Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231195

BiFormer: An End-to-End Deep Learning Approach for Enhanced Image-Based Photoplethysmography and Heart Rate Accuracy

Shaohua LIU [a], Yuchong YANG [a,1], XingJian JING [b], Bin LI [c], Hao LIU [d], Shangwei ZHU [a] and Chundong SHE [a]

[a] *School of Electronic Engineering, Beijing University of Posts and Telecommunications, China*
[b] *Department of Mechanical Engineering, City University of Hong Kong, China*
[c] *First Clinical Medical College, Guangzhou University of Chinese Medicine, China*
[d] *Peking University Shougang Hospital, China*

Abstract. Image-based photoplethysmography (IPPG) holds promise for applications like health surveillance and emotional state analysis. Despite recent progress in crafting deep learning-centric IPPG methodologies, which predominantly forge a correlation between spatiotemporal heart rate (HR) feature imagery and corresponding HR readings, these techniques encounter constraints in extended spatiotemporal comprehension and engagement. In this manuscript, we introduce the BiFormer architecture, an end-to-end solution integrating temporal difference convolution, multi-head self-attention transformer modules, and bidirectional long short-term memory networks to refine signal estimations and bolster the model's discernment prowess. Our framework was appraised through intra-database and inter-database evaluations on three accessible datasets, evidencing superiority over conventional IPPG strategies in HR accuracy metrics. Notably, assessments on the VIPL-HR dataset indicated a reduction in the average root mean square error to 7.24 beats per minute.

Keywords. Deep learning, neural network, long short-term memory network, self-attention mechanism, signal processing, heart rate measurement

1. Introduction

Heart rate (HR) serves as a crucial physiological indicator of emotional responses and is a fundamental marker of cardiovascular functions. Conventional approaches to measuring heart rate employ contact monitoring devices. Nevertheless, these methodologies might lead to discomfort and inconvenience for individuals.

In recent times, pre-trained models based on Transformer architectures have demonstrated proficiency in an array of tasks. Nonetheless, these models necessitate substantial

[1]Corresponding Author: Yuchong YANG,School of Electronic Engineering,Beijing University of Posts and Telecommunications,China; E-mail: y5131241997@bupt.edu.cn.

volumes of high-quality training data to yield precise outcomes in real-world scenarios. Furthermore, the robustness of the Transformer model is imperative for sustained IPPG measurement assignments.

To tackle these obstacles, this manuscript introduces a novel framework termed BiFormer. This framework amalgamates a bidirectional long short-term memory network, aiming to harness more comprehensive contextual information and thereby enhancing the efficacy of Transformer-based IPPG estimations. The proposed methodology is assessed using three openly accessible datasets, revealing that it surpasses most extant techniques for heart rate estimation, both within individual datasets and in a cross-dataset context. Our approach demonstrates significant robustness across varied datasets. To validate the potency of the BiFormer framework, we also conducted a series of ablation experiments.

2. Related Work

In recent times, several contemporary methodologies grounded in deep learning eschew the need for preprocessing datasets into relatively pristine and stable inputs, opting instead to leverage noise for achieving more consistent learning[1]. DeepPhys[14] utilized an end-to-end supervised learning approach executed via feedforward CNN, incorporating attention mechanisms to discern frame disparities rather than employing long short-term memory (LSTM) units for time information simulation. PhysNet[34] devised a model anchored on a 3D CNN and synergistically combined it with a 2D CNN model to glean spatiotemporal features and evaluate its efficacy.

Nonetheless, as underscored by Lee et al.[9] in their research concerning remote HR estimation via transductive meta-learner, the tangible efficacy of end-to-end supervised learning approaches[5][6][7][10] may be detrimentally impacted by alterations in data distribution between the phases of model training and deployment[3]. In response to this challenge, we advocate for an end-to-end adaptive Bidirectional LSTM (BiLSTM) transformer model[13], capable of directing global attention towards the amplification of quasi-periodic IPPG characteristics, culminating in the utilization of BiLSTM units for time-series information prediction.

3. Methodology

This manuscript introduces a novel learning paradigm termed BiFormer, depicted in Figure 1. Initially, a video sequence is fed into the model, and a region of interest (ROI)[15] encompassing the face is selected to compute the mean value of the RGB channel pixels within the designated ROI across all frames. Subsequently, the CNN Stem[16] extracts rudimentary local spatiotemporal features, facilitating swift convergence. In the subsequent step, a Transformer Block equipped with a multi-head self-attention mechanism processes the RGB facial videos, capturing both global and finely-tuned local features. Ultimately, a bidirectional long short-term memory network[17] refines and outputs the processed signal, treating signal estimation as a sequential regression challenge and yielding superior pulse waveforms. A comprehensive exposition of the entire methodology follows.

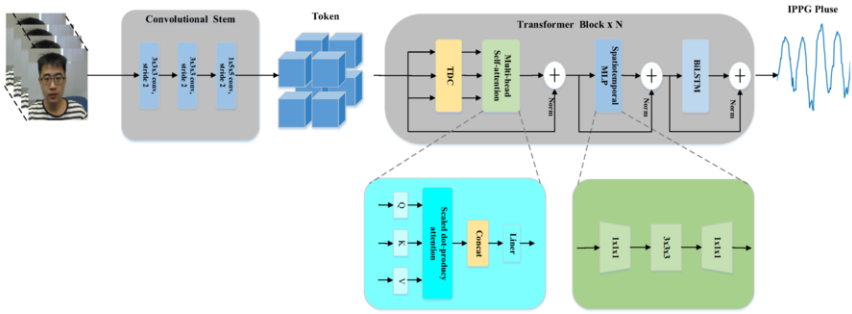

Figure 1. The framework of the BiFormer network is used for joint modeling of BVP prediction and HR estimation. The BiFormer architecture consists of a convolution stem for extracting local features, a transformer block containing temporal-difference convolution, multi-head self-attention mechanism, and spatiotemporal multi-layer perceptron, as well as a bidirectional long short-term memory network for estimating the processed signals and producing high-quality pulse waveforms to calculate more accurate physiological parameters.

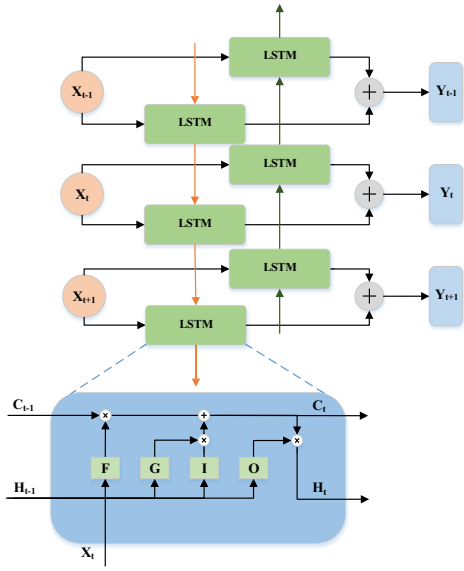

Figure 2. The basic structure of the BiLSTM unit in the BiFormer network.

3.1. Temporal Difference Convolution

Building upon the spatio-temporal central difference convolution (STCDC) detailed in [20], we introduce a variant focused solely on temporal central differences, termed temporal difference convolution (TDC). This design choice stems from the data's intrinsic characteristics. TDC is executed in two phases: sampling and aggregation.

3.2. Long Short-Term Memory Networks

Traditional practice involves solving time series problems with only a single layer of LSTM models[18] or by stacking multiple LSTM layers[19], but these strategies produce

suboptimal results.Hence, this study incorporates BiLSTM, depicted in Figure 2, where a cell's output is contingent on both antecedent and subsequent frames. This effectively melds the forward and backward LSTMs into a unified BiLSTM. Both LSTM and BiLSTM layers adeptly discern unidirectional and bidirectional long-term dependencies amid time steps in sequential data. Owing to BiLSTM's ability to simultaneously access information from both antecedent and subsequent time steps, it outperforms unidirectional LSTM in prediction accuracy. Post each BiLSTM step, the IPPG signal estimation is modeled as a multi-task output for ordinal regression. Specifically, as illustrated in Figure 2, given X_t as the input and Y_t as the BiLSTM cell output, Y_t denotes the estimated IPPG signal.

3.3. Dynamic Loss

The loss function for label distribution is defined based on Kullback-Leibler (KL) divergence as the dynamic loss:

$$\zeta_{KL} = KL\left(p, \text{Softmax}\left(p'\right)\right) \tag{1}$$

p' represents the power spectral density (PSD) of the predicted IPPG signal, allowing for efficient feature learning between adjacent labels with limited training data. $p \in R$ is the trained IPPG signal, and $p^t \in R$ is the true signal with accurate pulse peak positions.

In the specific IPPG training task, we define the loss function in the time domain using negative Pearson correlation. T is the number of frames in the input sequence, the loss function in the time domain is expressed as follows:

$$\zeta_T = 1 - \frac{T\sum_{i=1}^{T} p_i p_i^t - \sum_{i=1}^{T} p_i \sum_{i=1}^{T} p_i^t}{\sqrt{\left(T\sum_{i=1}^{T} p_i^2 - \left(\sum_{i=1}^{T} p_i\right)^2\right)\left(T\sum_{i=1}^{T} (p_i^t)^2 - \left(\sum_{i=1}^{T} p_i^t\right)^2\right)}} \tag{2}$$

Similar to signal-to-noise ratio loss [23], we treat HR estimation as a classification task in the frequency domain and provide the following formula for the classification loss:

$$\zeta_C = CE\left(p'(p), HR_t\right) \tag{3}$$

Here, CE represents the classic cross-entropy loss, and HR_t represents the true HR value. We combine exponential incremental strategy with dynamic supervision to gradually expand the frequency constraint, which is beneficial for intrinsic feature learning and can alleviate the overfitting problem. The dynamic loss can be expressed as follows:

$$\zeta_{\text{all}} = \underbrace{\alpha \cdot \zeta_T}_{\text{time}} + \underbrace{\beta \cdot (\zeta_C + \zeta_{KL})}_{\text{frequency}} \tag{4}$$

$$\beta = \beta_0 \cdot \left(\eta^{(n_i - 1)/n}\right) \tag{5}$$

Here, the hyperparameters α, β_0, and η are set to 0.1, 1.0, and 5.0, respectively. n_i represents the current training epoch, and n represents the total number of training epochs. With the dynamic loss, the training process can better perceive the signal trend at the beginning, which is beneficial for reinforcement learning in later stages.

4. Experimental Implementation

4.1. Evaluation Metrics

We used three common public data sets: VIPL-HR[24] dataset, PURE[25] dataset and UBFC[26] dataset.To assess the mean heart rate (HR), we adhered to established practices, employing prevalent evaluation metrics [37] for remote HR measurement: standard deviation (SD), mean absolute error (MAE), root mean squared error (RMSE), and the Pearson correlation coefficient (R). These metrics derive from the calculation of HR error, expressed as $H_d(i) = H_{\text{pre}} - H_{\text{true}}$.

4.2. Training Setup

For each video segment, an automatic cropping technique was employed, defaulting to a sequence of ROI frames. If a face is not detected in a frame, the facial region identified in the preceding frame is utilized. In the training phase, RGB sequences of dimensions $160 \times 128 \times 128$ ($H \times W \times C$) were randomly chosen as input, with a target pipeline size of $H_{\text{t}} \times W_t \times C_t = 4 \times 4 \times 4$. Each model was trained on a single NVIDIA GeForce RTX 3080 GPU, utilizing the Adam optimizer with a starting learning rate and weight decay of 1e-4 and 5e-5, respectively. The model was trained over 25 epochs, maintaining constant loss function weights of $\theta = 0.7$, $\alpha = 0.1$, and $\beta \in [1,5]$. The batch size was fixed at 4. For linear evaluation and ablation studies, datasets were randomly partitioned into five subsets, designating one as the test set and the remainder for training. During testing, the frame-level HR mean for each video was computed as the video-level average HR, ensuring a balanced evaluation by utilizing distinct samples in the training and test sets to minimize model dependency on individual samples.

4.3. Test Results

Evaluation on VIPL-HR for HR Estimation: Initially, our dataset was assessed using VIPL-HR, yielding 108,100 sample frame sequences based on the designated time window and sliding step configurations. Acknowledging instances of video loss, we arbitrarily selected 86,400 samples from the initial 80 subjects for training and allocated the residual 21,700 samples from the remaining subjects for testing. The outcomes, delineated in Table 1, reveal suboptimal performance by traditional methods such as SAMC[8], POS[27], and CHROM[4], sourced from open-source toolboxes[34]. In contrast, non-end-to-end deep learning approaches like CVD[14], RhythmNet[11], and Dual-GAN[35] showcased superiority, suggesting that deep learning techniques can efficiently extract information features for signal prediction and HR estimation[21][22]. Several methods were assessed end-to-end, with their results sourced directly from respective publications or citations due to implementation challenges. Some baseline studies did not provide standard deviation errors, denoted as - in this article. Overall, our BiFormer approach excelled across four evaluation metrics, evidencing its capacity for training on raw facial videos without dataset preprocessing, thereby ensuring convenience and continuous adaptation.

Table 1. The test results of the VIPL-HR dataset are presented herein, with the optimal outcome denoted in bold and the second-best outcome underlined for clarity.

Method	HR (bpm)			
	SD ↓	MAE ↓	RMSE ↓	R ↑
SAMC[8]	18.0	15.9	21.0	0.11
POS[27]	15.3	11.5	17.2	0.30
CHROM[4]	15.1	11.4	16.9	0.28
I3D[28]	15.9	12.0	15.9	0.07
PhysNet[34]	14.9	10.8	14.8	0.20
DeepPhy[12]	13.6	11.0	13.8	0.11
RhythmNet[11]	8.11	5.30	8.14	0.76
CVD[14]	7.92	5.02	7.97	0.79
Physformer[33]	7.74	4.97	7.79	0.78
Dual-GAN[35]	**7.63**	<u>4.93</u>	<u>7.68</u>	<u>0.81</u>
BiFormer	<u>7.64</u>	**4.48**	**7.24**	**0.88**

Table 2. The test results of the PURE dataset.

Method	HR (bpm)			
	SD ↓	MAE ↓	RMSE ↓	R ↑
CHROM[4]	-	2.07	9.92	0.97
PulseGAN-DAE[31]	-	3.24	5.97	0.97
PulseGAN[31]	-	2.09	4.42	<u>0.97</u>
Tsou[32]	-	**0.63**	<u>2.70</u>	0.83
BiFormer	**2.03**	<u>1.40</u>	**2.61**	**0.99**

Table 3. The cross-dataset testing results of UBFC dataset.

Method	HR (bpm)			
	SD ↓	MAE ↓	RMSE ↓	R ↑
GREEN[2]	20.2	7.50	14.41	0.62
ICA[29]	18.6	5.17	11.76	0.78
POS[27]	10.4	4.05	8.75	0.78
3D CNN[30]	8.55	5.45	8.64	-
Meta-rPPG[9]	7.12	5.97	7.42	0.53
CHROM[4]	-	2.37	4.91	0.89
PulseGAN[31]	-	**1.19**	**2.10**	**0.98**
BiFormer	**2.15**	<u>2.77</u>	<u>3.89</u>	**0.98**

Evaluation on PURE for HR Estimation: The efficacy of our BiFormer method was further substantiated through HR estimation on the PURE dataset. Adhering to the validation protocol from [11], we juxtaposed our approach against four baseline methods on PURE. As presented in Table 2, our method consistently surpassed most baseline methods, underscoring the robustness of the IPPG features discerned by our approach, even under less restrictive conditions.

Table 4. Ablation experiment results on the VIPL-HR dataset.

Method	HR (bpm) RMSE
BaseLine	21.37
MSA	14.13
MSA+TDC	13.49
MSA+TDC+MLP	7.79
BiFormer	**7.24**

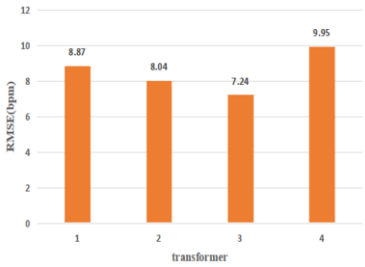

Figure 3. Results of ablation study on the number of transformers.

Cross-dataset HR Estimation Results using UBFC: Assessing model generalization in cross-dataset scenarios is pivotal for remote physiological parameter extraction. Consequently, we appraised our model's cross-database adaptability on the UBFC dataset. Utilizing VIPL-HR for training and UBFC for testing, the HR estimation outcomes, depicted in Table 3, indicate that our model surpassed the majority of baseline methods, showcasing commendable generalization capabilities in unfamiliar noise scenarios. This highlights the efficacy and resilience of our method, incorporating spatiotemporal attention and BiLSTM prediction.

4.4. Ablation Experiments

To scrutinize the impact of various elements on the efficacy of our approach, we conducted ablation studies using the VIPL-HR database, exclusively employing this dataset for the sake of brevity. As delineated in Table 4, we explicated the repercussions of modifying individual modules on the overall performance. Our findings indicate that the integration of the multi-head self-attention mechanism was most efficacious in diminishing RMSE. Likewise, the incorporation of the bidirectional BiLSTM module contributed to a notable reduction in RMSE. Each module was instrumental in augmenting the model's performance, underscoring their indispensability.

The impact of varying the quantity of transformers is illustrated in Figure 3. The performance was suboptimal with four transformers, while it improved with the use of either two or three transformers. Notably, employing three transformers led to a substantial reduction in RMSE. Analogous to ResNet [36], wherein distinct features are gleaned for efficacious representation learning via CNN layers, our model discerns diverse spatiotemporal representations by tokenizing with distinct kernel sizes for each transformer. Hence, the findings imply that the amalgamation of features across multiple scales through

Table 5. Results of LSTM ablation experiments on the VIPL-HR dataset.

LSTM category	RMSE
LSTM unidirectional single layer	8.29
LSTM unidirectional multilayer	8.12
BiLSTM	**7.24**

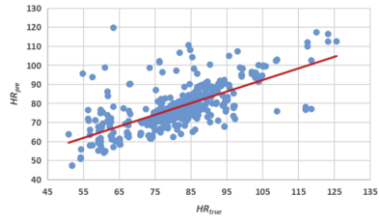

Figure 4. Scatter plot of correlation between ground truth HR and predicted HR in the VIPL database.

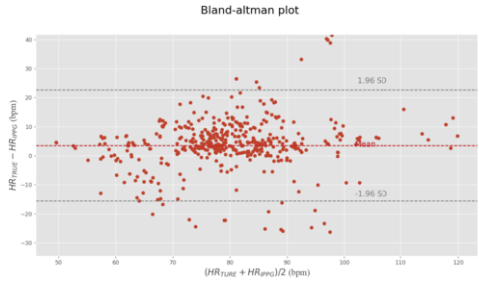

Figure 5. The Bland-Altman plot between ground truth HR and predicted HR in the VIPL database.

fusion can effectively facilitate the computation of reliable waveform-associated attributes. To validate the aptness of the BiLSTM approach for this framework, we conducted an analytical experiment contrasting it with traditional LSTM. Table 5 elucidates that our BiLSTM model surpasses other LSTM configurations in performance. This can be attributed to BiLSTM's proficiency in recognizing bidirectional patterns in IPPG signals, which correspond to certain nonlinear heartbeat characteristics (e.g., arrhythmia, frequency variability). The bidirectional context provided by BiLSTM potentially aids in mitigating external disturbances, thereby enhancing the model's robustness.

Furthermore, we evaluated the correlation between the predicted and actual HR as depicted in Figure 4. The results affirm that our proposed BiFormer method, grounded in end-to-end learning, is adept at yielding precise outcomes in video-based measurements, even in environments with constrained resources.

To further evaluate the proposed method, we used Bland-Altman analysis [38] to visualize the specific results, as shown in Figure 5. The positive and negative 1.96 SD represents the range of differences between the two measurement methods within a 95% confidence interval. Here, SD is the standard deviation of the differences, and 1.96 is a constant in statistics that represents the 95% confidence interval. If the difference between two values falls within this range, it can be considered that they have a certain consistency. The mean represents the average difference between the two values. In the figure, most of

the difference points are distributed on both sides of the mean line, and the mean line is close to the zero line, indicating that our BiFormer method has better consistency with the reference true heart rate.

5. Summary

The heart rate signal serves as a pivotal metric in evaluating human health, yet remote measurement of the BVP signal through IPPG is fraught with difficulties due to its feeble signal amplitude and susceptibility to noise. In this research, we employed a comprehensive BiFormer network architecture designed end-to-end, capable of assimilating extensive contextual information, thereby enhancing the efficacy of IPPG for procuring high-fidelity pulse waveforms. We substantiated our approach through experiments on three distinct datasets: VIPL, PURE, and UBFC. The outcomes revealed that the BiFormer network demonstrated commendable proficiency in both isolated and cross-database contexts, thereby augmenting the precision of camera-centric remote physiological assessments. While this research is nascent, these endeavors hold significant potential for broadening the scope of IPPG technology applications.

References

[1] Sun Y, Thakor N. Photoplethysmography revisited: from contact to noncontact, from point to imaging[J]. IEEE transactions on biomedical engineering, 2015, 63(3): 463-477.
[2] Wim V erkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. Optics express, 2008
[3] 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)
[4] De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rPPG. IEEETrans. Biomed. Eng. 60(10), 2878–2886 (2013)
[5] H.-Y . Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman. Eulerian video magnification for revealing subtle changes in the world. In ACM Transactions on Graphics,2012.
[6] Sungjun Kwon, Hyunseok Kim, and Kwang Suk Park. V alidation of heart rate extraction using video imaging on a builtin camera system of a smartphone. In International Conference of the IEEE Engineering in Medicine and Biology Society, pages 2174–2177, 2012
[7] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard.Non-contact, automated cardiac pulse measurements usingvideo imaging and blind source separation. Optics express,18(10):10762–10774, 2010.
[8] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, LijunYin, Jeffrey F Cohn, and Nicu Sebe.Self-adaptive matrix completion for heart rate estimation from face videos underrealistic conditions. In CVPR, 2016.
[9] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rPPG:Remote Heart Rate Estimation Using a Transductive MetaLearner. In ECCV, 2020.
[10] A Ni, A Azarang, N Kehtarnavaz.A review of deep learning-based contactless heart rate measurement methods.In Sensors, 2021.
[11] X Niu, S Shan, H Han, X Chen.Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation.In IEEE,2020
[12] H Qi, Q Guo, F Juefei-Xu, X Xie, L Ma, W Feng.Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms.In ACMDL,2020
[13] Weixuan Chen and Daniel McDuff. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In ECCV, 2018.
[14] Xuesong Niu, Zitong Yu, Hu Han,Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based Remote Physiological Measurement via Cross-verified Feature Disentangling.In ECCV, 2020

[15] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. In BMVC, 2019
[16] 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 ICCV, 2019
[17] Debidatta Dwibedi, Y usuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In CVPR, June 2020
[18] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. IEEE SPL,2020.
[19] M Singh, E Mintun, T Darrell, P Dollár. Early convolutions help transformers see better. In NeurIPS Proceedings, 2021.
[20] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. IEEE TIP, 2021
[21] Bin-Bin Gao, Hong-Y u Zhou, Jianxin Wu, and Xin Geng.Age estimation using expectation of label distribution learning. In IJCAI, 2018.
[22] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. IEEE Signal Processing Magazine, 2021
[23] Radim ˇSpetl´ık, V ojtech Franc, and Jir´ı Matas. Visual heart rate estimation with convolutional neural network. In BMVC,2018.
[24] Niu, X., Han, H., Shan, S., Chen, X.: VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In: Proceedings of the ACCV,2018
[25] Ronny Stricker, Steffen Müller, and Horst-Michael Gross.Non-contact video-based pulse rate measurement on a mobile service robot. In Proc. IEEE ISRHIC, pages 1056–1062,2014.
[26] S. Bobbia, R. Macwan, Y . Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography,"Pattern Recognit. Lett., vol. 124, pp. 82–90, 2019.
[27] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. IEEE Transactions on Biomedical Engineering, 2017
[28] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR,2017.
[29] Poh M Z, Mcduff D J, Picard R W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.Optics Express, 2010, 18(10):10762-10774.
[30] Bousefsaf, F., Pruski, A., Maaoui, C.: 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. Applied Sciences 9(20),4364 (2019)
[31] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Y u Liu,and Xun Chen. Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. IEEE J-BHI,pages 1–1, 2021
[32] Tsou Y Y , Lee Y A, Hsu C T, et al. Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos[C]. Proceedings of the 35th Annual ACM Symposium on Applied Computing. 2020: 2066-2073
[33] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Physformer: facial video-based physiological measurement with temporal difference transformer," in CVPR, 2022
[34] D. McDuff and E. Blackford, "IPHYS: An open non-contact imaging-based physiological measurement toolbox," in Proc. 41st Annu. Int. Conf.IEEE Eng. Med. Biol. Soc., Jul. 2019, pp. 6521–6524
[35] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement.In CVPR, 2021
[36] Wenxia Bao,Zhongyu Ma.Pose ResNet: 3D Human Pose Estimation Based on Self-Supervision.In MDPI,2023
[37] Hodson, T.O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geosci. Model Dev. 2022, 15, 5481–5487
[38] Giavarina, D. Understanding Bland Altman analysis. Biochem. Med. 2015, 25, 141–151.