



Hierarchical Attentive Upsampling on Input Signals for Remote Heart Rate Estimation

Pengfei Zhang, Xiang Li, Jianjun Qian^(✉), Zhong Jin, and Jian Yang

Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{zhangpengfei, xiang.li.implus, csjqian}@njust.edu.cn

Abstract. Heart Rate (HR) is one of the most important indicators reflecting the physiological state of the human body, and more researches have begun to focus on remote HR measurement in order to meet the challenging but practical non-contact requirements. Existing remote HR estimation methods rely on the high-resolution input signals constructed from low-resolution Spatial-Temporal Map (STMap) of facial sequences, but most of them use simple linear projection, which are difficult to capture the complex temporal and spatial relationships in between weak raw signals. To address this problem, we propose a Hierarchical Attentive Upsampling Module (HAUM) to obtain rich and discriminating input signals from STMap for accurate HR estimation. Our approach includes two parts: (1) a Hierarchical Upsampling Strategy (HUS) for progressively enriching the spatial-temporal information, and (2) an Attentive Space Module (ASM) to focus the model on more discriminating HR signal regions with clearer periodicity. The experiments performed on two public datasets VIPL-HR and MAHNOB-HCI show that the proposed approach achieves the state-of-the-art performance.

Keywords: Remote heart rate estimation · Spatial-Temporal map · Hierarchical attentive upsampling module

1 Introduction

Heart rate (HR) reflects the physiological information directly and is closely related to cardiovascular diseases [1, 9]. The traditional HR measurement techniques are based on the Electrocardiography (ECG) and Photoplethysmography (PPG), which are generally applied to professional medical diagnosis, such as wearable devices. However, these contact measurements are poorly performed for daily monitoring, especially under long-term human-computer interaction (HCI) situations.

This work was supported by the National Science Fund of China under Grant Nos. 61876083.

Due to these realistic demands, one of the measurement techniques based on remote Photoplethysmography (rPPG) has been developed rapidly [2, 16, 24, 34]. It has been proved that the measurement of HR can be realised by extracting subtle color variations of facial skin from a distance of several meters using cameras [30]. These researches aspire to get more accurate HR predicted results, in order to monitor the HR information at home and office by using abundant computing resources.

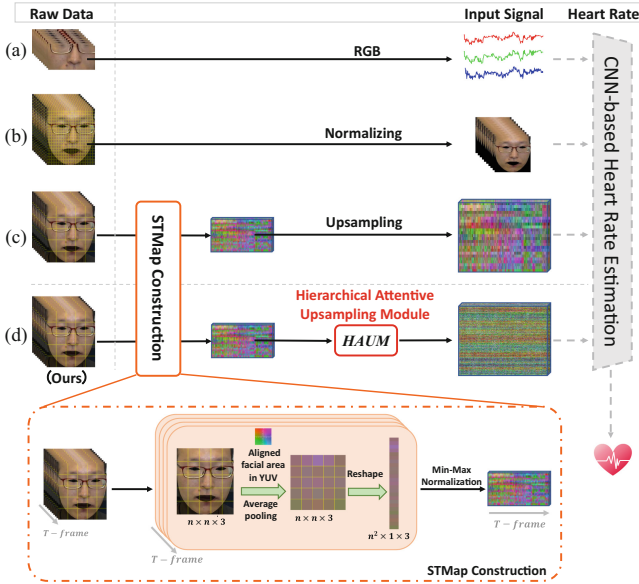


Fig. 1. Existing works ((a), (b), (c)) vs. our method (d) for obtaining discriminating high-resolution input signals. The construction of STMap is also illustrated in the dotted frame. (Color figure online)

Under the popular framework of Convolution Neural Network (CNN) [4, 27], most existing works based on rPPG focus on how to construct high-quality or high-resolution input signals with more discriminating spatial and temporal information, as shown in the Fig. 1. Similarly with the traditional methods [16, 24, 30], [6] (Fig. 1(a)) records the pixel intensity averages of the Red, Green and Blue (RGB) channels over the face region as the input signals, which, however, is too simple to capture the critical information. More other works [2, 29, 34] (Fig. 1(b)) normalize the face sequences into smaller size, where the excessively dense sampling of the face image will bring more noises into the signal representation. Figure 1(c) [19] illustrates a more reasonable approach named Spatial-Temporal Map (STMap) to represent input signals. With the appropriate face region allocation, the STMap preserves color statistical information while suppressing irrelevant noises, which effectively avoid the drawbacks of the aforementioned approaches. As depicted in the bottom of the Fig. 1, the face video

is compressed into STMap via space region division and time sequence concatenation. In addition, by means of linear color space transformation and average pooling of the face region, it also constructs color features with higher correlation between heart rhythm and pixel intensity.

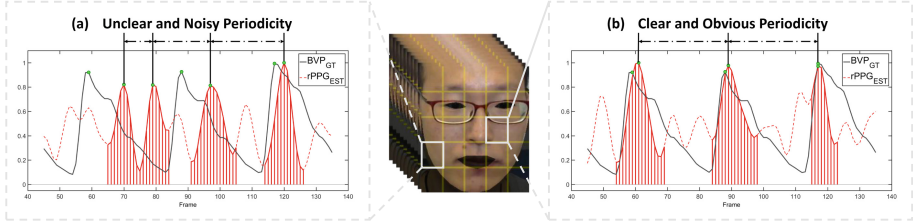


Fig. 2. Illustration of rPPG signal estimated from different face regions. Due to the uneven distribution of capillaries in skin regions, the contribution of each region to the rPPG estimation is not consistent. BVP_{GT} : the Ground-truth Blood Volume Pulse. $rPPG_{EST}$: the Estimated rPPG signal generated by [3].

However, the STMap-based method simply uses linear upsampling to construct the high-resolution input signal from the low-resolution STMap, making it difficult to extract the informative and complex spatial-temporal relationship. Moreover, as illustrated in Fig. 2, the rPPG signals estimated from various regions usually lack consistency due to the uneven distribution of micro-vessel density, leading to many ineffective areas with considerable noises. Therefore, the suppression of these noisy regions is very important for HR estimation considering the inconsistent contribution of each skin region.

To address these problems, we propose a Hierarchical Attentive Upsampling Module (HAUM) (see Fig. 1(d)), which consists of a Hierarchical Upsampling Strategy (HUS) and an Attentive Space Module (ASM), to obtain informative and discriminating high-resolution input signals. The Hierarchical Upsampling Strategy uses a multi-level approach to generate high-resolution input signal and extract complicated spatial-temporal information from the low-resolution STMap. The Attentive Space Module selects the regions with higher correlation with the target HR signals, aiming at reducing the influence of the noisy areas on the HR estimation.

In summary, our contributions include: (1) To the best of our knowledge, our work is the first to explore the discriminating features from the perspective of the original high-resolution input signal in an end-to-end fashion, greatly enhancing the low-resolution STMap. (2) We propose a novel Hierarchical Attentive Upsampling Module (HAUM), which can produce richer and discriminating input signal, especially highlight the regions with stronger heart rhythm signal. (3) Our method achieves state-of-the-art performance in most of the significant metrics on the public VIPL-HR dataset [19]. E.g., the mean absolute error between the estimated result and the ground-truth improves from 5.02 bpm to 4.67 bpm, which is 0.35 bpm superior than previous best record.

2 Related Work

rPPG-Based Remote HR Estimation. The rPPG-based remote HR estimation is founded on the fact that blood circulation causes periodic blood volume fluctuations in the micro-vascular tissue bed beneath the skin, which is induced by heartbeat, and then leads to the miniature variations in skin color. Although these subtle variations are not visible to the human eyes intuitively, they can be captured by the camera.

An early study proposed by Verkruysse [30] only used the green channel as the HR feature and analyzed its implicit physiological information. However, limited by its less-information, the methods after that usually used the average intensity of all facial skin pixels as the original features, such as Chrominance-based rPPG (CHROM) [3], Independent Component Analysis (ICA) [23], Plane-orthogonal-to-skin (POS) [31]. However, due to the difference in the density of facial capillaries, the signals in different facial regions may be distorted. There are also works [25,33] that only select a single face region, which may ignore the correlation between the regions. In recent years, a great deal of methods that use deep learning techniques to estimate HR are proposed, including [2]. These works use 2D CNN to extract spatial features from each frame of video, but do not take the relevance of temporal dimension into consideration. Another type of methods use 3D CNN to directly process the video sequences end-to-end [29,34]. They pay well attention to the features in temporal dimension, but these dense sampling approaches may easily introduce irrelevant noises caused by illumination and motion-induced artifacts.

Representation of Input Signal. How to represent HR signal is the most important component of the HR estimation. The earlier methods use the facial pixel intensity averages of the RGB channels as the input signals [23,24,30], and the green channel is widely used because of it featuring the strongest pulsatile signal [16,30]. However, the signal generated in this way is less informative and does not take the physiological differences of face regions into consideration. Later, [29,34] estimate rPPG signals from the video sequences, which are not robust enough since this intensive sampling method for direct processing of the face images is susceptible to complex scenarios, even if normalizing images to smaller size. [20] introduces a novel representation called STMap, which represents its physiological features by regional division with pixel intensity averaging and temporal concatenation with continuous video sequences. The methods mentioned above mainly focus on how to characterize the input signals, while our method pays more attention to make the input signals more discriminating.

Selection of Space Region. Theoretically, all skin areas exposed to light are employable to estimate HR. While, in practice, only the part of the face regions has a high correlation with physiological features such as HR. In the past few years, several works have discussed the influence of different face regions on the quality of rPPG signals [11,12,14]. Moreover, there are also works [2,17,22,34] proposed from the perspective of spatial dimension to improve the accuracy of HR estimation. [17] focuses on the super resolution of the original face image for

recovering physiological information. While, [2, 34] rely on the skin segmentation implemented by attention mechanism to select region of interest (ROI). Similarly, [22] advocates a spatial attention mechanism to the feature extraction network of ResNet [4], where the attention weights concentrate on both temporal and spatial dimensions of high-level convolutional features. Different from the above approaches, in our method, the selection of the skin area is performed on the initial STMap, which greatly reduces the risks of introducing more spatial noises into the subsequent network.

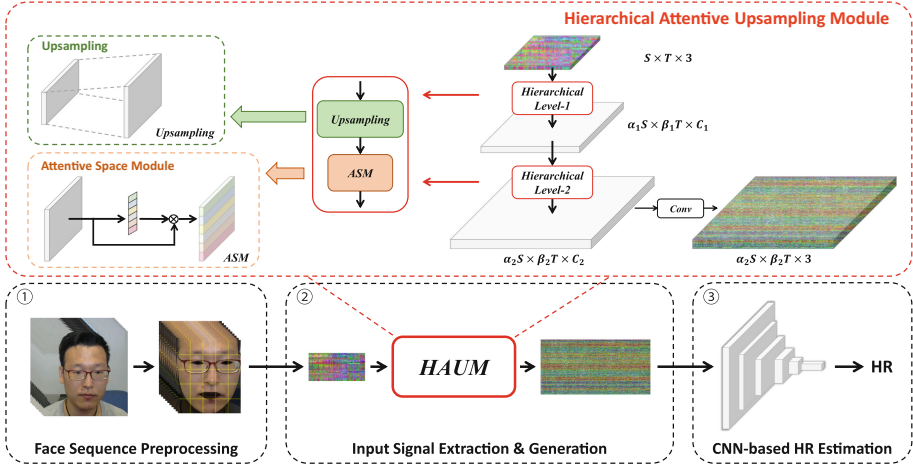


Fig. 3. The overview of our approach about rPPG-based remote heart rate estimation with Hierarchical Attentive Upsampling Module (HAUM).

3 Method

3.1 The Framework of HR Estimation

As depicted in Fig. 3, our approach is composed of three steps. The first step is the preprocessing of the video sequences. As described in [11, 14], the only useful information for HR estimation is heart-induced variation of skin color. Given the likely weakness of such signal compared with noises and artifacts, it is essential to select maximal pulsatile ROI. Consequently, we remove irrelevant pixels such as background, etc. based on facial landmarks.

Secondly, we generate initial low-resolution STMap, and then, apply the Hierarchical Attentive Upsampling Module (HAUM) to STMap on the purpose of constructing richer and discriminating input signals. The HAUM is made up of Hierarchical Upsampling Strategy (HUS) and Attentive Space Module (ASM) as illustrated in the top of Fig. 3. The details of HAUM will be explained in the following subsections.

In the third step, the popular framework of CNN is employed to estimate HR of each video sequence with the corresponding input signal, as commonly practiced in [6, 20].

3.2 Hierarchical Upsampling Strategy

With regards to the estimation HR from face video sequence, constructing effective representation of input signals is the majority component of the task. As mentioned in Sect. 2, previous HR measurement methods use too simple or excessively noisy statistics, which are not beneficial for HR estimation task since these input signals can not reflect the discriminating physiological information. Moreover, the situation becomes even worse with respect to the cases of varying illuminations and motions.

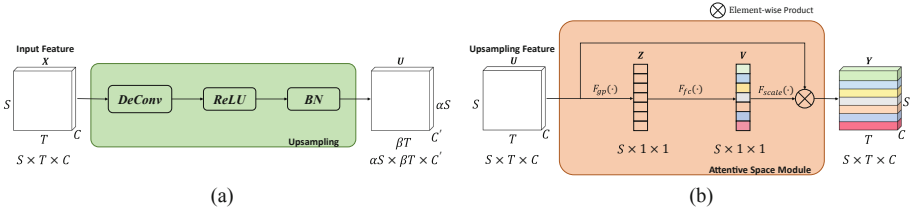


Fig. 4. (a)The Upsampling Layer. (b)The Attentive Space Module.

Compared with the previous work, the STMap proposed by Niu *et al.* [20] reasonably divides the skin region into an appropriate number of sub-regions, which can extract better statistical features of spatial-temporal signal. However, it applies the linear projection to the STMap for upsampling, which ignores the limitations of linear mapping in complex and varied scenarios.

Non-linear Spatial-Temporal Upsampling. For the sake of characterizing the more complex spatial-temporal data relationship of HR, and projecting the original color signal into a color space with stronger HR correlation, we use a non-linear spatial-temporal upsampling layer to construct the input feature. The upsampling layer is shown in Fig. 4(a). Specifically, for the input feature $\mathbf{X} \in \mathbb{R}^{S \times T \times C}$, where S , T , C represent space, time and channel dimensions respectively, we expand its spatial and temporal dimensions via $\mathcal{F}_{up}(\cdot)$ and map it to the output feature $\mathbf{U} \in \mathbb{R}^{\alpha S \times \beta T \times C'}$:

$$\mathbf{U} = \mathcal{F}_{up}(\mathbf{X}) = \mathcal{B}\left(\delta(\text{DeConv}(\mathbf{X}))\right), \quad (1)$$

where \mathcal{B} indicates the Batch Normalization [8], δ denotes the ReLU function [18], and $\text{DeConv}(\cdot)$ is a deconvolution layer with the $\times \alpha$ and $\times \beta$ expansion of space and time dimensions respectively. In addition, the channel dimension is changed from C to C' for the sake of diverse features. Although the non-linear spatial-temporal upsampling layer improves the discriminability of the input signal,

the hierarchical upsampling representation can model complex spatial-temporal relationships more accurately.

Hierarchical Upsampling. Previous works [3, 12, 24] assume that the rPPG signal extracted from face can be treated as a linear combination of the RGB channel signals. However, its performance significantly degrades when in challenging conditions [5]. On the contrary, [21] proposes the physiological encoder to get physiological feature implemented by several convolution layers, which confirms the spatial-temporal complexity of physiological signal from the side. Considering the limitation to one single layer of non-linear upsampling, we adopt the Hierarchical Upsampling Strategy to enable the non-linear upsampling layer model higher correlated representation with remote HR signal. Specifically, we propose to construct the high-resolution input signals by progressively expanding the spatial and temporal dimensions of the low-resolution STMap, aiming at modelling the underlying complex spatial-temporal relationship.

As illustrated in Fig. 3, with respect to the low-resolution STMap, a hierarchical upsampling block (e.g., with two levels) is applied for obtaining the high-resolution input signals. Following the sequential upsampling modules, the dimension of the channel is restored back to 3 through a convolution layer.

3.3 Attentive Space Module

Due to physiological factors such as uneven distribution of micro-vessel density in facial skin tissues, as well as external environmental elements such as facial rigid motion and external light changes, the HR information reflected by different facial regions is not equally important.

As shown in Fig. 2, The rPPG signal generated from the left part (a) of the face is noisy and the periodicity is not significant. The skin in this area does not face the camera and the skin illumination changes may not be obvious. In contrast, the periodicity of the rPPG signal shows a more regular shape on the right region (b), thus it can potentially provide more accurate HR estimations.

Based on the above observations and inspired by recent attention modules [7, 15], we propose an Attentive Space Module (ASM, see Fig. 4(b)) that aims to explicitly suppress the ineffective areas and leverage more information from the discriminating ones. Specifically, it consists of two parts: the *Spatial Information Aggregation* and *Selection*.

Spatial Information Aggregation. As stated above, the purpose of our method is to select discriminating regions for HR estimation according to the distribution of physiological information in face. The basic idea is to integrate all the features of different spatial regions. Therefore, we first aggregate the global information of all features along different spatial regions by using the global average pooling $\mathcal{F}_{gp}(\cdot)$ over both time and channel dimensions. Specifically, the s -th element of $z \in \mathbb{R}^{S \times 1 \times 1}$ is calculated by shrinking $\mathbf{U} \in \mathbb{R}^{S \times T \times C}$ through dimensions $T \times C$:

$$z_s = \mathcal{F}_{gp}(u_s) = \frac{1}{T \times C} \sum_{i=1}^T \sum_{j=1}^C u_s(i, j). \quad (2)$$

Spatial Information Selection. By using the spatial information aggregated from different *time-channel* dimensions, we create a vector $\mathbf{v} \in \mathbb{R}^{S \times 1}$ to selectively extract more remarkable features, which is achieved through two fully connected layers $\mathcal{F}_{fc}(\cdot)$. Thus, it can be written as:

$$v_s = \mathcal{F}_{fc}(z, \mathbf{W}) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 z)). \quad (3)$$

Here, σ denotes sigmoid activation, δ refers to ReLU function [18] and $\mathbf{W}_1 \in \mathbb{R}^{\frac{S}{\gamma} \times S}$, $\mathbf{W}_2 \in \mathbb{R}^{S \times \frac{S}{\gamma}}$. We also use a reduction ratio γ to limit the complexity of the model and facilitate information interaction between features. Then, the vector \mathbf{v} is dimensionally expanded and calculated with \mathbf{U} via $\mathcal{F}_{scale}(\cdot)$, which denotes space-wise multiplication:

$$Y = \mathcal{F}_{scale}(v_s, u_s) = v_s \cdot u_s, \quad (4)$$

finally, we get the output feature $\mathbf{Y} \in \mathbb{R}^{S \times T \times C}$ with the suppression of the noisy region and the prominent of the effective region.

Discussion. Although related, our ASM differs from the popular SE block [7] in both the *attentive targets and aggregated information*. While SE utilizes pure global spatial information to calibrate features in a channel-wise manner, our ASM makes use of both temporal and channel feature maps to selectively suppress or enhance signals from various regions, as illustrated in Fig. 6.

3.4 Implementation Details

Our HR estimation framework is based on the method of RhythmNet [20], but because the source code of RhythmNet is not publicly available, some details may be biased. Therefore, in this section, we mainly introduce some implementation details of STMap generation and subsequent deep convolutional networks.

Facial Skin Segmentation. First, we use SeetaFace¹ to detect the face region and get the facial landmarks of each frame, then perform median filtering on the sequence of each landmark to obtain stable and accurate face landmarks. After that, the face is aligned according to the eye center position of the landmarks, and the original image is cropped with the bounding box of $w \times 1.2h$, where w is the horizontal distance between the left and right borders of the face contour, and h is the vertical distance from the lower border to the mean value of the eyebrow landmarks. Finally, the skin area is segmented. For simplicity, we directly use landmarks to remove the pixels of irrelevant areas, such as background, eyes and mouth. Consequently, the rest area is highly correlated with HR information.

¹ <https://github.com/seetaface/SeetaFaceEngine>.

Spatial-Temporal Map Generation. After getting the the image with aligned face and segmented skin, we first convert the RGB color space of each frame to YUV color space, and then divide it into S sub-regions. Next, we calculate the average value of the C channel pixels of each sub-region, where only the pixels of skin are considered. By concatenating the features of the T -frame images, we can get an initial time series signal. After that, the min-max normalization is applied to the S -dimensional time series of the C channels with scaling the value into $[0, 255]$, which can reduce the influence of inconsistent environmental illumination and amplify the weak signal to some extent. Finally, the STMap is constructed with a size of $S \times T \times C$.

CNN-Based HR Estimation Network. The backbone network we used is ResNet-18 [4]. Following [22], the CBAM [32] module is integrated into building blocks of ResNet-18. The final HR estimation is obtained by the last fully connected layer. For each long face video, we use a sliding window with a fixed number T of frames to prepare a series of samples. During the training phase, each sample corresponds to a ground-truth HR. Whilst in the test phase, we average the estimated HR of all the samples from the same long video as the final predicted HR of the video. L_1 loss is used for measuring the gap between the estimated HR and the ground-truth one.

4 Experiment

4.1 Experiment Setup

Databases. Two popular public databases are adopted to verify the effectiveness of our method. Among them, VIPL-HR [19] is a large database for remote HR estimation. The database contains 9 scenarios of face videos for 107 subjects recorded by 3 different devices. The other database is MAHNOB-HCI [28], which contains 527 facial videos from 27 subjects. Following [20], we downsample the videos from 61 fps to 30.5 fps for efficiency. Moreover, the heart rate signal is measured by an ECG sensor. Following [16, 34], we use the EXG2 channel² signal to generate the corresponding HR value. Specifically, we use *qrs_detector* function from the MNE package³ to clean-up and calculate the HR based on the ECG sensor information provided. By referring to the previous work [2, 34], we use only a 30-second (frames 306 to 2135) clip of each video in our experiments.

Training Details. For VIPL-HR and MAHNOB-HCI database, we use a sliding window of $T = 300$ frames, the interval of the sliding window being 0.5s, the number of face region division being $S = 25$. The estimated HR is obtained according to its frame rate. The data augmentation method uses the strategy proposed in [22] as well as the random mask strategy to simulate the situation of missing face detection. Random horizontal flip and random crop are also applied to the input signals before fed into the network. Our method is implemented

² The position of ECG sensor is upper left corner of chest and under clavicle bone.

³ <https://github.com/mne-tools/mne-python>.

using the PyTorch⁴ framework, where the training uses the Adam [10] optimizer with the initial learning rate of 0.001, the batch size of 64, the maximum epoch number of VIPL-HR database being 50, and MAHNOB-HCI being 100.

Performance Metrics. There are a variety of different data evaluation metrics used to validate the performance of HR estimation approaches [16, 20]. Among them, we use six widely used metrics: the mean (Mean) and standard deviation (Std) of the error, the mean absolute error (MAE), the root mean squared HR error (RMSE), the mean of error rate percentage (MER), and Pearson’s correlation coefficients r .

4.2 Ablation Study

In this section, we conduct several ablation experiments to get a better understanding of the proposed HAUM. All experiments are performed on the VIPL-HR database.

Effectiveness of Attentive Space Module. We examine the significance of using Attentive Space Module (ASM) under two-hierarchical-level setting. As illustrated in the Table 1, the usage of ASM leads to a clear performance improvement in every evaluation metric.

Table 1. The effect of ASM for HR estimation. “H-Level” denotes the total Hierarchical Levels applied in HAUM block.

ASM	H-Level	MAE (bpm)	RMSE (bpm)	MER	r
✓	1	4.81	7.41	5.98%	0.81
	1	4.73	7.25	5.88%	0.82
✓	2	4.77	7.33	5.90%	0.81
	2	4.67	7.10	5.84%	0.82

Table 2. The HR estimation results by different reduction ratio γ of ASM on the VIPL-HR database.

Ratio γ	MAE (bpm)	RMSE (bpm)	MER	r
2	5.03	7.89	6.28%	0.78
4	4.89	7.62	6.12%	0.79
8	4.88	7.43	6.04%	0.81
16	4.67	7.10	5.84%	0.82
32	4.85	7.51	6.06%	0.80

⁴ <https://pytorch.org/>.

Settings for Attentive Space Module. From Eq. (3), the reduction ratio γ can potentially adjust the complexity of the module as it controls the hidden dimension of ASM. It would be reasonable to see how it affects the overall performance. Till this end, we conduct experiments with various reduction ratios, as shown in Table 2. It is observed that $\gamma = 16$ achieves best performance, therefore, due to its effectiveness, we apply the ASM in subsequent experiments.

Settings for Hierarchical Upsampling Strategy. In order to verify the effectiveness of the Hierarchical Upsampling Strategy (HUS), we study the effect of hierarchical levels by fixing the size of the output signal as $300 \times 600 \times 3$. It can be seen from Table 3 that the two-level is the best for HR estimation task.

4.3 Comparison with State-of-the-Arts

In this section, we compare our method with a variety of state-of-the-art approaches on two public datasets VIPL-HR [19] and MAHNOB-HCI [28]. For the sake of fairness, we use a five-fold subject-exclusive evaluation protocol to train the VIPL-HR database, and use the same train/test split as mentioned in [19]. The results of other methods are directly borrowed from [20]. All experimental results of the VIPL-HR database are shown in Table 4. Similarly, for the MAHNOB-HCI database, following [20, 34], we use a three-fold subject-independent cross-validation protocol. The experimental results of this database are shown in Table 5. It can be seen from Table 4 that our method achieves superior results over most of the significant metrics. It reflects that the enhancement of input signals is very crucial for HR estimation task, whilst most existing works fail to pay much attention on it. By introducing the proposed HAUM for boosted input signals, we obtain the new state-of-the-art performance on the VIPL-HR database, reaching a lower *MAE* of 4.67 and increasing the correlation *r* from 0.79 to 0.82. Table 5 shows that our method can generalize to other database, with a consistent improvement over most evaluation metrics, compared with the previous state-of-the-art methods.

Table 3. The HR estimation results w.r.t. different number of the Hierarchical Levels (H-Level) on the VIPL-HR database.

H-Level	MAE (bpm)	RMSE (bpm)	MER	<i>r</i>
0	4.84	7.41	6.02%	0.81
1	4.73	7.25	5.88%	0.82
2	4.67	7.10	5.84%	0.82
3	4.82	7.52	6.00%	0.80

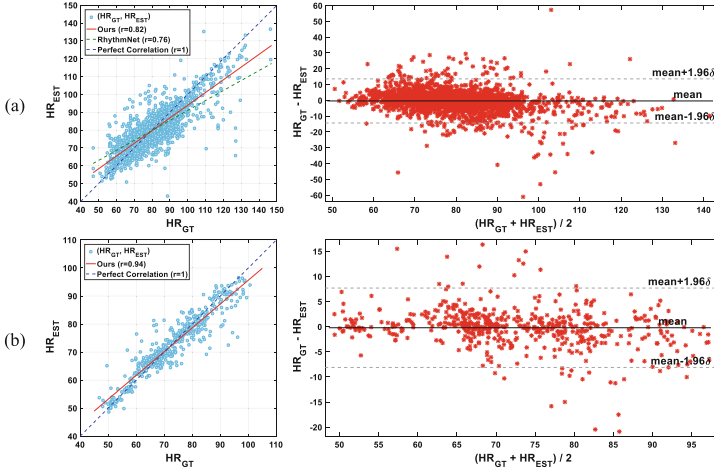


Fig. 5. The scatter plots and the Bland-Altman plot on the VIPL-HR (a) and MAHNOB-HCI (b) database. HR_{GT} : the Ground-truth HR. HR_{EST} : the Estimated HR by our method. The green dashed line in (a) is borrowed from the original paper [20]. The solid lines in the right plot represent the mean, and the dotted lines represent the 95% limits of agreement. (Color figure online)

4.4 Analysis

HUS and ASM Improve Accuracy. For the purpose of clearly illustrating the improvement of the data relevance of our method in HR estimation task, we draw scatter plots and the Bland-Altman plot on the VIPL-HR and MAHNOB-HCI database. The diagram is illustrated in Fig. 5. It can be observed that most of the estimated HRs are highly correlated with the ground-truth ones. Moreover, compared with RhythmNet, our method demonstrates a better linear correlation of data on the VIPL-HR database.

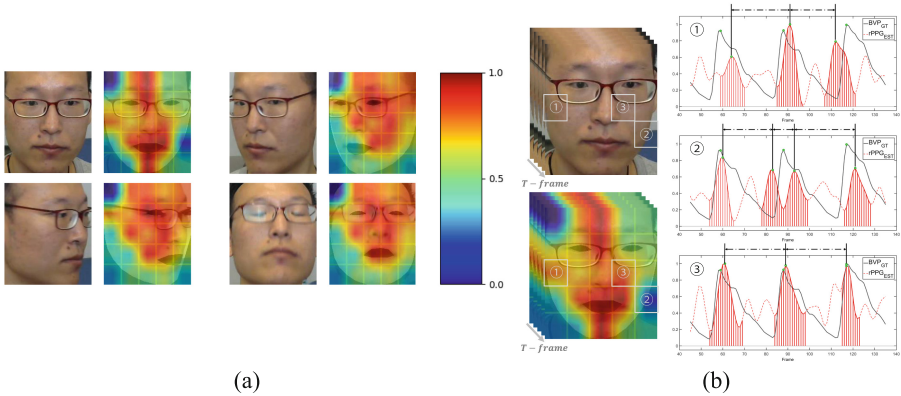
Table 4. Comparisons between state-of-the-art methods on VIPL-HR database. Best: **bold**; Second best: **blue**.

Method	Mean (bpm)	Std (bpm)	MAE (bpm)	RMSE (bpm)	MER	r
POS [31]	7.87	15.3	11.5	17.2	18.5%	0.30
CHROM [3]	7.63	15.1	11.4	16.9	17.8%	0.28
DeepPhy [2]	-2.60	13.6	11.0	13.8	13.6%	0.11
Niu2019 [22]	-0.16	7.99	5.40	7.99	6.70%	0.66
RhythmNet [20]	0.73	8.11	5.30	8.14	6.71%	0.76
CVD [21]	-	7.92	5.02	7.97	-	0.79
ours	0.35	7.08	4.67	7.10	5.84%	0.82

Table 5. Comparisons between state-of-the-art methods on MAHNOB-HCI database. Best: **bold**; Second best: **blue**.

Method	Mean (bpm)	Std (bpm)	MAE (bpm)	RMSE (bpm)	MER	r
Poh2011 [23]	2.04	13.5	–	13.6	13.2%	0.36
CHROM [3]	–2.89	13.67	–	10.7	12.9%	0.82
Li2014 [16]	–3.30	6.88	–	7.62	6.87%	0.81
rPPGNet [34]	–	5.57	4.03	5.93	–	0.88
RhythmNet [20]	0.41	3.98	–	4.00	4.18%	0.87
Meta-rPPG [13]	–	4.90	3.01	–	–	0.85
Deep-HR [26]	2.08	3.47	–	3.41	2.73%	0.92
ours	–0.14	3.96	2.68	3.98	3.70%	0.94

Visualization of ASM. In order to prove the effectiveness of ASM, we visualize several face images. From the highlighted regions in the Fig. 6(a), we observe that the network tends to prefer larger skin areas when selecting effective face areas, regardless of the angle of the face. Furthermore, Fig. 6(b) also clearly shows the detailed rPPG signal estimated from different regions. We observe that the quality of rPPG signal is relatively consistent with the focus by ASM: the more periodic the signal area, the more attention the network tends to pay. Please note that in those regions containing background pixels (e.g., ②), only facial part is considered in our data-processing as mentioned in Sect. 3.4.

**Fig. 6.** (a) Visualization of the discriminating face regions that the network focused on. (b) Illustration of rPPG estimated from different face regions.

5 Conclusion

In this paper, we propose a Hierarchical Attentive Upsampling Module (HAUM) for acquiring rich and discriminating input signals based on the Spatial-Temporal

Map for heart rate estimation. The Hierarchical Upsampling Strategy is used to progressively enrich the Spatial-Temporal information, and the Attentive Space Module is cascaded to select effective regions with clearer periodicity. The performances on two public datasets demonstrate the superiority of our approach.

References

1. Brüser, C., Antink, C.H., Wartzek, T., Walter, M., Leonhardt, S.: Ambient and unobtrusive cardiorespiratory monitoring techniques. *IEEE Rev. Biomed. Eng.* **8**, 30–43 (2015)
2. Chen, W., McDuff, D.: DeepPhys: video-based physiological measurement using convolutional attention networks. In: *Proceedings of IEEE ECCV*, pp. 349–365 (2018)
3. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rPPG. *IEEE Trans. Biomed. Eng.* **60**(10), 2878–2886 (2013)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of IEEE CVPR*, pp. 770–778 (2016)
5. Heusch, G., Anjos, A., Marcel, S.: A reproducible study on remote heart rate measurement. *arXiv preprint [arXiv:1709.00962](https://arxiv.org/abs/1709.00962)* (2017)
6. Hsu, G.S., Ambikapathi, A., Chen, M.S.: Deep learning with time-frequency representation for pulse estimation from facial videos. In: *2017 IEEE International Joint Conference on Biometrics*, pp. 383–389. IEEE (2017)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of IEEE CVPR*, pp. 7132–7141 (2018)
8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv preprint [arXiv:1502.03167](https://arxiv.org/abs/1502.03167)* (2015)
9. Jain, P.K., Tiwari, A.K.: Heart monitoring systems-a review. *Comput. Biol. Med.* **54**, 1–13 (2014)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)* (2014)
11. Kwon, S., Kim, J., Lee, D., Park, K.: ROI analysis for remote photoplethysmography on facial video. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4938–4941. IEEE (2015)
12. Lam, A., Kuno, Y.: Robust heart rate measurement from video using select random patches. In: *ICCV*, pp. 3640–3648 (2015)
13. Lee, E., Chen, E., Lee, C.Y.: Meta-RPPG: remote heart rate estimation using a transductive meta-learner. In: *Proceedings of IEEE ECCV* (2020)
14. Lempe, G., Zaunseder, S., Wirthgen, T., Zipser, S., Malberg, H.: ROI selection for remote photoplethysmography. In: Meinzer, H.P., Deserno, T., Handels, H., Tolxdorff, T. (eds.) *Bildverarbeitung für die Medizin 2013*, pp. 99–103. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36480-8_19
15. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: *Proceedings of IEEE CVPR*, pp. 510–519 (2019)
16. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: *Proceedings of IEEE CVPR*, pp. 4264–4271 (2014)
17. McDuff, D.: Deep super resolution for recovering physiological information from videos. In: *CVPR Workshops*, pp. 1367–1374 (2018)

18. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: ICML (2010)
19. Niu, X., Han, H., Shan, S., Chen, X.: VIPL-HR: a multi-modal database for pulse estimation from less-constrained face video. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11365, pp. 562–576. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20873-8_36
20. Niu, X., Shan, S., Han, H., Chen, X.: RhythmNet: end-to-end heart rate estimation from face via spatial-temporal representation. TIP **29**, 2409–2423 (2019)
21. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: Proceedings of IEEE ECCV (2020)
22. Niu, X., Zhao, X., Han, H., Das, A., Dantcheva, A., Shan, S., Chen, X.: Robust remote heart rate estimation from face utilizing spatial-temporal attention. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1–8. IEEE (2019)
23. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Eng. **58**(1), 7–11 (2010)
24. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt. Exp. **18**(10), 10762–10774 (2010)
25. Qiu, Y., Liu, Y., Arteaga-Falconi, J., Dong, H., El Saddik, A.: EVM-CNN: real-time contactless heart rate estimation from facial video. IEEE Trans. Multimed. **21**(7), 1778–1787 (2018)
26. Sabokrou, M., Pourreza, M., Li, X., Fathy, M., Zhao, G.: Deep-HR: fast heart rate estimation from face video under realistic conditions. arXiv preprint [arXiv:2002.04821](https://arxiv.org/abs/2002.04821) (2020)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
28. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **3**(1), 42–55 (2011)
29. Tsou, Y.Y., Lee, Y.A., Hsu, C.T., Chang, S.H.: Siamese-rPPG network: remote photoplethysmography signal estimation from face videos. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing, pp. 2066–2073 (2020)
30. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Opt. Express **16**(26), 21434–21445 (2008)
31. Wang, W., den Brinker, A.C., Stuijk, S., de Haan, G.: Algorithmic principles of remote PPG. IEEE Trans. Biomed. Eng. **64**(7), 1479–1491 (2016)
32. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: CBAM: convolutional block attention module. In: Proceedings of IEEE ECCV, pp. 3–19 (2018)
33. Yan, Y., Ma, X., Yao, L., Ouyang, J.: Noncontact measurement of heart rate using facial video illuminated under natural light and signal weighted analysis. Bio-Med. Mater. Eng. **26**(s1), S903–S909 (2015)
34. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In: ICCV, pp. 151–160 (2019)