FastBVP-Net: a lightweight pulse simulation network for measuring heart rhythm via facial videos

Jialiang Zhuang, Yuheng Chen, Yun Zhang, Xiujuan Zheng

Abstract—Remote photoplethysmography (rPPG) is a useful camera-based health monitoring method that can estimate the heart rate (HR) and heart rate variability (HRV) from facial videos. Many well-established deep learning models can provide highly accurate and robust results in heart rhythm measurement using rPPG. However, most previous models usually require enormous computational resources and a 30-second facial video, which significantly limits their applications in real-world scenarios. Hence, we propose a lightweight pulse simulation network named FastBVP-Net to measure heart rhythm via facial videos. In the FastBVP-Net, we designed a multi-frequency mode signal fusion mechanism to suppress the noise components and get stable blood volume pulse (BVP) signals quickly. Moreover, we developed an oversampling training strategy to solve the unbalanced distribution of HR in the dataset. Finally, we estimate the HR and HRV based on BVP signals derived by the FastBVP-Net. Comprehensive experiments are conducted on three benchmark datasets, our approach achieves competitive performance both on HR and HRV estimation from 30-second facial videos and HR estimation from 15-second facial videos comparing to the state-of-the-art methods. We also achieve promising results of computational speed and the number of parameters, demonstrating proposed network is a lightweight method.

Index Terms—remote photoplethysmography, lightweight, short-time, heart rate, heart rate variability.

I. INTRODUCTION

Blood volume pulse 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 hand-crafted algorithms, including independent component analysis (ICA) [1]–[3], principal component analysis (PCA) [4], Blund Altman and correlation analysis [5], [6] and adaptive filtering algorithms [7]–[9], 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 a better choice to overcome these problems. 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 [10]–[13] in the past five years. As we all know, if the network can perfectly simulate the blood volume pulse wave signal, the instantaneous heart rate can be calculated by only two cycles. However, if the calculated pulse wave signal has spurious waves and its peak position is deviated, the error of calculating the instantaneous heart rate by only two cycles can be very large. In the case of a large spurious blood volume pulse wave signal, it is necessary to reduce the error by calculating the average period of the signal over several periods. Therefore, we can conclude that the cleaner the pulse wave signal extracted by network is, the shorter the video length required to accurately calculate the heart rate is. When the network is highly robust to complex signals, the heart rate can be accurately calculated in a shorter video. For HR estimation, due to the lack of strong signal processing capabilities, most of current heart rate algorithms need to calculate heart rate of 41 video clips from a 30-second facial video. A large number of repeated and time-consuming calculations reduce the algorithm’s flexibility and timeliness. Meanwhile, deep-learning based algorithms also share a common problem, limited by a constrained training data set, prone easily to be over-fitting 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.

To break above these challenges, we propose lightweight pulse simulation network based on multi-frequency mode fusion mechanism, and use an oversampling training scheme. We first draw on the idea of empirical mode decomposition to effectively solve the problem that algorithms are vulnerable to complex interference noise. We propose a multi-frequency mode signal decompose module to decompose the input signal containing complex noise into multi-band signals (MBS) filtered by multiple frequency bands of interest. The MBS in each frequency band have strong similarities regardless of the target heart rate interval or strong interference, which helps to
model the network separately for different characteristics of each frequency band. To improve the neural network’s ability and robustness, we then place the individual band signals into a signal refinement and reconstruction network designed for the target task. The network refines effective signal features at multiple spatial-temporal resolutions, filters out noise, and recovers the target BVP in high definition. Therefore, we only need one calculation to get the heart rate value from a 30-second facial video or even a 15-second facial video, which avoids repeated calculations and saves a lot of time. Meanwhile, since the space size of the 2D spatial-temporal map we formed is 4, the network requires far fewer parameters than existing state-of-the-art algorithms, such as DualGAN [10] and CVD [14]. In addition, to further reduce the potential overfitting for a specific heart rate range, we also design an oversampling training scheme which is combined with data enhancement strategies. It can maximize the training effect of limited datasets. Results on multiple datasets for different physiological measurement tasks as well as the cross-database testing and the ablation study show the effectiveness of our method.

The contributions of this work include:

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) Propose a spectrum-based attention mechanism which automatically locates specific locations of high noise in spatio-temporal features.
3) Propose a multi-frequency mode signal fusion mechanism which considers the difference of mode information from different spectrum band and commonality in the same frequency band for different samples, greatly improving network robustness against noise.
4) Propose an oversampling training strategy for mitigating over-fitting phenomena.

II. RELATED WORK

A. 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: ballistocardiogram (BCG) and photoplethysmography (PPG). At first, some study [15], [16] 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 BCG of the video by tracking the head motion 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. After that, rPPG based method was proposed, which has demonstrated that traditional blind source separation methods, including independent component analysis (ICA) [3] and principal component analysis (PCA) [9], [17], and time-domain filter (TDF) [2], can effectively extract BVP precisely [6]. Many previous studies have been conducted to improve the robustness, for example, Sun et al. [8] 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. [9] 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 [18], [19]. 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 and correlation analysis to process the original signal, while De Haan et al. [20] 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 adaptively filtering methods is the minimum-mean adaptive filter [7], [21]. Zhang et al. [22] 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. [23] 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.

B. 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. [16] used 3D convolution networks to process the video signal directly to calculate heart rate values. In previous work [24], 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. [25] 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. [13] 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. [14] designed a cross-verified feature sep-
aration strategy to separate physiological features from non-physiological features and then performed robustly multitask physiological measurements using the extracted physiological features. Lu et al. [10] proposed a DualGAN-based remote physiological measurement algorithm, jointly modeling the BVP predictor and noise distribution to suppress the noise mixed together with the physiological information.

C. Attention mechanism

Recently, there have been an amount of works to incorporate attention processing to improve the performance of CNNs, which helps the network perform well 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) [26] pioneered the algorithm of aggregating global context to each query location to capture long-term dependencies. Cao et al. [27] proposed to apply a simplified design similar to the squeezed excitation network (SENet) [28] to the above network structure, which can design a three-step generic framework for modeling global context more effectively. In addition, Woo et al. [29] created a lightweight generic module that performs adaptive feature refinement along the channel and spatial dimensions. The core constructs in the well-known Transformer [30] 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 [26], [27] 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 [28] 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

In this section, we give detailed explanations of the proposed multi-frequency mode signal fusion strategy for capturing waveform information quickly from face videos. Fig. 1 gives an overview of the proposed method, which includes the data preprocessing, the multi-frequency mode signal decompose module, the signal refinement and reconstruction network and multi-frequency mode signal compose module.

A. Data preprocessing

The physiological information reflecting the blood flow in blood vessels of face video is weak, and the lighting variation and face movements will generate noise to contaminate the effective signal. Therefore, it is important to highlight the physiological component in face videos.

Data preprocessing part firstly compresses the input video clip into initial spatial-temporal map. Specifically, the mediapipe is used to perform landmark detection on the $i^{th}$ 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. Then the color space conversion (CSC) is used to convert the initial spatial-temporal map to the YUV color space. After that Time-domain normalization module (TDN) is performed to obtain the normalized spatial-temporal map. TDN 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. Finally, we get the spatial-temporal map by adding white noise to the normalized spatial-temporal map. The white noise 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 mode signal fusion mechanism

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], [9]. 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. Moreover, blind source separation algorithms including PCA and ICA are only suitable for processing linearly correlated mixed signals. However, remote photoelectric volume pulse wave signals are interspersed with complex noise, and blind source separation algorithms can struggle in the face of this situation. Therefore, 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 blind source signal separation method.

In order to handle complex signals with non-linear correlation, we extract signal features in each frequency band adaptively. This paper proposes the multi-frequency mode signal fusion mechanism to break this challenge. Specifically, as shown in Fig. 1 we firstly use discrete cosine transform (DCT) and inverse discrete cosine transform (IDCT) to obtain the initial multi-band signals. Then we exploit 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 mode signal decompose module: In order to extract multi-frequency mode components from the STM in frequency domain, multi-frequency mode signal decompose module is introduced. 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 STM:

$$f(n) = STM(i)$$  \hspace{1cm} (1)
The cosine discrete transform is then applied to each region of the input time domain signal \( f(n) \) with the following formula:

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

Where

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

Based on the boundary defined by \( FM(k) \), we cut part of the frequency spectrum and do zero-padding to get adjusted frequency spectrum \( FT(u) \), which is then transformed to the filtered time domain signal by discrete cosine inverse transform:

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

Where

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

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

2) Signal refinement network: For effectively extracting signal features at multiple resolutions, we design a signal refinement network. The STM and the MBS is fed into the network, and a corresponding high-dimensional frequency (HDF) signal is the output. The pipeline of the signal refinement network is designed as shown in Figure 2, which consists of temporal multi-scale 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 of TMSC-module to \( 3 \times 1, 5 \times 1, 7 \times 1 \) similar as the settings in [31]. And we concatenate the output of the three convolution blocks for fusing multi-scale features and

---

**Algorithm 1: Multi-frequency mode signal decompose module**

**Input:** Spatial-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:** Multi-band signals \( MBS(k) \)

1. For each region \( i \) of the face in STM do
2. Get the time domain signals \( f(n) = STM(i) \);
3. Do discrete cosine transform to obtain frequency spectrum \( F(u) \);
4. For each frequency band \( k \) do
5. Based on the boundary defined by \( FM(k) \), intercepting part of the frequency band and do zero-padding to get adjusted frequency spectrum \( FT(u) \);
6. Use inverse discrete cosine transform to obtain filtered signal \( f'(n) \);
7. Concatenate all the region of the face in \( f'(n) \) to obtain multi-band signal \( MBS(k) \) of each frequency band;
8. return \( MBS(k) \);
obtain a better representation to highlight the physiological information.

3) Signal reconstruction network: The signal reconstruction network, including the spectrum self-attention module (SAA-module), deconvolution layer, activation and BN layer, uses 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 the frequency domain features of different periods should be similar in a signal waveform. 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) Multi-frequency mode signal compose module: We fuse multiple high-dimensional intrinsic mode signals to extract the effective component and obtain the target blood volume pulse wave signal.

C. 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 several 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 data preprocessing module to maximize the limited training data set to strengthen the network fitting ability.

IV. EXPERIMENTS AND RESULTS

A. Datasets and Experimental Settings

1) Datasets: We evaluate our method on three widely-used publicly-available physiological measurement datasets, i.e., VIPL-HR, UBFC-rPPG and MMSE-HR.

VIPL-HR contains 2378 less-constrained RGB videos from 107 individuals, which include a variety of scenarios such as head movements, lighting changes, etc. The frame rate of all videos varies from 25 fps to 30 fps. We use this dataset for Cross-dataset testing.

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 our experiments, UBFC-rPPG was used in Intra-dataset testing and Ablation study. We train the network on the first 30 subjects and test the remaining 12 subjects.

MMSE-HR is a large dataset of remote heart rate tests. It has 102 videos taken from 40 people. In addition, it contains...
HRV estimation results by our method and several state-of-the-art methods on the UBFC-rPPG dataset.

| Method          | LF-(u,n) | HF-(u,n) | LF/HF |
|-----------------|----------|----------|-------|
|                 | Std↓ RMSE↓ r↑ | Std↓ RMSE↓ r↑ | Std↓ RMSE↓ r↑ |
| POS [32]        | 0.17 0.169 0.479 | 0.17 0.169 0.479 | 0.405 0.399 0.518 |
| CHROM [20]      | 0.243 0.240 0.159 | 0.243 0.240 0.159 | 0.655 0.645 0.226 |
| Green [35]      | 0.186 0.186 0.280 | 0.186 0.186 0.280 | 0.405 0.405 0.518 |
| CVD [14]        | 0.053 0.065 0.740 | 0.053 0.065 0.740 | 0.169 0.168 0.812 |
| Dual-GAN [10]   | 0.034 0.035 0.891 | 0.034 0.035 0.891 | 0.131 0.136 0.881 |
| FastBVP(30s)    | 0.030 0.030 0.921 | 0.030 0.031 0.921 | 0.101 0.101 0.895 |

* Notes: Bold indicates best performance, italic indicates second best performance.

The number of parameters and FLOPs in several state-of-the-art methods.

| Method          | Params | FLOPs   |
|-----------------|--------|---------|
| CVD [14]        | 42k    | 1.1 × 10^{10} |
| DualGAN [10]    | 66k    | 9.9 × 10^{9}  |
| FastBVP(Ours)   | 11k    | 1.3 × 10^{8}  |

HR estimation results of our method and several state-of-the-art methods on the UBFC-rPPG dataset.

| Method          | MAE↓ RMSE↓ r↑ |
|-----------------|--------------|
| POS [32]        | 8.35 10.00 0.24 |
| CHROM [20]      | 8.20 9.92 0.27 |
| Green [33]      | 6.01 7.87 0.29 |
| SynRhythm [11]  | 5.59 6.82 0.72 |
| PulseGAN [32]   | 1.19 2.10 0.98 |
| Dual-GAN [10]   | 0.44 0.67 0.99 |
| FastBVP(30s)    | 0.75 1.12 0.99 |
| FastBVP(15s)    | 1.53 2.04 0.98 |

Cross-dataset HR estimation results by the proposed approach and several state-of-the-art methods on the MMSE-HR dataset.

| Method          | MAE↓ RMSE↓ r↑ |
|-----------------|--------------|
| LI2014 [15]     | 20.02 19.95 0.38 |
| CHROM [20]      | 14.08 13.97 0.55 |
| SAMC [35]       | 12.24 11.37 0.71 |
| RhythmNet [24]  | 6.98 7.33 0.78 |
| CVD [14]        | 6.06 6.04 0.84 |
| FastBVP(30s)    | 5.6 5.75 0.90 |
| FastBVP(15s)    | 6.23 6.36 0.84 |

Ablation study of our FastBVP-Net in terms of OSS, TMSC, SSA and MFF for HR estimation on UBFC-rPPG.

| Method          | MAE↓ RMSE↓ r↑ |
|-----------------|--------------|
| Without OSS     | 1.55 3.13 0.92 |
| Without TMSC    | 0.85 1.24 0.98 |
| Without SSA     | 0.92 1.35 0.98 |
| Without MFF     | 2.05 4.01 0.89 |
| FastBVP(30s)    | 0.75 1.12 0.99 |

B. Analysis of the number of parameters and FLOPs

We will use the data provided in the detailed network structure diagram in CVD [14] and DualGan [10] for analysis. Only considering the testing phases, the related module of CVD includes the physiological encoder Ep, non-physiological encoder En, the decoder D, and the physiological estimator, whose parameters are 42k and FLOPs is 1.1 × 10^{10}. Similarly, the number of parameters in the BvpEstimator of DualGAN are approximate 66k and FLOPs is 9.9 × 10^{9}. In contrast, the network that we provide contains only 11k parameters in total as shown in Table II. The number of parameters is only one-fourth of CVD [14] and one-sixth of DualGAN [10], which is sufficient to show that the network structure proposed in this paper is lightweight and has excellent advantages for cloud deployment. In terms of FLOPs, the FLOPs of proposed network is 1.3 × 10^{8}, which is smaller than those two networks in order of magnitude.

Moreover, if we consider the monitoring task of 30 seconds video, the proposed network in this paper can be completed by only one calculation, but CVD [14] can only calculate the heart rate in 10 seconds interval for each input, and it needs a total of 41 calculations to complete the whole task. The situation of DualGAN [10] is similar. So both of excellent existing methods need much higher computational complexity than our proposed network to complete the 30 seconds monitoring task.

C. Intra-dataset testing

1) HR estimation on UBFC-rPPG: We evaluate the average HR estimation on UBFC-rPPG. State-of-the-art methods including hand-crafted methods (SAMC [36], POS [32], CHROM [20]) and deep learning based methods (SynRhythm [37], PulseGAN [34], DualGAN [10]) are used for videos of multiple facial expressions and head movements recorded at 25 fps. We use MMSE-HR dataset for testing in Cross-dataset testing.

2) Evaluation Metrics: Various metrics are used for evaluation. For the task of average HR estimation, we use the metrics including the standard deviation of the error (Std), the mean absolute error (MAE), the root mean square error (RMSE), and the Pearson’s correlation coefficients (r). For the task of HRV estimation, we follow existing methods [14, 17] and use low frequency (LF), high frequency (HF), and LF/HF ratio in terms of Std, RMSE, and rare reported.
comparison. We can directly take the results of these state-of-the-art methods from previous work. The results of the proposed method and the state-of-the-art methods are given in Table III. They all slice multiple segments of 30 seconds of videos for heart rate estimation, after which the average heart rate value was used as the final result, while we used the 30 seconds of video to generate a separate BVP to calculate heart rate values. From results, we can see that, in the case of single heart rate calculation, the proposed method achieves promising results with an RMSE of 1.12 bpm, an MAE of 0.75 bpm and a ρ of 0.99. It outperforms all the state-of-the-art traditional and deep learning methods which need 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 a more accurate heart rate value can be obtained in just one calculation.

2) HRV estimation on UBFC-rPPG: Following the protocol in [10], we use the first 30 subjects for training, and the remaining 12 subjects for testing. For the task of HRV estimation, we use several state-of-the-art methods including POS [32], CHROM [20], Green [33], CVD [14], DualGAN [10] for comparison, the results of which are taken from [10]. The HRV estimation results are shown in Table IV. 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-dataset testing

Besides the intra-dataset testing on the UBFC-rPPG dataset. Following [14], we train our method on VIPL-HR 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 [35], CHROM [20], SAMC [56], Rhythm-Net [24], CVD [14] are from [10]. From the Table IV we 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

We also provide the results of ablation studies for the proposed method for HR estimation on the UBFC-rPPG dataset. All the results are shown in Table V.

1) Effectiveness of oversampling training strategy: In order to validate the effectiveness of oversampling strategy, we train out network with oversampling strategy as well as common sampling. From the results in Table V we can see that when we put all training set samples into network training in one epoch, the 30s HR estimation results is worse than using oversampling strategy. The MAE is reduced from 1.55 bpm to 0.75 bpm and the RMSE is reduced from 3.13 to 1.12 bpm. It indicates 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 the TMSC-module: we then test the effectiveness of TMSC-module. The results from the Table V suggest that the multi-scale convolution kernel in the TMSC-module helps the net to improve the time-domain field of perception, allowing neurons to base their decisions on more comprehensive information. Extracting multi-scale features in the time-domain via TMSC-module is helpful for improving estimation accuracy. As a result, MAE decreases from 0.85 to 0.75 bpm and RMSE decreases from 1.24 to 1.12 bpm.

3) Effectiveness of the SSA-module: We further evaluate the effectiveness of combination of SSA-module. The results in the Table V show that the robustness of the network is further enhanced by using TMSCM and SSAM instead of the traditional residual convolution block, with MAE decreasing from 0.92 to 0.75 bpm and RMSE decreasing from 1.35 to 1.12 bpm. This validates our conception that SSA-module can significantly improve the filtering ability of the physiological task against noise.

4) Effectiveness of MFF mechanism: From the Table V we can see that when our network loses the MFF mechanism, the results fall off a cliff, with MAE dropping from 0.75 bpm to 2.05 bpm and RMSE and MAE dropping by 2.89 bpm and 1.3 bpm, respectively. The pyramid-like structure of the multi-scale feature segmentation fusion mechanism allows the network to learn the high-dimensional features of the signal from multiple scale modes in all directions, taking into account both the global and local information of the signal. In addition, signals with complex noise are decomposed into multi-frequency modes, which contain less noise and are more traceable, validating the importance of our proposed MFF for improving the robustness of the network.

F. 15s-HR estimation task

As mentioned in the introduction, the accuracy of heart rate extraction from short-time face videos reflects the signal processing capability of the network, so we purposely conducted a 15-second heart rate prediction experiment to verify the superiority of our proposed method from this aspect.

Table III demonstrates that when we perform 15s-HR estimation on UBFC-rPPG with one calculation, the method also achieves comparable results, MAE only worse than PulseGAN [34] and DualGAN [10] which needs 30s video and multiple calculations. Besides 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. Besides, we also use cross-validation to evaluate our proposed short-time method on MMSE-HR. From Table V we can see that the proposed approach achieves a promising result with a RMSE of 6.36 bpm when only using 15s face video. Compared with other existing 30s methods, although CVD [14] is better than ours, we only have a relatively small difference, with an RMSE difference of 0.29 and a Std difference of 0.17.
V. CONCLUSION

This article proposes a novel method for fast and stable computation of BVP at the cost of small computational resources. The MFF and overfitting sampling mechanisms proposed in this can also be applied to other signal processing tasks. In the future, we need to further improve the network’s ability to adapt to complex noise.

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. 53, no. 3, pp. 566–568, 2006.

[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] 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.

[9] E. Pinheiro, O. Postolache, and P. Silva Giral, “Empirical Mode Decomposition and Principal Component Analysis implementation in processing non-invasive cardiovascular signals,” Measurement, vol. 45, pp. 175–181, 02 2012.

[10] H. Lu, H. Han, and S. K. Zhou, “Dual-GAN: Joint BVP and Noise Modeling for Remote Physiological Measurement.”

[11] 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.

[12] W. V. Chen and D. McDuff, “DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks,” ArXiv, vol. abs/1805.07888, 2018.

[13] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, “Remote Heart Rate Measurement from Highly Compressed Facial Videos: an End-to-end Learning Solution with Video Enhancement,” 2019.

[14] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, “Video-based Remote Physiological Measurement via Cross-verified Feature Disen-tangling,” 2020.

[15] 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.

[16] G. Balakrishnan, F. Durand, and J. Guttag, “Detecting Pulse from Head Motions using DCT,” in 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol. 3, pp. 118–124, 2014.

[17] 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/CFVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1390–13906, 2018.

[18] 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.

[19] 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.

[20] G. de Haan and V. Jeanne, “Robust Pulse Rate From Chrominance-Based rPPG,” IEEE Transactions on Biomedical Engineering, vol. 60, no. 10, pp. 2878–2886, 2013.

[21] 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.

[22] 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.

[23] N. K. Lakshminarasimha Murthy, P. C. Madhusudana, P. Suresha, V. Periyasamy, and P. K. Ghosh, “Multiple Spectral Peak Tracking for Heart Rate Monitoring from Photoplethysmography Signal During Intensive Physical Exercise,” IEEE Signal Processing Letters, vol. 22, no. 12, pp. 2391–2395, 2015.

[24] 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.

[25] Z. Yu, X. Li, and G. Zhao, “Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks,” 2019.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” 2017.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 2015.

[32] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, “Algorithmic Principles of Remote PPG,” IEEE Transactions on Biomedical Engineering, vol. 64, no. 7, pp. 1479–1491, 2017.

[33] W. Verkruysse, L. Svasaand, and J. Nelson, “Remote photoplethysmographic imaging using ambient light,” Optics express, vol. 26, pp. 14354–14355, 12 2008.

[34] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, “PulseGAN: Learning to Generate Realistic Pulse Waveforms in Remote Photoplethysmography,” IEEE Journal of Biomedical and Health Informatics, vol. 25, pp. 1373–1384, 2021.

[35] L. Shi, 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.

[36] S. Tulyakov, X. Alamda-Pineda, E. Ricci, L. Yin, J. Cohn, and N. Sebe, “Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions,” 06 2016.
[37] A. Lam and Y. Kuno. “Robust Heart Rate Measurement from Video Using Select Random Patches,” in 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3640–3648, 2015.