Learning Motion-Robust Remote Photoplethysmography through Arbitrary Resolution Videos

Jianwei Li[1*], Zitong Yu[2*], Jingang Shi[1†]

[1]School of Software Engineering,Xi'an Jiaotong University
[2]Great Bay University
3121158008@stu.xjtu.edu.cn, zitong.yu@ieee.org, jingang@xjtu.edu.cn

Abstract

Remote photoplethysmography (rPPG) enables non-contact heart rate (HR) estimation from facial videos which gives significant convenience compared with traditional contact-based measurements. In the real-world long-term health monitoring scenario, the distance of the participants and their head movements usually vary by time, resulting in the inaccurate rPPG measurement due to the varying face resolution and complex motion artifacts. Different from the previous rPPG models designed for a constant distance between camera and participants, in this paper, we propose two plug-and-play blocks (i.e., physiological signal feature extraction block (PFE) and temporal face alignment block (TFA)) to alleviate the degradation of changing distance and head motion. On one side, guided with representative-area information, PFE adaptively encodes the arbitrary resolution facial frames to the fixed-resolution facial structure features. On the other side, leveraging the estimated optical flow, TFA is able to counteract the rPPG signal confusion caused by the head movement thus benefits the motion-robust rPPG signal recovery. Besides, we also train the model with a cross-resolution constraint using a two-stream dual-resolution framework, which further helps PFE learn resolution-robust facial rPPG features. Extensive experiments on three benchmark datasets (UBFC-rPPG, COHFACE and PURE) demonstrate the superior performance of the proposed method. One highlight is that with PFE and TFA, the off-the-shelf spatio-temporal rPPG models can predict more robust rPPG signals under both varying face resolution and severe head movement scenarios. The codes are available at https://github.com/LJW-GIT/Arbitrary_Resolution_rPPG.

Introduction

Heart rate (HR) is an important physiological signal which is widely used in many circumstances, especially for healthcare or medical purposes. Electrocardiography (ECG) and Photoplethysmograph (PPG)/Blood Volume Pulse (BVP) are the two most common methods of measuring heart activities. However, these sensors need to be attached to body parts, limiting their usefulness and scalability. Due to the inconvenience of long-term monitoring and discomfort for the users, traditional ways limit the application scenarios such as driver status assessment and patient health monitoring. To solve this problem, non-contact HR measurement, which aims to measure heart activity remotely, has become an increasingly popular research problem in physiological signal measurement in recent years.

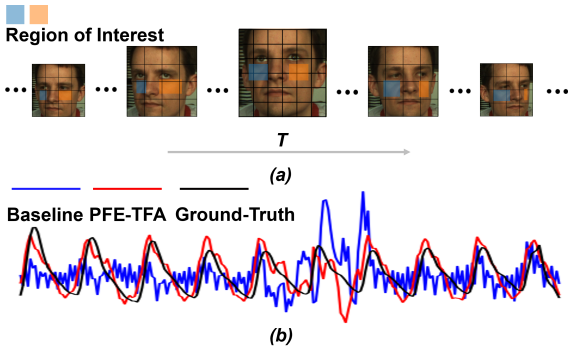

Figure 1: rPPG measurement from arbitrary resolution videos with head movements. (a) The ROIs might have different spatial size and shape across the temporal dimension. (b) Compared with the baseline PhysNet (Yu, Li, and Zhao 2019), the proposed PFE-TFA can predict more accurate rPPG signals with the ground truth BVP signals.

Most existing non-contact HR measurement approaches are based on the facial video-based remote Photoplethysmograph (rPPG) technique. The rPPG method uses digital cameras to record variations of reflected ambient light on facial skin, which contains information on cardiovascular blood volume and pulsation. However, the rPPG measurement is very susceptible and vulnerable to the quality of video recording and head motions. In the early stage, handcrafted features based methods (Takano and Ohta 2007; Verkruysse, Svaasand, and Nelson 2008) require an exhausted multi-stage process (preprocessing, filtering and post-processing) and are with low robustness to head motions and illumination changes. Thus, they are usually tested and deployed under controlled lab environment scenarios.

With the rapid development of deep learning, neural network models have also been widely applied in the rPPG field. Recent spatio-temporal representation map (Niu et al.

*Equal contribution
†Corresponding author

2020; Lu, Han, and Zhou 2021) based and end-to-end (Yu, Li, and Zhao 2019; Chen and McDuff 2018) models utilize Convolutional Neural Networks (CNNs) to learn the spatio-temporal rPPG cues from facial videos with fixed resolution, which have shown superior performance compared with traditional approaches. However, most of the existing rPPG approaches rarely consider the real-world practical situation with arbitrary resolution face videos (e.g., the distance of the participants varies by time, see Figure 1(a) for visualization). Previous methods (McDuff 2018) simply use the spatial interpolation method to scale the face frames with arbitrary resolutions to a fixed size to adapt the model input, where the vital pixels of the region of interest (ROI) might be confused by the interpolation method, thus harming the accuracy of rPPG measurement. Meta-SR (Hu et al. 2019) and LIIF (Chen, Liu, and Wang 2021) bring the scale-arbitrary super-resolution problem into the horizon. To our best knowledge, there is still no solution proposed yet to counter the problem of rPPG measurement from arbitrary face resolution videos.

Besides arbitrary face resolution, another noteworthy issue of rPPG measurement is the motion-robustness. Due to the limited spatio-temporal receptive field of CNN with a weak capacity of spatial contextual ROI localization, both rigid and non-rigid head motions usually have serious impacts on rPPG measurement. Existing end-to-end models (Yu, Li, and Zhao 2019) do not take initiative (e.g., face alignment operation) to describe the head movement to compensate the motion artifacts, resulting in the vulnerability to severe head movement. Few works passively adopt the ROI tracking (Niu et al. 2018) or utilize normalized frame difference motion representation as input (Chen and McDuff 2018; Liu et al. 2020). However, these methods have limited stability and are hard to directly plug in the off-the-shelf end-to-end spatio-temporal rPPG models.

Motivated by the discussions above, we propose two plug-and-play blocks (i.e., Physiological Signal Feature Extraction block (PFE) and Temporal Face Alignment block (TFA)) to capture resolution- and motion-robust rPPG features. To learn the similarity of rPPG signals from arbitrary resolution frames, we design a new cross-resolution constraint using a dual-resolution framework, which further helps PFE learn resolution-robust facial rPPG features. As shown in Figure 1(b), compared with the vanilla PhysNet (Yu, Li, and Zhao 2019), the proposed PFE and TFA blocks can benefit more accurate rPPG measurement from arbitrary resolution videos with head movements. The contributions of this work are as follows:

- To our best knowledge, we provide the first solution to plug-and-play modules on robust rPPG measurement from facial videos with arbitrary fixed resolution or varying face resolution.
- We propose the PFE block to adaptively encode the arbitrary resolution facial frames to the fixed-resolution facial structure features. Besides, we propose to train the model with a cross-resolution constraint using a two-stream dual-resolution framework, which further helps PFE learn resolution-robust facial rPPG features.
- We propose the TFA block to counteract the rPPG signal confusion caused by the head movement via wrapping facial frames by the estimated optical flow, which benefits the motion-robust rPPG signal recovery and alleviates the influence of head movement.
- We conduct extensive experiments on benchmark datasets to demonstrate the superior performance of the proposed method under both arbitrary face resolution and severe head movement scenarios.

Related Work

Remote Photoplethysmography Measurement

Plenty of handcrafted rPPG measurement methods have been proposed since the researches (Takano and Ohta 2007; Verkruysse, Svaasand, and Nelson 2008) show the feasibility of recovering physiological signals through a digital camera. Some early works take traditional signal processing methods into consideration, which contain matrix transformation (Tulyakov et al. 2016; Shi et al. 2020), Least Mean Squares (Li et al. 2014), and Blind Source Separation (BSS) (Poh, McDuff, and Picard 2010a,b). In recent years, with the boost of deep learning methods, Deephys (Chen and McDuff 2018) and PhysNet (Yu, Li, and Zhao 2019) firstly introduce end-to-end CNN framework to this field. Meanwhile, spatio-temporal signal map based methods (Niu et al. 2020; Lu, Han, and Zhou 2021) also attract more attention due to their excellent performance. Towards efficient rPPG measurement, Auto-HR (Yu et al. 2020) and EfficientPhys (Liu et al. 2022) search and design lightweight end-to-end models. Recently, PhysFormer (Yu et al. 2022, 2023) gains progress via temporal difference transformers to explore the long-range spatio-temporal relationship for rPPG representation. Besides supervised learning with labeled facial videos, unsupervised learning has also been validated in (Gideon and Stent 2021) to achieve rPPG measurement. The vulnerability of rPPG models has also been discussed recently, such as phase difference (Mironenko et al. 2020; Moço et al. 2018), camera rolling (Moço et al. 2018; Zhan et al. 2020) and video compression format (McDuff, Blackford, and Estepp 2017).

Face Resolution and its Impact for RPPG

Face super-resolution (Shi et al. 2018; Shi and Zhao 2019) is widely used to amplify the resolution of low-quality face images. In real-world applications, the distances between the camera and participants are variant, resulting in arbitrary resolution on the facial region. Some recent works explore rPPG extraction from low-resolution videos with fixed sizes. They use super-resolution models through the end-to-end (McDuff 2018) or two-stage (Song et al. 2020; Yue et al. 2021) network to recover the high-quality face videos and corresponding rPPG signals simultaneously. However, the target of super-resolution is mainly toward the recovery of visual performance (Shi et al. 2022), but not to maintain the quality of rPPG signals. Furthermore, it is still a challenge to obtain reasonable rPPG signals in the arbitrary face resolution scenario, which is more practical in the real-world application.

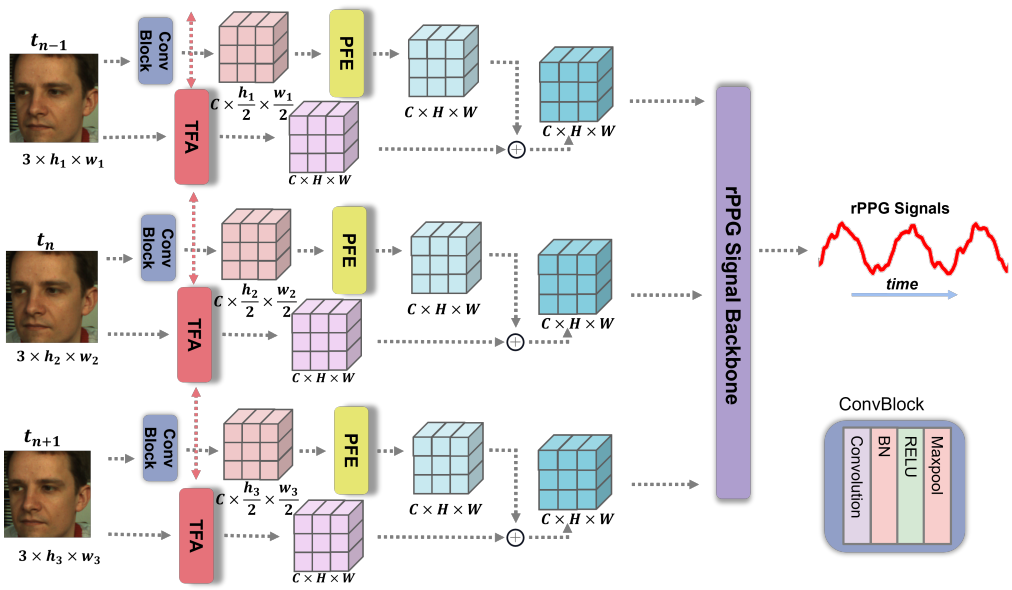

Figure 2: Overall framework of the proposed method. The frames in the sequence could have arbitrary resolutions. In the spatial stream, each arbitrary resolution face frame in the sequence forwards the PFE block, which maps the arbitrary size features to the fixed size facial structure features. In the temporal stream, the TFA block interpolates the frames to the same shape to generate the temporal aligned features. The facial structure and temporal aligned features are added to form the facial structure-motion features. Finally, the facial structure-motion features forward a rPPG Signal backbone to predict the rPPG signals.

Methodology

Overall Framework

As shown in Figure 2, given an arbitrary resolution face sequence $X = [x_1, x_2, ..., x_i], x_i \in \mathbb{R}^{3\times h_i \times w_i}, i \in \{1, ..., T\}$ as input, the proposed method forwards the two-stream pathway including the Physiological Signal Feature Extraction block (PFE) and the Temporal Face Alignment block (TFA) to form the facial structure-motion features. Then the rPPG backbone (e.g., PhysNet (Yu, Li, and Zhao 2019)) is used for rPPG signals prediction. Note that the arbitrary width h_i and height w_i of each frame can be different.

Before the PFE stream, we adopt a ConvBlock to extract features $X_{ar} = [x^1_{ar}, x^2_{ar}, ..., x^i_{ar}]$, $x_{ar} \in \mathbb{R}^{C\times \frac{h_i}{2} \times \frac{w_i}{2}}$ from the arbitrary resolution face sequence X. Specifically, the ConvBlock is formed by a convolutional block with kernel size ($1 \times 5 \times 5$) cascaded with batch normalization (BN), $RELU$, and $MaxPool$, where the pooling layer halves the spatial dimension. Then X_{ar} forwards the PFE block to generate facial structure features $X_{st} \in \mathbb{R}^{T\times C\times H\times W}$. T, C, H, W indicate constant clip length, channel, height, and width, respectively.

In term of the TFA stream, the TFA block first interpolates the arbitrary resolution sequence X to $\hat{X} \in \mathbb{R}^{T\times 3\times H\times W}$, which has the same height and width as X_{st}. Then, TFA uses the bi-directional optical flow (forward and backward) from successive frames to obtain temporal face alignment features $X_{mo} \in \mathbb{R}^{T\times C\times H\times W}$.

The output of PFE X_{st} and the output of TFA X_{mo} are summed to form the facial structure-motion features $X_{st-mo} \in \mathbb{R}^{T\times C\times H\times W}$. Finally, the rPPG signal backbone predicts the 1D rPPG signal $Y \in \mathbb{R}^T$ from X_{st-mo}.

Physiological Signal Feature Extraction (PFE)

The PFE block is used in the spatial dimension for each face frame. As shown in Figure 3, the proposed PFE contains two parts. The upper and lower branches are devised for facial information and position information respectively. For upper branch, the features x_{ar} from arbitrary-resolution frame are first interpolated to constant resolution features $x_{cr} \in \mathbb{R}^{C\times H\times W}$. To exploit the facial features, **receptive field expansion** is conducted as Eq.(1) to obtain the expanded features $\hat{x}_{cr} \in \mathbb{R}^{(n^2C)\times H\times W}$. Meanwhile, in the lower branch, **Representative area encoding (RAE)** is employed as Eq.(2) to record the mapping relationship of pixel positions between x_{ar} and x_{cr}. The relationship is described as coordinate tensor $x_{size} \in \mathbb{R}^{2\times H\times W}$, where two channels represent the scaling ratio on the height and width accordingly. Then, the expanded features $\hat{x}_{cr}$ together with coordinate tensor x_{size} are fed into the **facial feature encoding** as Eq.(3) to produce the facial structure features $x_{st} \in \mathbb{R}^{C\times H\times W}$.

Receptive field expansion. To enrich and mine the contextual information contained in the facial structure features x_{cr}, we unfold the facial structure features x_{cr} first, and then expand its receptive field via concatenating the $n \times n$ neighboring features to obtain $\hat{x}_{cr}$. Formally, the receptive field expansion is defined as

$$\hat{x}_{cr}(i, j) = Concat(\{x_{cr}(i + n, j + n)\}_{n\in Neighbor}), \quad (1)$$

where i and j indicate the spatial position of the features. The number of neighbors n is set to 3.

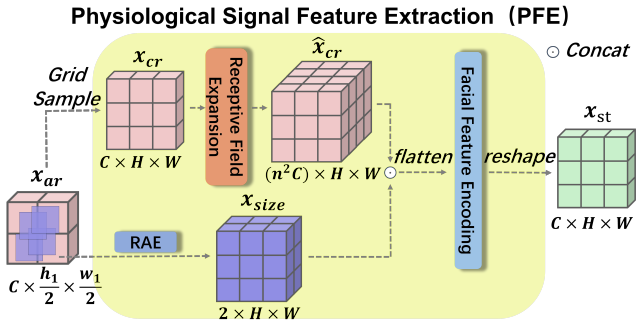

Figure 3: The structure of the PFE block.x_{ar} means the arbitrary resolution feature maps. x_{st} means the facial structure feature maps.

Representative area encoding (RAE). As the arbitrary size features x_{ar} have different spatial sizes with the structure features x_{st}, spatial positions in x_{st} correspond to different areas from x_{ar}. It is important to explicitly describe the representative area for each spatial position. Here we formulate the representative area information $x_{size} \in \mathbb{R}^{2\times H\times W}$ as

$$x_{size}\,(i,j) = [\sigma_H, \sigma_W], \\ \sigma_H = \frac{h}{H}, \sigma_W = \frac{w}{W}, \tag{2}$$

where σ_H and σ_W mean the scaling ratio on the height and width dimensions when x_{ar} is transformed to x_{st}.

Facial feature encoding. We concatenate $\hat{x}_{cr}$ and x_{size} along the channel dimension. A shallow facial feature encoding function is designed to mine the semantic facial structure features, which is simply parameterized as an MLP. Facial feature encoding takes the form:

$$x_{st} = Reshape\,(MLP\,(Flatten(Concat\,(\hat{x}_{cr}, x_{size})))) \tag{3}$$

After extracting the facial structure features $x_{st} \in \mathbb{R}^{C\times H\times W}$ from each frame, we merge them in temporal dimension to form the $X_{st} \in \mathbb{R}^{T\times C\times H\times W}$.

Temporal Face Alignment (TFA)

The facial structure features X_{st} from the PFE block have rich representation capacity on the arbitrary resolution condition. However, in practice, head movement influences end-to-end rPPG measurement significantly. For example, the huge-angle rotation could make partial facial features out of the scope of the facial structure features X_{st}. Previous works use landmark detection methods such as OpenFace (Baltrušaitis, Robinson, and Morency 2016) to extract the face landmarks for facial ROI alignment. However, the robustness of rPPG measurement is highly dependent on the accuracy of face landmarks. Here, three problems are noted for face alignment:

1. Landmarks status. The position of face landmarks could change dramatically because of head motion, which induces the inaccurate detection of ROIs in the facial clips.

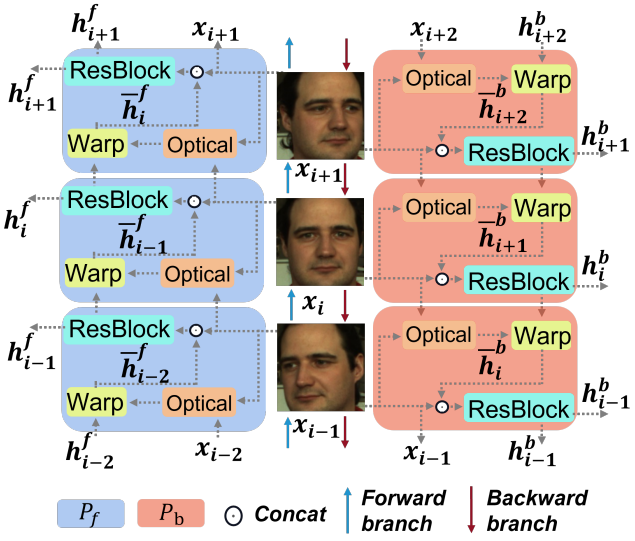

Figure 4: The structure of TFA block. P_f and P_b mean the forward and backward optical flow blocks for face alignment.

2. Interpolation. The shape of ROI might be different, and interpolation is usually used to keep the consistency (Hu et al. 2021). However, interpolation may corrupt the color change of pixels, and eliminate the rPPG cues.
3. Lost landmarks. When the head movement encounters huge-angle rotation, partial face may disappear from the frame. In this case, the predicted landmarks would mark some regions randomly.

Head rotation is a continuous process, and the state of each frame is correlated with the forward and backward states. According to these observations, we propose the temporal face alignment (TFA) block to leverage optical flow to describe the facial motion and wrap the facial structure features.

As shown in Figure 4, TFA adopts a typical bidirectional recurrent network. The video sequence $\hat{X}$ forwards **optical flow face alignment** as Eq.(4) to get head motion features $H^{b,f} = \left[h_1^{b,f}, h_2^{b,f}, ..., h_i^{b,f}\right]$, $h_i^{b,f} \in \mathbb{R}^{C\times H\times W}$. Then, h^b and h^f are fed into **bidirectional aggregation** as Eq.(5) to produce the facial structure features x_{st}.

Optical flow face alignment. The video sequence $\hat{X}$ is first fed into 'Optical' to calculate the optical flow s_i by SPyNet (Ranjan and Black 2017), which is constituted by six convolution blocks with six cascaded 7×7 convolutions with RELU in each block. Then, the optical flow s_i is utilized to 'Warp' the head motion features h_{i-1} of previous frame to get aligned features $\bar{h}_{i-1}$ for frontalizing faces. Notice status h_0 is initialized by features of all zeros. The aligned features together with the recent frame are then passed to 15 basic residual blocks for obtaining h_i, where each block is constructed by cascaded 3×3 convolution, RELU, and 3×3 convolution with residual connection. Op-

tical flow face alignment takes the form:

$$\begin{aligned} s_i^{b,f} &= Optical\left(x_i, x_{i\pm1}\right), \\ \bar{h}_{i\pm1} &= Wrap\left(h_{i\pm1}, s_i^{b,f}\right), \\ h_i^{b,f} &= ResBlock\left(Concat\left(x_i, \bar{h}_{i\pm1}\right)\right), \end{aligned} \tag{4}$$

The head motion features h_i can be represented from two temporal directions (i.e., forward and backward). Thus, we use h_i^f and h_i^b to represent the h_i via forwarding and backwarding x_i.

Bidirectional aggregation. To aggregate the backward and forward head motion features, we concatenate h_i^b and h_i^f along the channel dimension and introduce a convolutional layer to maintain the number of channels. Formally, the bidirectional aggregation is defined as

$$x_{mo} = F\left(Concat\left(h_i^b, h_i^f\right)\right) \tag{5}$$

where $F(\cdot)$ represents an 1×1 convolutional layer and $x_{mo} \in \mathbb{R}^{C\times H\times W}$ represents generated temporal face alignment features.

Finally, the facial structure features x_{st} are added to x_{mo} to obtain the facial structure-motion features $x_{st-mo} \in \mathbb{R}^{C\times H\times W}$.

Cross-Resolution Constraint and Loss Functions

Despite we have designed the PFE block to tackle the arbitrary resolution problem, it is still hard to learn resolution-invariant rPPG features with the traditional Negative Pearson loss $\mathcal{L}_{time}$ (Yu, Li, and Zhao 2019) and frequency cross-entropy loss $\mathcal{L}_{fre}$ (Niu et al. 2020). Here we design a novel cross-resolution constraint $\mathcal{L}_{crc}$ which forces the model to learn consistent rPPG predictions between two resolution views. Specifically, as shown in Figure 5, we sample video clip into different resolutions as $X_1 \in \mathbb{R}^{T\times 3\times h1\times w1}$ and $X_2 \in \mathbb{R}^{T\times 3\times h2\times w2}$. The two sampled clips forward the unshared PFE and shared TFA blocks first, and then go through a shared rPPG backbone model to predict the corresponding rPPG signals $Y_1 \in \mathbb{R}^{T\times 1}$ and $Y_2 \in \mathbb{R}^{T\times 1}$. The cross-resolution constraint $\mathcal{L}_{crc}$ can be formulated via calculating the L1 distance between two predicted signals. The overall loss function $\mathcal{L}_{overall}$ can be formulated as

$$\begin{aligned} \mathcal{L}_{crc} &= \| Y_1 - Y_2 \|_1, \\ \mathcal{L}_{overall} &= \mathcal{L}_{time} + \mathcal{L}_{fre} + \alpha \cdot \mathcal{L}_{crc}, \end{aligned} \tag{6}$$

where hyperparameter α equals to 0.1. The loss function avoids the model to pay attention to the similarity of low-level features from different resolution. In other words, $\mathcal{L}_{crc}$ focuses on the consistency of predicted rPPG signals, instead of the feature-level consistency, which determines the performance measurement and provides the direct supervision signals for the model learning.

Experiment

We first conduct experiments of rPPG-based HR measurement on three benchmark datasets with their original protocols and normal setting. Then, the UBFC-rPPG (Bobbia et al. 2019) dataset is used for performance evaluation on arbitrary-resolution facial videos and ablation studies.

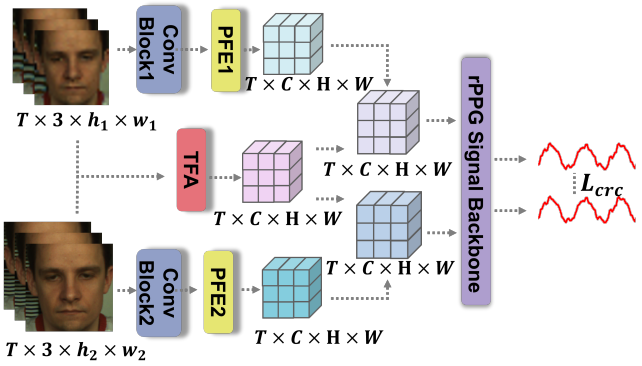

Figure 5: The framework of the Cross-Resolution Constraint. It is calculated on the two input sequences of arbitrary resolutions.

UBFC-rPPG. The UBFC-rPPG dataset (Bobbia et al. 2019) includes 42 videos, which are about 2 minutes long and recorded. The bio-signal ground-truth was recorded by a pulse oximeter with a 60 Hz sampling rate.

PURE. The PURE dataset (Stricker, Müller, and Gross 2014) contains 60 videos from 10 subjects performing six different head motion tasks: steady, talking, slow translation, fast translation, small rotation, and medium rotation.

COHFACE. The COHFACE dataset (Heusch, Anjos, and Marcel 2017) is consisted of 160 one-minute videos from 40 healthy individuals, captured under studio and natural light. The videos are heavily compressed using MPEG-4 Visual, which was noted by (McDuff, Blackford, and Estepp 2017) to potentially cause corruption of the rPPG signal.

Implementation Details

Experimental settings. For each video clip, we use the MTCNN (Zhang et al. 2016) to crop the enlarged face area and resize each frame to 128×128 pixels. And then we downsample the face image ranging from 1.0 to 4.0 times to get the arbitrary scale frames. The facial video clip with arbitrary sizes would be mapped to the fixed size H=W=64 after PFE and TFA blocks while the number of channels is set to C=16. Random horizontal flipping is used for clip-level spatial data augmentation. The proposed method is trained with batchsize 2 on RTX3090 GPU with PyTorch. The Adam optimizer is used and the learning rate is set as 1e-4. The weight decay is 5e-5.

Metrics and evaluation. Following (Comas, Ruiz, and Sukno 2022), we calculate the root mean squared error (RMSE) and mean absolute error (MAE) between the predicted average HR versus the groundtruth HR. We first forward the models using 160-frame clips without overlapping to predict clip-level HR. The comparisons on whole-videos are calculated via averaging the clip-level predictions.

Comparison on Normal Face Resolution Setting

For fair comparison, we train the baseline model PhysNet (Yu, Li, and Zhao 2019) and our methods using the same

Method	UBFC		PURE		COHFACE	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
CHROM	3.44	4.61	2.07	2.5	-	-
POS	2.44	6.61	3.14	10.57	-	-
HR-CNN	-	-	1.84	2.37	8.10	10.8
DeepPhys	2.90	3.63	1.84	2.31	-	-
Zhan et al.	2.44	3.17	1.82	2.29	-	-
Gideon et al.	3.60	4.60	2.30	2.90	2.30	7.60
AND-rPPG	2.67	4.07	-	-	-	-
TDM	2.32	3.08	1.83	**2.33**	-	-
PhysNet	2.38	3.19	2.16	2.7	5.38	10.76
PFE-TFA(Ours)	**0.76**	**1.62**	**1.44**	2.50	**1.31**	**3.92**

Table 1: HR estimation results (bpm) on UBFC, PURE, and COHFACE datasets.

recipe to avoid bias. We compare our method with eight state-of-the-art methods (CHROM (De Haan and Jeanne 2013), POS (Wang et al. 2016), HR-CNN (Špetlík, Franc, and Matas 2018), DeepPhys (Chen and McDuff 2018), Zhan et al. (Zhan et al. 2020), Gideon et al. (Gideon and Stent 2021), AND-rPPG (Lokendra and Puneet 2022), TDM (Comas, Ruiz, and Sukno 2022)) in Table 1.

Results on UBFC-rPPG. It can be seen from Table 1 that the vanilla 3DCNN-based PhysNet performs worse than the TDM method. When assembled with the proposed PFE and TFA blocks, the PhysNet+PFE+TFA achieves the best performance on UBFC, outperforming the TDM by 1.56 bpm on MAE. Compared to the PhysNet performance, the proposed PFE and TFA blocks improve the baseline results by 1.62 bpm on MAE and 1.57 bpm on RMSE for UBFC dataset, indicating the effectiveness of robust rPPG features representation via PFE-TFA blocks.

Results on PURE. As shown in Table 1, compared with the baseline PhysNet, the proposed PFE and TFA blocks improve the MAE performance from 2.16 bpm to 1.44 bpm on PURE. It indicates that PFE-TFA is able to represent more motion-robust rPPG features as there are plenty of hard samples with severe head movement in PURE. As for the RMSE metric, the proposed method performs slightly worse than TDM (+0.17 bpm), which is mainly caused by the difference of rPPG backbones (two extra differential temporal convolutions are used in TDM). Please note that the proposed PFE and TFA block could also be plugged into TDM to further improve performance.

Results on COHFACE. Compared with UBFC and PURE, the face videos are highly compressed on COHFACE, resulting in obvious compression artifacts. As can be seen from the last column of Table 1, existing supervised CNN-based methods (HR-CNN, Gideon et al., and PhysNet) perform poorly (>5 bpm RMSE) due to the low face video quality. Meanwhile, the proposed method achieves state-of-the-art performance with 1.31bpm on MAE, which outperforms previous methods by a large margin.

Comparison on Arbitrary Face Resolution Settings

In the following experiments, we downsample the face images from the UBFC dataset with two settings: fixed face resolution and varying face resolution. The former describes the long-distance scenario while the latter mimics the varying face-camera-distance scenario.

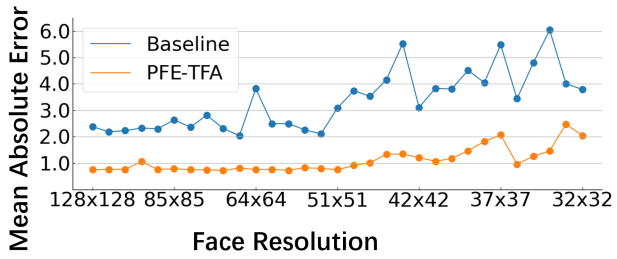

Figure 6: HR estimation results (bpm) on UBFC-rPPG under different fixed face resolution settings.

Model \ Resolution	128-64	64-32	128-64-128
Baseline	10.73	8.14	11.85
PFE	3.49	3.84	3.29
TFA	5.44	5.04	5.14
PFE-TFA	**1.86**	**1.97**	**1.85**

Table 2: MAE results (bpm) of the PFE and TFA blocks on UBFC under varying face resolution settings. '128-64' means that the face resolution gradually varies from 128×128 to 64×64 in a video clip.

Results on fixed face resolution. It can be seen from Figure 6 that the baseline PhysNet is easily influenced by the face resolution. When the fixed face resolutions are smaller than 50×50, the performance of PhysNet drops sharply (e.g., MAE>4 bpm). In contrast, when assembled with the proposed PFE and TFA blocks, it can predict accurate rPPG-based HR (MAE<2 bpm) in most face solution settings.

Results on varying face resolution. rPPG measurement from varying face resolution video is challenging due to the complex temporal contextual interference. The results of varying face resolution on UBFC are shown in Table 2. Compared with the vanilla PhysNet, the proposed PFE and TFA blocks benefit the facial feature alignment and refinement among consecutive frames, thus improving the MAE

Model \ Resolution	128×128	96×96	64×64
Baseline	2.38	2.64	3.82
PFE w/o RAE	3.79	3.77	3.76
PFE w/o $\mathcal{L}_{crc}$	2.87	2.85	2.84
PFE	2.40	2.39	2.33
TFA	8.73	8.30	8.29
PFE-TFA(single)	2.28	2.29	2.44
PFE-TFA	**0.76**	**0.78**	**0.76**

Table 3: MAE results (bpm) of the PFE and TFA blocks on UBFC under different fixed face resolution settings.

performance (8.87 bpm, 6.17 bpm, and 10.00 bpm) in three scenarios, respectively.

Ablation Study

We also provide the ablation studies of the proposed modules under arbitrary face resolution and severe head movement scenarios on the UBFC dataset.

Efficacy of the cross-resolution constraint. In the default setting, the models with PFE are trained in two views with different face resolution frames using a cross-resolution constraint $\mathcal{L}_{crc}$. In this study, we consider how $\mathcal{L}_{crc}$ impacts the PFE. As shown in Table 3, 'PFE' outperforms 'PFE w/o $\mathcal{L}_{crc}$' by a convincing margin (0.4 to 0.5 bpm MAE) for almost all different face resolution settings. It indicates such cross-resolution constraint benefits the PFE learning resolution-robust rPPG cues.

Efficacy of the PFE block on arbitrary face resolution. Here we investigate the impacts of the representative area encoding (RAE) of PFE on fixed face resolution UBFC first. It can be seen from the results 'PFE w/o RAE' and 'PFE' in Table 3 that the PFE could not achieve well results without RAE under high-resolution cases. When equipped PFE with RAE, it can achieve robust HR estimation under all different fixed face resolution settings. Besides, under more challenging varying face resolution scenarios, we can also find the consistent conclusion from Table 2 that PFE significantly improves the baseline performance (reducing 8.87 bpm MAE for the varying resolution of 128-64 scenario).

Efficacy of the TFA block on arbitrary face resolution. As shown in the result of 'TFA' in Table 3, the baseline with only TFA block performs even worse than the baseline itself. It is because when estimating the optical flow, all clips are interpolated to a fixed resolution, which makes the TFA block weak in describing rPPG-aware color area (Xue et al. 2019). We can find from the result 'PFE-TFA' that the best performance can be achieved for all different fixed face resolution settings when assembling the baseline with both PFE and TFA blocks. A similar conclusion can be also drawn from the varying face resolution setting in Table 2. Moreover, we also consider the online testing case that the temporal alignment state from backward frames is not available. In this case, we design a TFA block with a single forward direction for facial feature alignment. It can be seen from the result 'PFE-TFA(single)' in Table 3 that the MAE

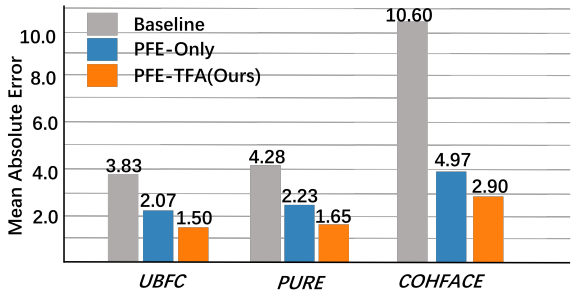

Figure 7: HR estimation results (bpm) on the samples with severe head movement and huge face rotation.

Model	Params	RTX 3090	AGX Orin
Physnet	0.75MB	10ms	0.25s
Physnet+PFE	0.78MB	17ms	3.63s
Physnet+PFE+TFA	1.31MB	41ms	4.82s

Table 4: The number of parameters for the model and inference time on edge devices.

performance is degraded by about 1.5 bpm compared to bi-directional TFA, but it still improves the performance of the model compared to baseline PhysNet.

Efficacy of the PFE and TFA blocks on severe head movement. To confirm the efficacy of the PFE and TFA blocks on severe head movements with large angle rotation, we also conduct studies on the carefully selected videos with large angle rotation from UBFC, PURE, and COHFACE. Specifically, the participants in COHFACE quickly rotate their heads at an average angle of 80°, while the participants in UBFC and PURE rotate their heads very slowly at an average angle of 35°. The results are shown in Figure 7. We can find that 1) compared with baseline PhysNet, the PFE reduces MAE by 1.76 bpm, 2.05 bpm, and 5.63 bpm on UBFC, PURE and COHFACE, respectively; 2) the proposed PFE-TFA further decreases the MAE by 0.57 bpm, 0.58 bpm, and 2.07 bpm on these datasets due to the excellent motion-robust capacity of TFA.

Edge Deployment. Considering the deployment on edge devices, we provide a detailed analysis of complexity as well as the inference times on different devices (i.e., Nvidia RTX 3090 and Jetson AGX Orin) in Table 4. The additional parameters induced by PFE are 37KB, while TFA causes an increase of 539KB. It shows the proposed model achieves the target by a moderate additional burden.

Conclusion

In this paper, we propose two plug-and-play blocks, namely PFE and TFA, for remote physiological measurement. With the above two proposed blocks, the baseline model is able to achieve superior performance on benchmark datasets with arbitrary resolution. Future directions include: 1) designing lightweight network architecture to achieve reasonable performance; 2) discussion on the influence of variant video qualities and compression rate.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62002283, 62071380), and the Fundamental Research Funds for the Central Universities.

References

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. IEEE.

Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; and Dubois, J. 2019. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124: 82–90.

Chen, W.; and McDuff, D. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, 349–365.

Chen, Y.; Liu, S.; and Wang, X. 2021. Learning Continuous Image Representation with Local Implicit Image Function. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8624–8634.

Comas, J.; Ruiz, A.; and Sukno, F. 2022. Efficient Remote Photoplethysmography with Temporal Derivative Modules and Time-Shift Invariant Loss. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2181–2190.

De Haan, G.; and Jeanne, V. 2013. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10): 2878–2886.

Gideon, J.; and Stent, S. 2021. The Way to my Heart is through Contrastive Learning: Remote Photoplethysmography from Unlabelled Video. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3975–3984.

Heusch, G.; Anjos, A.; and Marcel, S. 2017. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*.

Hu, C.; Zhang, K.-Y.; Yao, T.; Ding, S.; Li, J.; Huang, F.; and Ma, L. 2021. An End-to-end Efficient Framework for Remote Physiological Signal Sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2378–2384.

Hu, X.; Mu, H.; Zhang, X.; Wang, Z.; Tan, T.; and Sun, J. 2019. Meta-SR: A Magnification-Arbitrary Network for Super-Resolution. *Conference on Computer Vision and Pattern Recognition*, 1575–1584.

Li, X.; Chen, J.; Zhao, G.; and Pietikainen, M. 2014. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4264–4271.

Liu, X.; Fromm, J.; Patel, S.; and McDuff, D. 2020. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33: 19400–19411.

Liu, X.; Hill, B. L.; Jiang, Z.; Patel, S.; and McDuff, D. 2022. EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Vitals Measurement. *IEEE/CVF Winter Conference on Applications of Computer Vision*.

Lokendra, B.; and Puneet, G. 2022. AND-rPPG: A novel denoising-rPPG network for improving remote heart rate estimation. *Computers in biology and medicine*, 141: 105146.

Lu, H.; Han, H.; and Zhou, S. K. 2021. Dual-GAN: Joint BVP and Noise Modeling for Remote Physiological Measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12404–12413.

McDuff, D. 2018. Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1367–1374.

McDuff, D. J.; Blackford, E. B.; and Estepp, J. R. 2017. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *12th IEEE International Conference on Automatic Face & Gesture Recognition*, 63–70. IEEE.

Mironenko, Y.; Kalinin, K.; Kopeliovich, M.; and Petrushan, M. 2020. Remote photoplethysmography: Rarely considered factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 296–297.

Moço, A.; Stuijk, S.; van Gastel, M.; and de Haan, G. 2018. Impairing factors in remote-ppg pulse transit time measurements on the face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1358–1366.

Niu, X.; Han, H.; Shan, S.; and Chen, X. 2018. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In *Asian conference on computer vision*, 562–576.

Niu, X.; Yu, Z.; Han, H.; Li, X.; Shan, S.; and Zhao, G. 2020. Video-based remote physiological measurement via cross-verified feature disentangling. In *European Conference on Computer Vision*, 295–310.

Poh, M.-Z.; McDuff, D. J.; and Picard, R. W. 2010a. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1): 7–11.

Poh, M.-Z.; McDuff, D. J.; and Picard, R. W. 2010b. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10): 10762–10774.

Ranjan, A.; and Black, M. J. 2017. Optical Flow Estimation Using a Spatial Pyramid Network. In *Conference on Computer Vision and Pattern Recognition*, 2720–2729.

Shi, J.; Alikhani, I.; Li, X.; Yu, Z.; Seppänen, T.; and Zhao, G. 2020. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8): 2781–2795.

Shi, J.; Liu, X.; Zong, Y.; Qi, C.; and Zhao, G. 2018. Hallucinating face image by regularization models in high-resolution feature space. *IEEE Transactions on Image Processing*, 27(6): 2980–2995.

Shi, J.; Wang, Y.; Dong, S.; Hong, X.; Yu, Z.; Wang, F.; Wang, C.; and Gong, Y. 2022. Idpt: Interconnected dual pyramid transformer for face super-resolution. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1306–1312.

Shi, J.; and Zhao, G. 2019. Face hallucination via coarse-to-fine recursive kernel regression structure. *IEEE Transactions on Multimedia*, 21(9): 2223–2236.

Song, R.; Zhang, S.; Cheng, J.; Li, C.; and Chen, X. 2020. New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method. *Computers in biology and medicine*, 116: 103535.

Špetlík, R.; Franc, V.; and Matas, J. 2018. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, 1–12.

Stricker, R.; Müller, S.; and Gross, H.-M. 2014. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 1056–1062.

Takano, C.; and Ohta, Y. 2007. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8): 853–857.

Tulyakov, S.; Alameda-Pineda, X.; Ricci, E.; Yin, L.; Cohn, J. F.; and Sebe, N. 2016. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2396–2404.

Verkruysse, W.; Svaasand, L. O.; and Nelson, J. S. 2008. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26): 21434–21445.

Wang, W.; Den Brinker, A. C.; Stuijk, S.; and De Haan, G. 2016. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7): 1479–1491.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8): 1106–1125.

Yu, Z.; Li, X.; Niu, X.; Shi, J.; and Zhao, G. 2020. AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching. *IEEE Signal Processing Letters*, 27: 1245–1249.

Yu, Z.; Li, X.; and Zhao, G. 2019. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. *Proceedings of the British Machine Vision Conference (BMVC)*, 1–12.

Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Cui, Y.; Zhang, J.; Torr, P.; and Zhao, G. 2023. PhysFormer++: Facial Video-based Physiological Measurement with SlowFast Temporal Difference Transformer. *International Journal of Computer Vision*, 1–24.

Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Torr, P. H.; and Zhao, G. 2022. PhysFormer: facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4186–4196.

Yue, Z.; Ding, S.; Yang, S.; Yang, H.; Li, Z.; Zhang, Y.; and Li, Y. 2021. Deep super-resolution network for rPPG information recovery and noncontact heart rate estimation. *IEEE Transactions on Instrumentation and Measurement*, 70: 1–11.

Zhan, Q.; Wang, W.; Wang, W.; and Haan, G. d. 2020. Analysis of CNN-based remote-PPG to understand limitations and sensitivities. *Biomedical Optics Express*, 11(3): 1268–1283.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.