Robust Remote Heart Rate Estimation from Face Utilizing Spatial-temporal Attention

Xuesong Niu[1,3,†], Xingyuan Zhao[5,†], Hu Han[1,2], Abhijit Das[6], Antitza Dantcheva[6], Shiguang Shan[1,2,3,4], and Xilin Chen[1,3]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

[2] Peng Cheng Laboratory, Shenzhen, China

[3] University of Chinese Academy of Sciences, Beijing 100049, China

[4] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China

[5] Zhejiang University, Zhejiang, China

[6] Inria, Sophia Antipolis, France

xuesong.niu@vipl.ict.ac.cn, xyzhao1@zju.edu.cn, {hanhu, sgshan, xlchen}@ict.ac.cn, {abhijit.das, antitza.dantcheva}@inria.fr

***Abstract*— In this work, we propose an end-to-end approach for robust remote heart rate (HR) measurement gleaned from facial videos. Specifically the approach is based on remote photoplethysmography (rPPG), which constitutes a pulse triggered perceivable chromatic variation, sensed in RGB-face videos. Consequently, rPPGs can be affected in less-constrained settings. To unpin the shortcoming, the proposed algorithm utilizes a spatio-temporal attention mechanism, which places focus on the salient features included in rPPG-signals. In addition, we propose an effective rPPG augmentation approach, generating multiple rPPG signals with varying HRs from a single face video. Experimental results on the public datasets VIPL-HR and MMSE-HR show that the proposed method outperforms state-of-the-art algorithms in remote HR estimation.**

I. Introduction

Computer vision based human health monitoring has attracted increased research attention due to the benefits such monitoring techniques offer. Recently, remote rPPG technologies, which allow for non-intrusive, efficient and low-cost measurements of a plethora of fundamental health signs such as heart-rate, heart-rate variability, and respiration, have shown great potentiality in many applications such as elderly care and mental state analysis [1], [2] (see Fig. 1).

The HR signal captured using rPPG technology is based on pulse-induced subtle color variations of human skin using RGB video data. Specifically, the pulsatile blood propagation in the human cardiovascular system changes the blood volume in skin tissue. The oxygenated blood circulation leads to fluctuations in the amount of hemoglobin molecules and proteins thereby causing variations in the optical absorption and scattering across the light spectrum [3], [4]. Such information captured by visible cameras can be instrumental to identify the phase of the blood circulation based on minor changes in skin reflections and hence among other the information of HR.

† Equal contribution. This work was done when X. Zhao was an intern at ICT, CAS.

978-1-7281-0089-0/19/$31.00

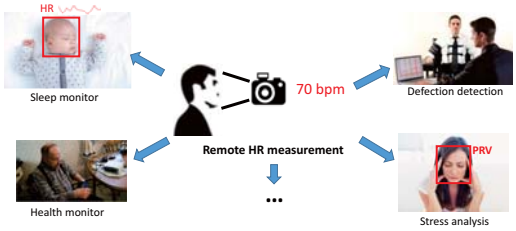

Fig. 1. Vision based remote HR measurement has various applications, such as sleep and health monitoring, defection detection, and stress analysis.

While rPPG based HR measurement has shown promising results under controlled conditions [5], [6], [7], [8], less-constrained settings such as non-homogeneous illumination and the movement of the human pose remaining challenges in this context. At the same time, data driven methods, especially deep learning methods, have shown great modeling power and have achieved great success in many other applications such as object detection [9], image classification [10], as well as face recognition [11]. Several works have successfully leveraged the strong modeling ability of deep neural networks to the task of remote HR estimation [12], [13], [14].

However, all existing methods predominantly focused on learning a mapping function from the representation of face videos to the ground-truth HR, and failed to take the characteristics of rPPG based HR signal into consideration. As stated in Fig. 2, rPPG signals can produce biased result due to face movement and illumination lighting variations. This biasness will introduce great noise and greatly influence the learning producer. An automatic mechanism that can help to remove these polluted HR signals is required.

To mitigate the aforementioned gap in this paper, we propose an end-to-end approach for remote HR measurement from faces utilizing a *channel and spatial-temporal attention mechanism*. We first utilize the spatial-temporal

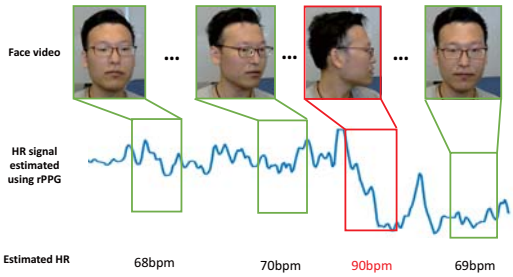

Fig. 2. For remote HR estimation under less-constrained situations, the estimated HR signal from a face video is highly influenced by various conditions such as head movements, illumination changes, etc. Such noises in the extracted HR signal (indicated with red rectangles) will lead to poor HR estimation accuracy. If the HR estimator can focus on the stable part of the signal, the HR estimation accuracy can be improved.

map proposed in [13] to get the effective representation of the HR signals from face videos. Then all spatial-temporal maps are utilized for training the HR estimator based on the attention mechanism. With the attention mechanism, we are able to filter the salient features in video sequences and hence learn a rich representation targeted to extract the substantial visual characteristics of rPPG-based for HR measurement. At the same time, in order to overcome the limitation of the training database size, we further propose an effective data augmentation method specifically designed for face video based remote HR estimation.

The main contributions of this work are three-fold: i) we propose a novel channel and spatial-temporal attention mechanism for the robust HR estimator to focus on the informative part of the video; ii) we design a novel data augmentation method to overcome the limitation of training data; and iii) we achieve the state-of-the-art performance on the public domain databases (VIPL-HR and MMSE-HR).

II. Related Work

In this section, we briefly review existing works on remote HR estimation and deep representation learning utilizing attention.

A. rPPG-based HR Estimation

The possibility of performing PPG analysis from videos captured with commodity cameras was firstly proposed by Verkruysse *et al.* [15]. In recent years, tremendous progress has been made in improving the robustness of rPPG based HR measurement, which can be grouped in three main categories: (a) blind signal separation (BSS) based methods, (b) optical model based methods, and (c) data-driven methods.

Blind signal separation (BSS) based methods were firstly introduced by Poh *et al.* [5], [6] and employ independent component analysis (ICA) to temporally filter red, green, and blue (RGB) color channel signals to seek for heartbeat-related signal, which they assumed is one of the separated independent components. In a following work [16], a patch-level-based HR signal calculation using ICA was performed and achieved the state-of-the-art result on the public MAHNOB-HCI HR measurement database [17].

Optical model based methods focus on leveraging the prior knowledge of the skin optical model to perform remote HR estimation. Haan and Jeanne [18] firstly proposed a skin optical model considering different color channels under the motion condition and computed a chrominance signal using the combination of RGB signals to reduce the motion noise. In a later work of [8], pixel-wise chrominance features are computed and used for HR estimation. A detailed discussion of different skin optical model used for rPPG-based HR estimation was summarized in [19], and the authors also proposed a new projection method for the original RGB signals in order to extract better HR signals. In [20], Niu *et al.* further improved the chrominance feature representation [18] to perform continuous HR estimation.

Data-driven methods aim at leveraging big training data to perform remote HR estimation. Tulyakov *et al.* [21] divided the face into multiple regions of interest (ROI) to obtain a matrix of temporal representation and used a matrix completion approach to purify the HR signals. Hsu *et al.* [14] generated time-frequency maps from different color signals and used them to learn an HR estimator. Although the existing data-driven approaches attempted to build HR estimator via learning, they failed to build an end-to-end estimator. Besides, the features they used remain hand-crafted, which may not be optimum for the HR estimation task. In [12], Niu *et al.* proposed a general-to-specific learning strategy to obtain a HR signal representation encoding both spatial and temporal cues, and solved the problem of insufficient training data.

B. Representation Learning Utilizing Attention

Attention is a pertinent mechanism in human perception [22]. It has been shown that the *human visual system* is able to sense the general information of a scene by combining several glimpses of areas, instead of processing the whole scene at once [23]. Recently, great efforts have been made to utilize such a similar attention mechanism in different computer vision tasks, such as image classification [24], [25], [26], and object detection [24]. In this context, attention aims to discover the most informative parts of an image, which are then processed in detail.

Spatial attention is the most widely used attention mechanism. Mnih *et al.* [27] proposed an attention algorithm for automatically selecting a sequence of regions that will be processed in detail. Chen *et al.* [28] applied attention mechanism to softly weigh features of different scales, and achieved the best performance in pixel-level classification tasks such as image segmentation. Wang *et al.* [25] proposed the Residual Attention Network, which is an hourglass style module, inserted into a deep residual network. This network can generate an attention map covering every pixel location and perform well in image classification.

Deviating from spatial attention, more recent methods focus on *channel-wise attention*. Hu *et al.* [26] proposed a Squeeze-and-Excitation module for modeling channel-wise relationships and enhancing the feature representation ability. Woo *et al.* [24] considered jointly spatial attention and

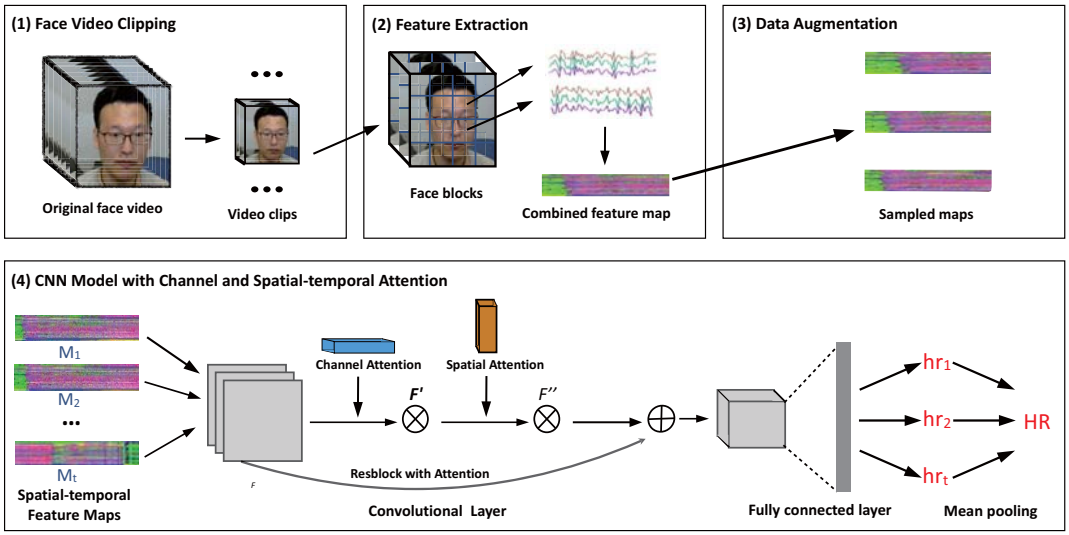

Fig. 3. Overview of the proposed end-to-end trainable approach for rPPG based remote HR measurement via representation learning with spatial-temporal attention.

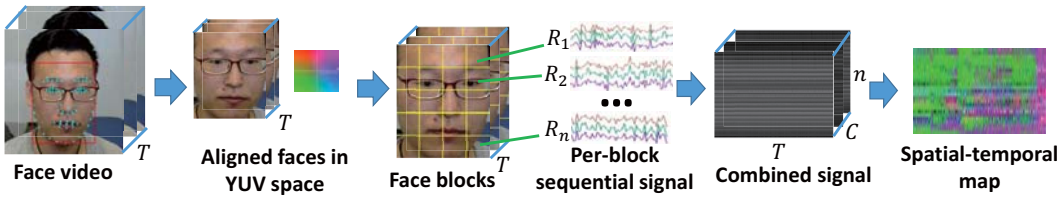

Fig. 4. An illustration of the spatial-temporal map generation procedure that was proposed in [13].

channel-wise attention and proposed a convolutional block attention module (CBAM) for generating visual features.

While a number of attention based feature learning methods are available, their effective in handling very weak visual signal such as HR is not known.

III. Proposed Approach

In the section we proceed to describe the proposed spatial-temporal map generation with attention mechanism, the network architecture, and data augmentation.

A. Spatial-temporal Map Generation

The spatial-temporal map was firstly introduced by Niu *et al*. [13] for representing the HR signals of a short video clip. As illustrated in [13], Niu *et al*. first utilized the face detector provided by the open source SeetaFace[1] to obtain the face location and 81 facial landmarks. A moving average filter was applied to the 81 landmark points to get more stable landmark localization across video frames. After that, face alignment was performed based on the two eye centers. The bounding box with the size of $w \times 1.2h$ was defined using the horizontal distance w between the outer cheek border points and the vertical distance h between chin location and eye center points. Skin segmentation was then applied to the pre-defined ROI to remove the non-face area such as the eye region and the background area.

[1] https://github.com/seetaface/SeetaFaceEngine

After the face was aligned and cropped from every frame, the face images were then divided into n grids and transformed to YUV color space. The average of the pixel values of each grid was calculated, and then concatenated into a sequence of T for C channels. The n grids were directly placed into rows, and the final spatial-temporal map was in the size of $n \times T \times C$. A detailed spatial-temporal map generation producer is illustrated in Fig. 4.

B. Network Architecture

Our proposed network architecture is shown in Fig. 3. We firstly generate small video clips sampled from the original video using a fixed sliding window. Then, the spatial-temporal feature maps of each video clip are computed, and the data augmentation method is applied. We use the augmented spatial-temporal feature maps $M_1, M_2, \cdots, M_t$ as the input of the network to estimate the HR per video clip. The convolutional layers are based on ResNet-18 [29]. The ResNet-18 architecture includes four blocks comprising convolutional layers and a residual link, one convolutional layer, and one fully connected layer for the final results prediction. The feature map of each convolution output in each block goes through the channel attention blocks followed by spatial-temporal attention sequentially. Next, fed into the fully connected layer to regress the HR.

The output of the network for each spatial-temporal map is a single HR value $hr_1, hr_2, \cdots, hr_t$ regressed by the

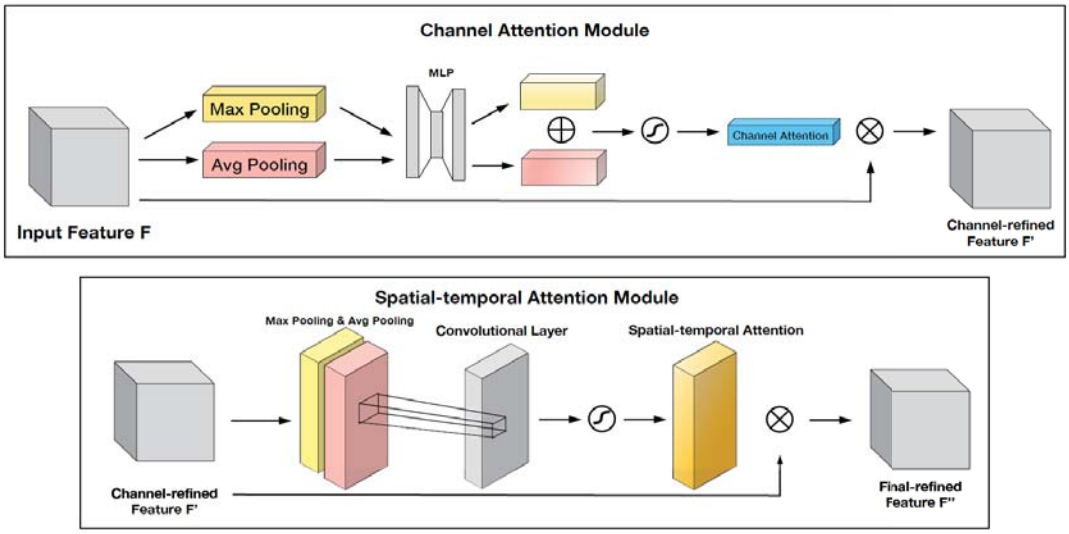

Fig. 5. An illustration of channel attention module (top) and the spatial-temporal attention module (bottom).

fully connected layer and then mean pooled to get the final estimated HR per video. $L1$ losses are used for measuring the residual between the estimated HR (HR_{est}) and ground truth HR (HR_{gt}).

C. Channel and Spatial-temporal Attention

The information that is expected to be used by rPPG based HR measurement is the color variation of the face in the captured video. However, illumination variation, occlusion, head movement can act as a noise and affect the precise signal. In an attempt to overcome these noises, previous methods typically rely heavily on pre-processing methods such as skin segmentation and ROI selection algorithms. However, these pre-processing techniques might not be sufficient to mitigate the effect of the noises. Therefore, in our end-to-end learning model, we propose a HR estimator based on the channel and spatial-temporal attention mechanism. Specifically, we add an attention module to the input spatial-temporal maps in both spatial and channel dimensions (inspired by CBAM [24]) to filter out the channel and spatial level salient information from the captured data.

Specifically, we first apply the *channel attention module*, which determines the pertinent channel in the convolution output of the input feature map. Accordingly, the output of the input feature map from the convolution is passed through global max pooling and global average pooling respectively. Two different spatial context descriptors are then generated: F_{max} and F_{avg}, which denote max-pooled features and average-pooled features, respectively. Then we forward these two descriptors to a shared network composed of multi-layer perception (MLP) to produce the channel attention map M_c. The output feature of MLP is subjected to an element-wise addition operation, and then a sigmoid activation operation is performed to generate the final channel attention feature map. The channel attention feature map and input feature map are subjected to an element-wise multiplication operations to obtain the input features for the spatial attention module i.e the channel refine feature.

Spatial attention module reveals the pertinent spatial part from the input. We firstly apply both max pooling and average pooling operations across the channels refined feature to generate two 2D maps. Then we concatenate and convolve these two maps to produce a 2D spacial attention map M_s. A sigmoid function is used here as the activation function. Finally, the spatial attention map is multiplied with the input features of this module to obtain the final refined feature representation of the HR signal. Figure 5 illustrates the process of the channel attention module and the spatial-temporal attention module respectively.

The overall processing of our attention based representation learning can be summarized as:

$$\begin{aligned} F'_n &= M_c(F_n) \otimes F_n, \\ F''_n &= M_s(F_n) \otimes F'_n, \\ F_{n+1} &= F_n + F''_n, \end{aligned} \tag{1}$$

where F_n denotes the input feature map of the convolutional block, and $\otimes$ denotes element-wise multiplication. During multiplication, the attention feature map values are broadcasted accordingly: channel attention values are broadcasted along the spatial dimension, and vice versa. F_n is the final refined output. The sum of F_n and F_n would be the output of this convolutional block, which is then passed to the next convolutional block.

D. Data Augmentation

Existing databases are usually of limited sizes (less than 100 subjects) accompanied by unbalanced data (w.r.t. the HR distribution). However, the normal human HR distribution ranges from 60 to 135 bpm[2]. To tackle the above challenges of lacking data, we propose an effective data augmentation

[2]https://en.wikipedia.org/wiki/Heart_rate

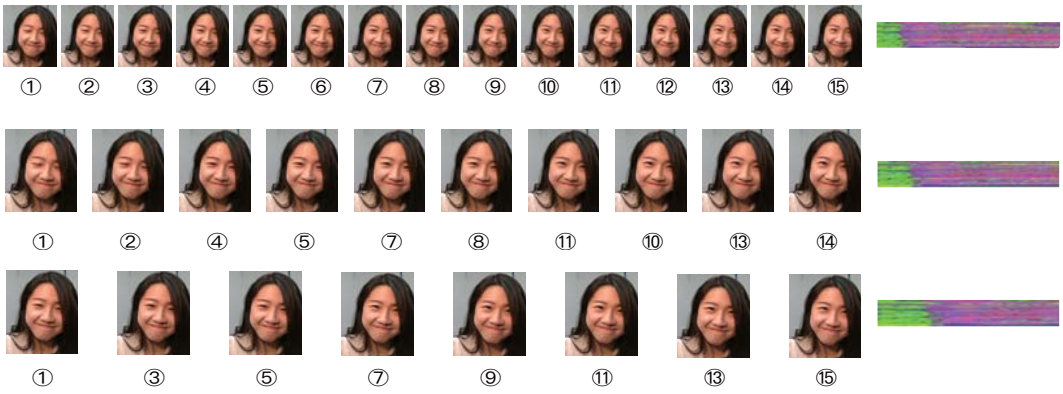

Fig. 6. An example of HR video augmentation by the proposed approach. The middle row is the original video sequence with a ground-truth HR of 89 bpm. The top row illustrates a temporally down-sampled video sequence with a sampling rate of 0.67 and thus with an increased HR of 133 bpm. The bottom row is a temporally up-sampled video sequence with a sampling rate of 1.5 and thus a decreased HR of 71.2 bpm.

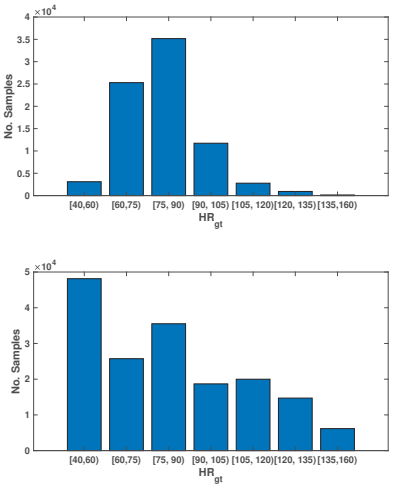

Fig. 7. The ground-truth HR distribution (top) and the augmented HR distribution (bottom) of the VIPL-HR database [13].

approach which samples an original video and can generate multiple spatial-temporal maps with varying HRs.

Since the heart rate is estimated by determining the frequency of the color changes in the face video, we first change the color frequency of one video by either down-sampling or up-sampling frames from the original video sequence, leading to new video clips with high / low HRs w.r.t. to the original video. An example of the data augmentation process is in Fig. 7. Then we generated the spatial-temporal maps for the augmented video clips and use all these spatial-temporal maps for training.

By adopting the proposed augmentation technique we do not perform the data augmentation in the image space; instead, we perform it on the spatial-temporal space since it preserves the HRs distribution. We observed the statistic of the ground-turth HRs of the large scale VIPL-HR database [13], and find that the majority of the data is associated to the ground-truth HRs ranging from 74 bpm to 136 bpm, with a small fraction of data having a ground-truth HR smaller than 74 bpm or larger than 136 bpm (see Fig. 7).

Based on this observation, we perform the following operations for augmenting the original data. The videos for which the ground-truth HRs range between 60 bpm and 110 bpm, we first perform the down-sampling processing with a sampling rate of 0.67 (e.g., sampling a 300-frame video using 450 video frames), and then generate the spatial map from the sampled video. The respective HR is then calculated as 1.5 times the original HR since the sampled video is treated as a normal video when training, and the frame rate of the sampled video is 1.5 times the normal one. Following the same idea, we then perform the up-sampling processing on the videos with a ground-truth HRs range between between 70 bpm and 85 bpm. The sampling rate is 1.5 (e.g., we first select 200 frames per video to generate the spatial map, and then resize the temporal demision of the spatial-temporal map to 300.), and the corresponding HR is thus decreased to 0.67 times the original HR.

Thereby we change the relevant data proportionally. An illustration of the data augmentation can be found in Fig. 6, and the HR distributions of VIPL-HR database [13] before and after our data augmentation can be found in Fig. 6. From the figure we can see that such a data augmentation method allows us to obtain a more diverse dataset with wide HR distribution covering both low and high HRs, thus balance the HR distribution.

IV. Experimental Results

A. Datasets and Settings

In this paper, we choose the widely used public domain MMSE-HR database [21], as well as the newly proposed large-scale less-constrained VIPL-HR database [13] for evaluation. Details about these databases are summarized in Table I. The ground truth HRs associated to the VIPL-HR database are calculated as the average HR for the whole video. For the MMSE-HR database, the ground truth HRs have been computed from electrocardiography (ECG) signals provided in the database using the OSET ECG Toolbox[3].

[3]http://www.oset.ir

TABLE I

A SUMMARY OF THE FACE VIDEO DATABASES USED FOR REMOTE HR MEASUREMENT IN OUR EXPEIRMENTS.

	No. Subj.	No. Vids.	Video Length	Protocol
MMSE-HR [21]	40	102	30s	cross-database
VIPL-HR [13]	107	2,378	30s	five-fold

Different evaluations have been proposed for HR estimation methods, such as the mean, standard deviation (HR_{me} and HR_{std}) of the HR error, the mean absolute HR error (HR_{mae}), the root mean squared HR error (HR_{rmse}), the mean of error rate percentage (HR_{mer}), and associated Pearson's correlation coefficients r [7], [21]. In this paper, we apply named evaluation metrics for experiments on both, the VIPL-HR databases and the MMSE-HR database.

Both intra-database testing and cross-database testing are conducted to evaluate the effectiveness of the proposed method. For the intra-database testing, since the VIPL-HR database is sufficiently large for training an HR estimator, we perform a participant independent validation on VIPL-HR database. Specifically $1,923$ videos of 85 subjects are used for training and the other 455 videos of 22 subjects are used for testing. This results in $62,717$ original spatial-temporal maps for training and further augment the number of spatial-temporal maps to $129,105$. For the cross-database testing, we first train our model on the augmented VIPL-HR database, and then directly test the results on the MMSE-HR database. All the 102 videos in the MMSE-HR database are involved for cross-database testing. A sliding window size of 300 frames were employed for both datasets for spatial-temporal maps generation, and the HR estimation of each video is calculated as the average of the prediction results of all the clip sequences.

For the proposed approach, we use a face rectangle ROI of 200×200, and divide it into 5×5 blocks. A random mask strategy proposed in [13] is applied to improve the general ability of the HR estimator to the situations where it failing detecting faces.

The model is implemented based on the PyTorch[4] framework. The number of maximum iteration epochs employed is 50, and the batch size is 100. We first use the augmented dataset to train the network without attention mechanism using the Adam solver [30] with an initial learning rate of 0.001. Then we use it as the pretrained model to train the channel and spatial-temporal attention CNN model and reset the initial learning rate to 0.0015.

B. Intra-database Testing

We first perform the intra-database testing on the VIPL-HR database [13]. Several state-of-the-art methods, including the hand-crafted methods (De Haan and Jeanne [18] and Wang *et al.* [19]) and data-driven methods (Tulyakov *et al.* [21], Niu *et al.* [13]) are taken into consideration for state-of-the-art comaprision. The results corresponding to the works of De Haan and Jeanne [18], Tulyakov *et al.* [21], Wang *et al.* [19], Niu *et al.* [13] are taken from [13]. All the related results are shown in Table II.

[4]https://pytorch.org/

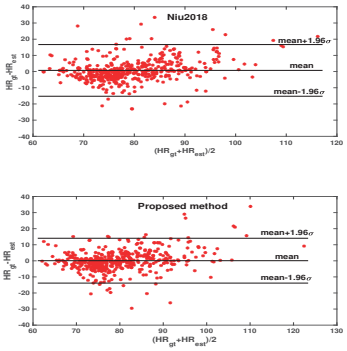

Fig. 8. The Bland-Altman plots for the results predicted using Niu2018 [13] (top) and the proposed method (bottom) for intra-database testing on the VIPL-HR database.

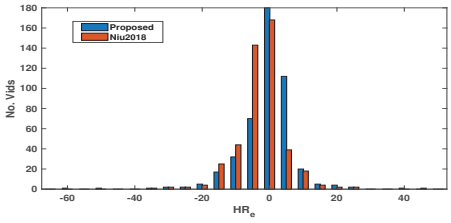

Fig. 9. Comparison of the HR estimation error distributions of the proposed approach and Niu2018 [13].

From the results, we can observe that the proposed method achieves promising results on all evaluation metrics. The proposed method achieves the best results of $HR_{std}, HR_{mae}, HR_{rmse}$ and HR_{mer}, which outperforms not only the hand-crafted methods, but also the data-driven methods. In addition, we achieve a promising Pearson's correlation coefficients r, indicating the good consistency between the ground truth and the predicted HR. All these results indicate that our method is very effective and performs well on the less-constrained settings of the VIPL-HR database and thus reflect the success of the proposed attention model.

We use the Bland-Altman plot (see Fig. 8) to further evaluate the correspondence between ground-truth HR and the predicted HR. The results suggest that the proposed attention mechanism, as well as the data augmentation strategy contribute to a better consistency and smaller standard deviation in challenging less-constraint settings.

We also present the statistics of the predicted errors for all the videos in Fig. 9. From the results we can see that most videos (80%) are predicted with an error less than 5 bpm, while the percentage of Niu *et al.* [13] is 76%. These results indicate that our method is very effective for remote HR estimation under the less-constrained situations.

TABLE II

THE INTRA-DATABASE TESTING RESULTS OF HR ESTIMATION ON COLOR FACE VIDEOS ON VIPL-HR DATABASE.

Method	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{mae} (bpm)	HR_{rmse} (bpm)	HR_{mer}	r
Haan2013 [18]	7.63	15.1	11.4	16.9	17.8%	0.28
Tulyakov2016 [21]	10.8	18.0	15.9	21.0	26.7%	0.11
Wang2017 [19]	7.87	15.3	11.5	17.2	18.5%	0.30
Niu2018 (ResNet-18) [13]	1.02	8.88	5.79	8.94	7.38%	**0.73**
ResNet-18 + DA*	**-0.08**	8.14	5.58	8.14	6.91%	0.63
Proposed	-0.16	**7.99**	**5.40**	**7.99**	**6.70%**	0.66

*DA stands for data augmentation;

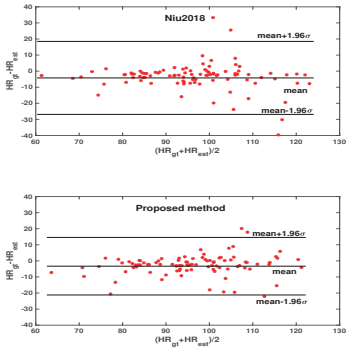

Fig. 10. The Bland-Altman plots for the results predicted using Niu2018 [13] (top) and the proposed method (bottom) for cross-database testing on the MMSE-HR database.

TABLE III

THE RESULTS OF AVERAGE HR ESTIMATION PER VIDEO USING DIFFERENT METHODS ON THE MMSE-HR DATABASE.

Method	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer}	r
Li2014 [7]	11.56	20.02	19.95	14.64%	0.38
Haan2013 [18]	9.41	14.08	13.97	12.22%	0.55
Tulyakov2016 [21]	7.61	12.24	11.37	10.84%	**0.71**
Niu2018 [13]	**-2.26**	10.39	10.58	**5.35%**	0.69
Proposed	-3.10	**9.66**	**10.10**	6.61%	0.64

C. Cross-database Testing

The cross-database experiments are then conducted based on the VIPL-HR and MMSE-HR databases. Specifically, the VIPL-HR database is used for training and all videos in the MMSE-HR database are directly used for testing. Following the testing protocol in Niu *et al.* [13], HR_{mae} is not used for evaluation. The baseline methods we use for comparisons are from Li *et al.* [7], Haan *et al.* [18], Tulyakov *et al.* [21] and Niu *et al.* [13], and the results are taken from [13]. All results are summarized in Table III.

From the results, we can see that the proposed method achieves a promising performance with an HR_{rmse} of 10.10 bpm, even when we directly test our VIPL-HR pre-trained model on MMSE-HR. We also present the Bland-Altman plots [31] for the results of cross-validation on the MMSE-HR database in Fig. 10, and we illustrate additionally the results of Niu *et al.* [13] for comparison. We can see that our method achieves a good consistency under the unseen conditions. These results indicate that the proposed network has good generalization ability to unseen scenarios with the help of the diverse information contained in VIPL-HR.

D. Ablation Study

We also perform the ablation study of our proposed method. Our work is based on the method of [13] with two key components different from [13]: the data augmentation strategy and the spatial-temporal attention mechanism. We remove these two key modules step by step to verify their effectiveness for HR estimation on the VIPL-HR database. All the results can also be found in Table II.

We firstly remove the spatial-temporal attention module and directly train our model using ResNet-18 on the augmented dataset. From Table II we can see that the HR estimation error is increased from a HR_{rmse} of 7.99 bpm to 8.14 bpm. The difference looks not that big. This is because not all face videos in the VIPL-HR database contain such kind of noises. To better illustrate the effectiveness of our attention module in reducing the influence of noise, we further calculate the results under the head movement scenario of the VIPL-HR, in which big head movements during talking are expected to bring more noises to HR spatial-temporal maps. The HR_{rmse} of the proposed method is 7.85 bpm while the HR_{rmse} without using the attention module is 8.27 bpm. The results suggest that the proposed spatial-temporal attention module does improve the HR measurement robustness under less-constrained scenarios.

We further remove the data augmentation module and directly train our model on the VIPL-HR database using ResNet-18, which has been applied in [13]. The HR_{rmse} is further increased to 8.94 bpm, indicating that our data augmentation strategy is very effective for training the HR estimator. In order to further illustrate the effectiveness of our data augmentation, we calculate the HR_{rmse} for videos with ground-truth HRs in different HR ranges, and the statistical results can be found in Fig. 11. We can see that with our data augmentation, the prediction errors of the videos with

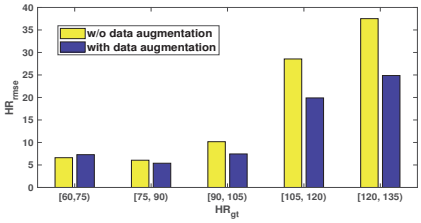

Fig. 11. The estimation error (HR_{rmse}) of videos with ground-truth HRs in different HR ranges for the method with and without our data augmentation module.

unusual ground-truth HRs (ground-truth HR larger than 105 bpm) are reduced by a large margin. These results indicate that our data augmentation could effectively augmented to the unusual ground-truth situations and solve the problem of imbalance.

V. Conclusions and Future Work

HR is a pertinent for the related plethora applications such as health monitoring, press analysis and defection detection. While both, contact and non-contact HR measurement methodologies are available, in this work, considering the convenience of remote measurement, we proposed an end-to-end approach for HR estimation, which is robust to unconstrained scenarios. In order to capture substantial visual characteristics of HR from unconstrained face videos, we propose an end-to-end learning network based on channel and spatial-temporal attention. At the same time, we also design an effective video augmentation method to overcome the limitation of training data. Experimental results on the VIPL-HR and MMSE-HR datasets show the effectiveness of the proposed method. Future work includes the expansion of the work onto continuous HR measurement. Additionally, remote measurement of further physical signals, such as breath rate, as well as heart rate variability, will be studied.

Acknowledgment

This research was supported in part by the National Key R&D Program of China (grant 2017YFA0700804), Natural Science Foundation of China (grants 61390511, 61672496, and 61650202), and External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843). A. Das was supported by the research program FER4HM funded by Inria and CAS. A. Dantcheva was supported by the French Government (National Re-search Agency, ANR) under grant agreement ANR-17-CE39-0002.

References

[1] 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.

[2] 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.

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

[4] 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.

[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, pp. 10 762–10 774, 2010.

[6] ——, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, 2011.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] X. Niu, H. Han, S. Shan, and X. Chen, "Synrhythm: Learning a deep heart rate estimator from general to specific," in *Proc. IAPR ICPR*, 2018.

[13] ——, "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video," in *Proc. ACCV*, 2018.

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

[15] 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.

[16] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proc. IEEE ICCV*, 2015, pp. 3640–3648.

[17] 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, 2012.

[18] 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.

[19] 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.

[20] X. Niu, H. Han, S. Shan, and X. Chen, "Continuous heart rate measurement from face: A robust rppg approach with distribution learning," in *Proc. IJCB*, 2017, pp. 642–650.

[21] 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.

[22] R. A. Rensink, "The dynamic representation of scenes," *Visual cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[23] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," in *Proc. NeurIPS*, 2010, pp. 1243–1251.

[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. ECCV*, 2018.

[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE CVPR*, 2017, pp. 3156–3164.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, 2017.

[27] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Proc. NeurIPS*, 2014, pp. 2204–2212.

[28] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE CVPR*, 2016, pp. 3640–3649.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] D. G. Altman and J. M. Bland, "Measurement in medicine: the analysis of method comparison studies," *The statistician*, pp. 307–317, 1983.