Heart Rate Estimation From Facial Videos Using a Spatiotemporal Representation With Convolutional Neural Networks

Rencheng Song, *Member, IEEE*, Senle Zhang, Chang Li, *Member, IEEE*, Yunfei Zhang, Juan Cheng, *Member, IEEE*, and Xun Chen, *Senior Member, IEEE*

***Abstract*— Remote photoplethysmography (rPPG) is a kind of noncontact technique to measure heart rate (HR) from facial videos. As the demand for long-term health monitoring grows, rPPG attracts much attention from researchers. However, the performance of conventional rPPG methods is easily degenerated due to noise interference. Recently, some deep learning-based rPPG methods have been introduced and they revealed good performance against noise. In this article, we propose a new rPPG method with convolutional neural networks (CNNs) to build a mapping between a spatiotemporal HR feature image to its corresponding HR value. The feature map is constructed in a time-delayed way with noise-contaminated pulse signals extracted from existing rPPG methods. The CNN model is trained using transfer learning where images built from synthetic rPPG signals are taken to train the model first in order to generate initials for the practical one. The synthetic rPPG signals are interpolated from blood volume pulses or electrocardiograms through a modified Akima cubic Hermite interpolation. The proposed method is tested in both within-database and cross-database configurations on public databases. The results demonstrate that our method achieves overall the best performance compared to some other typical rPPG methods. The mean absolute error reaches 5.98 beats per minute and the mean error rate percentage is 7.97% in the cross-database testing on MAHNOB-HCI data set. Besides, some key factors that affect the performance of our method are also discussed which indicates potential ways for further improvements.**

***Index Terms*— Convolutional neural network, heart rate estimation, remote photoplethysmography, spatiotemporal representation, transfer learning.**

Manuscript received February 4, 2020; accepted March 17, 2020. Date of publication March 30, 2020; date of current version September 15, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61922075, Grant 81571760, and Grant 41901350 and in part by the Fundamental Research Funds for the Central Universities under Grant JZ2019HGTA0049 and Grant JZ2019HGBZ0151. The Associate Editor coordinating the review process was Dr. Zheng Liu. *(Corresponding author: Juan Cheng.)*

Rencheng Song, Senle Zhang, Chang Li, and Juan Cheng are with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: rcsong@hfut.edu.cn; zhangsenle@mail.hfut.edu.cn; changli@hfut.edu.cn; chengjuan@hfut.edu.cn).

Yunfei Zhang is with Senturing Technologies Ltd., Vancouver, BC V6T 1Z1, Canada, and also with Shenzhen ViWiStar Technologies Ltd., Shenzhen 518133, China (e-mail: yunfeizhang0616@gmail.com).

Xun Chen is with the Department of Electronic Engineering & Information Science, University of Science and Technology of China, Hefei 230026, China (e-mail: xunchen@ustc.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2020.2984168

I. Introduction

HEART rate (HR) is an important physiological indicator that reflects physical and mental status of human body. In general, HR can be measured by electrocardiography or photoplethysmography, which needs to employ specific sensors to contact with participant's skin. However, this may cause inconvenience and discomfort to patients especially for a long-term monitoring. In recent years, there is a growing interest from researchers to investigate noncontact HR measurement techniques [1]–[3]. The remote photoplethysmography (rPPG) is such a kind of video-based HR monitoring method, which detects pulsation from invisible facial color change caused by cardiac activity [2], [4].

Although the potentials of rPPG are promising, there are still many challenges. It is well known that the rPPG pulsation signal is very weak. The intensity or color spectrum changes of reflected light due to noise interference can easily dominate the pulsation signal. On the other hand, the conventional rPPG methods are usually designed under some assumption to simplify the noise reduction. For example, the blind source separation methods [5], [6] commonly assume that the sources should satisfy some statistical principles such as the independence. Similarly, the model-based methods [4], [7] are usually introduced based on a simplified skin optical reflection model under prior assumptions. However, the noise contaminated in practical applications are more complicated and diverse. This can break the assumptions of conventional rPPG methods and lead to unstable results.

In recent years, deep learning (DL) technique has achieved a great success in the fields of computer vision [8] and natural language processing [9]. Since DL is a data-driven method, the network is fit with a large amount of training data covering various real scenarios. In general, this characteristic ensures that the DL method is robust and flexible for practical applications, which indicates that it is very attractive to solve the rPPG problem under a DL framework. But many factors may affect the performance of DL methods, such as the design of input–output mapping, the network structure, and the selection of training data.

Inspired by the success of DL technique, a new rPPG method is introduced with convolutional neural networks (CNNs) to build the mapping between a spatiotemporal HR feature image to its corresponding HR value. The spatiotemporal

feature image is constructed with pulse signals which are extracted from conventional rPPG methods. The pulse is standardized and clipped into sections in a time-delayed way to assemble a feature map. This representation can capture both the morphological and chronological features of pulse signals. Therefore, the generated feature images have a specific and unified structure to be learned by a deep CNN. A revised ResNet-18 [10] is taken to map the spatiotemporal image to its corresponding HR value. Our approach is considered to integrate the advantages of conventional and DL methods.

We also evaluate some key factors that affect the performance of the proposed method. First, we introduce a transfer learning to train the model. The purpose is to reduce the demand of a large amount of training data. The model is firstly trained with images from synthetic rPPG signals and then refined with feature maps from real data. We construct synthetic rPPG signals by interpolating electrocardiogram (ECG) or blood volume pulse (BVP) signals through a modified Akima cubic Hermite interpolation [11]. The synthetic rPPG signals retain the heart rate variation (HRV) information of the real pulses. Second, we evaluate the benefit of using color feature images instead of the gray ones. The color feature images are constructed with rPPG signals extracted from three different skin regions of interest. The inherent spatial correlation of HR information is implied in the feature map. Third, we consider the issue of data imbalance. Namely, most HR distribution concentrate within the range of [60 90] beats per minute (bpm). This may degrade the prediction accuracy of HR out of this range. We take a balancing treatment to improve the prediction accuracy. Finally, we also compare the impact of constructing feature images using different types of rPPG signals. The feature images with less noise contaminated are expected to improve the accuracy of HR value prediction. By adopting all the above techniques, the proposed method achieves the state-of-the-art results in both within-database and cross-database tests.

We indicate that part of the materials of this article have been covered in our proceeding article [12] accepted by IEEE CIVEMSA 2019 [2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)]. The purpose of [12] is to validate the feasibility of the spatiotemporal representation. So Zhang *et al*. [12] only addresses the results of CNN model trained and tested with synthetic rPPG signals. The current article extends the conference paper in both methods and results. For methods, the rPPG pulse is extracted with the chrominance-based signal processing (CHROM) method [7] from real data instead of using synthetic ones. This article further constructs color feature image instead of the gray one and employs a novel transfer learning to train the model. For results, the proposed method is fully tested on public databases considering both within- and cross-database scenarios. The results are compared with other typical rPPG methods.

In summary, the main contributions of this article are threefold: 1) A novel spatiotemporal representation is introduced to construct color feature images with rPPG pulse signals extracted from conventional methods. The image reverses morphological and chronological features of pulse signals which are nice to be learned with CNN; 2) A new transfer learning scheme is proposed to pretrain the HR estimator. The training images are generated with synthetic rPPG signals which are interpolated from real ECG or BVP data through a modified Akima cubic Hermite interpolation; and 3) Factors that affect the performance of our method have also been evaluated. The intrinsic spatial correlation of HR from multiple skin regions of interest can improve the predicting accuracy through building multichannel images. The imbalance of HR distribution can degrade the performance of HR estimator. This can be mitigated through a consolidation of data from multiple databases with complementary HR distributions. High-quality rPPG signals can reduce the noise of feature images, thereby further improving the prediction accuracy of HR.

II. Related Work

A. Conventional Methods

In 2008, Verkruysse *et al.* [13] firstly evaluated the possibility of measuring HR remotely from facial videos. Since then, many researchers have devoted their efforts in rPPG research. Among them, a large class of methods are based on blind source separation which assume the sources are statistically independent. Poh *et al.* [5] applied independent component analysis (ICA) to demix pulse signal from raw RGB signals. In their follow-up work [14], they applied temporal filters to further improve the signal quality. Cheng *et al.* [15] introduced an independent vector analysis to eliminate illumination artifacts by extracting the common components of facial and background regions. Another major kind of methods are based on the skin optical reflection model. In order to overcome the motion artifacts, de Haan *et al.* [7] proposed a CHROM method. The RGB channels were projected into the chrominance subspace where the motion component was greatly eliminated. Wang *et al.* [4] introduced a different projection orthogonal to the skin tone to extract pulse.

These conventional methods have made outstanding contributions to rPPG development. However, they are usually designed for certain scenarios or under strong assumptions which may not be realistic in practical environment. On the other hand, although the performance of conventional methods may be suboptimal, they still remove part or most of noise from the original RGB channels. Therefore, it is worth of taking these methods as a preprocessing tool to simplify the complexity of mapping in DL-based methods.

B. DL-Based Methods

Recently, some articles have applied DL technique for rPPG-based HR estimation. Chen *et al.* [16] firstly proposed an end-to-end system to establish a mapping from a video frame contrast to the derivative of pulse. A soft-attention mask is learned simultaneously to improve the measurement quality. Niu *et al.* [17] introduced a novel spatiotemporal representation of cardiac information with RGB channels from multiple regions of interest. The spatiotemporal image is then mapped by a CNN to its HR value. In [18], they further improved their work with an attention mechanism to enhance the salient

features of rPPG signals. In another research, Qiu *et al.* [19] proposed a framework by estimating HR from spatial and temporal filtering feature maps with a CNN. The facial color change was magnified using Eulerian video magnification (EVM) [20] in order to enhance the signal-to-noise ratio. Špetlík *et al.* [21] designed an end-to-end HR estimator with a two-step neural network. The model was tested on three public databases and was proven to outperform the state-of-the-art methods. Different from the previous DL-based methods to estimate average HR value, Yu *et al.* [22] tried to extract the rPPG signal directly from raw video sequences with an end-to-end deep spatiotemporal convolutional network.

In short, the existing DL-based rPPG methods can be roughly divided into two categories, the feature-decoder methods and the end-to-end methods. Articles [17]–[19] belong to the feature-decoder methods which need to define handcrafted features. The performance of this kind of methods depends heavily on the quality of feature maps. For example, in [17], raw RGB signals from different ROIs are directly combined into a feature image. In [19], the authors obtain the feature image through a bandpass temporal filtering of a concatenated ROI subimages. Differently, the end-to-end methods [16], [21], [22] learn features directly by the network itself. However, this may require more training data to fit the network. Meanwhile, the resulting model of end-to-end method is often a black box which is difficult to interpret. On the other hand, some of the existing methods train the model with private databases [16], while others [17]–[19], [21], [22] use public databases as the sources of training data. Only references [16], [22], and [18] tested the model generalization capability in a cross-database way, while the others [17], [19], and [21] tested the proposed model within the same database.

This article intends to introduce a new feature-decoder method which integrates the advantages of conventional and DL methods. We propose a novel spatiotemporal representation with rPPG pulse signals extracted from the CHROM method. The generated feature map shows clear structures to be learned by CNN. It is considered to be more motion-resistant compared to the ones directly built from RGB channels. We also introduce a new transfer learning scheme to pretrain the HR estimator. Different from [17] where the transfer learning was conducted with synthetic pulses defined from some superimposed sinusoidal waves, we use standardized pulses which are interpolated from real ECG or BVP signals. This is more realistic to simulate true pulse signal because they retain the same HRV information. We test our method on public databases in both within- and cross-database settings to demonstrate its accuracy and generalization capability.

III. Method

The framework of the proposed method is illustrated in Fig. 1. Our method follows a feature-decoder model. Spatiotemporal feature images are firstly constructed using noise contaminated pulse signals extracted from conventional methods. Next, the feature maps are fed into a CNN to get a single predicted HR value corresponding to the ground truth of the pulse. Here ResNet-18 is chosen as the CNN model which is initialized with ImageNet pretraining. The ResNet-18

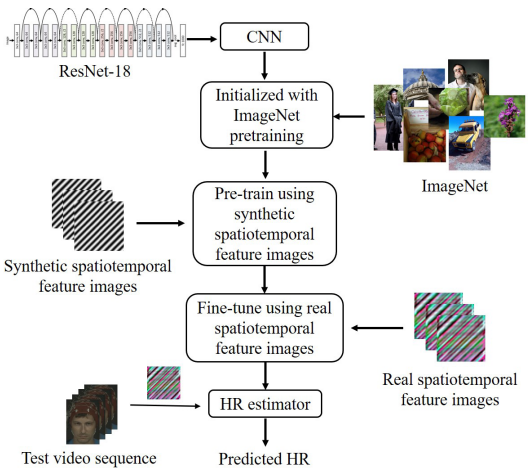

Fig. 1. The framework of the proposed HR estimation method using synthetic and real spatiotemporal feature images with CNN.

is trained in two steps by a transfer learning approach: 1) the model is firstly trained with spatiotemporal images built from synthetic rPPG pulses; 2) the spatiotemporal images corresponding to true rPPG pulses are taken to further refine the model. Finally, the obtained HR estimator is used to map the real spatiotemporal image to corresponding HR value in testing. The details of the method are introduced as follows.

A. Spatiotemporal Feature Map Construction

As shown in Fig. 2, a noise contaminated pulse can be extracted by conventional rPPG methods from each region of interest (ROI). We can construct a spatiotemporal feature map through a time-delayed way for each pulse signal. Suppose the signal $P = (p_1, p_2, \ldots, p_K)$ totally has K points and K is an even number. For rPPG applications, K is determined as the product of the length of processing window and the video frame rate. We take the first $K/2$ points and put them into the first row of a matrix. And in turn the second row is from the second point to the $(K/2+1)$th point, and so on. Therefore, a square Toeplitz matrix I with the size equal to $K/2$ is obtained, which satisfies

$$I = \begin{pmatrix} p_1 & p_2 & \cdots & p_{K/2} \\ p_2 & p_3 & \cdots & p_{K/2+1} \\ \vdots & \vdots & \cdots & \vdots \\ p_{K/2} & p_{K/2+1} & \cdots & p_{K-1} \end{pmatrix}.$$

The matrix I can be directly converted into a gray image. Since the input signal is quasiperiodic, the generated image has a clear structure as shown in Fig. 2. As we can see, periodic information is revealed in the vertical direction of the stripes. This indicates that the Toeplitz representation can illustrate the periodicity of a 1D signal in a 2D format. On the other hand, the Toeplitz representation is simple and it retains all the morphological and chronological information of the 1D signal. Therefore, the CNN can extract correct HR values from consistent input feature images.

Since a single-channel gray image can be created for each ROI, a color feature image is finally synthesized with all

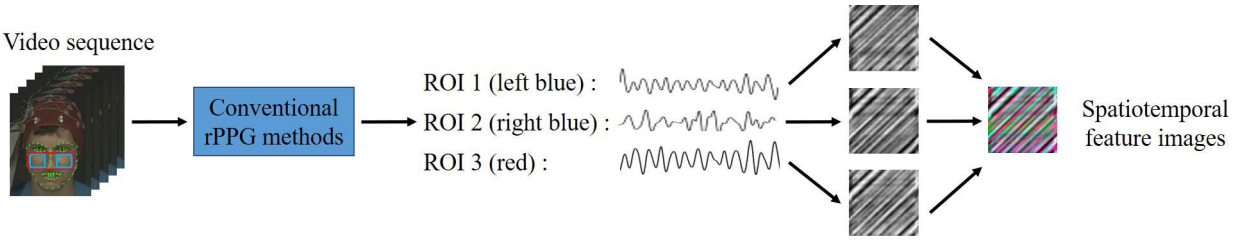

Fig. 2. Diagram of spatiotemporal feature map generation from video sequence.

gray images generated within the same processing window. Although more channels can be included in the feature image, here we choose three channels in our method considering the following two aspects. First, the selection of proper ROI has great impact on the quality of rPPG signals [23]. The ROI should be chosen with rich pulsation information and be less disturbed by nonrigid motions. Second, we need to take a balance between the spatial correlation of HR information and the amount of network parameters of CNN. Therefore, if more channels are included in the feature image, it may introduce extra noise and increase the difficulty of training the network. The three-channel color image already retains the intrinsic correlation information among different ROIs. The neural network can learn this property and predict the HR value more accurately. We will prove this through experiments later. If we repeat the feature map construction by processing videos from different time windows, a feature image data set can be created. For each image, the ground truth label is the average HR value in the same time window.

B. rPPG Pulse Extraction

Many conventional methods can be used to extract the rPPG pulse signals. Considering the efficiency and motion-resistant property of CHROM [7], we use this method as the pulse extractor. The CHROM method can be simply divided into four steps: (1) determine ROIs; (2) get RGB traces by averaging pixel values within each ROI; (3) derive chrominance signals from RGB traces; and (4) extract the pulse with alpha tuning.

According to [23], the cheek region contains rich pulsation information and it is less affected by nonrigid motion such as smile or wink. Hence, we choose the cheek area to extract the raw rPPG signals. A 68-point facial landmarks detection algorithm [24] is applied to accurately locate this area. Considering the spatial correlation of HR information, we define three ROIs based on the landmarks as shown in Fig. 2, where the left and right blue ones indicate ROI 1 and 2, respectively, and the red one represents ROI 3.

For each ROI, the raw RGB signals can be determined through a pixel averaging. According to the optical reflection model defined in [7] and [4], the observed raw RGB signals are linear mixtures of the motion-induced intensity signal, the pulse signal, and the specular reflection signal. A pair of orthogonal chrominance signals (X_s and Y_s) are then defined through a projection on the RGB signals to eliminate the motion-induced intensity signal. So the chrominance signals

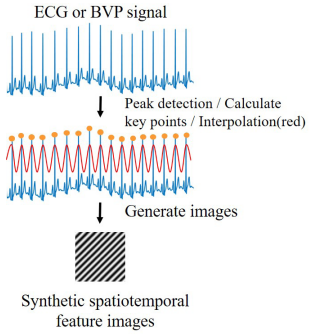

Fig. 3. The workflow of generating synthetic spatiotemporal feature images from real ECG or BVP signals.

are considered to be motion-resistant. After taking an alpha tuning on X_s and Y_s, the pulse signal is finally extracted as $S = X_f - \alpha Y_f$, where X_f and Y_f are bandpassed version of X_s and Y_s, and $\alpha = \sigma(X_f)/\sigma(Y_f)$. Here $\sigma(\cdot)$ indicates the standard deviation of $(\cdot)$. The readers can refer to [7] and [4] for more details. The extracted pulse is normalized to a $[-1, 1]$ range for further usage in constructing the spatiotemporal image.

C. Synthetic Pulse Generation

Since the proposed rPPG method is a data-driven method, a large amount of paired facial videos and pulse data are needed to train the CNN model. However, there are few available public databases including the desired data. Even for the qualified database the sample size is usually insufficient. Inspired by [17] and [25], synthetic data can be used to pretrain the model. Considering there are many ECG or BVP signals available in public databases we try to generate synthetic rPPG pulse by interpolating real ECG or BVP signals.

The procedure is as follows. Let us take an ECG signal for example. As shown in Fig. 3, we first perform a peak detection (to get N yellow nodes) on the ECG signal within a predefined processing window. The inter beat interval (IBI) sequence can then be obtained by calculating the difference of adjacent peaks, recorded as $[A_1, A_2, \ldots, A_{N-1}]$. Let $C^1 = [A_1/2, A_1, \ldots, A_{N-1}, (A_{N-1})/2]$. The top positions X of synthetic rPPG signal can be defined as a cumulative sum of C^1. Namely,

$$X_{i-1} = \sum_{j=1}^{i-1} C_j^1, \quad i = 2, \ldots, N+2 \tag{1}$$

where C_j^1 is the jth component of C^1.

Let $C^2 = 1/2[A_1, A_1 + A_2, A_2 + A_3, \ldots, A_{N-2} + A_{N-1}]$. Similarly, we can define the bottom Y of rPPG signal as a cumulative sum of C^2, i.e.,

$$Y_{i-1} = \frac{A_1}{2} + \sum_{j=1}^{i-1} C_j^2, \quad i = 2, \ldots, N \tag{2}$$

where C_j^2 is the jth component of C^2.

Next, the interpolation node sequence of a synthetic rPPG signal can be determined as

$$Z = \begin{pmatrix} 0 & X_1 & Y_1 & \cdots & X_{N-1} & Y_{N-1} & X_N & X_{N+1} \\ a_0 & 1 & -1 & \cdots & 1 & -1 & 1 & a_1 \end{pmatrix} \tag{3}$$

where the values at top and bottom positions are set as 1 and -1, respectively, a_0 and a_1 are two random numbers between -1 and 1.

The synthetic rPPG signal (denoted as a red curve in Fig. 3) is interpolated at Z with a modified Akima cubic Hermite interpolation method. Finally, the curve is resampled at a desired sampling rate to construct the synthetic spatiotemporal feature image. The whole workflow is as demonstrated in Fig. 3.

Different from [17], the synthetic rPPG pulse here is interpolated from real physiological signals instead of a combination of several sinusoidal waves. It is considered to retain the HRV information more accurately. During the training process, we did not introduce any artificial noise into the synthetic data. We hope to use the ideal synthetic feature data to pretrain the network parameters first. The real feature data contaminated by noise are then taken to fine-tune the model. This transfer learning strategy is employed to ensure the convergence of the training process. In short, the synthetic rPPG signal used here has following advantages: 1) it maintains the heart rate variability (HRV) information as the real physiology signals; 2) it standardizes pulse signals from different modalities (BVP or ECG); and 3) it can be used to pretrain neural networks in a transfer learning task.

D. HR Estimation Using CNN

In this article, the remote HR estimation is regarded as a regression problem. The residual neural network, more specifically, the ResNet-18 is selected as the CNN model. We replace the last layer of the network to predict a single HR value. The loss function is defined as a L_1 loss in the following equation:

$$\text{Loss} = \frac{1}{T} \sum_{i=1}^{T} |\text{HR}_{\text{predict}}(p_i) - \text{HR}_{\text{label}}(i)| \tag{4}$$

where T is the total number of samples, p_i is the ith spatiotemporal feature image, $\text{HR}_{\text{predict}}$ is the HR value predicted by CNN, and $\text{HR}_{\text{label}}(i)$ is the ground truth HR corresponding to p_i. The stochastic gradient descent (SGD) algorithm with momentum is employed as the optimizer.

As illustrated in Fig. 1, the parameters of ResNet-18 are initialized from an ImageNet pretraining to accelerate the convergence of iterations. The model is then trained in two steps through a transfer learning method. First, the synthetic spatiotemporal feature images are taken to train the CNN model. The neural network is guided to learn the mechanism between the feature map and the related HR value. Second, the model is further fine-tuned using real spatiotemporal feature images to bridge the gap from reality. The learned model is finally taken as a HR estimator for testing.

IV. Experiments

In this section, we will test the proposed method in following aspects: 1) evaluate the accuracy of the HR estimator with the condition-controlled MAHNOB-HCI database [26] in a within-database way-**Task 1**; 2) evaluate the accuracy of the HR estimator with the challenging realistic ECG-Fitness database [21] in a within-database way-**Task 2**; 3) evaluate the accuracy of the HR estimator with the MAHNOB-HCI database in a cross-database way-**Task 3**; and 4) evaluate the influence of some crucial factors in our method-**Ablation study**.

A. Experimental Settings

The experiment is conducted with public databases for both training and testing purposes. Four databases are involved including the MAHNOB-HCI database [26], the ECG-Fitness database [21], the VIPL-HR database [27], and the UBFC-RPPG database [28]. All the four databases are composed of synchronized videos and physiological signals. Particularly, the MAHNOB-HCI database provides only ECG signals, the VIPL-HR database and the UBFC-RPPG database provide only BVP signals, and the ECG-Fitness database provides both.

We take a 5-s window to process all video and physiological data. The adjacent windows are defined using the forward sliding with a one second step. To eliminate the differences among data sets, the sampling rates of rPPG-based pulse and ground-truth physiological signal are all resampled to 30 Hz. The frequency range of bandpass filtering is set to [0.7, 3] Hz. All the generated synthetic and real feature images are up-sampled to 224×224 before being fed to the CNN. The details of each data set and the corresponding training/testing samples are introduced next.

1) Databases: The MAHNOB-HCI database consists of 527 videos in total under a well-controlled environment, where the subjects are required to be stationary in most times. The 15 female and 12 male participants are involved with ages varying between 19 and 40 years old. We took the first 60 s of each video and its corresponding ECG signal to generate feature maps. A total of 26520 spatiotemporal images were constructed from the real and synthetic rPPG pulses, respectively. In task 1, part of the samples were used as the training data and the other were used as the testing data. According to whether there are repeated subjects between the training and testing data, the experiment is further divided into two cases, i.e., the subject-dependent and the subject-independent cases. We compared the performance of the proposed method with some existing ones for both cases in task 1. All the samples of MAHNOB-HCI database were used as the testing data in task 3.

The ECG-Fitness database [21] is a very challenging data set for remote HR measurement task, in which the subjects

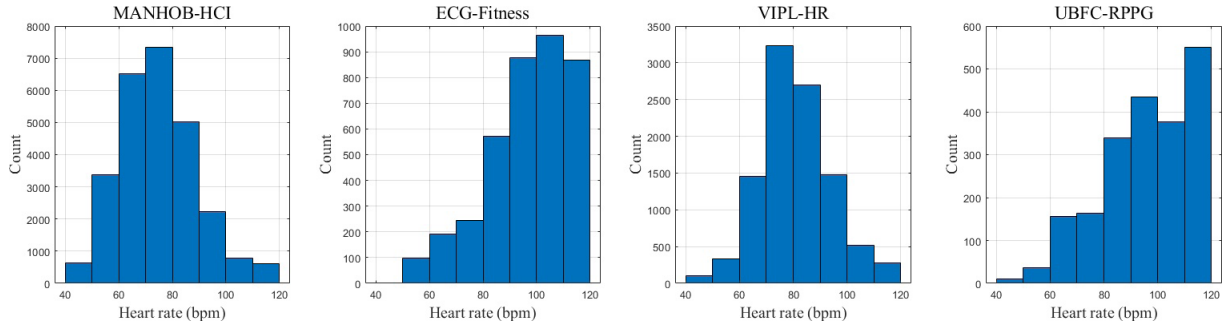

Fig. 4. The HR distributions of ground truth pulses in MANHOB-HCI, ECG-fitness, VIPL-HR, and UBFC-RPPG databases.

performed physical activities on fitness machines under different lighting setups. There are 17 subjects (14 male and 3 female) with ages ranging from 20 to 53 years. They performed four different activities such as speaking, rowing, and riding on a stationary bike and on an elliptical trainer. Each scene was recorded by two Logitech C920 web cameras. We only use 96 videos captured by the one positioned on a tripod to achieve a better video stabilization. According to the processing window and sliding setting, we totally got 4506 spatiotemporal images from the videos, and 5144 synthetic spatiotemporal images from the BVP signals in this data set. All the samples in the ECG-Fitness database were taken for a subject-independent test within the database in task 2.

The VIPL-HR database is a recently released database for remote HR measurement. It covers several typical rPPG application scenarios and therefore is a good candidate to train the model. There are 2378 visible light videos and 752 near-infrared videos from 107 subjects (79 males and 28 females, the ages are between 22 and 41 years old) under diverse situations. Only visible light videos were used in the experiments. Most videos from VIPL-HR database are very short, even less than 30 s. According to the processing window and sliding setting, we totally got 10 282 spatiotemporal images from videos, and 17 546 synthetic spatiotemporal images from BVP signals in this data set. The reason that the number of synthetic samples is much more than that of the real ones is because the BVP signals were recorded in a much longer time than the videos in this data set. All the samples in VIPL-HR database were taken as training data in task 3.

The UBFC-RPPG database is also publicly released for rPPG analysis, which includes two scenarios, the simple and realistic situations, with a total of 50 videos (8 for simple and 42 for realistic situations, age information is absent). We only use the 42 videos from the realistic situation. The subjects were required to play a time-sensitive mathematical game in order to make their heart beat change. According to the same processing window and sliding settings, 2217 real feature maps were generated from videos, while a total of 2272 synthetic ones were constructed from BVP signals. All the samples in UBFC-RPPG database were taken as training data in task 3.

The HR distributions of ground truth pulses in the four databases are shown in Fig. 4. As we can see, most of the samples in the MAHNOB-HCI and VIPL-HR databases are concentrated within 60–90 bpm, while most of the HR values of the ECG-Fitness and UBFC-RPPG databases fall into the range from 80 to 120 bpm.

2) Metrics: Several quality metrics are employed in evaluation as below.

1) the standard deviation $\mathrm{HR_{sd}}$:

$$\mathrm{HR_{sd}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{HR}_e^{(i)} - \overline{\mathrm{HR}_e}\right)^2} \tag{5}$$

where $\mathrm{HR}_e^{(i)} = \mathrm{HR}_{\mathrm{predict}}^{(i)} - \mathrm{HR}_{\mathrm{label}}^{(i)}$ is the error of HR for the ith sample, and $\overline{\mathrm{HR}_e}$ indicates the mean value of the HR_e vector;

2) the mean absolute error $\mathrm{HR_{mae}}$:

$$\mathrm{HR_{mae}} = \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{HR}_{\mathrm{predict}}^{(i)} - \mathrm{HR}_{\mathrm{label}}^{(i)}\right| \tag{6}$$

3) the root mean square error $\mathrm{HR_{rmse}}$:

$$\mathrm{HR_{rmse}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{HR}_{\mathrm{predict}}^{(i)} - \mathrm{HR}_{\mathrm{label}}^{(i)}\right)^2} \tag{7}$$

4) the mean error rate percentage $\mathrm{HR_{mer}}$:

$$\mathrm{HR_{mer}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\mathrm{HR}_{\mathrm{predict}}^{(i)} - \mathrm{HR}_{\mathrm{label}}^{(i)}\right|}{\mathrm{HR}_{\mathrm{label}}^{(i)}} \times 100\% \tag{8}$$

5) Pearson's correlation coefficient r:

$$r = \frac{\sum_{i=1}^{n}\left(X^{(i)} - \overline{X}\right)\left(Y^{(i)} - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X^{(i)} - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y^{(i)} - \overline{Y}\right)^2}} \tag{9}$$

where $X^{(i)}$ indicates the $\mathrm{HR}_{\mathrm{label}}^{(i)}$, $Y^{(i)}$ represents the $\mathrm{HR}_{\mathrm{predict}}^{(i)}$, and the symbol $\overline{X}$ is the mean value of X vector.

B. HR Estimation Results

Unless specified, all subsequent studies trained the model following the flowchart in Fig. 1. The learning rate was set as 10^{-3} in the pre-training with synthetic feature images for the first 10 epochs and then 10^{-4} for the next 30 epochs. The model was further fine-tuned with real spatiotemporal feature images using the same learning rate as 10^{-4} for another 40 epochs to get the final HR estimator.

1) Task 1: We first evaluate the performance of our method on the well-controlled MAHNOB-HCI database in a within-database way. Namely, the training and testing data are both taken from the same database.

For the subject-dependent case, where the subjects in the training set can be the same with those in the testing set,

TABLE I

THE SUMMARY OF AVERAGED HR RESULTS FOR MAHNOB-HCI DATA SET: A WITHIN-DATABASE CASE

Method	HR_{sd}(bpm)	HR_{mae}(bpm)	HR_{rmse}(bpm)	HR_{mer}	r
EVM-CNN (subject-dependent) [19]	**2.79**	-	3.26	3.67%	0.95
Proposed (subject-dependent)	3.19	1.53	**3.23**	**2.22%**	**0.97**
SynRhythm (subject-independent) [17]	10.88	-	11.08	12.26%	-
Proposed (subject-independent)	**5.57**	4.61	**5.70**	**5.67%**	0.86

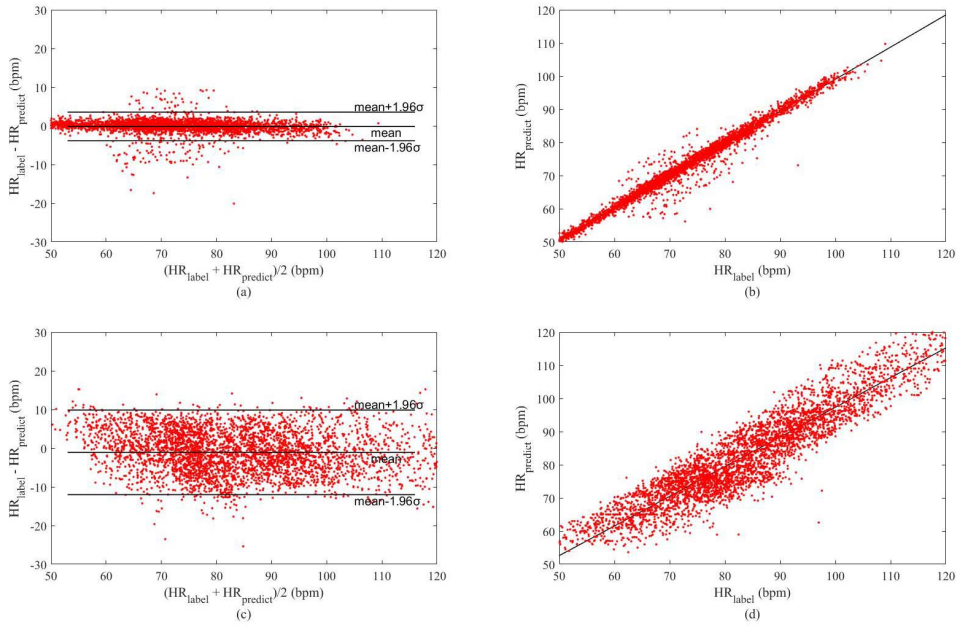

Fig. 5. Bland-Altman plots (left) and scatter plots (right) between the predicted $\text{HR}_{\text{predict}}$ and the ground truth HR_{label}: a within-database case. The (a) and (b) correspond to the subject-dependent case. The (c) and (d) correspond to the subject-independent case.

TABLE II

THE SUMMARY OF AVERAGED HR RESULTS FOR ECG-FITNESS DATA SET: A WITHIN-DATABASE CASE

Method	HR_{sd}(bpm)	HR_{mae}(bpm)	HR_{rmse}(bpm)	HR_{mer}	r
HR-CNN [21]	-	14.48	19.15	-	0.50
Proposed	12.75	**10.34**	**12.99**	9.85%	**0.57**

we compare our method with the state-of-the-art EVM-CNN in [19]. The training and testing data were selected using the same way as that of EVM-CNN. Specifically, half of the video sequences were randomly chosen as the training set and all video sequences were employed for testing. The comparison results are shown in Table I, where the best results are highlighted in bold. It can be observed that the metrics derived from the proposed model are comparable to those of EVM-CNN.

For the subject-independent case, where the subjects in the training set are different from those in the testing set, we take a threefold cross validation similar to [17] on all samples of MAHNOB-HCI database to test the performance of our method. The results of our method are also listed in Table I to compare with the SynRhythm in [17]. It can be observed that our method consistently outperforms the SynRhythm algorithm.

The Bland-Altman and scatter plots of the within-database testing are illustrated in Fig. 5. It shows that the predicted HR and the corresponding ground truth have a good consistency in all HR distributions, especially for the subject-dependent case.

2) Task 2: Considering that most of the subjects in the MAHNOB-HCI database are stationary, we further test the proposed method on another more challenging database ECG-Fitness [21], where the subjects performed physical activities under different lighting conditions. In order to compare, we used a similar experimental setup as [21], where 72 videos of 12 subjects were taken for training and 24 videos from another 4 subjects were used for testing. The experimental results are shown in Table II. We can see that the performance of the proposed method outperforms the HR-CNN in [21]. We should point out that the results of HR-CNN were obtained using videos from both cameras, while we only use those videos from the camera fixed on the tripod. Although our method achieves better results, there is still a big room for further improvement in such realistic scenario.

We also show the statistical histogram of errors in Fig. 6 to demonstrate the effectiveness of the nonlinear mapping approximated by neural network. It is observed that the

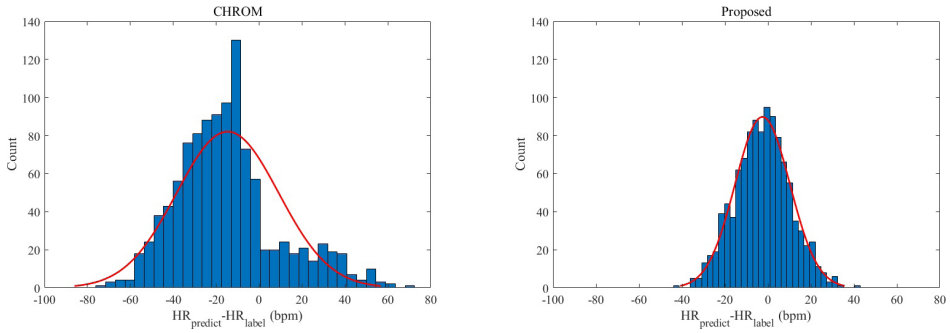

Fig. 6. The statistical histograms of errors of predicted HR on the ECG-Fitness database. The figure on the left is the result obtained by the CHROM algorithm. The figure on the right is the result obtained by the proposed method.

TABLE III
THE SUMMARY OF AVERAGED HR RESULTS FOR MAHNOB-HCI DATABASE: A CROSS-DATABASE CASE

Method	HR_{sd}(bpm)	HR_{mae}(bpm)	HR_{rmse}(bpm)	HR_{mer}	r
Poh *et al.* [5]	14.01	10.56	14.08	14.86%	0.64
De Haan *et al.* [7]	9.91	7.15	8.96	10.26%	0.72
Wang *et al.* [4]	10.51	8.24	9.51	11.70%	0.71
Verkruysse *et al.* [13]	20.08	15.57	20.09	21.82%	0.54
Balakrishnan *et al.* [29]	19.42	14.86	18.51	20.88%	0.55
DeepPhys [16]	-	**4.57**	-	-	-
HR-CNN [21]	-	7.25	9.24	-	0.51
PhysNet128 [22]	8.75	6.85	8.76	-	0.69
Proposed	**7.31**	5.98	**7.45**	**7.97%**	**0.75**

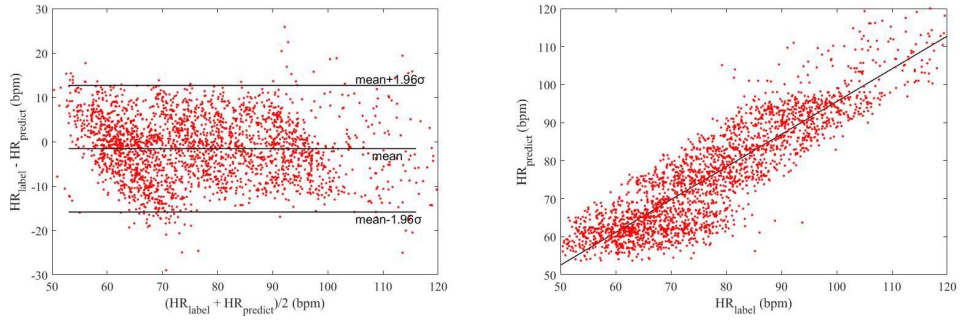

Fig. 7. Bland-Altman plot (left) and scatter plot (right) between the predicted $\mathrm{HR}_{\mathrm{predict}}$ and the ground truth $\mathrm{HR}_{\mathrm{label}}$: a cross-database case.

proposed method has clearly smaller errors compared to the CHROM. Hence, our approach can predict the HR value more accurately than the input rPPG signal. It proves that the CNN effectively builds the mapping between the input feature image and its corresponding ground truth HR value.

3) Task 3: To fully evaluate the generalization capability of the proposed method, we further took a cross-data set testing. The samples in VIPL-HR and UBFC-RPPG were taken as the training data, while the samples from MAHNOB-HCI data set were used as the testing data. Several conventional methods [4], [5], [7], [13], [29] as well as the DL-based rPPG methods [16], [21], [22] were compared. The implementation of conventional methods borrowed the code from the MATLAB toolbox iPhys [30] developed by McDuff. The same 5-s processing window was used in these conventional methods. The results of DL-based methods were directly taken from corresponding articles due to the complexity of implementations. Therefore, the comparison of DL-based methods in Table III may not be fair since their settings are not exactly the same. However, it still can partially indicate the performance of the proposed method. The full comparison results are listed in Table III.

From the results, we can see that our method achieves state-of-the-art performance compared to other methods except the convolutional attention networks (CAN) method in [16]. The $\mathrm{HR}_{\mathrm{mae}}$ of CAN in [16] is 4.57 bpm, which is a little better than 5.98 bpm of our method. The Bland-Altman and scatter plots are illustrated in Fig. 7, from which we can observe a good match between the predicted and the ground truth HR values. To avoid excessive data overlap, it is worth noting that only part of the samples (10%) were demonstrated in Fig. 7, which were evenly selected among the total 26 520 spatiotemporal images.

To demonstrate the stability of our method for a continuous HR monitoring, four 60-s predicting curves (red) are shown in Fig. 8 to compare with their ground truth curves (blue).

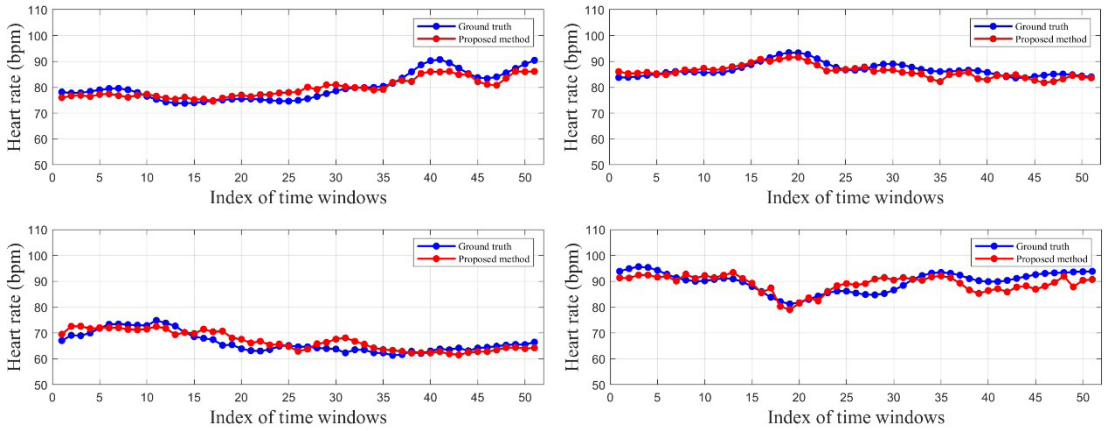

Fig. 8. Continuous HR monitoring of four 1-min sequences. The blue curve represents the ground truth HR values and the red curve indicates the predicted HR values by CNN.

The horizontal axis represents the indices of processing windows and the vertical axis indicates the HR values. The four examples were taken from different subjects with diverse HR distributions. We can see that the predicted HR values closely follow the trend of ground truth HR, even in some challenging situations.

C. Ablation Study

In this section, we further analyze four essential factors that may affect the performance our proposed method, including the balancing treatment on HR distribution of training samples (factor 1), the use of color feature map instead of the gray one (factor 2), the employment of transfer learning with synthetic data (factor 3), and the use of CHROM signal instead of the green one to generate feature maps (factor 4). Correspondingly, the following configurations will be tested, where the bold highlights the default settings of our method:

1) **Include**/remove UBFC samples in the training set;
2) **Use color**/gray feature map;
3) **Employ**/remove pretraining in the model training process;
4) **Use CHROM**/green signals to generate feature maps.

Suppose $(1, 2, 3, 4)$ means to use all the four default settings in the proposed method, while the absence of one of them indicates to use an alternative setting. For example, $(1, 2, 4)$ means the method is tested without using synthetic samples for pretraining. If not specified, other settings remain the same except the above four factors. All the following evaluations are only taken under a cross-data set testing as task 3.

As we can see in Fig. 4, the HR distributions are not uniform in the four data sets. To assess factor 1, we compare the results of our method using configurations $(2, 3, 4)$ and $(1, 2, 3, 4)$, respectively. The results are shown in Table IV. It can be seen that the results degrade if we remove UBFC samples from the training data. To understand the reason in detail, we further calculated the measurement error HR_{mae} in different HR ranges, as shown in Fig. 9. Obviously, the use of UBFC training data clearly improves the error in the range of 100–120 bpm, which is consistent with the HR distribution of UBFC samples. This verifies that the imbalance of training data will affect the performance of our method. We can improve this by a balancing treatment to prepare training data with a more uniform HR distribution.

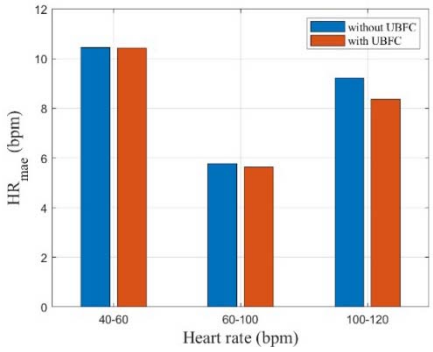

Fig. 9. The comparison of HR estimation accuracy (HR_{mae}) with and without using the UBFC samples in training data.

Next, we analyze the impact of using color feature image. We selected the ROI 3 in Fig. 2 to generate grayscale feature images, while all three ROIs were used to construct color feature images. The color feature image implies the spatial correlation characteristic of HR information. To assess factor 2, the results of using grayscale $(1, 3, 4)$ and color feature maps $(1, 2, 3, 4)$ were compared accordingly. The experimental results are shown in Table IV.

From the results, we can clearly observe that the results using color images are better than those with grayscale ones. This verifies that incorporating spatial correlation information of HR in the feature map will help to improve the results. The neural network learns the consistency of HR from the depth of color image, thereby increasing the prediction accuracy.

We further evaluate the influence of pretraining with synthetic data in the training process. To access factor 3, we compare the results of our method using configurations $(1, 2, 4)$ and $(1, 2, 3, 4)$, respectively. The results are shown in Table IV. It can be seen that the performance degrades if the pretraining is removed. Theoretically, the pretraining can guide the neural network to learn the decoder mapping between the feature

TABLE IV
THE EVALUATION OF THE PROPOSED METHOD ON MAHNOB-HCI DATABASE WITH A CROSS-DATABASE TESTING

Method	HR_{sd}(bpm)	HR_{mae}(bpm)	HR_{rmse}(bpm)	HR_{mer}	r
(1,2,3)	11.09	7.83	10.36	12.11%	0.67
(1,2,4)	8.09	6.35	8.17	8.11%	0.74
(1,3,4)	8.51	6.89	8.78	8.36%	0.73
(2,3,4)	8.02	6.57	8.11	9.34%	0.74
(1,2,3,4)	**7.31**	**5.98**	**7.45**	**7.97%**	**0.75**

Fig. 10. Comparison of feature images generated by different signals: (a) BVP signal, (b) raw green signal, and (c) CHROM signal.

image and its ground truth HR. The training with real ones further refines this mapping when noise is contaminated. Therefore, the pretraining reduces the risk of trapping in local minima and improves the solution quality.

Finally, we evaluate the influence of different input rPPG signals in the proposed method. To access factor 4, we also generated feature images using the green signals instead of the default CHROM ones. All other settings were kept consistent. The results are shown in Table IV. Obviously, the results with CHROM signals as feature sources outperform those from green ones.

To show more details, we also compare the generated feature images using the green and CHROM signals, respectively, as shown in Fig. 10. The figures were generated using a video clip of subject 27 from the UBFC data set. In this video clip, rigid motion (head rotation) and nonrigid motion (pout and wink) are both contaminated. Therefore, the original RGB signals are distorted by motion artifacts. Compare to the feature image in Fig. 10(b) generated by the green signal, the one in Fig. 10(c) from the CHROM signal is much more similar to that of Fig. 10(a) by the BVP signal. It indicates that the feature map generated by the CHROM signal is more motion-resistant compared to the one with raw RGB signals. Hence, it can reduce the difficulty to train the network. The comparison results in Table IV also verify this.

V. CONCLUSION

In this article, we presented a new rPPG method based on CNN for remote HR estimation from facial videos. The method takes a feature-decoder framework to map the HR feature image to the corresponding HR value through a ResNet-18 network. The spatiotemporal feature images are constructed in a time-delayed way using pulse signals extracted from conventional rPPG methods. The CNN model was firstly trained with synthetic feature images derived from ECG or BVP signals. It is then further refined with real feature images generated from noise contaminated rPPG pulses. We have taken both within-database and cross-database studies to fully investigate the accuracy and generalization capability of the proposed method. Experimental results demonstrate that our method gets overall state-of-the-art results on the public MAHNOB-HCI and ECG-Fitness databases compared to conventional and DL-based methods. We also evaluated some key factors that may affect the performance of our method. It indicates that the balancing treatment of training samples, the use of color feature maps, the pretraining with synthetic feature maps, and the adoption of high-quality rPPG signals as feature map sources are all necessary to improve the HR prediction accuracy of the proposed method.

REFERENCES

[1] J. Kranjec, S. Begus, J. Drnovsek, and G. Gersak, "Novel methods for noncontact heart rate measurement: A feasibility study," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 4, pp. 838–847, Apr. 2014.
[2] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, "Video-based heart rate measurement: Recent advances and future prospects," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3600–3615, Oct. 2019.
[3] Y.-C. Lin, Y.-J. Wang, J. C.-H. Cheng, and Y.-H. Lin, "Contactless monitoring of pulse rate and eye movement for uveal melanoma patients undergoing radiation therapy," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 2, pp. 474–482, Feb. 2019.
[4] 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, Jul. 2017.
[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, no. 10, p. 10762, May 2010.
[6] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3640–3648.
[7] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.
[8] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Feb. 2018.
[9] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[11] H. Akima, "A new method of interpolation and smooth curve fitting based on local procedures," *J. ACM*, vol. 17, no. 4, pp. 589–602, Oct. 1970.
[12] S. Zhang, R. Song, J. Cheng, Y. Zhang, and X. Chen, "A feasibility study of a video-based heart rate estimation method with convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. (CIVEMSA)*, Tianjin, China, Jun. 2019.
[13] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Opt. Express*, vol. 16, no. 26, pp. 21434–21445, Dec. 2008.
[14] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in non-contact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[15] J. Cheng, X. Chen, L. Xu, and Z. J. Wang, "Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 5, pp. 1422–1433, Sep. 2017.

[16] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 349–365.

[17] X. Niu, H. Han, S. Shan, and X. Chen, "SynRhythm: Learning a deep heart rate estimator from general to specific," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3580–3585.

[18] X. Niu *et al.*, "Robust remote heart rate estimation from face utilizing spatial-temporal attention," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.

[19] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, and A. E. Saddik, "EVM-CNN: Real-time contactless heart rate estimation from facial video," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1778–1787, Jul. 2019.

[20] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, Aug. 2012.

[21] R. Špetlík, V. Franc, J. Čech, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proc. Brit. Mach. Vis. Conf., Brit. Mach. Vis. Assoc. (BMVA)*, 2018, pp. 3–6.

[22] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," 2019, *arXiv:1905.02419*. [Online]. Available: http://arxiv.org/abs/1905.02419

[23] S. Kwon, J. Kim, D. Lee, and K. Park, "ROI analysis for remote photoplethysmography on facial video," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 4938–4941.

[24] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 94–108.

[25] J. Tremblay *et al.*, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 969–977.

[26] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[27] 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. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 562–576.

[28] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognit. Lett.*, vol. 124, pp. 82–90, Jun. 2019.

[29] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3430–3437.

[30] D. McDuff and E. Blackford, "IPhys: An open non-contact imaging-based physiological measurement toolbox," 2019, *arXiv:1901.04366*. [Online]. Available: http://arxiv.org/abs/1901.