	ConvB_2	$16 \times \frac{T}{2} \times \frac{W}{4} \times \frac{H}{4}$	[3×3×3]×2
	ConvB_3	$64 \times \frac{T}{4} \times \frac{W}{8} \times \frac{H}{8}$	[3×3×3]×2
	ConvB_4	$64 \times \frac{T}{8} \times \frac{W}{16} \times \frac{H}{16}$	[3×3×3]×2
	ConvB_5	$128 \times \frac{T}{8} \times \frac{W}{32} \times \frac{H}{32}$	[3×3×3]×2
Decoder	DConv_1	$128 \times \frac{T}{4} \times \frac{W}{32} \times \frac{H}{32}$	4×1×1
	DConv_2	$128 \times \frac{T}{2} \times \frac{W}{32} \times \frac{H}{32}$	4×1×1
	DConv_3	$64 \times T \times \frac{W}{32} \times \frac{H}{32}$	4×1×1
Smoothing	SGAP	$64 \times T \times 1 \times 1$	None
	Smooth1d	$32 \times T \rightarrow 1 \times \ T$	5/3

3.2. Network Architecture

In order to strengthen the temporal and spatial correlation of the extracted features from video sequence, the encoder-decoder network structure is built based on 3D convolutions, which is shown in Figure 1: Backbone Network (a).

The encoder structure contains five 3D convolution blocks $ConvB\_i$ ($i \in [1, 5]$). The convolution block $ConvB\_1$ includes one 3D convolution layer and each $ConvB\_2 - 5$ contains two 3D convolution layers, which is shown in detail in table 1. At the last layer of $ConvB\_i$, the generated 3D feature maps are downsampled by average pooling to reduce the redundant information ($T \downarrow \frac{1}{2}, W \downarrow \frac{1}{2}, H \downarrow \frac{1}{2}$, where T is frame number, W and H are the width and height of the feature map). The 3D residual constant block contains a 3D convolution with kernel size 2 and stride 2, which is used to transfer loss information during feature reduction as shown in Figure 1: Residual Structure (b). The decoder structure includes three deconvolutionx. Due to the rPPG signal is only determined by the temporal characteristics of shallow color transformation, so the generated temporal feature maps restore to the original length of the video sequence through multiple deconvolution operation ($T \uparrow 2$), and the spatial texture features finally are converted to the related rPPG signals through adaptive average pooling and dimension compression operations. Finally, two one-dimensional convolutions filters



(a) 3D-S/T

(b) 3D-S-T

Figure 3: Two different temporal attention processes, (a) 3D-S/T separates the temporal dimensions and performs operations on the feature maps of each frame. (b) 3D-S-T processes the entire spatial-temporal 3D feature maps as a unit.

with kernel sizes of 5 and 3 are used to improve the accuracy of the generated waveform.

3.3. Spatial-temporal fusion attention mechanism

The proposed encoder structure generates facial features at different spatial and temporal dimensions. We need to figure out which information are more meaningful for the rPPG signal reconstruction. Therefore, two attention models are built to strengthen the correlation between the generated spatial-temporal 3D features F_{3D} as shown in Figure 3(a) and Figure 3(b). 3D-S/T attention model: the generated 2D spatial features $F_{2D}^{(W\times H)_t}$ of each frame $t \in T$ are stacked on channels C to form 3D spatial sequence $F_{3D-S/T}^{(C\times W\times H)_t}$. Then, the 3D channel attention (C_{3D}) and 3D spatial attention (S_{3D}) is preformed in each $F_{3D-S/T}^{(C\times W\times H)_t}$ respectively. 3D-S-T attention model: the generated 2D features $F_{2D}^{(W\times H)_c}$ of each channel $c \in C$ are stacked on frame sequence T to form 3D temporal feature maps $F_{3D-S-T}^{(T\times W\times H)_c}$. Then, the 3D channel attention (C_{3D}) and 3D spatial attention (S_{3D}) is preformed in entire $F_{3D-S-T}^{(C\times T\times W\times H)}$. Finally, the generated feature maps through the C_{3D} and S_{3D} are multiplied by the input feature maps in order.

C_{3D}: the feature maps through global max pooling and average pooling to generate $FM(\cdot)$ and $FA(\cdot)$. Then those feature maps are fed into a two layer structure which is composed of one convolution layer with kernel size 1 and ReLU. Finally, an element-wise operation is performed and the weight matrix of each feature map is obtained through the sigmoid function.

S_{3D}: focuses on features that are highly correlated with rPPG signal in spatial dimensions. The pooling operation of S_{3D} is similar with C_{3D}, and the produced features are fused at the channel dimensions and transformed into one channel.

The 3D-S/T strengthens the correlation of channel features generated by each in the spatial dimension, while 3D-S-T extract the feature relations in the temporal dimension. The 3D-S/T is added after the first convolution block $ConvB\_1$ to focus on shallow features from each frame, and the 3D-S-T is used after the $ConvB\_2$.

3.4. Model training settings

To restore the accurate rPPG signal and measure HR from facial video, we chose BVP signals collected from finger skin as the ground truth to guide the model. Moreover, negative Pearson (NP) is used to learn the trend and peak position of the BVP signal to minimize the linear correlation error, the NP function L_{NP} is defined as

$$L_{NP} = 1 - \frac{T\sum_{i=1}^{T} x_i y_i - \sum_{i=1}^{T} x_i \sum_{i=1}^{T} y_i}{\sqrt{\left(T\sum_{i=1}^{T} x_i^2 - \left(\sum_{i=1}^{T} x_i\right)^2\right)\left(T\sum_{i=1}^{T} (y_i)^2 - \left(\sum_{i=1}^{T} y_i\right)^2\right)}}. \tag{2}$$

where T denotes the length of the video sequence, and x and y represent the ground-truth BVP signal and the predicted rPPG signal.

Due to the deep layers of decoder and smoothing, there are fluctuations in parameter tuning during training. Branching loss is introduced as an auxiliary structure in the training phase, which is set at the end of the encoder to help adjust the model parameters and thus improve the stability of the model. The predicted rPPG signals which are produced by middle stage are used as L_{NP}^{mid} loss. The output of the decoder structure is regarded as the L_{NP}^{last} loss of the proposed model, which can be seen in Figure 1: Branch Loss (c). The final loss function is defined as

$$L_{final} = \gamma L_{NP}^{mid} + (1-\gamma) L_{NP}^{last}. \tag{3}$$

where $\gamma \in [0, 1]$, the γ is a balance parameter.

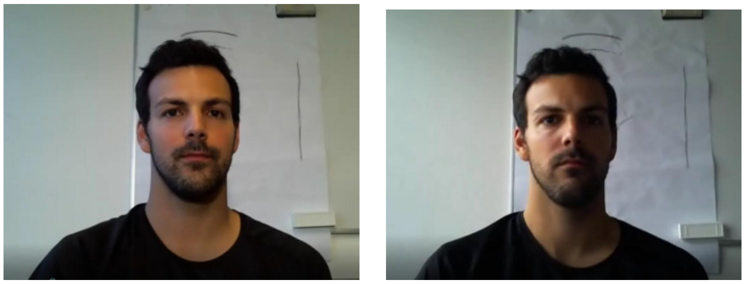
4. Experiments and results

We evaluate the performance of the proposed network on four dataset: OBF [42], UBFC [43], COHFACE [44], PHY-100.

4.1. Datasets

4.1.1. OBF The OBF dataset [42] contains 200 high-quality face RGB videos collected from 100 subjects(61%M, 39%F). The video is captured by an HD camera with a resolution of 1920 × 1080PX^2, 60 fps and video length is 5 minutes. Each subject has videos of exercise and calm scenes. The ECG, respiration rate and BVP signals are collected using Biosignal NX-series sensors simultaneously.

4.1.2. UBFC UBFC dataset [43] consists of 43 uncompressed face videos from 43 subjects. The videos are captured by a low-cost webcam with a resolution of 640 × 480PX^2, the frame rate is 30 fps and video length is 1 minute. The subjects are asked to play time-sensitive digital games. The HR and BVP signals are synchronously obtained using the Medical CMS50E finger clip sensor.



(a) Studio Light (b) Natural Light

Figure 4: The example of illumination variation on the COHFACE dataset.

4.1.3. COHFACE COHFACE dataset [44] consists of 160 highly compressed videos from 40 subjects (70%M, 30%F). Videos of each subject are recorded in two different lighting environments, each subject are recorded two videos for each scenario with a resolution of $640 \times 480\text{PX}^2$, the frame rate of video is 20 fps with a length of about 1 minute. 1. Studio light: the subject's face has bright illumination; 2. Natural light: one-sided natural light is used from an open window; the facial contour of each subject can only be clearly visible on the side of the light source. The sample of the different scenarios is shown in Figure 4. The BVP signal and respiration rate are synchronously obtained by Thought Technologies (BVP model SA9308M, belt model SA9311M).

4.1.4. PHY-100 Although the existing datasets have considered the simultaneous capture of face video and physiological data, the frame rate of the video capture devices is not sufficiently stable, which will lead to shift in the face video and physiological signals and make data alignment difficult. Therefore, in order to comprehensively verify the performance of the proposed method, we built our own dataset with high quality face video and physiological signals. Our dataset contains 100 face videos from 100 participants(60%M, 40%F) named PHY-100. The video is captured by an Blackmagic G2 with resolution $1920 \times 1080\text{PX}^2$, 30 fps, each video length is 1 minute. The physiological data is synchronously obtained from BIOPAC MP160, including respiration rate, blood oxygen, ECG and BVP.

4.2. Implementation details

4.2.1. Preprocessing For each frame in the video sequence, the face detector ‡ is used to locate the rough face position of the of first frame. Then, a skin segmentation algorithm § is used to preserve the skin region for better removal of background interference (i.e., skin pixels that are closely related to rPPG signal are preserved in face image, while non-skin pixels, such as the obscured areas of hair and clothing, and the background behind the person, are removed.), and the detection region is resized to $100 \times 100\text{PX}^2$.

‡ https://github.com/seetaface/SeetaFaceEngine

§ https://github.com/nasir6/rPPG

The physiological signal is synchronised with video frames by downsampling (one video frame corresponds to one normalized BVP signal). The PPG signals are filtered by a low-pass filter with 5.5hz to eliminate redundant interference in PHY-100 and OBF datasets. While the PPG signals in UBFC and COHFACE datesets have been precessed, so we use the label directly as the true value.

4.2.2. Training settings We select 160 frames (approximately 5 s) as the input to the network and the HR is calculated from the the generated rPPG signals through PSD algorithm. The evaluation of our proposed network is carried on an Intel(R) Core(TM) i7-9700 computer with 32 GB RAM and NVIDIA GeForce RTX 2080Ti GPU using PyTorch. Adam is used as the optimizer with an initial learning rate 1e-4, the maximum training epochs and batchsize are set to 30 and 8 respectively, and the hyperparameter γ is 0.2.

4.2.3. Performance Metrics The MAE(Mean Absolute error), RMSE(Root Mean Square Error) and r(Pearson correlation coefficient) are used to measure the estimation accuracy of the rPPG signals and HR. MAE and RMSE represent the measurement errors of HR, the r indicates the correlation between predicted rPPG signal and ground truth.

4.3. Intra-dataset Testing

We first performed intra-dataset testing on two public datasets (OBF and COHFACE) and compare the performance with the state-of-the-art methods, the dataset is divided into a train set, validation set, and test set in the ratio of 7:1:2. Then we validated the data augment strategy on our private dataset (PHY-100), data is divided into high and low-frequency data close to a 1:1 ratio and cross-test on different frequency data.

4.3.1. OBF The OBF dataset contains two types of subjects (rest and exercise), which have good lighting conditions and clear facial contour. The experimental results are presented in Table 2. The combination of six different structures are compared to verify the performance of the proposed network, including backbone net (bn), residual structure (rs), branch loss (bl), 3D-S/T and 3D-S-T, and the fusion attention mechanism (fa). The *bn* structure performs well compared with others, which illustrates the design of the feature scale transformation is effective for timing-related rPPG feature extraction. With the continuous optimization of the auxiliary structure, the measurement accuracy is improved gradually. To further demonstrate the performance of the proposed network, the quantitative comparison with the state-of-the-art network, i.e., ROI_green [42], CHROM [18], Physnet [25], POS [19], rPPGnet [34], CVD [32] are listed in Table 2. Our network achieves HR$_{mae}$ 0.742 and HR$_{rmse}$ 1.19, which over performs other methods.

Table 2: The HR estimation results of different models on OBF dataset.

Methods	HR_{mae}	HR_{rmse}	r
ROI_green[42]	-	2.162	0.99
CHROM[18]	-	2.773	0.98
Physnet[25]	-	1.812	0.992
POS[19]	-	1.906	0,991
rPPGNet[34]	-	1.8	0.992
CVD[32]	-	1.26	0.996
bn	0.98	1.77	0.992
bn+rs	0.95	1.57	0.994
bn+rs+bl	0.936	1.47	0.994
bn+rs+bl+3D-S/T	**0.821**	**1.24**	**0.997**
bn+rs+bl+3D-S-T	0.825	1.36	0.997
bn+rs+bl+fa	**0.742**	**1.19**	**0.998**

Table 3: The HR estimation results of different models on COHFACE dataset.

Method	HR_{mae}	HR_{rmse}	r
HR-CNN[21]	8.1	10.78	0.29
CHROM[18]	7.8	12.45	0.26
Li2014[15]	19.98	25.59	-0.44
2SR[37]	20.98	25.84	-0.32
DeeprPPG[45]	3.07	7.06	0.86
HU-Convlstm[29]	7.31	11.88	0.36
POS[19]	11.43	17.05	-4.47
ICA[13]	12.24	15.67	0.24
physnet[25]	8.59	11.6	0.36
DeepPhys[22]	6.89	13.89	0.34
ETA-rPPGNet[30]	4.67	6.65	0.77
AND-rPPG [38]	3.82	5.10	0.79
Ours	**2.788**	**4.926**	**0.892**

4.3.2. COHFACE Due to data compression and dramatic changes in illumination, some physiological signals are missing, which make the accurate face ROIs localization more difficult. Therefore, the performance of the existing methods on COHFACE dataset is not very good. In order to improve the accuracy of the proposed network, the data with good illumination is selected for pre-training, and then all the training data are loaded into the model for optimization. As shown in Table 3, our model achieves the best performance compared with existing methods. The experimental results show that the proposed multiscale spatial-temporal convolution network has a good ability to adapt complex scenes, and the proposed attention method improves the feature extraction capability of the network.

Table 4: The performance of our method on the collected dataset, where "high" and "low" represents high-frequency and low-frequency data, respectively. "x_adjust" means the training set x is roughly adjusted to the same distribution at the test set, and "x→y" denotes model trains on x and tests on y; Finally, "(x)" represents the HR distribution graph of the training set is shown on Figure 5(x) and the prediction comparison image corresponds to Figure 6(x).

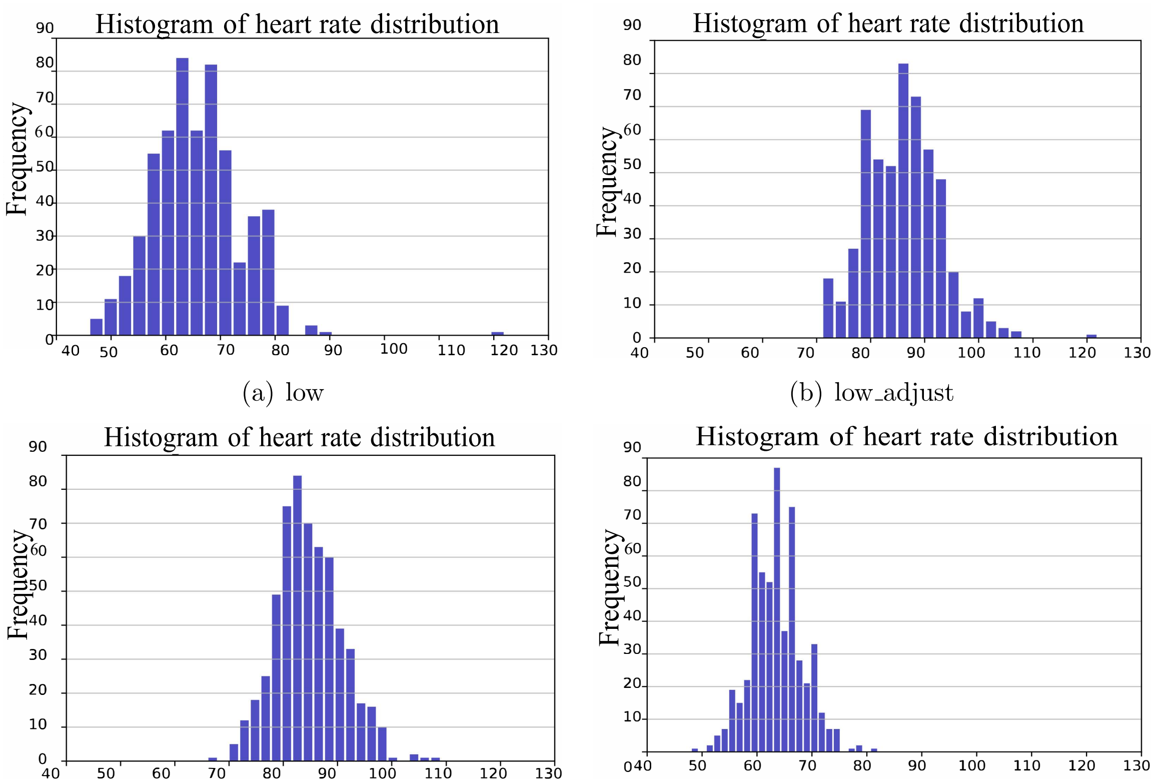
Cross-test	HR_{mae}	HR_{rmse}	r	Prec-3	Prec-5
(a)low→high	3.53	7.68	0.491	0.823	0.878
(b)low_adjust→high	**1.75**↓	**3.02**↓	**0.886**↑	**0.900**↑	**0.960**↑
(c)high→low	5.19	11.37	0.34	0.767	0.835
(d)high_adjust→low	**2.42**↓	**5.03**↓	**0.796**↑	**0.847**↑	**0.918**↑

4.3.3. PHY-100 To intuitively verify that proposed data augmentation strategy can solve the boundary effect effectively, the dataset is divided into high and low frequency by threshold TR, and TR is determined by the average HR of subjects in the first training batch. E.g., if $TR = 80$, the HR higher than 80 bpm is regarded as high-frequency data, the opposite is low-frequency data. The original data distribution is shown in Figure 5(a) and (c). We first use the high and low frequency data to train and cross-test. Then, the data augmentation strategy is used to adjust the HR distribution in the training set. The adjusted data is shown in Figure 5(b) and (d). The cross-test results obtained before and after changing the distribution are listed in Table 4. The result illustrates that the performance of the network has been greatly improved after data augmentation strategy. This is due to the proposed method learns skin color variation at different frequency domains, the changes of HR distribution helps improve the network performance. Another visualized result is shown in Figure 6, which further illustrates that the prediction errors of the un-adjusted distribution are larger than adjusted. We also perform data augmentation strategy on three public datasets to homogenize the HR distribution of the training set. The results on OBF dataset (intra-test) are not significantly due to the large data scale. The measurement accuracy is improved on smaller COHFACE(intra-test) dataset and crpss-database PHY-100(test in UBFC)datasets. The RMSE of COHFACE is 4.52 and UBFC is 3.26, which indicates the effectiveness of the data augmentation strategy.

In addition, we notice that the difference between the front and rear frames is increased when using the down-sampling strategy, which make the data-driven network model enhance the adaptability to a large range of head motion.

4.4. Cross-dataset Testing

We conduct a cross-dataset testing on the small-scale physiological dataset for evaluating the robustness of model. The model is trained on our private dataset (PHY-100) and tested on the UBFC.



(a) low

(b) low_adjust

(c) high

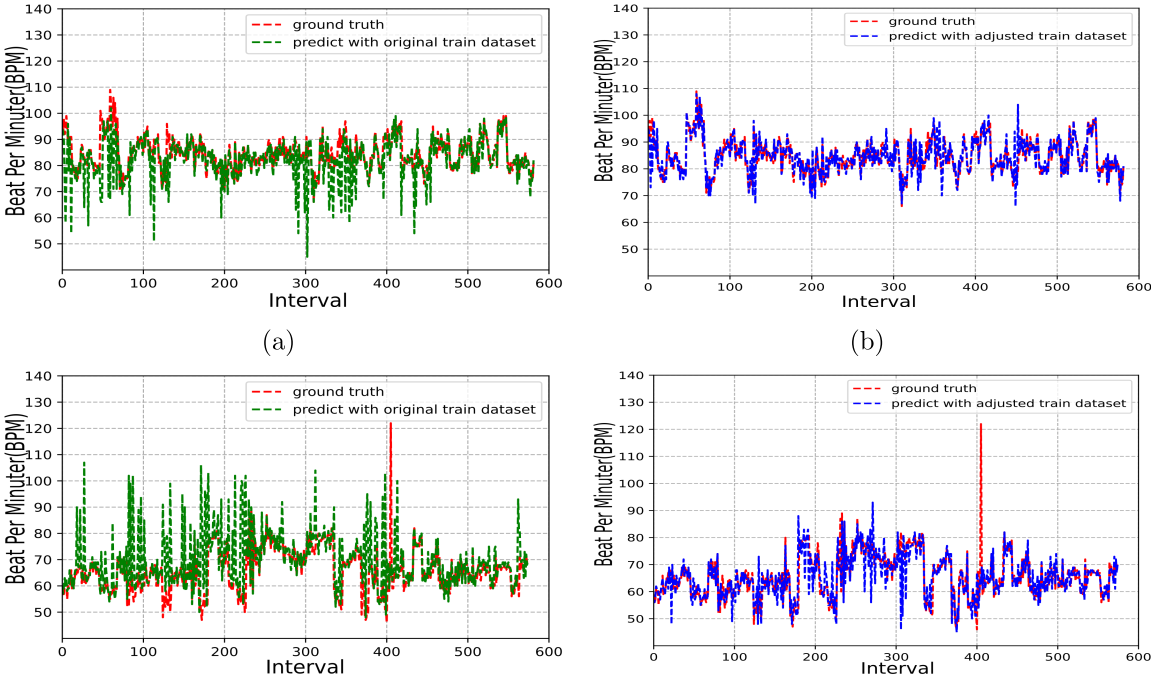
(d) high_adjust

Figure 5: The frequency of HR on separated and adjusted data, where (a)(c) is separated data and (b)(d) is adjusted data whose HR distribution is similar to (a)(c), respectively.

4.4.1. UBFC UBFC is a challenging public dataset and the HR of each subject changes significantly in the time-sensitive game scene. Table 5 presents the results of the proposed method compared with the existing method. The proposed method achieves the best performance and far surpassed the existing methods, which the RMSE is 3.62, MAE is 1.97, and r is 0.95. Compared to other methods that require face ROI detection for each frame, our method build a smooth feature associations between adjacent frames based on the face region in the first frame, which achieves higher processing speed and measurement accuracy. In addition, the proposed network only needs 5s video, which has more training batches to optimize parameters and obtain a robust generalization model in train set. The experiment results demonstrate that our method is more adaptable in small-scale datasets.

5. Discussion

Our main work is to design a robust method for short-time remote heart rate prediction. The performance of the three attention strategies on the challenging COHFACE dataset is compared in Section 5.1. From Figure 7, 3D-S/T + 3D-S-T can better adapt to scenes with dim light sources. In Section 5.2, the performance of the proposed method



(a)

(b)

(c)

(d)

Figure 6: Comparison of ground truth HR and predicted HR on the cross-test of the original and adjusted data, (a)(c) represents model trained on original data, (b)(d) corresponds to train on adjusted data.

Table 5: The cross-database HR estimation results of different models on UBFC dataset.

Method	HR$_{mae}$	HR$_{rmse}$	r
ICA[13]	3.51	8.64	0.91
CHROM[18]	3.44	4.61	0.97
POS[19]	2.44	6.61	0.94
ROI_green[42]	10.2	20.6	0.41
3DCNN[46]	8.35	10.45	0.76
Phynet[25]	3.63	5.29	0.92
rPPGnet[34]	3.89	6.75	0.89
CVD[32]	4.27	8.17	0.79
Ours	**1.97**	**3.62**	**0.95**

is discussed at different time lengths and compared with other methods, our method is more stable as Table 6. The effect of different signal preprocessing strategies on model training is discussed in Section 5.3. As shown in Table 8, the prediction performance is better with the use of low-pass(3.0hz). In Section 5.4, the limitations of the current work and the goals of future work are briefly described.

5.1. Spatial-Temporal attention

To verify the fusion attention mechanism can better improve the performance of the network, we experiment on the challenging COHFACE dataset with a complex lighting scene by using the following three strategies: 1. without attention, 2. 3D-S/T, 3. 3D-S/T + 3D-S-T.

The comparison between measured HR and ground truth is shown in Figure 7. The Bland-Altman plots are continuously centralized and scatter plots gradually become closer to a linear relationship from (a) to (c). The experiment results show that 3D-S/T strengthen the correlation of shallow features and performance of the proposed network is improved. 3D-S/T + 3D-S-T combines the two characteristics and correlates shallow spatial features with temporal features, illustrating the effectiveness of the fusion attention mechanism.

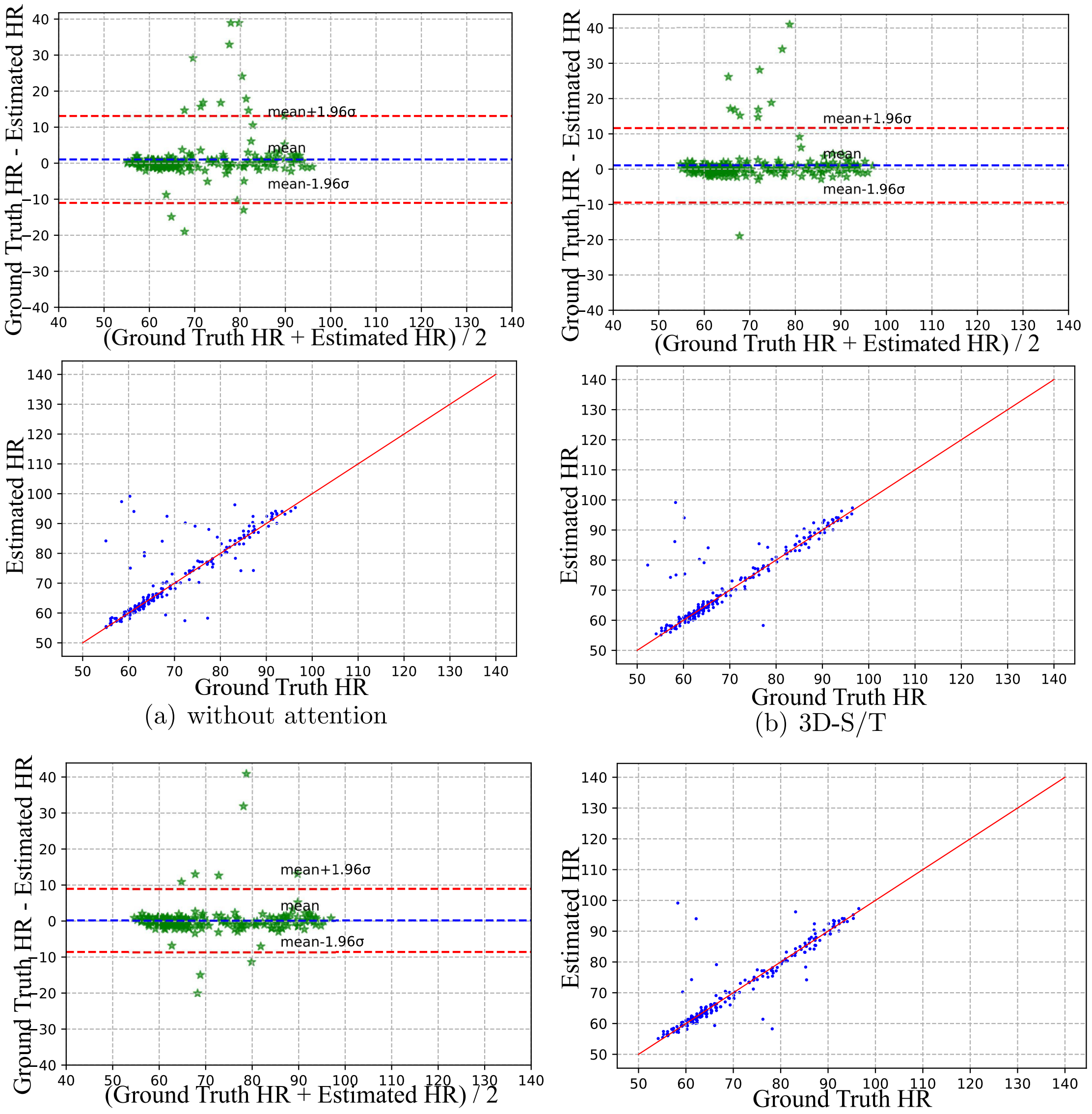
5.2. Short-Time HR Estimation

To fairly compare the performance of the network in short-time HR estimation, the proposed and other existing methods are carried out on the UBFC dataset by using three different non-overlapping time windows(i.e., 4s 6s 8s). Proportion of absolute deviation in 3 (Prec-3) or 5 (Prec-5) bpm is used, which show the performance of the proposed model in different time. The experimental results are presented in Table 6. Our method achieves the best results for three different time windows.

In addition, the accuracy of HR measurement decrease dramatically in all three methods when the time window below 4s. This is because the measurement time window become shorter, rPPG signal and HR are more sensitive to interference information. The rPPG fitting accuracy of our method is higher than others in different time window, which indicates the proposed method has greater robustness for rPPG signals reconstruction and HR measurement. Considering model parameters, accuracy and user experience, we finally chose 160 frames (approximately 5s) for rPPG and HR measurements.

5.3. Effect of signal pre-processing

In the current heart rate prediction task, the main focus is on the periodic relationship of the extracted signals, but other important indicators in medical diagnosis are neglected, such as the location of peak points associated with heart rate variability, and the morphological features of signals potentially associated with continuous blood pressure signals (systolic peak and diastolic peak) [47]. In this regard, we set up three different signal preprocessing strategies in the same experimental setting (PHY-100 dataset intra-testing) to discuss their effects on training time and prediction accuracy. The three strategies are (a) raw signal, (b) low-pass (5.5hz) and (c) low-pass (3.0hz), the corresponding PPG signal are shown in Figure 8. The corresponding experimental results are shown in Table 7. As we can see from the table, it is very difficult for the



(a) without attention

(b) 3D-S/T

(c) 3D-S/T + 3D-S-T

Figure 7: Bland-Altman plots and scatter plot on COHFACE datasets.

model to extract information from the face that is consistent with the (a) raw PPG signal, and it takes a long time and obtains the worst results. In addition, (c) low-pass (3.0hz) its filters out the features of diastolic peak, the learning process of the model is fast, and although it is effective for the heart rate prediction task, other features are lost. Finally, (b) low-pass (5.5hz) filters out the interference information and retains the features of diastolic peak, so its training time is slightly slower and its loss is larger, but it has a more detailed morphology and the best accuracy that will help in future studies on heart rate variability and other related parameters.

Table 6: Short-term HR estimation intra-test results: comparison on UBFC dataset.

(a) window size:4s

Method	HR_{mae}	HR_{rmse}	r	Prec-3	Prec-5
Physnet[25]	1.84	4.70	0.91	0.870	0.912
rPPGnet[34]	1.80	4.59	0.91	0.844	0.922
Ours	1.72	3.84	0.92	0.894	0.941

(b) window size:6s

Method	HR_{mae}	HR_{rmse}	r	Prec-3	Prec-5
Physnet[25]	1.73	2.82	0.95	0.912	0.943
rPPGnet[34]	1.68	3.90	0.95	0.908	0.938
Ours	1.34	1.70	0.97	0.94	0.975

(c) window size:8s

Method	HR_{mae}	HR_{rmse}	r	Prec-3	Prec-5
Physnet[25]	1.45	2.42	0.96	0.928	0.952
rPPGnet[34]	1.48	2.50	0.97	0.934	0.961
Ours	0.95	1.68	0.99	0.956	0.989

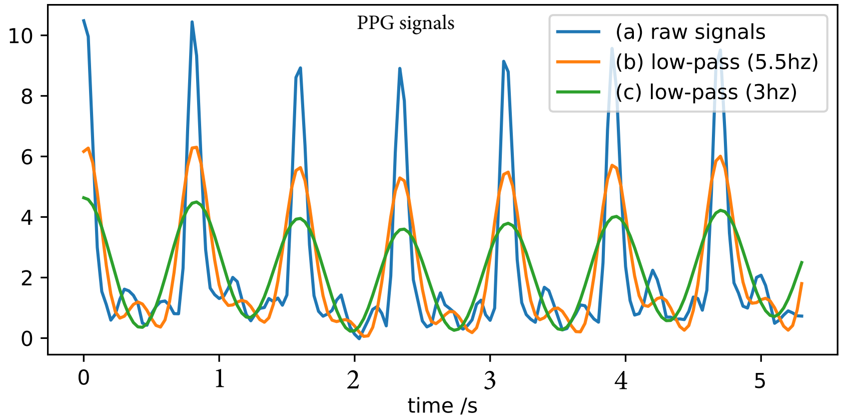


Figure 8: PPG signals image with different pre-processing strategies

Table 7: The training time and prediction accuracy of different pre-processing strategies on PHY-100 dataset.

pre-processing strategies	Time /hour	L_{NP}^{last}	HR_{mae}	HR_{rmse}
(a) raw signal	3.9	0.735	4.47	17.9
(b) low-pass (5.5hz)	2.9	0.75	1.21	3.89
(c) low-pass (3.0hz)	2.5	0.66	1.07	2.96

5.4. Limitations and future works

Unlike traditional medical heart rate monitoring methods (e.g., PPG or ECG) or wearable devices (e.g., health watches), rPPG technology is a non-contact heart rate measurement method. For other wireless (e.g. wifi or radar), rPPG technology is a convenient and low cost measurement method. Although rPPG technology is maturing in heart rate monitoring with increasing accuracy, there is a lack of methods to effectively recover and construct rPPG morphological features to further analyze respiration rate and oxyhemoglobin saturation. The proposed method can be tried in the future to construct new loss functions to recover more detailed morphological features. Most of the current non-contact physiological signal detection focuses on heart rate detection, including some research on heart rate variability, but the rPPG signal reflects the human cardiovascular activity, in addition to blood pressure, respiration and other important human health indicators, through real-time observation of these indicators can also provide real-time feedback on the human health state, the full use of these indicators is also one of the future development directions of non-contact physiological signal detection.

6. Conclusion

In this paper, we propose an end-to-end non-contact physiological signal measurement method, which can accurately measure rPPG signal and HR from only 5s video. The experimental results on four face-rPPG datasets show that our method overperforms the state-of-the-art methods and require less video frames. The innovation of our proposed network are: (1) a spatial-temporal fusion attention mechanism is proposed to enhance the relationship between spatial information and contextual semantics. The experimental results show that our approach achieves good performance on publicly available datasets and can be adapted to smaller-scale data and complex lighting scenarios; (2) The residual structure and branch loss are introduced as auxiliary modules to transfer the lost information in scale transformation and retain sufficient features; (3) a data augmentation strategy based on sampling is designed to change the HR values to balance data distribution in different frequency domain ranges. The improvements brought by proposed structures have been proved by comprehensive experiments on four datasets.

7. Acknowledgment

The author's have confirmed that any identifiable participants in this study have given their consent for publication. The dataset collection has been approved by the local ethics committee (Shaanxi Provincial People's Hospital), and all the subjects signed an informed consent form. This work was supported by the China Postdoctoral Science Foundation (Program No. 2020M683696XB); The Xi'an Key Laboratory of Intelligent

Perception and Cultural Inheritance (No. 2019219614SYS011CG033); Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2021JQ-455).

8. Reference

[1] A. Hammer, M. Scherpf, M. Schmidt, H. Ernst, H. Malberg, K. Matschke, A. Dragu, J. Martin, O. Bota, Camera-based assessment of cutaneous perfusion strength in a clinical setting, Physiological Measurement 43 (2) (2022) 025007.

[2] G. Du, S. Long, H. Yuan, Non-contact emotion recognition combining heart rate and facial expression for interactive gaming environments, IEEE Access 8 (2020) 11896–11906.

[3] Z. Yu, X. Li, P. Wang, G. Zhao, Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection, IEEE Signal Processing Letters 28 (2021) 1290–1294.

[4] M. Chen, X. Liao, M. Wu, Pulseedit: Editing physiological signals in facial videos for privacy protection, IEEE Transactions on Information Forensics and Security (2022).

[5] J. Allen, D. Kulin, S. Zanelli, M. Bachler, K. Pilt, B. Paliakaite, P. Charlton, J. Allen, D. Kulin, S. Zanelli, Assessing hemodynamics from the photoplethysmogram to gain insights into vascular age: A review from vascagenet, American Journal of Physiology (322-4 Pt.2) (2022).

[6] J. Allen, Photoplethysmography and its application in clinical physiological measurement, Physiological measurement 28 (3) (2007) R1.

[7] H. Liu, J. Allen, D. Zheng, F. Chen, Recent development of respiratory rate measurement technologies, Physiological Measurement 40 (7) (2019) 07TR01–.

[8] R. Macwan, Y. Benezeth, A. Mansouri, Heart rate estimation using remote photoplethysmography with multi-objective optimization, Biomedical Signal Processing and Control 49 (2019) 24–33.

[9] G.-S. Hsu, A. Ambikapathi, M.-S. Chen, Deep learning with time-frequency representation for pulse estimation from facial videos, in: 2017 IEEE international joint conference on biometrics (IJCB), IEEE, 2017, pp. 383–389.

[10] Y. Zhao, B. Zou, F. Yang, L. Lu, A. N. Belkacem, C. Chen, Video-based physiological measurement using 3d central difference convolution attention network, in: 2021 IEEE International Joint Conference on Biometrics (IJCB), 2021.

[11] H. Lu, H. Han, S. K. Zhou, Dual-gan: Joint bvp and noise modeling for remote physiological measurement, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[12] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light, Optics Express 16 (26) (2008) 21434–45.

[13] M.-Z. Poh, D. J. McDuff, R. W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation, Optics express 18 (10) (2010) 10762–10774.

[14] L. Qi, H. Yu, L. Xu, R. S. Mpanda, S. E. Greenwald, Robust heart-rate estimation from facial videos using projectica, Physiological Measurement 40 (8) (2019) 085007.

[15] X. Li, J. Chen, G. Zhao, M. Pietikainen, Remote heart rate measurement from face videos under realistic situations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 4264–4271.

[16] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, N. Sebe, Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2396–2404.

[17] M. Das, T. Choudhary, M. Bhuyan, L. Sharma, Non-contact heart rate measurement from facial video data using a 2d-vmd scheme, IEEE Sensors Journal (2022).

[18] De, Haan, Gerard, Jeanne, Vincent, Robust pulse rate from chrominance-based rppg., IEEE Transactions on Biomedical Engineering (2013).

[19] W. Wang, A. C. den Brinker, S. Stuijk, G. De Haan, Algorithmic principles of remote ppg, IEEE Transactions on Biomedical Engineering 64 (7) (2016) 1479–1491.

[20] G. Balakrishnan, F. Durand, J. Guttag, Detecting pulse from head motions in video, in:

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3430–3437.

[21] R. Špetlík, V. Franc, J. Matas, Visual heart rate estimation with convolutional neural network, in: Proceedings of the british machine vision conference, Newcastle, UK, 2018, pp. 3–6.

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

[23] Z.-K. Wang, Y. Kao, C.-T. Hsu, Vision-based heart rate estimation via a two-stream cnn, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 3327–3331.

[24] X. Niu, H. Han, S. Shan, X. Chen, Synrhythm: Learning a deep heart rate estimator from general to specific, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 3580–3585.

[25] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, arXiv preprint arXiv:1905.02419 (2019).

[26] B. Huang, C.-L. Lin, W. Chen, C.-F. Juang, X. Wu, A novel one-stage framework for visual pulse rate estimation using deep neural networks, Biomedical Signal Processing and Control 66 (2021) 102387.

[27] J. Allen, H. Liu, S. Iqbal, D. Zheng, G. Stansby, Deep learning based photoplethysmography classification for peripheral arterial disease detection: a proof-of-concept study, Physiological Measurement 42 (5) (2021) –.

[28] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, X. Chen, Robust remote heart rate estimation from face utilizing spatial-temporal attention, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–8.

[29] M. Hu, D. Guo, X. Wang, P. Ge, Q. Chu, A novel spatial-temporal convolutional neural network for remote photoplethysmography, in: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2019.

[30] M. Hu, F. Qian, D. Guo, X. Wang, L. He, F. Ren, Eta-rppgnet: Effective time-domain attention network for remote heart rate measurement, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–12.

[31] X. Niu, S. Shan, H. Han, X. Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, IEEE Transactions on Image Processing 29 (2019) 2409–2423.

[32] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, G. Zhao, Video-based remote physiological measurement via cross-verified feature disentangling, in: European Conference on Computer Vision, Springer, 2020, pp. 295–310.

[33] E. Lee, E. Chen, C.-Y. Lee, Meta-rppg: Remote heart rate estimation using a transductive meta-learner, in: European Conference on Computer Vision, Springer, 2020, pp. 392–409.

[34] Z. Yu, W. Peng, X. Li, X. Hong, G. Zhao, Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 151–160.

[35] M. Lewandowska, J. Ruminski, T. Kocejko, J. Nowak, Measuring pulse rate with a webcam - a non-contact method for evaluating cardiac activity, in: Federated Conference on Computer Science and Information Systems - FedCSIS 2011, Szczecin, Poland, 18-21 September 2011, Proceedings, 2011.

[36] W. Wang, S. Stuijk, G. De Haan, Exploiting spatial redundancy of image sensor for motion robust rppg, IEEE transactions on Biomedical Engineering 62 (2) (2014) 415–425.

[37] W. Wang, S. Stuijk, G. De Haan, A novel algorithm for remote photoplethysmography: Spatial subspace rotation, IEEE transactions on biomedical engineering 63 (9) (2015) 1974–1984.

[38] B. Lokendra, G. Puneet, And-rppg: A novel denoising-rppg network for improving remote heart rate estimation, Computers in biology and medicine 141 (2022) 105146.

[39] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (1) (2019) 1–48.

[40] G. Batista, R. C. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, Acm Sigkdd Explorations Newsletter 6 (1) (2004) 20–29.

[41] H. He, Y. Bai, E. Garcia, S. A. Li, adaptive synthetic sampling approach for imbalanced learning. ieee international joint conference on neural networks, in: 2008 (IEEE World Congress On Computational Intelligence), 2008.

[42] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, G. Zhao, The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 242–249.

[43] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters 124 (2019) 82–90.

[44] G. Heusch, A. Anjos, S. Marcel, A reproducible study on remote heart rate measurement, arXiv preprint arXiv:1709.00962 (2017).

[45] S.-Q. Liu, P. C. Yuen, A general remote photoplethysmography estimator with spatiotemporal convolutional network, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 481–488.

[46] F. Bousefsaf, A. Pruski, C. Maaoui, 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video, Applied Sciences 9 (20) (2019) 4364.

[47] H. Liu, J. Allen, S. G. Khalid, F. Chen, D. Zheng, Filtering-induced time shifts in photoplethysmography pulse features measured at different body sites: the importance of filter definition and standardization, Physiological Measurement 42 (7) (2021) 074001–.