Time Lab's approach to the Challenge on Computer Vision for Remote Physiological Measurement

Yuhang Dong, Gongping Yang*, Yilong Yin
School of Software, Shandong University, China
dongyuhang@mail.sdu.edu.cn, gpyang@sdu.edu.cn, ylyin@sdu.edu.cn

Abstract

Computer vision for remote physiological measurement is novel and uniquely challenging task, which enables non-contact monitoring of the blood volume pulse (BVP) using a commonly accessible camera. This paper introduces Time Lab's approach presented at the 2nd challenge on Remote Physiological Signal Sensing (RePSS) organized within ICCV2021. We propose an end-to-end rPPGNet for remote photoplethysmographyraphy (rPPG) signals estimation. A improved design of spatial-temporal map is also made, which is an an efficient representation of the rPPG signal by removing most of the irrelevant background content. Furthermore, our approach achieved first place on the 2nd RePSS Challenge Track 1 and has outperformed the methods of other participants as we have achieved M_IBI = 117.25(4.51% improvement compared to the challenge top-2 result), R_HR = 0.62(8.77% improvement). The codes are publicly available at https://github.com/yuhang1070/2nd_RePSS_Track1_Top1_Solution.

1. Introduction

The 2nd Challenge of Remote Physiological Signal Sensing in ICCV2021 was organized by X.Li *et al.* Remote measurement of physiological signals from face videos is an emerging, challenging and promising topic. Hence, both scholars and companies have paid more attention to this topic and the number of published papers is growing every year.

However, many previous studies[21] only focused on the measurement of average heart rate (HR) from face videos, which is not sufficient for many medical applications (e.g., atrial fibrillation detection). Thus, more detailed information such as heart rate variability (HRV) features are needed, which requires accurate measurement of the time location of each heartbeat, i.e., the IBI curve. 2nd RePSS Challenge Track1 requires participants to reconstruct the IBI curve from raw face videos, which can be then processed to achieve detailed cardiac activity analysis. Raw face videos and corresponding BVP/ECG curves will be provided for training.

The training set of Track1 contains 2500 pieces of 10s videos of 500 persons, sampled from VIPL-HR-V2 database[13]. VIPL-HR-V2 database is a large-scale multi-modal database for remote HR estimation from face videos, the second version of VIPL-HR database[16]. This dataset was collected under less-constrained situations same as before. The testing set of Track 1 contains two parts, OBF[12] and VIPL-HR, with 1000 videos of 200 subjects in total. OBF is provided by the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland.

2. Related Work

Conventional Methods. rPPG is the monitoring of blood volume pulse from a camera at a distance. Verkruysse *et al.* [24] proved, for the first time, that plethysmography (PPG) signals can be measured remotely (>1m) from human face videos using ambient light. After that, many scholars have devoted their efforts in this challenging hot topic. Poh *et al.* [20] introduced a new methodology which applied independent component analysis (ICA) to reconstruct rPPG signals from raw RGB facial videos. Similarly, Lewandowska *et al.* [11] proposed a new rPPG signals estimation method based on principal component analysis (PCA). In comparison to ICA, PCA reduces computational complexity greatly. For improving motion robustness, Haan *et al.* [5] proposed a CHROM method, which linearly combines the RGB channels to separate pulse signal from motion-induced distortion. Wang *et al.* [26] proposed a "plane-orthogonal-to-skin"(POS) method. Both CHROM and POS are based on the skin reflection model.

Deep-learning based Methods. In order to overcome the conventional methods' limitations, scholars have tried to employ deep learning technology for remote physiolog-

*Corresponding author.

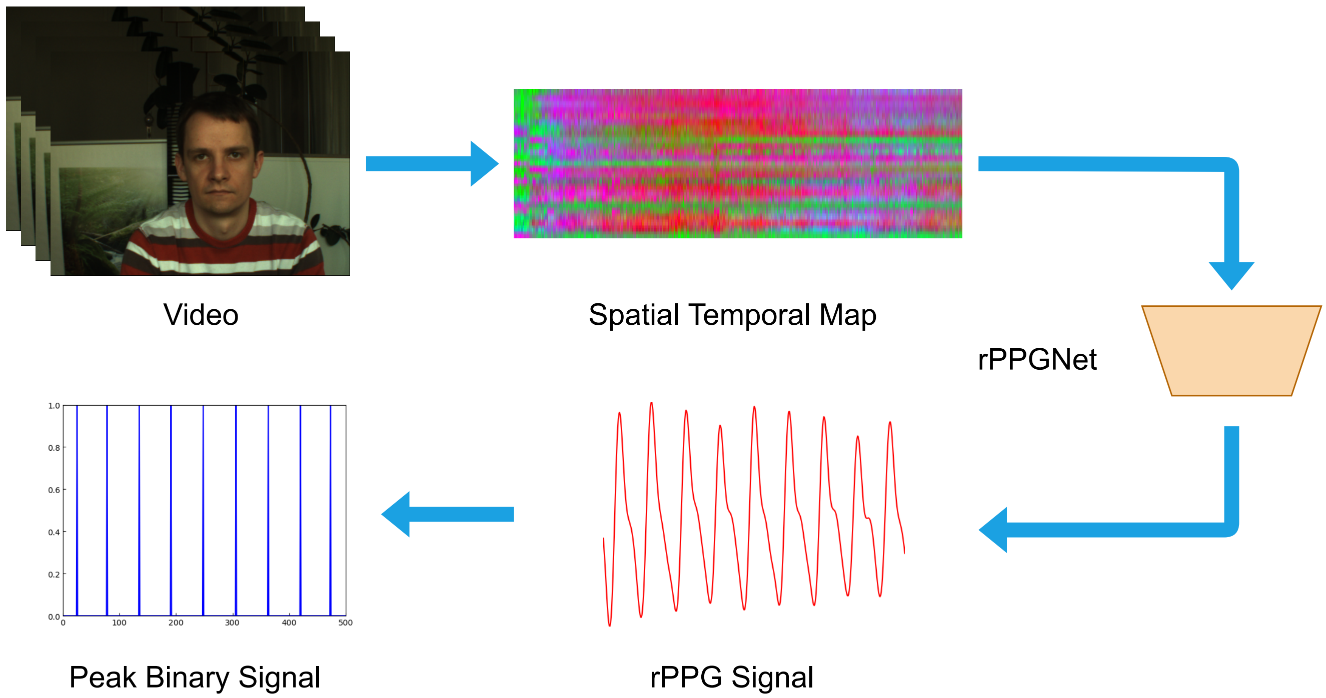

Figure 1. Time Lab's solution pipeline.

ical measurement in recent years. The first deep-learning based remote physiological measurement method is DeepPhys, which was originally proposed by Chen *et al.* [3]. DeepPhys is an end-to-end convolutional neural network (CNN) for video-based heart rate measurement. Spetlik *et al.* [22] proposed the HR-CNN which predicts remote HR from aligned face images using a two-step CNN. Niu *et al.* [17] designed a novel and efficient spatio-temporal map, which is mapped by a CNN to its HR value.

3. Methodology

As shown in Fig. 1, our pipeline can be divided into three steps: STmap generation, deep learning-based rPPG signal estimation and post-processing. In this section, we elaborate on each in detail.

3.1. Spatial Temporal Map

Many previous methods[10, 27] focused on direct applying CNNs to the human facial videos with good results. However, due to the low PSNR of rPPG signals in facial videos, those methods are expensive and time-consuming. In order to avoid high computational complexity and time-consuming, we choose to use Spatial Temporal Map (STMap) as the input of CNN. STmap is an efficient representation of the pulse signal by removing most of the irrelevant background content.

Unlike the previous design of STmap[17], we have made the following improvements: (1) We detect faces using RetinaFace[6] with MobileNet[8] backbone, which can get more precise face landmarks. (2) We appropriately reduced the region of interest (ROI) area by discarding non-skin facial areas such as eyes and mouth region. An example of ROI is shown in the Fig. 2. (3) Skin segmentation is applied to the defined ROI to remove the non-skin area such as hair region and background area by open source Bob[1] with threshold=0.05.

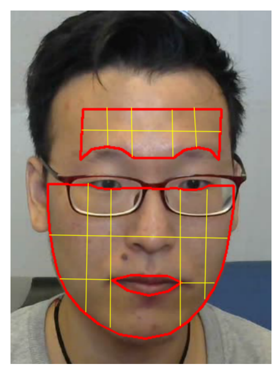

Figure 2. An example of ROI visualisation.

3.2. rPPGNet

The network architecture of rPPGNet is shown in Fig. 5. In order to balance computational complexity and performance, we adapt the strategy of appropriately reducing the number of channels and increasing the number of net-

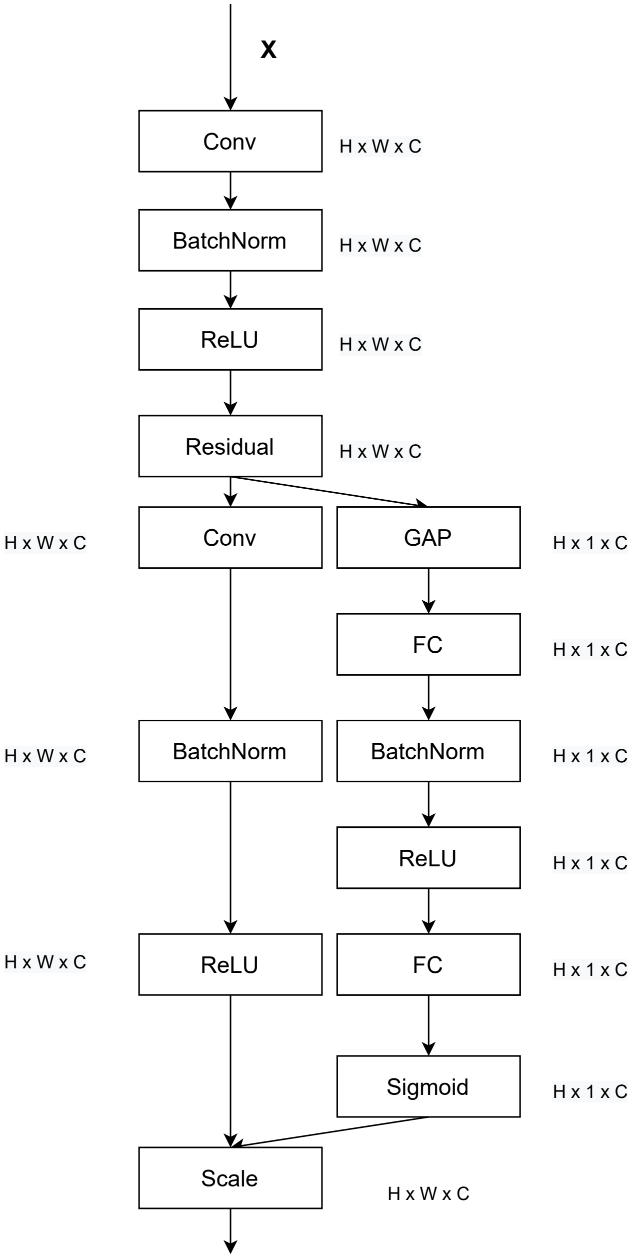

Figure 3. Attention Block. "Conv" denotes one convolution layer. "GAP" denotes global averaging pooling. "FC" denotes one linear layer.

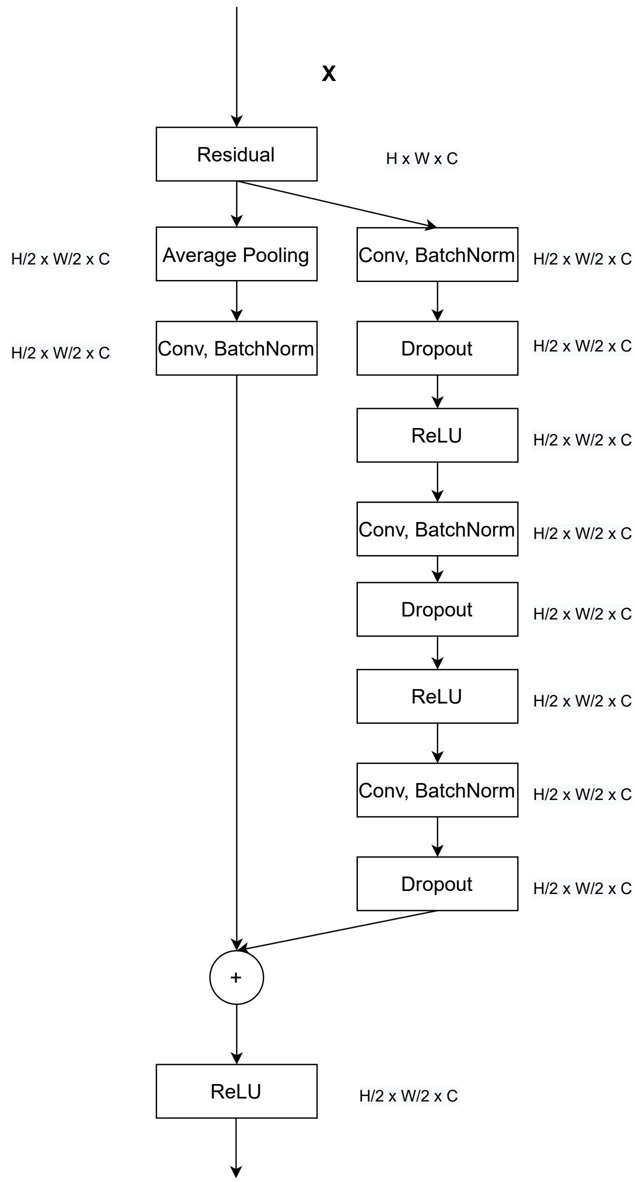

Figure 4. Basic Block. "Conv" denotes one convolution layer.

work layers. The architecture of Attention Block and Basic Block are shown in Fig. 3 and Fig. 4 respectively. In our experiments, all dropout rates of rPPGNet are set to 0.2.

3.3. Loss function

In this article, the rPPG signal estimation is regarded as a regression problem. The following three loss functions are used to constrain the relationship between the predicted rPPG signals and the real rPPG signals.

Negative Pearson correlation coefficient loss[27] proved to be an effective loss function for rPPG signal prediction, which is calculated between the ground truth rPPG signals and the estimated rPPG signals.

$$\mathcal{L}_p = 1 - \frac{\sum_{i=1}^{n}(X^{(i)} - \overline{X})(Y^{(i)} - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X^{(i)} - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y^{(i)} - \overline{Y})^2}} \quad (1)$$

where X denotes the ground-truth rPPG signals and Y denotes the estimated rPPG signals.

L1 loss[14] is also used for rPPG signal estimation in our

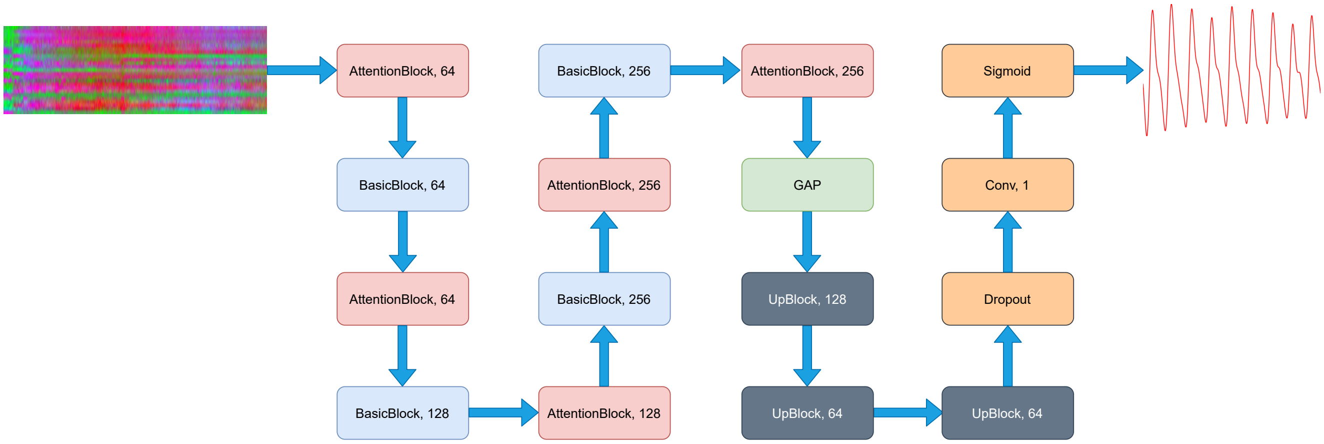

Figure 5. Network Architecture of rPPGNet. "UpBlock" denotes one transposed convolution layer followed by Batch Normalization and ELU activation[4]. "Conv" denotes one convolution layer. "GAP" denotes global averaging pooling.

method.

$$\mathcal{L}_{l1} = \frac{1}{n}\sum_{i=1}^{n}|rPPG_{es}^{(i)} - rPPG_{gt}^{(i)}| \tag{2}$$

where $rPPG_{es}$ indicates the estimated rPPG signal and $rPPG_{gt}$ indicates the ground-truth rPPG signal.

SNR loss[18] is a frequency domain loss constraining the relationship between the predicted rPPG signals and the ground-truth heart rate values.

$$\mathcal{L}_{fre} = CE(PSD(rPPG_{es})), HR_{gt}) \tag{3}$$

where $PSD(\cdot)$ indicates the power spectral density of $rPPG_{es}$, HR_{gt} indicates the ground-truth heart rate, and $CE(\cdot)$ indicates the cross-entropy loss.

The overall loss function of our rPPG signal estimation pipeline is

$$\mathcal{L} = \mathcal{L}_p + \lambda_{l1}\mathcal{L}_{l1} + \lambda_{fre}\mathcal{L}_{fre} \tag{4}$$

For our experiments, we set $\lambda_{l1} = 1$ and $\lambda_{fre} = 1$.

3.4. Training procedure

First, we pre-processed the ground truth rPPG signal using a 4th-order Butterworth band-pass filter with cutoff frequency [0.6, 3] Hz for restricting outliers like [25]. Then, we normalized the ground truth rPPG signal to have a minimum value of zero and a maximum value of 1.

After that, we train rPPGNet for 20 epochs, using kaiming initialization[7]. Adam optimizer[9] is used while learning rate is set to 0.01 and batch size is set to 256. In order to make our model more robust, we adapt following four data enhancement strategies: 1) randomly erase part of STmap; 2) randomly add random noise to part of STmap; 3) randomly reverse STmap and the ground truth rPPG signal at the same time; 4) randomly flip facial video horizontally.

Finally, the network was trained on 1 NVIDIA GeForce GTX 3090 GPU. Our rPPG signal estimation pipeline was implemented using PyTorch framework[19].

3.5. Post-processing

We post-processed the estimated rPPG signal using a 4th-order Butterworth band-pass filter with cutoff frequency [0.6, 3] Hz for restricting outliers. Then, $scipy.signal.find\_peaks$ was used to find peaks of rPPG signal.

4. Experiments

4.1. Datasets

Three external datasets were used for training(VIPL-HR, PURE, UBFC-rPPG).

Before generating STmap, all face videos and the corresponding rPPG signals were resampled to 30 fps using cubic spline interpolation like [15].

UBFC-rPPG dataset[2] is a database for remote heart rate estimation, which contains 42 uncompressed RGB videos. The videos were recorded with a low cost webcam at 30 frames per second. The ground-truth heart rate values and rPPG signals are provided, which were collected by a pulse oximeter finger clip sensor. In order to make this dataset cover a wider range of heart rate values, all subjects were asked to play a time sensitive mathematical game that supposedly raises their heart rate.

VIPL-HR dataset[16] is a challenging large-scale multi-modal database, which contains 2,378 visible light facial videos of 107 subjects. In order to simulate real world conditions as realistic as possible, this dataset was collected under less-constrained scenarios, which contains various variations such as different head movements, illumination condition variations, and acquisition device changes. Due to different recording scenarios and devices, the frame rates of the videos vary from 25 fps to 30 fps. In addition, the ground-truth HR is recorded using a CONTEC CMS60C BVP sensor (a FDA approved device).

PURE dataset[23] is a public available database for remote heart rate estimation, which comprises 60 RGB videos from 10 subjects(8 male, 2 female) in 6 different setups. The videos were recorded using an eco274CVGE camera at 30 fps and a resolution of 640 × 480. The ground-truth rPPG signals were captured using a finger clip pulse oximeter (pulox CMS50E).

4.2. Evaluation Metrics

The following five metrics were used to evaluate the performance of our approach on the 2nd RePSS Challenge Track1 test dataset.

1. M_IBI(mean of IBI error)

   For two IBI curves $R_1(t)$ and $R_2(t)$, the IBI error and M_IBI can be defined as,

$$AE = \sum_{t=0}^{T} |R_1(t) - R_2(t)| \tag{5}$$

$$M\_IBI = \frac{1}{K} \sum_{k=0}^{K} AE_k \tag{6}$$

   where T is the time length of the IBI curves, and K is the number of videos.

2. SD_IBI(standard deviation of IBI error)

$$SD\_IBI = \sqrt{\frac{1}{K} \sum_{k=0}^{K} (AE_k - M\_IBI)} \tag{7}$$

3. MAE_HR(mean absolute error of heart rate)

$$MAE\_HR = \frac{1}{n} \sum_{i=1}^{n} |HR_{predict}^{(i)} - HR_{label}^{(i)}| \tag{8}$$

   where $HR_{predict}$ is the estimation of HR and HR_{label} is the ground-truth of HR.

4. RMSE_HR(root mean squared error of heart rate)

$$RMSE\_HR = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (HR_{predict}^{(i)} - HR_{label}^{(i)})^2} \tag{9}$$

5. R_HR(Pearson correlation coefficient of heart rate)

$$R\_HR = \frac{\sum_{i=1}^{n} (X^{(i)} - \overline{X})(Y^{(i)} - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X^{(i)} - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y^{(i)} - \overline{Y})^2}} \tag{10}$$

   where $X^{(i)}$ denotes $HR_{predict}^{(i)}$, $Y^{(i)}$ denotes $HR_{label}^{(i)}$, $\overline{X}$ denotes the mean value of X vector, $\overline{Y}$ denotes the mean value of Y vector.

Table 1. Public Leaderboard of 2nd RePSS Challenge Track1

Rank	Team Name	M_IBI	SD_IBI	MAE_HR	RMSE_HR	R_HR
1	**TIME**	**117.25**	**153.18**	**7.31**	**11.44**	**0.62**
2	Dr. L	122.80	153.91	7.29	11.05	0.57
3	The Anti-Spoofers	168.08	162.82	11.84	14.51	0.02
4	shankejinjiboy	224.41	163.98	15.44	18.75	-0.05
5	ZJUT-WTCrPPG	273.53	171.13	23.89	27.96	-0.03
6	ZJUT-ASTrPPG	295.70	175.24	29.24	33.69	-0.10

4.3. Results

As shown in Table 1, our team(Team Name TIME) achieved first place on the 2nd RePSS Challenge Track 1.

Our approach has outperformed the methods of other participants as we have achieved M_IBI = 117.25(4.51% improvement compared to the challenge top-2 result), R_HR = 0.62(8.77% improvement).

5. Conclusion

In this paper, we have proposed Time Lab's approach to IBI estimating from facial video. We propose an novel and efficient rPPGNet for rPPG signals estimation and a improved design of spatial-temporal map. IBI data estimated with this method was submitted for the 2nd Challenge Track1 on RePSS organized within ICCV2021. Our method achieved first place on the 2nd RePSS Challenge Track 1. Due to the limited time available for this challenge, we didn't perform well enough. Our method still has a lot to improve.

6. Acknowledgments

This work was supported in part by the NSFC-Xinjiang Joint Fund under Grant U1903127 and in part by the Natural Science Foundation of Shandong Province under Grant ZR2020MF052.

References

[1] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel. Continuously reproducing toolchains in pattern recognition and machine learning experiments. In *International Conference on Machine Learning (ICML)*, Aug. 2017. 2

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 4

[3] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 2

[4] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential

linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 4

[5] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 4

[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[10] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *European Conference on Computer Vision*, pages 392–409. Springer, 2020. 2

[11] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jkedrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011. 1

[12] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 242–249. IEEE, 2018. 1

[13] Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. The 1st challenge on remote physiological signal sensing (repss). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 314–315, 2020. 1

[14] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020. 3

[15] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12404–12413, 2021. 4

[16] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018. 1, 4

[17] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 2

[18] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *European Conference on Computer Vision*, pages 295–310. Springer, 2020. 4

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 4

[20] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 1

[21] Philipp V Rouast, Marc TP Adam, Raymond Chiong, David Cornforth, and Ewa Lux. Remote heart rate measurement using low-cost rgb face video: a technical literature review. *Frontiers of Computer Science*, 12(5):858–872, 2018. 1

[22] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, pages 3–6, 2018. 2

[23] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014. 5

[24] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 1

[25] Wenjin Wang, Albertus C Den Brinker, and Gerard De Haan. Single-element remote-ppg. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2018. 4

[26] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 1

[27] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 151–160, 2019. 2, 3