LFPS-Net: a lightweight fast pulse simulation network for BVP estimation

Jialiang Zhuang, Yun Zhang, Yuheng Chen, Xiujuan Zheng

Abstract—Heart rate estimation based on remote photoplethysmography plays an important role in several specific scenarios, such as health monitoring and fatigue detection. Existing well-established methods are committed to taking the average of the predicted HRs of multiple overlapping video clips as the final results for the 30-second facial video. Although these methods with hundreds of layers and thousands of channels are highly accurate and robust, they require enormous computational budget and a 30-second wait time, which greatly limits the application of the algorithms to scale. Under these circumstances, we propose a lightweight fast pulse simulation network (LFPS-Net), pursing the best accuracy within a very limited computational and time budget, focusing on common mobile platforms, such as smart phones. In order to suppress the noise component and get stable pulse in a short time, we design a multi-frequency modal signal fusion mechanism, which exploits the theory of time-frequency domain analysis to separate multi-modal information from complex signals. It helps proceeding network learn the effective features more easily without adding any parameter. In addition, we design a oversampling training strategy to solve the problem caused by the unbalanced distribution of dataset. For the 30-second facial videos, our proposed method achieves the best results on most of the evaluation metrics for estimating heart rate or heart rate variability compared to the best available papers. The proposed method can still obtain very competitive results by using a short-time (15-second) facial video.

Index Terms—photoplethysmography, lightweight, short-time.

I. INTRODUCTION

Blood volume pulse (BVP) is an essential physiological signal with great potential in health care for humans. Although many excellent works have proven that the contact-based method can comprehensively and accurately measure cardiac activities, the sensors attached to the human bodies will bring discomfort and inconvenience. In recent years, an increasing number of methods have been proposed for remote estimation of heart rate or pulse signals, most of which can only estimate heart rate but not pulse signals precisely. In the field of camera-based physiological parameters monitoring, the early handcrafted algorithms, including independent component analysis (ICA) [1]–[3], principal component analysis (PCA) [4], Bland Altman and correlation analysis [5], [6] and adaptive filtering algorithms [7]–[10], were proposed to derive the effective component of signals from the facial videos based on the remote photoplethysmography (rPPG) or ballistocardiography (BBG).

Nevertheless, these methods are still limited by several manually calibrated parameters. In addition, they are challenging to deal with complex noise signals due to the changing ambient light conditions and face movements. In this case, deep learning is widely considered as the perfect choice to accomplish this work. The powerful semantic feature extraction capability of neural networks, which can effectively learn adaptively and deal with complex noise conditions, has led to a large number of deep learning-based heart rate computation and achieve great success [11]–[14] in the past five years. However, deep learning-based algorithms share a common problem, limited by a constrained training data set, prone to be overfitting under complex lighting conditions. As the target heart value has a wide range, from 50 to 120 beats per minute (bpm), and the unbalance distribution of training data results in the algorithm having a "tendency" for a specific heart rate range. Besides, for HR estimation, most of current heart rate algorithms need to calculate heart rate of 41 video clips from a 30-second facial video, which are overlapping, which takes too long and reduces the algorithm's flexibility and user experience.

For these reasons, this work proposes lightweight Fast Pulse Simulation Network (LFPS-Net) based on multi-frequency modal fusion mechanism. Computational resources are precious if one wants to perform physiological signal computation on the mobile platforms, and lightweight networks are of greater significance for the large-scale popularization of non-contact physiological algorithms. The network designed in this study benefits from the idea of 2D TDC [1] and requires far fewer parameters than the top algorithms in its class, such as DualGAN [11] and CVD [15]. To effectively solve the problem that algorithms are vulnerable to complex interference noise, this work draws on the idea of empirical modal decomposition to decompose the input signal into multiple intrinsic mode function and finally selects the practical components for fusion. A multi-frequency modal signal decompose module (MFD) is proposed, which decompose the input signal containing complex noise into an initial multi-band signal (IMBS) filtered by multiple frequency bands of interest. Therefore, the IMBS in each frequency band have strong similarities regardless of the target heart rate interval or strong interference. On this basis, learning their features is beneficial to improve the neural network’s ability and robustness. Individual band signals are placed into a signal refinement and reconstruction network designed for the target task, which refines effective signal features at multiple spatial-temporal resolutions, filters out noise, and recovers the target BVP in high definition. To further reduce the potential overfitting for a specific heart rate
range, an oversampling training scheme is designed in this work, which combined with data enhancement strategies, can maximize the training effect of limited data sets. We have built a time-frequency hybrid loss function to learn the waveform mode accurately.

1) Propose a powerful temporal multiscale convolution module with temporal difference convolution based on 2D spatial-temporal feature, intending to capture waveform Information in limited computational and time budget.

2) Spectrum-based attention mechanism automatically locates specific locations of high noise in spatio-temporal features.

3) A multi-frequency modal signal fusion mechanism considering the difference of modal information from different spectrum band and commonality in the same frequency band for different samples.

4) An oversampling training strategy for mitigating over-fitting phenomena.

II. RELATED WORK

1) Hand-crafted based remote physiological measurement: The technology of extracting physiological information through the camera has developed rapidly in the last ten years. Traditional non-contact heart rate measurement is divided into two main technical lines: BCG and PPG. Some study [16], [17] et al. experimentally found that the periodic motion of blood from the heart to the head through the abdominal aorta and carotid artery causes periodic head motion. Therefore they extracted the pulse waveform of this motion by tracking the features of the head as a way to calculate the heart rate. However, the periodic motion signal of the face is submerged in complex other motion signals, such as the motion of the subjective consciousness of the human body, and the algorithm’s effectiveness is drastically reduced. The most classical technical route is rPPG based method, which has demonstrated that traditional blind source separation methods, including independent component analysis (ICA) [5] and principal component analysis (PCA) [10], [18], and time-domain filter (TDF) [2], can effectively extract BVP precisely [6]. Many previous studies have been conducted to improve the robustness, for example, Xuxue Sun et al. [9] designed a combined empirical modal decomposition (EMD) and blind source separation algorithm to help the algorithm sort out the heart rate-related components of the chaotic original signal more efficiently. To make this method work more efficiently, Eduardo et al. [10] used principal component analysis (PCA) as an integration mechanism to evaluate the reshaped target waveform. Besides, some works put their hopes on improving the hardware performance, proposing multi-band cameras to make the obtained more robust signals [19], [20]. However, such solutions are demanding for hardware devices and difficult to apply on a large scale. To eliminate the influence of head movement, B.S. et al. [2] proposed Bland-Altman correlation analysis to process the original signal, while De Haan et al. [21] offered a chromaticity-based color space projection (CHROM) to enhance the rPPG signal in the original signal. However, signals induced by the environment or spontaneous by physiological processes are non-stationary, and it isn’t easy to obtain a priority statistical knowledge of the information to be extracted. In this case, adaptive filtering can directly use the observed data to continuously and recursively update the processing parameters to obtain real-time optimal processing results. The most representative of these methods is the minimum-mean adaptive filter [7]. [8]. [22]. Zhang et al. [23] proposed a general framework called TROIKA, which consists of signal decomposition for denoising, sparse signal reconstruction for high-resolution spectrum estimation, and spectral peak tracking with verification. Murthy et al. [24] yield several estimates of the heart rate trajectory from the spectrogram of the denoised PPG signal, which is finally combined using a novel measure called trajectory strength.

2) Deep learning based remote physiological measurement: Extracting physiological information via deep learning has come a long way, which is divided into two main categories, heart rate calculating and BVP simulation. Radim et al. [1] used 3D convolution networks to process the video signal directly to calculate heart rate values. In previous work [25], a spatial-temporal representation was designed as the input of CNN, which contains temporal and spatial features simultaneously. As a result, the stability and accuracy of heart rate estimation were significantly improved. All of the above algorithms calculate the heart rate directly, and although they work well, they are not able to extract the BVP precisely, missing a lot of important information. Yu et al. [26] proposed a 3D spatiotemporal network for extracting BVP from video sequences, which firstly offered Pearson correlation coefficients as the loss function for pulse wave simulation, significantly improving network performance. To recover rPPG signals from highly compressed facial videos, Yu et al. [14] proposed a video enhancement network. In [1], a feature extracting algorithm based on temporal difference convolution (TDC) was constructed using neural architecture search (NAS) for the first time. Niu et al. [15] designed a cross-verified feature separation strategy to separate physiological features from non-physiological features and then performed robustly multitask physiological measurements using the extracted physiological features. Lu et al. proposed a Dual-GAN-based remote physiological measurement algorithm [11], jointly modeling the BVP predictor and noise distribution to suppress the noise mixed together with the physiological information.

3) Attention mechanism: Recently, there have been an amount of works to incorporate attention processing to improve the performance of CNNs, which help the network not only performs well but also to be robust to noisy inputs. Many existing well-established attention mechanisms modeling in multiple dimensions have achieved tremendous success in classification task. Among them, the non-local network (NLNet) [27] pioneered the algorithm of aggregating global context to each query location to capture long-term dependencies. Cao et al. proposed [28] to apply a simplified design similar to the squeezed excitation network (SENet) [29] to the above network structure, which can design a three-step generic framework for modeling global context more effectively. In
addition, Woo et al. created a lightweight generic module that performs adaptive feature refinement along the channel and spatial dimensions [30]. The core constructs in the well-known Transformer [31] network is the multi-headed self-attentive mechanism, which shows us that the self-attentive tool for extracting global features has great potential in improving algorithmic capabilities. Our approach falls under the attention mechanism based approach, but with several significant differences compared to existing methods:

1. Different from [27, 28] requiring high-dimensional semantic features as the input, we convert the signal features into spectrum in the time domain dimension, which is then fed into attention module.

2. Unlike [29], let the network automatically learn the weight of each dimension feature, we mainly locate the noise and calculate its weight by looking for the difference of signals in different periods.

### III. Method

![Fig. 1. Overview of the proposed method. Firstly, Data pre-processing module cuts the face image by means of landmark points localized by mediapipe, which are converted to YUV color space and normalized, then white noise is added to generate spatial-temporal map. Next, multi-frequency modal signal decompose module extracts multi-frequency modal components from the spatial-temporal map in frequency domains, obtaining the initial multi-band signal IMBS. Finally, Multi-frequency modal signal compose module integrates features gained from multiple frequency and output the target BVP.](image)

#### A. Data pre-processing

The physiological information reflecting the blood flow in blood vessels of face video is weak. Still, the noise generated by lighting variation and face movements contaminate the effective signal. Therefore, to highlight the physiological component, this work firstly compresses the input video clip into initial spatial-temporal map (ISTM). Specifically, the mediapipe is used to perform landmark detection on the $t^{th}$ frame of the face video, and the set of average pixel values of the four regions is obtained by cutting the face regions with these landmarks. After that, the color space is converted to the YUV color space. Time-domain normalization module (TDN), which has been proven in numerous papers to be effective in helping networks recover physiological signals accurately in the face of large amounts of noise interference is then performed to obtain the normalized spatial-temporal map (NSTM). Finally, we get spatial-temporal map (STM) by adding white noise to the NSTM for enhancement to effectively simulate various potential noisy signals caused by movements such as harsh ambient light conditions and motion expressions in the actual scene during the training phase.

#### B. Multi-frequency modal signal fusion mechanism (MFF)

![Fig. 2. Architecture of the signal refinement and reconstruction network. The signal refinement network consists of four tandem modules, each consisting of a solution layer, a BatchNormalization layer, a Relu layer, a TMSC-module and a pooling layer. The signal reconstruction network consists of three tandem modules, each consisting of an inverse convolution layer, a BatchNormalization layer, a Relu layer and a SAA-module.](image)

Blind source signal separation algorithms have been shown in many works to filter noise and extract blood volume pulse from reflected light signals from faces videos [3, 10]. Various time-domain filters and frequency-domain filters can be considered as signal separators, performing the task of signal separation. Most blind source signal separation algorithms use the hand-crafted method such as independent component analysis, which decomposes the received mixed signal into several separate components, as an approximate estimate of the source signal. However, in the blood volume pulse extraction task, we are faced with an unknown distribution of the composite signal mixed with a large amount of noise, which significantly limits the effectiveness of the independent component analysis method.

This paper proposes a multi-frequency modal signal fusion mechanism to break this challenge. We first use the multi-frequency mode signal decompose module to obtain the initial multi-band signal by discrete cosine transform of the target signal. We then exploit the signal refinement and reconstruction network (SRRN) to extract multi-scale features from each component of the multi-frequency mode signal, obtaining the high-dimensional intrinsic mode signal. Finally, using the multi-frequency mode signal compose module, we get the target blood volume pulse (BVP) signal by fusing the practical components of the high-dimensional intrinsic mode signal.

1) **Multi-frequency modal signal decompose module(MFD):**

We set $K$ frequency bands of interest and $I$ face regions, the whole process is shown in Algorithm 1. First, we extract the time-domain signal $f(n)$ on a specific region from the STFM:

$$f(n) = STFM(i)$$ (1)

The cosine discrete transform is then applied to each region of the input time domain signal $f(i)$ with the following formula:

$$F(u, n) = c(u) \sum_{i=0}^{N-1} f(n) \cos \left(\frac{(2n + 1)u \pi}{2N}\right)$$ (2)
Where \( c(u) = \begin{cases} \sqrt{\frac{1}{N}}, & u = 0 \\ \sqrt{\frac{2}{N}}, & u \neq 0 \end{cases} \), \( u = 0, 1, \ldots, N-1 \).

Based on the frequency band set of interest \( FM \), we cut the spectrum and do zero padding to get \( FT(u, n, k) \), which is transformed to the time domain signal by discrete cosine inverse transform:

\[
f'(n) = \sum_{u=0}^{N-1} c(u)F'(u, n) \cos \left( \frac{(2n+1)u\pi}{2N} \right) \tag{3}
\]

Where \( c(u) = \begin{cases} \sqrt{\frac{1}{N}}, & u = 0 \\ \sqrt{\frac{2}{N}}, & u \neq 0 \end{cases} \), \( u = 0, 1, \ldots, N-1 \).

Finally, we concatenate all the region of the face in \( f'(n) \) and obtain initial multi-band signal \( IMBS(k) \) of each frequency band in the end.

Algorithm 1: Multi-frequency modal signal fusion mechanism (MFFM)

**Input:** Spat-temporal mapping STM, set of frequency bands \( FM(k) \), set of regions of face \( ROI(i) \), where \( k = 1, \ldots, K \), and \( i = 1, \ldots, I \)

**Output:** Predicted blood volume pulse BVP

1. for each region \( ROI(i) \) of the face in STM do
2. Extract the time domain signals \( f(n) \);
3. Do discrete cosine transform(DCT) to obtain \( F(u, n) \);
4. for each frequency band \( FM(k) \) do
   5. Cut out the frequency band of interest \( Fc(u, n, k) \);
   6. zero padding to get \( FT(u, n, k) \);
   7. inverse discrete cosine transform to obtain \( f'(n) \);
   8. Concatenate all the region of the face in \( f'(n) \) to obtain initial multi-band signal \( IMBS(k) \) of each frequency band ;
9. return \( IMBS(k) \);

2) signal refinement network: For effectively extracting signal features at multiple resolutions, we design a signal refining network. The IMBS is fed into the network, and a corresponding HDF signal is the output. The pipeline of the signal refining network is designed as shown in Figure 2, which consists of temporal multiscale convolution module (TMSC-module) convolution layer, activation layer, BN layer and pooling layer. It aims to gradually map the original signal to relatively pure potential stream shapes at lower resolutions, enhancing the network’s ability to adaptively filter out noisy signals. In this network, TMSC-module is used as the main structure that can improve the ability of network while keeping the computational budget constant. Besides, in order to capture multi-scale features in time domain, we set the convolution kernel sizes to \( 3 \times 1, 5 \times 1, 7 \times 1 \) similar as the settings in [32].

3) signal reconstruction network: The signal reconstruction network, including the Spectrum self-attention module (SAA-module), deconvolution layer, activation and BN layer, combines the multi-scale features in the feature extraction stage to perform the work of feature signal reconstruction. The main purpose of this design is to improve the SNR of the reconstructed signal and help the network to locate the position of each peak more precisely. We assume that in a signal waveform, the frequency domain features of different periods should be similar. If there is a large difference in the waveform features in a certain period compared with other periods, We can confirm the presence of loud noise signals in this time. We can get the self-attention weights of each filter through this mechanism and then put these weights into the channel attention module to globally group the individual filter features to further reduce the pulse wave noise. The spectrum-based attention mechanism can be abstracted into three processes: (a) time-domain discrete cosine transform. The signal is segmented in the time domain and then subjected to discrete cosine transform to obtain the frequency features. (b) frequency domain self-attention, using 1x1 convolution to reduce the number of channels to one dimension and extract the self-attention weight features for each period; (c) feature aggregation, using multi convolution layers, combining global contextual features, aggregating the original signal and self-attention weights to reconstruct the filtered signal.

4) signal reconstruction network: We fuse multiple high-dimensional intrinsic mode signals to extract the effective component and obtain the target snow volume pulse wave signal.

A. Oversampling training scheme

Due to the highly unbalanced distribution of training samples in terms of heart rate, putting all samples directly into the network training will inevitably overfit a certain heart rate range. Therefore, for the network to effectively learn the pulse wave signals corresponding to all heart rate intervals, this experiment uses an overfitting training scheme that divides the training samples into \( H \) groups according to the heart rate range. In each batch, the number of samples in each group is guaranteed to be in the same proportion, together with the data enhancement described in Part I to maximize the limited training data set to strengthen the network fitting ability.

V. EXPERIMENTS AND RESULTS

A. Database, metrics and tasks

ASPD2 contains 2166 RGB videos captured from 2166 individuals. We used this dataset for training with MMSEHR for cross-dataset testing and ablation study. Following previous work, we use the metrics including the standard deviation of the error (Std), the mean absolute error (MAE), the root mean squared error (RMSE), and the Pearson’s correlation coefficient (r).

UBFC-rPPG is a challenging dataset of remote physiological measurements with sunlight and indoor lighting conditions. It contains 42 RGB videos, all captured at 30 frames per second using a Logitech C920 HD Pro webcam, and corresponding actual BVP signals acquired using a CMS50E. In addition, UBFC-rPPG was used to perform heart rate and heart rate variability tests. We train the network on the first 30
subjects and test the remaining 12 subjects. For HR estimation, we report MAE, Std, RMSE, and r. For the task of HRV and RF estimation, follow existing methods \[15\], \[18\]. low frequency (LF), high frequency (HF), and LF/HF ratio in terms of Std, RMSE, and rare reported.

The MMSE-HR is a large dataset of remote heart rate tests. It has 102 videos taken from 40 people. In addition, it contains videos of multiple facial expressions and head movements recorded at 25 fps. Metrics.

### B. Analysis of the number of parameters

We will use the data provided in the detailed network structure diagram in CVD \[15\] without considering the training and testing phases, the physiological and non-physiological encoder, Ep and En, the decoder D, and the physiological estimator contain a total of 42k counts. Similarly, we only consider the number of parameters included in the BvpEstimator of DualGAN \[11\], which consists of a total of approximately 66k parameters. In contrast, the network we provide contains only 11k parameters in total as shown in Table I which is only one-fourth of CVD \[15\] and one-sixth of DualGAN \[11\], which is sufficient to show that the network structure proposed in this paper is lightweight and has excellent advantages for cloud deployment.

| Method       | Params |
|--------------|--------|
| CVD          | 42k    |
| DualGAN      | 66k    |
| Proposed     | 11k    |

### C. Intra-database testing

1) HR estimation on UBFC-rPPG: We evaluate the average HR-estimation on UBFC-rPPG.Following \[11\], state-of-the-art methods including hand-crafted methods (SAMC \[33\], POS \[34\], CHROM \[21\]) and deep learning based methods (SynRhythm \[35\], PulseGAN \[36\], DualGAN \[11\]) are used for comparison. We directly take the results of these state-of-the-art methods from \[11\]. The results of the proposed method and the state-of-the-art methods are given in Table II They both used 30 seconds of video for multiple segments, after which the average value was used as the final result, while we used the 30 seconds of video and the first 15 seconds of video to generate a separate BVP to calculate the heart rate values. From the results, we can see that, in the case of single calculation, the proposed 30s method achieves promising results with an RMSE of 1.12 bpm, an MAE of 0.75 bpm, an MER of 0.83 percent and a r of 0.99, outperforming all the state-of-the-art traditional and deep learning methods which needs to get the average result of multiple overlapping video clips except for Dual-GAN, this indicates that the waveform generated by our proposed algorithm is of higher quality and that a more accurate heart rate value can be obtained in just one calculation. On the other hand, the proposed 15s method also achieves comparable results, only worse than PulseGAN \[36\] and DualGAN \[11\]. They both have separate heart rate calculation modules for calculating heart rate values. Since our heart rate results are calculated from the pulse waveform, it is more difficult to calculate an accurate heart rate for short duration pulse waves, which further validates the effectiveness of our algorithm in reconstructing the BVP signal.

2) HRV and RF estimation on UBFC-rPPG: Following the protocol in \[11\], we used the first 30 subjects for training, and the remaining 12 subjects for testing. For the task of HRV and RF estimation, we use several state-of-the-art methods including POS \[34\], CHROM \[21\], Green \[37\], CVD \[15\], DualGAN \[11\] for comparison, the results of which are taken from \[11\]. The HRV and RF estimation results are shown in Table III, we use low frequency (LF), high frequency (HF), LF/HF, respiration frequency(RF) as evaluation metrics. We can see that the proposed approach outperforms all the existing state-of-the-art methods for HRV estimation under all measures. This shows that the architecture proposed in this paper can achieve accurate recovery of blood volume pulse wave shapes in a network with fewer parameters.

### D. Cross-database testing

Besides the intra-database testings on the VIPL-HR and UBFC-rPPG database, Following \[15\], train our method on ASPD2 and test it on MMSE-HR. The results of our proposed method and other existing state-of-the-art methods are given in Table IV, in which Li2014 \[38\], CHROM \[21\], SAMC \[33\], RhythmNet \[25\], CVD \[15\] are from \[11\]. From the Table IV we can see that our proposed 30s method achieves promising results compared with other state-of-the-art methods, which indicates the proposed method generalizes well in the unconstrained scenarios.

### E. Ablation study
TABLE III
RF and HRV estimation results by our method and several state-of-the-art methods on the UBFC-rPPG database.

| Method       | LF-(u.n) Std ↓ | RMSE ↓ | r ↑ | HF-(u.n) Std ↓ | RMSE ↓ | r ↑ | LF/HF Std ↓ | RMSE ↓ | r ↑ | RF-(Hz) Std ↓ | RMSE ↓ | r ↑ |
|--------------|----------------|--------|----|----------------|--------|----|-------------|--------|----|--------------|--------|----|
| POS [11]     | 0.17           | 0.169  | 0.479 | 0.17           | 0.169  | 0.479 | 0.405       | 0.399  | 0.518 | 0.109        | 0.107  | 0.087 |
| CHROM [21]   | 0.243          | 0.240  | 0.159 | 0.243          | 0.240  | 0.159 | 0.055       | 0.645  | 0.226 | 0.086        | 0.089  | 0.102 |
| Green [31]   | 0.186          | 0.186  | 0.280 | 0.186          | 0.186  | 0.280 | 0.405       | 0.399  | 0.518 | 0.087        | 0.086  | 0.111 |
| CVD [35]     | 0.053          | 0.065  | 0.740 | 0.053          | 0.065  | 0.740 | 0.169       | 0.168  | 0.812 | 0.017        | 0.018  | 0.252 |
| Dual – GAN   | 0.034          | 0.035  | 0.891 | 0.034          | 0.035  | 0.891 | 0.131       | 0.136  | 0.881 | 0.010        | 0.010  | 0.395 |
| Proposed – 30s| 0.030         | 0.030  | 0.921 | 0.030          | 0.031  | 0.921 | 0.101       | 0.101  | 0.895 | 0.041        | 0.041  | 0.163 |

TABLE IV
The cross-database HR estimation results by the proposed approach and several state-of-the-art methods on the MMSE-HR database.

| Method       | MAE↓ | RMSE↓ | r↑ |
|--------------|------|-------|----|
| Li2014 [38]  | 20.02| 19.95 | 0.38|
| CHROM [7]    | 14.08| 13.97 | 0.55|
| SAMC [33]    | 12.24| 11.37 | 0.71|
| ResNet-MNet [25]| 6.58 | 7.33  | 0.78|
| CVD [15]     | 6.06 | 6.04  | 0.84|
| Proposed-7s  | 6.23 | 6.36  | 0.84|
| Proposed-30s | 5.6  | 5.75  | 0.90|

TABLE V
Results of BP measurements.

| Method        | MAE↓ | Std↓ | RMSE↓ | r↑ |
|---------------|------|------|-------|----|
| Without OSS   | 4.65 | 7.94 | 7.97  | 0.78|
| Without MFF   | 4.75 | 9.89 | 10.21 | 0.63|
| Without SSA   | 3.89 | 6.35 | 6.41  | 0.87|
| Without TMSC  | 3.53 | 6.11 | 6.13  | 0.89|
| Proposed-30s  | 3.36 | 5.61 | 5.75  | 0.90|

1) Effectiveness of oversampling training strategy: In order to validate the effectiveness of oversampling strategy, we train our network with oversampling strategy as well as common sampling. From the results in Table [7] we can see that when we put all training set samples into network training in one epoch, the 30s HR estimation results are worse than using oversampling strategy. The MAE is reduced from 3.36 bpm to 4.65 bpm and the RMSE is reduced from 9.9 to 7.5 bpm, indicating that different heart rate intervals have similar temporal and spatial characteristics. Putting them equally into the network for training will help the network have better generalization performance and denoising ability in the whole heart rate interval.

2) Effectiveness of MFMSFM: From the Table [11] we can see that when our network loses the MFMSFM module, the results fall off a cliff, with MAE dropping from 3.36 bpm to 4.75 bpm and Std and RMSE dropping by 4.38 bpm and 4.46 bpm, respectively. This once again validates the importance of our proposed MFMSFM for improving the robustness of the network.

3) Effectiveness of the SSA-module: We further evaluate the effectiveness of combination of SSA-module. The results in the Table [15] show that the robustness of the network is further enhanced by using TMSCM and SSAM instead of the traditional residual convolution block, with RMSE decreasing from 6.05 to 5.75 and Std decreasing from 6.05 to 5.61. This validates our conception that SSA-module can significantly improve the filtering ability of the physiological task against noise.

4) Effectiveness of the TMSC-module: Finally, we test the effectiveness of TMSC-module. The results from the Table [12] suggest that extracting multi-scale features in the time-domain via TMSC-module is helpful for estimation accuracy. As a result, RMSE decreases from 6.13 to 5.75 and Std decreases from 6.11 to 5.61.

VI. ACKNOWLEDGMENTS
This work is supported by the Sichuan Science and Technology Program under Grant 2022YFS0032 to Xiujuan Zheng. The authors would like to thank the engineers of Xi’an Singularity Fusion Information Technology Co. Ltd for their supports of data collection and experimental procedures.

DECLARATION OF COMPETING INTEREST
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES
[1] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, “Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching,” IEEE Signal Processing Letters, vol. 27, pp. 1245–1249, 2020.
[2] B. Kim and S. Yoo, “Motion artifact reduction in photoplethysmography using independent component analysis,” IEEE Transactions on Biomedical Engineering, vol. 62, no. 2, pp. 415–425, 2015.
[3] R. Krishnan, B. Natarajan, and S. Warren, “Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data,” IEEE Transactions on Biomedical Engineering, vol. 57, no. 8, pp. 1867–1876.
[4] W. Wang, S. Stuijk, and G. de Haan, “Exploiting spatial redundancy of image sensor for motion robust rppg,” IEEE Transactions on Biomedical Engineering, vol. 62, no. 2, pp. 415–425, 2015.
[5] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation,” Opt. Express, vol. 18, pp. 10762–10774, May 2010.
[6] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in non-contact, multiparameter physiological measurements using a webcam,” IEEE Transactions on Biomedical Engineering, vol. 58, no. 1, pp. 7–11, 2011.
[7] K. Chan and Y. Zhang, “Adaptive reduction of motion artifact from photoplethysmographic recordings using a variable step-size lms filter,” in SENSORs, 2002 IEEE, vol. 2, pp. 1343–1346 vol.2, 2002.
[8] K. J. Han H., “Artifacts in wearable photoplethysmographs during daily life motions and their reduction with least mean square based active noise cancellation method,” Comput Biol Med, vol. 42, no. 4, pp. 387–93.
[9] X. Sun, P. Yang, Y. Li, Z. Gao, and Y.-T. Zhang, “Robust heart beat detection from photoplethysmography interlaced with motion artifacts based on empirical mode decomposition,” in Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, pp. 775–778, 2012.

[10] E. Pinheiro, O. Postolache, and P. Silva Girão, “Empirical mode decomposition and principal component analysis implementation in processing non-invasive cardiovascular signals,” Measurement, vol. 45, pp. 175–181, 02 2012.

[11] H. Lu, H. Han, and S. K. Zhou, “Dual-gan: Joint bvp and noise modeling for remote physiological measurement,”

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

[13] W. V. Chen and D. McDuff, “Deepleys: Video-based physiological measurement using convolutional attention networks,” ArXiv, vol. abs/1805.07888, 2018.

[14] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, “Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement,” 2019.

[15] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, “Video-based remote physiological measurement via cross-verified feature disentangling,” 2020.

[16] G. Balakrishnan, F. Durand, and J. Guttag, “Detecting pulse from head motions in video,” in 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3430–3437, 2013.

[17] R. Irani, K. Nasrollahi, and T. B. Moeslund, “Improved pulse detection from head motions using dct,” in 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol. 3, pp. 118–124, 2014.

[18] C. Tang, J. Lu, and J. Liu, “Non-contact heart rate monitoring by combining convolutional neural network skin detection and remote photoplethysmography via a low-cost camera,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1390–13906, 2018.

[19] D. McDuff, S. Gontarek, and R. W. Picard, “Improvements in remote cardiopulmonary measurement using a five band digital camera,” IEEE Transactions on Biomedical Engineering, vol. 61, no. 10, pp. 2593–2601, 2014.

[20] S. S. Wang W, den Brinker AC and de Haan G, “Robust heart rate from fitness videos,” Physiol Meas., vol. 38, no. 6, pp. 1023–1044, 2017.

[21] G. de Haan and V. Jeanne, “Robust pulse rate from chrominance-based rrg,” IEEE Transactions on Biomedical Engineering, vol. 60, no. 10, pp. 2878–2886, 2013.

[22] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, “Remote heart rate measurement from face videos under realistic situations,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 4264–4271, 2014.

[23] Z. Zhang, Z. Pi, and B. Liu, “Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise,” IEEE Transactions on Biomedical Engineering, vol. 62, no. 2, pp. 522–531, 2015.

[24] N. K. Lakshminarasimha Murthy, P. C. Madhusudana, P. Suresha, V. Periyasamy, and P. K. Ghosh, “Multiple spectral peak tracking for heart rate monitoring from photoplethysmogram signal during intensive physical exercise,” IEEE Signal Processing Letters, vol. 22, no. 12, pp. 2391–2395, 2015.

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

[26] Z. Yu, X. Li, and G. Zhao, “Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks,” 2019.

[27] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[28] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Gcnet: Non-local networks meet squeeze-excitation networks and beyond,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Oct 2019.

[29] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141, 2018.

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in Proceedings of the European Conference on Computer Vision (ECCV), September 2018.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.