# BVPNet: Video-to-BVP Signal Prediction for Remote Heart Rate Estimation

Abhijit Das[1,*], Hao Lu[2,*], Hu Han[2], Antitza Dantcheva[3,4], Shiguang Shan[2] and Xilin Chen[2]
[1] Thapar Institute of Engineering & Technology, Thapar University, Patiala, India
[2] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, China
[3] Inria, Sophia Antipolis, France [4] Université Côte d'Azur, France

***Abstract*— In this paper, we propose a new method for remote photoplethysmography (rPPG) based heart rate (HR) estimation. In particular, our proposed method BVPNet is streamlined to predict the blood volume pulse (BVP) signals from face videos. Towards this, we firstly define ROIs based on facial landmarks and then extract the raw temporal signal from each ROI. Then the extracted signals are pre-processed via first-order difference and Butterworth filter and combined to form a Spatial-Temporal map (STMap). We then propose to revise U-Net, in order to predict BVP signals from the STMap. BVPNet takes into account both temporal and frequency domain losses in order to learn better than conventional models. Our experimental results suggest that our BVPNet outperforms the state-of-the-art methods on two publicly available datasets (MMSE-HR and VIPL-HR).**

## I. INTRODUCTION

Traditional heart rate (HR) measurement methods such as electrocardiograph (ECG) and contact photoplethysmography (PPG) are not instrumental in portable scenarios. Therefore, non-contact HR measurement-based on rPPG has received increasing attention in recent years [1], [2], [3]. Applications of rPPG include gauging HR variability, monitoring respiration and emotion state [4], [5]. The principle behind remote HR measurement using rPPG is based on the variation in optical absorption by the skin owing to the periodic change in blood volume. Thus, the frequency of the periodic skin color changes can indicate the HR.

While significant progress has been made in rPPG based HR measurement [6], [7], [8], [9], the studies were mainly limited to controlled settings. Less-constrained settings including large variations of lighting, expression, movement, occlusion remain challenging. Currently, machine learning methods, and specifically deep learning methods based on Convolutional Neural Networks (CNNs), have shown outstanding modelling power and achieved remarkable success in many research problems such as object detection [10], image classification [11], as well as face recognition [12]. Similarly, several works have employed the strong modelling ability of CNNs in remote HR estimation [13], [14], [15], [16], [17].

However, as shown in Fig. 1, we note that existing methods are not robust in extracting the BVP signal from video, and are being particularly challenged in unconstrained scenarios including variations of illumination, and pose. In order to tackle such challenges, we propose a new method to estimate BVP signals (*BVPMap*) from raw temporal signals with noises using constraints from both, temporal and frequency domains. Such a method is more effective in dealing with noise incorporated in temporal signals under unconstrained conditions.

In addition, the inclusion of non-skin areas of the face such as hair and other objects has been found to introduce large noise that is able to impact model robustness. A mechanism that can help to remove this noise is ROI selection. Incidentally, previous works considered the whole face region while extracting BVP, whereas only the *exposed skin area* is naturally the most suitable ROI. Hence, we propose a ROI selection method based on facial landmarks.

Unstable camera sampling frequency has been an additional unresolved challenge in rPPG based HR estimation. Towards tackling this, we here propose a novel preprocessing technique (including first-order difference, cubic spline interpolation and band-pass filter), in order to obtain normalized physiological signals. Next, we utilize this signal as input of U-Net aimed at producing a refined STMap.

The main contributions of this work for efficient HR estimation can be summarized as follows.

- We propose a technique, which directly computes HR based on raw temporal signals from videos, and then employs different filters to improve the PSNR. We leverage the strong modeling capacity and robustness of U-Net with temporal and frequency constraints to predict accurate BVP signals from raw temporal signals.
- We propose an ROI selection method based on facial landmarks to reduce the influence of non-skin face areas to improve the robustness.
- We propose a pre-processing method based on cubic spline interpolation across frames to solve the problem of unstable camera sampling frequency and improve the HR estimation accuracy.

The rest of the paper is organized as follows. In Section II, we summarize the conventional and deep learning-based methods for remote heart rate estimation. In Section III, we provide details pertained to BVPNet. We provide experimen-

* Equal contribution.

This research was supported in part by the National Key R&D Program of China (grant 2018AAA0102501), Natural Science Foundation of China (grant 62176249), and Youth Innovation Promotion Association CAS (grant 2018135).

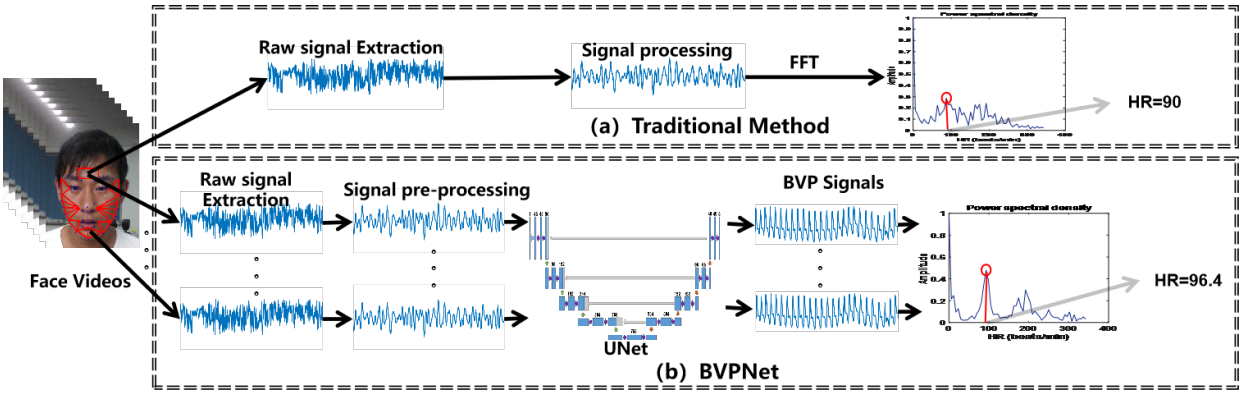


Fig. 1. Comparison of traditional rPPG based methods and the proposed method (BVPNet). (a) Traditional methods usually extract the periodical temporal signal from a face video (filtering may also be used to improve PSNR) and then perform FFT to compute the HR. (b) Our BVPNet aims to predict a more accurate BVP signal via U-Net, and then perform FFT to get the HR. Moreover, we also introduce a region of interest (ROI) selection to improve the robustness.

tal evaluations in Section IV and summarize this work in Section V.

## II. RELATED WORK

In this section, we briefly review scientific works on remote HR estimation, which can be generally classified into traditional approaches and deep learning based approaches.

### A. Traditional HR Estimation

Verkruysse *et al.* verified the possibility of performing rPPG analysis from videos with commercial cameras [18]. What followed were a number of traditional signal analysis based algorithms aimed at estimating HR from face videos. For example, Balakrishnan *et al.* proposed a heart rate estimation algorithm based on principal component analysis (PCA) [19]. However, this method was based on the assumption that heartbeat can cause minor head shakes. In reality, maintaining a still head position is challenging; instead, various rigid head movements are common. This hinders accurate HR estimation based on the named method. Most following studies have attempted to estimate HR based on changes in facial brightness. For example, independent component analysis (ICA) [20] was employed as a blind signal separation (BSS) method, to separate time-domain signals from three color channels into independent signal sources to retrieve the pulse [21]. The chrominance signals were linearly combined to reduce the noise in the time series [21]. Later, pixel-wise chrominance was proposed to improve the accuracy of heart rate estimation [9]. [22] studied the characteristics of rPPG signals at different wavelengths and then distinguished between noise and motion brightness changes caused by heartbeat.

### B. Deep Leaning based HR Estimation

Deep learning (DL) entails powerful nonlinear fitting capabilities and has been widely used in various applications. In recent years, DL based methods have been studied for remote heart rate estimation [23], [17], [16], [24], [25], [15], [26], [13], [3]. Many DL methods are proposed for extracting BVP signals, in order to obtain HR values. Specifically, the BVP signal is learned from a continuous video sequence without prior knowledge. DeepPhys was proposed to extract physiological signals by two consecutive video frames, and then obtain the HR from the signal's spectrum in [17]. Vspetlik *et al.* used two successive CNNs to extract BVP signals with spectrum-based loss, and estimated the heart rate from BVP [16]. LSTM was used to predict BVP signals from face video in [24]. A two-stream method was proposed in [25] to estimate HR, by using a low-rank constraint loss to derive reliable features. The 3D convolutional network (named Spatial-Temporal Net) was used to obtain the BVP signal directly from video [16]. Yu *et al.* attempted to extract physiological signals from highly compressed images [15]. These methods have achieved good performance attributed to the strong modelling capacity of CNNs.

HR estimation from face videos does not constitute naive learning, as datasets for training are relatively small. Hence, DL based HR estimation methods have predominantly investigated following techniques to improve network convergence.

1) *Transfer learning and data augmentation*. There are various sources of video noise, and the amount of data containing the HR and BVP signals labels is small, whereas the data for training the depth model is insufficient. Hence, using ImageNet or synthetic noise signals for pre-training is common, in order to migrate drawbacks [23], [24]. In [26] Niu *et al.* proposed an upsampling and downsampling strategy for augmenting the dataset, which improved accuracy for more extensive and smaller valued HR estimation.

2) *New feature representation methods*. In DL, the input format of the network is essential. It is essential that it contains sufficient information about the heart rhythm, as well as possess a concise form, with removed background information. HR-CNN used aligned face images to predict HR [16]. Time-frequency representation was proposed as the input of CNN and had exhibited state-of-the-art results for HR estimation on the constrained MAHNOB-HCI database [13]. Spatial-temporal representation was designed as the input of ResNet in [3], which led to improved performance over state-

of-the-art methods on MAHNOB-HCI, MMSE-HR [27] and VIPL-HR databases [14].

## III. Proposed Approach

As shown in Fig. 2, we here propose BVPNet, consisting of four main components: (a) facial landmarks detection and ROIs determination; (b) the raw time signals extracted from ROIs are then preprocessed and then combined into a spatiotemporal map (STMap); (c) a revised U-Net is designed to predict the BVP signals from STMap; (d) computation of the final HR estimate via trimmed mean.

### A. ROI Extraction

Towards reducing the environmental impact and placing attention to the change of skin brightness in the face, we propose an ROI extraction method. In the first step of the ROI extraction, we employ Seetaface[1] for landmark detection. There are following two issues faced with landmark detection.

1. Face landmarks fluctuate randomly, hence the face positions between frames may become inconsistent, so we use an average filter with a window size of 5 to smooth the landmark detection process across adjacent frames [26].

2. A facial landmark may be located outside the face area, so we move all landmarks toward the tip of the nose in a particular proportion.

Hence, we refine the landmark coordinates as follows:

$$P_i^{'}(x,y) = P_i(x,y) + k(P_n(x,y) - P_i(x,y))$$
$$P_n^{'}(x,y) = P_n(x,y)$$
$$i \in \{\, 1, 2, ...81 \,|\, i \neq n \,\},$$

where $P_i(x,y)$ denotes the $i$-th detected landmark that has been processed on average. $n$ represents the index of the landmark of the nose. $P_n(x,y)$ is the landmark on the tip of the nose, which is kept unchanged during the landmark refinement. $P_i^{'}(x,y)$ denotes the $i$-th refined landmark. $k$ denotes the proportionality coefficient, which is fixed as 0.05 in our experiments. Based on the refined landmarks, we define 15 ROIs as shown in Fig. 3.

### B. Spatial-Temporal Map Representation

**Raw Signal Preprocessing.** The raw signals/ROIs extracted from the video have the following three characteristics that go through the corresponding proposed processing.

1) The rigid movement of the head and the change in light source can cause low-frequency components with large amplitude [28], whereas the brightness change of skin owing to the volume and oxygen saturation of the blood is of low amplitude. So, the first-order difference is used to segregate them, solving this problem.

2) The high-frequency noise (such as the non-rigid motion of the face and the shake of landmarks) renders the extracted original signal noisy. The normal HR ranges within [40, 160] beats/min, or [0.67, 2.67] Hz. Hence towards normalization, the Butterworth filter is used to extract the frequency band of interest.

3) The sampling frequency of the ordinary camera is unstable, but the acquisition frequency of BVP signal in video is highly stable. Therefore, the signal obtained from the video sequence cannot correspond to the ground-truth BVP signal. So, we use cubic spline interpolation to solve the above problem in this paper.

Moreover, the information extracted directly in the RGB color space is not necessarily the best choice [29], hence the ROIs in the RGB space domain is converted to the YUV color space.

In summary, our preprocessing technique comprises of the following steps: 1) ROIs is converted from RGB to YUV, 2) The first-order difference is used to reduce low-frequency components, 3) The sampling frequency of signals is normalized by cubic spline interpolation to 30Hz. 4) Butterworth filter with parameter [0.4,10] Hz is used as a band-pass filter to retain the frequency band of interest, related to HR.

To convert the **Signal to Spatial-temporal Map**, after preprocessing, $R \times C$ signals of length $T$ can be extracted from $C$ colour channels of $R$ ROIs. In order to use DL to predict BVP signals, we design a spatial-temporal map and BVPMap. We construct the STMap [3], as shown in Fig. 4. To obtain the **ground-truth BVPMap** for training the model, we copy and extend the ground-truth BVP signals into different rows to establish 2D-BVPMap. Fig. 5 shows the BVPMap construction process in detail.

### C. BVP Prediction via Revised U-Net

Based on the properties of the STMap, we propose a revised U-Net to predict the BVP signal. The structure of U-Net allows feature extraction on different scales of the signal, which is equivalent to different filtering and enhancement methods for different frequencies of the signal. Besides, cascading them at the same scale endows U-Net the ability to retain features at different scales. In order to customize U-Net for our task, we change the pooling technique. In the classic U-Net, which is designed for segmentation, max pooling was used, which we replace by average pooling in our scenario [17]. We note that max pooling ignores the sampling theorem, small input translations or shifts of signals can cause drastic changes in the output. In contrast, average pooling can be normalized owing to its mean property. The architecture and parameters of the network are listed in Table I.

It is challenging to accurately restore the BVPMap using traditional reconstruction loss, e.g., L1 and MSE loss. The reason behind this is that the input signal and ground truth BVP signal have a random phase difference. Hence, we define a loss based on the Pearson correlation coefficient [15]:

$$L_p = \frac{\sum_{d=1}^{D}\sum_{j=1}^{C}(1 - corr(S(d_j), S_g(d)))}{D \times C},$$

[1] https://github.com/seetafaceengine/SeetaFace2

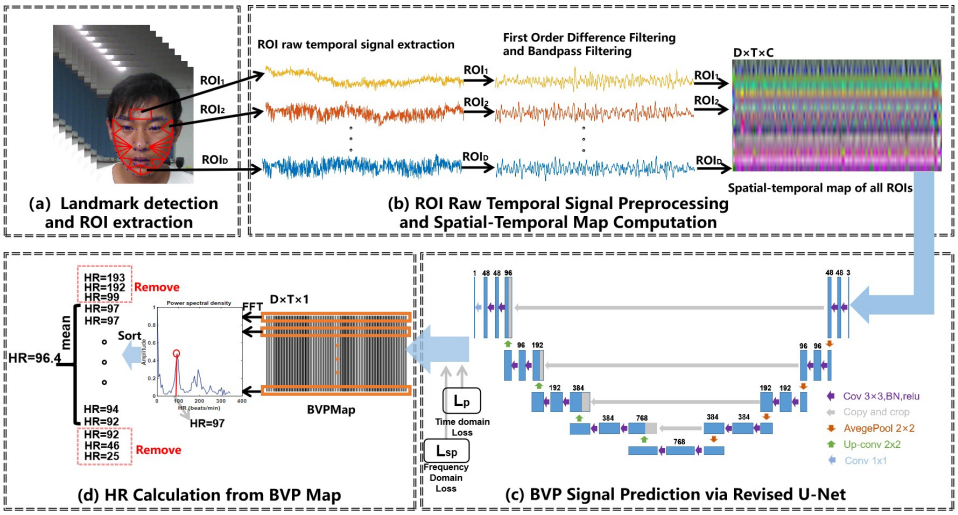


Fig. 2. The framework of BVPNet: (a) Facial landmark detection and ROIs extraction based on detected landmarks; (b) Raw temporal signal extraction from each ROI followed by preprocessing via first-order difference, cubic spline interpolation and band-pass filtering. The preprocessed temporal signals are combined into a $D \times T \times C$ spatial-temporal map (STMap); (c) A revised U-Net is designed to predict more accurate BVP signals from STMap; (d) The three largest and smallest HRs computed from all the spectral distributions are removed, and the remaining ones are averaged to obtain the final HR estimate.

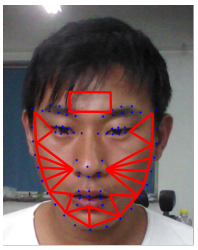

Fig. 3. The ROIs (red triangles) used in our approach are defined based on 81 facial landmarks.

where $d_j$ denotes the $d$-th ROI and the $j$-th color channel, $D$ is the number of ROI and $C$ is the number of color channels. The $corr(S(d_j), S_g(d))$ is Pearson coefficients of $S(d_j)$ and $S_g(d)$ i.e., the predicted signal and ground-truth. Hence, the prediction loss $L_p$ ranges from 0 to 2. When the two signals are identical, $L_p$ is 0.

The HR estimate is very sensitive to frequency domain information, in particular to peaks in the frequency domain. Therefore, we define a new loss in the frequency domain. In the first step, Fast Fourier Transform (FFT) followed by absolute operation converts rPPG signals $SP(r, f)$ by revised U-Net, and ground truth BVP signals $SP_g(r, f)$ to spectrum squared. Then, we select the frequency band of interest, and normalize operations as follows.

$$SP^{'}(d_j, f_i) = \frac{SP_g(d_j, f_i)}{\sum_{f_n}^{f_m} SP(d, f_i)},$$

$$SP_g^{'}(d, f_i) = \frac{SP_g(d, f_i)}{\sum_{f_n}^{f_m} SP_g(d, f_i)},$$

where $f_n$ and $f_m$ represent the maximum and minimum values of the HR range, with the value of 180 and 21 beats/min,

TABLE I

ARCHITECTURE OF THE REVISED U-NET FOR BVP PREDICTION FROM RAW TEMPORAL SIGNALS.

| Layer | input size | output size | content |
|---|---|---|---|
| inc | $D \times T \times 3$ | $D \times T \times 48$ | DualConv |
| down1 | $D \times T \times 48$ | $\frac{D}{2} \times \frac{T}{2} \times 96$ | AvgPool DualConv |
| down2 | $\frac{D}{2} \times \frac{T}{2} \times 96$ | $\frac{D}{4} \times \frac{T}{4} \times 192$ | AvgPool DualConv |
| down3 | $\frac{D}{4} \times \frac{T}{4} \times 192$ | $\frac{D}{8} \times \frac{T}{8} \times 384$ | AvgPool DualConv |
| down4 | $\frac{D}{8} \times \frac{T}{8} \times 384$ | $\frac{D}{16} \times \frac{T}{16} \times 384$ | AvgPool DualConv |
| up1 | $\frac{D}{16} \times \frac{T}{16} \times [384 \times 2]$ | $\frac{D}{8} \times \frac{T}{8} \times 192$ | Upsample DualConv |
| up2 | $\frac{D}{8} \times \frac{T}{8} \times [192 \times 2]$ | $\frac{D}{4} \times \frac{T}{4} \times 96$ | Upsample DualConv |
| up3 | $\frac{D}{4} \times \frac{T}{4} \times [96 \times 2]$ | $\frac{D}{2} \times \frac{T}{2} \times 48$ | Upsample DualConv |
| up4 | $\frac{D}{2} \times \frac{T}{2} \times [48 \times 2]$ | $D \times T \times 48$ | Upsample DualConv |
| outc | $D \times T \times 48$ | $D \times T \times 1$ | Conv |

1. $D$ and $T$ represent the number of ROIs and the length of the signal, respectively. 2. DualConv includes Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d, and ReLU in this order. 3. The input of 'up' includes the output of the previous layer and the same-scale output of 'down'. Hence, the feature map size has doubled.

respectively. Both high-frequency and low-frequency components are removed. While the energy of the low-frequency component is often very high, it is usually the DC component that is irrelevant to HR. At the same time, the energy of high-frequency components is often very low and meaningless. Hence, the final loss in the frequency domain is as follows

$$L_{sp} = \frac{\sum_{d=1}^{D} \sum_{j=1}^{C} \sum_{i=n}^{m} |SP^{'}(d_j, f_i) - SP_g^{'}(d, f_i)|}{D \times C},$$

where $L_{sp}$ denotes the difference between predicted signals

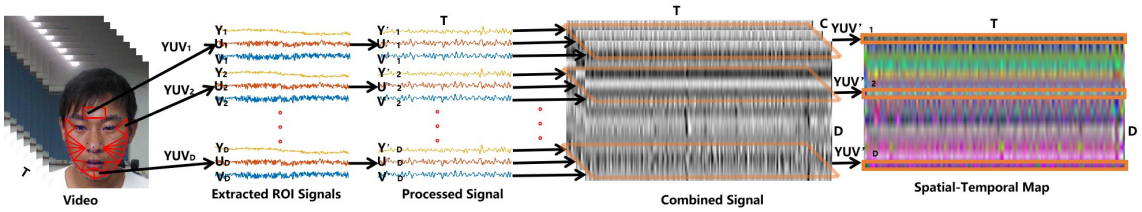


Fig. 4. The spatial-temporal map (STMap) calculation process: Firstly, face video is converted from RGB to YUV. Then, we extract three signals (for each of R, G, B respectively) of length T from each ROI (after first-order difference and Butterworth filter). The signals from all $n$ ROIs are further combined into a $D \times T \times 3$ matrix, which can be visualized as a 3-channel color image. We call such a $D \times T \times 3$ matrix as a spatial-temporal map containing the heart rhythm signal of the input face video, which is expected to retain most of the information for the rhythm signal, while suppressing the background information irrelevant to HR estimation.

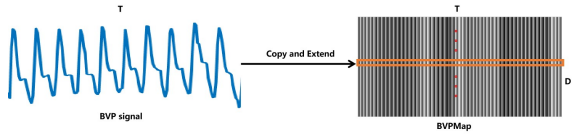


Fig. 5. The ground-truth BVP signal is copied and extended into a 2D BVPMap, with every row the same as the ground-truth BVP signal.

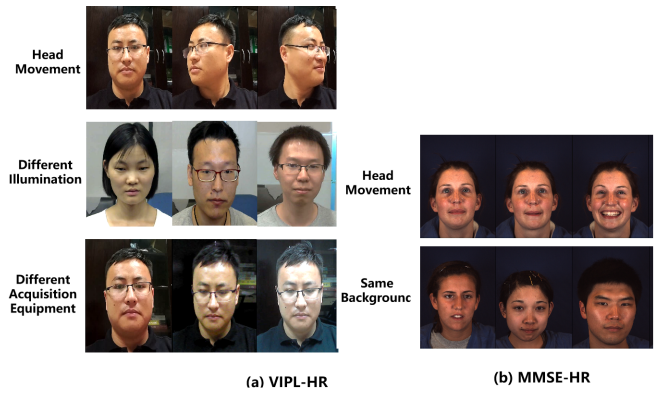


Fig. 6. Examples of the VIPL-HR [14] and MMSE-HR [27] dataset: (a) The VIPL-HR dataset was captured in a less-constrained condition, i.e., with different acquisition equipment, large head movement and different lighting conditions. (b) The MMSE-HR dataset was collected under less-constrained conditions, i.e., with small head movements and the same background.

and ground truth in the frequency domain, ranging between 0 and 2. Hence, the final loss constitutes the sum of $L_p$ and $L_{sp}$.

### D. From Spectrum to HR

The fast Fourier transform (FFT) transforms of each row of the predicted BVPMap is computed to obtain a spectrum. The frequency with the largest amplitude in this spectrum is the HR value. Then, we can obtain $D$ number of HRs. We use a trimmed mean to compute the final estimated HR. Specifically, the three largest and smallest HRs among the D number of HRs are dropped, and the remaining ones are averaged to obtain the final HR.

## IV. Experiments

In this section, we evaluate the proposed approach on two public-domain datasets (VIPL-HR [14] and MMSE-HR [27]), and provide comparisons with state-of-the-art methods. We also provide an ablation study to validate the effectiveness of individual components in our approach.

### A. Datasets and Experimental Settings

**Datasets:** We tested our algorithm on the following two databases:

(1) VIPL-HR [14] database is a challenging database for remote HR estimation, containing 2,378 RGB videos and 752 near-infrared videos of 107 subjects captured under nine different conditions. It is a dataset in a real environment, as shown in Fig. 6. Large head movements, different lighting conditions, as well as different acquisition equipment have brought to the fore great challenges for HR estimation. In addition, a considerable challenge is that the sampling frequency of the camera is unstable.

(2) MMSE-HR [27] database is widely used for remote HR estimation, which contains 101 video segments and provides real-time blood pressure signals for each video. In this dataset, head movements are weak, whereas the background is consistent for all videos.

**Data augmentation:** In order to improve the generalization ability of our algorithm, we propose the following three data augmentation strategies.

(1) We use a 2% probability that the positions of the two random rows of the STMap are randomly exchanged, i.e., to simulate the vibration of the ROIs across the aligned face images;

(2) We refer to [26], and perform temporal down-sampling and up-sampling to the original face video with sampling rates from 0.67 to 1.5, to enrich the HR distribution of the training dataset.

(3) For every epoch of the training, we use a window of $D \times T$ to crop the STMap, which reduces the effect of the phase difference of input videos.

**Parameter settings:** Our BVPNet is implemented in PyTorch, and the FFT algorithm is implemented by Numpy. We use the Adam solver as an optimizer with an initial learning rate of 0.001, and the batch size is 64. On both the MMSE-HR database and VIPL-HR database, we use a 5-fold cross-validation protocol on the shuffled database; HR ground truth is obtained from the ground truth BVP signal collected by FDA approved sensor. Specifically, the frequency of the amplitude peak of the spectrum of BVP

signal is the HR value; hyperparameters $D$, $T$ and $C$ are empirically set to 15, 256 and 3 respectively.

**Performance Metrics:** We follow existing methods [8], [3], and use several different measures to report the performance, i.e., the mean error (ME), standard deviation (Std) and root mean squared (RMSE), mean absolute HR error (MAE), mean of error rate percentage (MER) and Pearson's correlation coefficients ($\rho$).

### *B. Pre-processing Details and Result Discussion*

As mentioned previously, in order to better predict the BVP signal, we pre-process the raw temporal signals obtained from the video to make the model easier to converge by using first-order difference and Butterworth filter. We show the effect pre-processing step and similarity with the ground truth BVP signal in Fig. 7 using one video form MMSE-HR.

In Fig. 7 (a), we observe that the original signal has a DC trend component and the high-frequency noise, which is very different from the ground truth BVP signal. The DC component may be caused by a rigid movement of the head or the change in ambient lighting, while the high-frequency component may be caused by the non-rigid movement of the face or face landmarks shake. The first-order difference effectively suppresses the tendency of the original signal as shown in Fig. 7 (b), and the Butterworth filter removes high-frequency noise as shown in Fig. 7 (c). After the preprocessing, the signal becomes more similar to the ground-truth BVP signal. This suggests that the proposed preprocessing step is instrumental in the final HR estimation task.

There is an irreversible gap between the timestamp of the real BVP signal and the timestamp of our pre-processed signal. The final HR value is derived from the spectrum of the signal. Therefore, we propose a loss in the frequency domain. Towards better defining the frequency domain loss, we analyze the spectral characteristics of the BVP signal. As shown in Fig. 8, the blue line is the spectrum of the BVP signal. Along the X-axis and Y-axis are the HR in beats per minute and the normalized signal amplitude, respectively.

As shown in Fig. 8, the spectrum has three characteristics: (i) The energy of the low-frequency component is very high, which is often a DC trend component irrelevant to HR, which should be removed. (ii) There are often higher harmonics than HR frequency in the spectrum that are multiples of the HR frequency. (iii) The high-frequency components after the triple HR frequency harmonics are very small. Therefore, we only select the frequency band of interest to define the loss. The physical meaning of the frequency of the spectrum peak refers to the HR value. Therefore, performing a square operation on the obtained spectrum renders peaks more pronounced and helps to estimate HR accurately.

### *C. Results on VIPL-HR*

VIPL-HR has a large number of RGB videos for remote HR estimation, which contains nine complex environments and large head movements. In order to test the effectiveness of our proposed method, we compare it with state-of-the-art methods including deep learning based methods, namely I3D [30], DeepPhy [17] and RhythmNet [23] and RhythmNet+Attention [26] and several hand-crafted methods (Tulyakov [27], POS [31] and Haan2013 [21]. The deep learning methods are tested on VIPL-HR datasets using a 5-fold cross-validation [26]. A comparison of these algorithms is shown in Table II.

TABLE II

THE HR ESTIMATION RESULTS PERTAINED TO OUR METHOD AND STATE-OF-THE-ART METHODS ON THE VIPL-HR DATABASE.

| Method | ME | Std | MAE | RMSE | MER | $\rho$ |
|---|---|---|---|---|---|---|
| SAMC [27] | 10.8 | 18.0 | 15.9 | 21.0 | 26.7% | 0.11 |
| POS [31] | 7.87 | 15.3 | 11.5 | 17.2 | 18.5% | 0.30 |
| Haan2013 [21] | 7.63 | 15.1 | 11.4 | 16.9 | 17.8% | 0.28 |
| I3D [30] | 1.37 | 15.9 | 12.0 | 15.9 | 15.6% | 0.07 |
| DeepPhy [17] | -2.60 | 13.6 | 11.0 | 13.8 | 13.6% | 0.11 |
| RhythmNet [14] | 1.02 | 8.88 | 5.79 | 8.94 | 7.38% | **0.73** |
| RhythmNet+DA +Attention [26] | **-0.16** | 7.99 | 5.40 | 7.99 | 6.70% | 0.66 |
| **BVPNet** | -1.15 | **7.75** | **5.34** | **7.85** | **6.62%** | 0.70 |

TABLE III

HR ESTIMATION RESULTS PERTAININT TO OUR METHOD AND STATE-OF-THE-ART METHODS ON THE MMSE-HR DATABASE.

| Method | ME | Std | RMSE | MER | $\rho$ |
|---|---|---|---|---|---|
| Li2014 [8] | 11.56 | 20.02 | 19.95 | 14.64% | 0.38 |
| Haan2013 [21] | 9.41 | 14.08 | 13.97 | 12.22% | 0.55 |
| Tulyakov2016 [27] | 7.61 | 12.24 | 11.37 | 10.84% | 0.71 |
| **BVPNet** | **1.57** | **7.27** | **7.47** | **4.96%** | **0.79** |

From Table II, we can see that our BVPNet performs the best with respect to Std, MAE, RMSE, and MER. RhythmNet [14] and RhythmNet+DA+Attention [26], as end-to-end deep learning-based algorithms, perform very well for the indicator ME and $\rho$. However, for other indicators, our algorithm has achieved better performance. At the same time, our method also outperforms methods reconstructing signals, i.e., Haan2013 [21], Tulyakov2016 [27], POS [31] and DeepPhy [17], by a large margin.

### *D. Results on MMSE-HR*

In order to test the generality of our algorithm, we also perform experiments on the MMSE-HR dataset. MMSE-HR contains 101 videos, out of which we remove one video, as the video length is too short. Three of the state-of-the-art methods were chosen as a comparison of our method, including Li2014 [8], Haan2013 [21], Tulyakov2016 [27]. The evaluation results of these algorithms are shown in Table III.

From Table III, we can see that the results of our BVPNet are the best in terms of all measures, proving the superiority of our approach. Compared to the best of the baseline methods, i.e., the method of Tulyakov $etal.$ [27], our method reduces the HR estimation errors by a large margin, e.g., 2.39 for Std, 2.63 for RMSE, 0.39 for MER, and increases the $\rho$ by 0.0567. Despite the MMSE-HR dataset being small, our algorithm still performs very well.

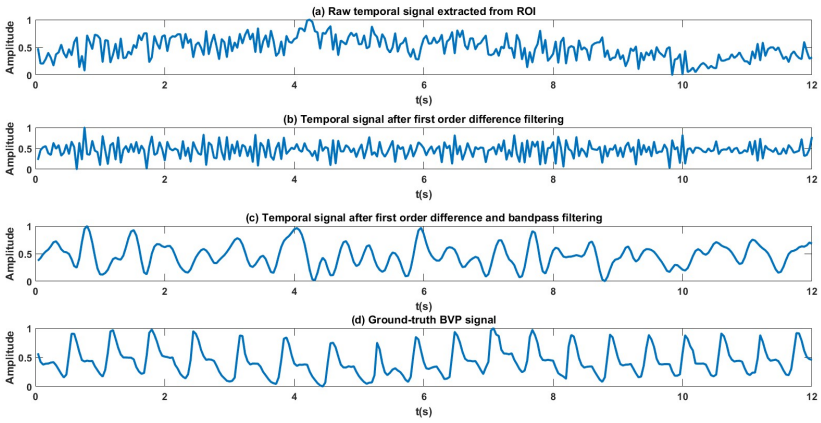


Fig. 7. Visualization of the signal pre-processing process: (a) raw temporal signal extracted from ROI, (b) temporal signal after first-order difference filtering, (c) temporal signal after both first-order difference and Butterworth filtering, and (d) corresponding ground-truth BVP signal.

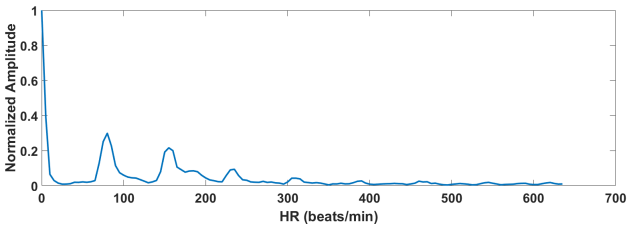


Fig. 8. The frequency spectrum computed from the predicted BVP signal by our BVPNet.

TABLE IV
THE ABLATION STUDY RESULTS OF BVPNET ON VIPL-HR.

| Method | ME | Std | MAE | RMSE | MER | $\rho$ |
|---|---|---|---|---|---|---|
| BVPNet w/o $L_p$ | 1.32 | 8.51 | 6.38 | 10.05 | 7.85% | 0.63 |
| BVPNet w/o $L_{sp}$ | 1.55 | 8.12 | 5.98 | 9.91 | 7.45% | 0.65 |
| BVPNet w/o RE | 2.01 | 7.92 | 5.65 | 8.43 | 7.03% | 0.69 |
| **BVPNet** | -1.15 | 7.75 | 5.34 | 7.85 | 6.62% | 0.70 |

### *E. Ablation Study*

To discuss the effect of each part of our algorithm, we perform ablation experiments. Time-domain loss($L_p$), frequency-domain loss($L_{sp}$) and ROI extraction (RE) are the pertinent parts of our algorithm, and are tested on the VIPL-HR separately. The test results are shown in Table IV.

We discuss the results in Table IV from two aspects:

**Time-domain and frequency-domain loss:** When either of the two losses is removed, the algorithm evaluation index will become worse. In other words, it is effective to use both time-domain and frequency-domain loss. Time-domain loss can make the shape of the predicted BVP signal similar to the ground truth BVP signal. The frequency-domain loss guarantees that each frequency component of the predicted BVP signal is the same as the real BVP signal. Nevertheless, using frequency-domain loss alone can ensure that the spectrum of the predicted signal is similar to the true value as a whole, but it does not guarantee the peak position of the spectrum perfectly. This suggests that it is effective to use both time and frequency domain loss.

**ROI extraction:** When only the original STMap method [14] is used without the ROI extraction method, the accuracy of the algorithm decreases significantly. The ROI region extracted according to landmarks can effectively reduce the motion impact, which is the main reason for improving the robustness of the algorithm.

### *F. Visualization and Discussion*

To discuss the effects of the revised U-Net, we visualize the spectrum of the physiological signal. We select three videos of the test dataset of the VIPL-HR dataset. The first column of each row in the BVPMap is predicted. The first row in each predicted BVPMap is selected for visualization, i.e., the BVP signal predicted from the first ROI. Its time-frequency plot is drawn, and the sliding window is 128, and its step is 64. In order to prove the effectiveness of the revised U-Net, we also plot the frequency spectrum of the pre-processed signal and the ground truth BVP signal (GT) accordingly. As shown in Fig. 9, the first column denotes the input - the video sequence(Fig. 9 v1, v2, v3), the second to the fourth columns are the frequency spectrum of the pre-processed signals(Fig. 9 a1, b1, c1), our prediction signals (Fig. 9 a2, b2, c2) and the real ground truth BVP signals (Fig. 9 a3, b3, c3). In the time-frequency diagram, the x-and y-coordinates denote the time and frequency, respectively. Different colors represent different values.

As shown in Fig.9, all time-frequency diagrams in the low-frequency component appear in yellow to indicate a very high value, and the high-frequency component appears blue to indicate a low amplitude. Except for the low-frequency DC component, the frequency with the highest amplitude in the spectrum is the HR value. We observe that the pre-processed signal spectrum (Fig. 9 a1, b1, c1) does not have apparent peaks, and the predicted signal spectrum (Fig. 9 a2, b2, c2) and GT spectrum (Fig. 9 a3, b3, c3) have obvious peaks. This proves that the revised U-Net can learn effective functions in the frequency domain. In addition, we see that there are apparent peaks near the second and third octave of the HR frequency. This phenomenon can be further studied to improve the HR estimation accuracy.

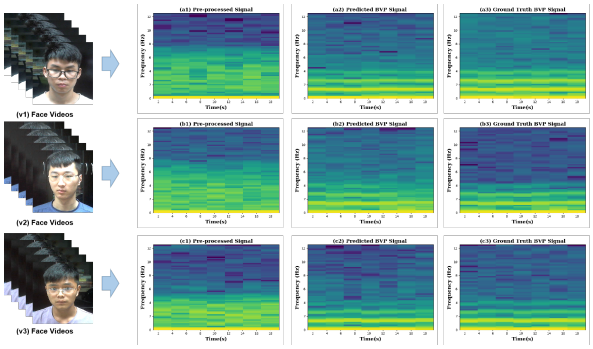


Fig. 9. Examples of three face videos (v1, v2, v3) with associated preprocessed temporal signals extracted from ROIs (a1, b1, c1), the predicted BVP signals by revised U-Net (a2, b2, c2), and the ground-truth BVP signals (a3, b3, c3).

## V. Conclusions and Future Work

Non-contact heart rate estimation is a challenging task due to variations in lighting, head movements, expression changes, occlusion, and frequency domain information variation. In this paper, we propose BVPNet streamlined to predict BVPMap from face video to overcome the challenge of frequency domain information variation. In addition, we propose an ROI selection method aimed at overcoming occlusion and discarding non-skin face areas. In BVPNet, we propose raw signal preprocessing (includes color space conversion, first-order difference, cubic spline interpolation, and Butterworth filtering) on the data according to prior knowledge. Then a spatial-temporal map is created in a revised U-Net, designed to predict BVP signals from the spatial-temporal map using both, time and spectral domain supervision information. Our algorithm achieves better performance than several state-of-the-art methods. In future work, we will improve rPPG based heart rate estimation by (1) further analyzing the frequency domain characteristics of BVP signals, as well as the noise signals, in order to improve the stability of the algorithm; (2) more robust modeling of head movements, facial expressions, changes in ambient lighting, etc.; (3) designing and testing with more efficient deep learning modules for BVP signal prediction from STMap; (4) reducing the amount of depth model parameters for real-time application.

## References


[1] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, p. R1, 2007.

[2] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Robust heart rate from fitness videos," *Physiol. Meas.*, vol. 38, no. 6, p. 1023, 2017.

[3] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Trans. on Image Process.*, pp. 2409–2423, 2020.

[4] Y. Sun and N. Thakor, "Photoplethysmography revisited: from contact to noncontact, from point to imaging," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 463–477, 2016.

[5] A. Sikdar, S. K. Behera, and D. P. Dogra, "Computer-vision-guided human pulse rate estimation: a review," *IEEE Rev. Biomed. Eng.*, vol. 9, pp. 91–105, 2016.

[6] 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, no. 10, pp. 10 762–10 774, 2010.

[7] M. Poh, J. McDuff, and W. Picard, "Advancements in noncontact multiparameter physiological measurements using a webcam," vol. 58, no. 1, pp. 7–11, 2011.

[8] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE CVPR*, 2014, pp. 4264–4271.

[9] W. Wang, S. Stuijk, and G. De Haan, "Exploiting spatial redundancy of image sensor for motion robust rppg." *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 415–425, 2015.

[10] R. Girshick, "Fast r-cnn," in *Proc. IEEE ICCV*, 2015, pp. 1440–1448.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[12] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. IEEE CVPR*, 2018, pp. 5265–5274.

[13] M.-S. C. Gee-Sern Hsu, ArulMurugan Ambikapathi, "Deep learning with time-frequency representation for pulse estimation," in *Proc. IJCB*, 2017, pp. 642–650.

[14] X. Niu, H. Han, S. Shan, and X. Chen, "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video," in *Proc. Springer ACCV*, 2018, pp. 562–576.

[15] 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," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 151–160.

[16] R. Špetlík, V. Franc, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proc. BMVC*, 2018, pp. 3–6.

[17] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proc. IEEE ECCV*, 2018, pp. 349–365.

[18] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Opt. Express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[19] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proc. IEEE CVPR*, 2013, pp. 3430–3437.

[20] G. R. Tsouri, S. Kyal, S. A. Dianat, and L. K. Mestha, "Constrained independent component analysis approach to nonobtrusive pulse rate measurements," *Journal of biomedical optics*, vol. 17, no. 7, p. 077011, 2012.

[21] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, 2013.

[22] G. De Haan and A. Van Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiol. Meas.*, vol. 35, no. 9, p. 1913, 2014.

[23] X. Niu, H. Han, S. Shan, and X. Chen, "Synrhythm: Learning a deep heart rate estimator from general to specific," in *Proc. IEEE ICPR*, 2018, pp. 3580–3585.

[24] M. Bian, B. Peng, W. Wang, and J. Dong, "An accurate lstm based video heart rate estimation method," in *Proc. Springer PRCV*, 2019, pp. 409–417.

[25] Z.-K. Wang, Y. Kao, and C.-T. Hsu, "Vision-based heart rate estimation via a two-stream cnn," in *Proc. IEEE ICIP*, 2019, pp. 3327–3331.

[26] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, and X. Chen, "Robust remote heart rate estimation from face utilizing spatial-temporal attention," in *Proc. IEEE FG*, 2019, pp. 1–8.

[27] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. IEEE CVPR*, 2016, pp. 2396–2404.

[28] X. Niu, H. Han, J. Zeng, X. Sun, S. Shan, Y. Huang, S. Yang, and X. Chen, "Automatic engagement prediction with gap feature," in *Proc. ACM ICMI*, 2018, p. 599–603.

[29] G. R. Tsouri and Z. Li, "On the benefits of alternative color spaces for noncontact heart rate measurements using standard red-green-blue cameras," *J. Biomed. Opt.*, vol. 20, no. 4, p. 048002, 2015.

[30] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE CVPR*, 2017, pp. 6299–6308.

[31] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, 2017.