Non-Contact Heart Rate Estimation From Photoplethysmography Using EEMD and Convolution-Transformer Network

1st Kehong Liu
School of Artificial Intelligence
Xidian University
Xi'an, China
kehongliu@stu.xidian.edu.cn

2nd Shuo Wu
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, China
s.wu@bupt.edu.cn

3rd Tianhuan Li
School of Software
Beihang University
Beijing, China
tianhuanlee@foxmail.com

4th Shuiping Gou
School of Artificial Intelligence
Xidian University
Xi'an, China
shpgou@mail.xidian.edu.cn

5th Xinlin Wang
School of Artificial Intelligence
Xidian University
Xi'an, China
wangxinlin@xidian.edu.cn

6th Zhang Guo
School of Artificial Intelligence
Xidian University
Xi'an, China
guozhang@xidian.edu.cn

***Abstract*—Remote heart rate measurement aims to use the remote photoplethysmography (rPPG) technology to extract the heart rate information from face videos in a non-contact manner. Most of the existing methods are susceptible to environmental factors and the movements of the subjects. To address this challenge, we propose a lightweight convolutional neural network for real-time heart rate detection under realistic conditions. Here, the rPPG signals are decomposed using Ensemble Empirical Mode Decomposition (EEMD) technology to extract the Intrinsic Mode Functions (IMFs) that contain valuable heart rate features. Then, feed the extracting IMFs into a parallel CNN-Transformer network, which consists of two branches: the CNN branch is used to extract local features of the rPPG signal, while the Transformer branch is employed to capture the global representation of the rPPG signal. And then, heart rates are calculated using power spectral density. Extensive experiments are conducted on two public datasets – the UBFC-rPPG dataset and PURE dataset. In terms of the experiment results, the proposed method achieves 0.87 MAE (bpm), 1.72 RMSE (bpm) and 0.99 r value of Pearson's correlation coefficient on UBFC-rPPG dataset, and 0.81 MAE (bpm), 1.65 RMSE (bpm) and 0.99 r value of Pearson's correlation coefficient on PURE dataset.**

***Index Terms*—Remote Photoplethysmography (rPPG), Non-Contact, Heart Rate Estimation, Ensemble Empirical Mode Decomposition (EEMD), Convolutional Neural Network (CNN)**

I. Introduction

According to the 2023 report of the World Health Organization [1], Cardiovascular diseases are the leading cause of global death, taking an estimated 17.9 million lives each year. Heart rate refers to the number of the heart beats per minute and plays an irreplaceable role in human health, the diagnosis and prevention of cardiovascular diseases. Changes in heart rate can provide information about an individual's health and cardiovascular system. Therefore, monitoring heart rate variations is crucial for maintaining heart health and diagnosing potential disease.

Existing heart rate monitoring technologies can be divided into two categories based on whether directly contact with the skin [2]. The contact-based heart rate measurement devices used in modern medicine include pressure-based heart rate monitors, electrocardiogram-based heart rate monitors, and photoplethysmography-based heart rate monitors. These methods are highly accurate and advanced, but they are not suitable for infants, individuals with skin burns, or those who has skin disease and also unapplicable for situations such as underwater activities.

Based on Remote Photoplethysmography (rPPG), non-contact heart rate monitoring [3] has become an attractive research direction due to its outstanding convenience and consistency. Unlike traditional methods, remote photoplethysmography does not require direct skin contact. Therefore, some traditional signal processing methods have been proposed to extract physiological signals from visual data, such as Blind Source Separation (BSS) [4] and Principal Component Analysis (PCA) [5]. Nevertheless, these methods are limited by their strong dependence on specific types of cameras or devices and require high computational resources and time but have low accuracy.

With the development in deep learning and computer vision technologies, deep learning-based heart rate estimation methods have gained significant research and application in recent years, offering a potential solution to overcome some of the limitations of traditional methods. Early methods relied on manual feature representations for rPPG prediction, such as time-frequency maps [6] and spatiotemporal maps [7]. Chen and McDuff introduced a convolutional attention network that utilized normalized differential frames as input to predict derivatives of the pulse wave signal [8]. In addition to traditional CNNs, PulseGAN [9] used a generative model with

adversarial learning to post-process estimated rPPG signals. These methods are typically more robust and capable of handling a wider range of scenarios and subject types.

Although there has been extensive research on non-contact heart rate detection, there are still rooms for further lightweighting of models and improvement in accuracy in resource-constrained devices or scenarios. The method we proposed employs EEMD for the decomposition of rPPG signals and set up a parallel CNN-Transformer network based on the feature fusion of Transformer and CNN for the IMF extracted after EEMD containing heart rate information. The algorithm uses two different types of networks for feature fusion, which allows it to better handle various features and noise in signals, thereby enhancing the robustness of the model and improving the accuracy of heart rate measurement.

The paper is organized as follows. Section II introduces related work about remote photoplethysmography. In Section III, we elaborate on the proposed method. Section IV presents experimental results on two physiological benchmark datasets to demonstrate the superiority of our proposed method. In Section V, we summarize the work with some concluding remarks.

II. Related Works

With the extensive research and application of deep learning in computer vision and natural language processing, many researchers use deep learning methods to address issues in remote heart rate estimation based on facial videos, such as Synrhythm [10], DeepPhys [11], HR-CNN [12], and remote heart rate estimation based on 3D CNNs. The authors of DeepPhys, Chen and McDuff, were the first to introduce end-to-end system for obtaining heart rate and respiration rate. They designed a convolutional neural network (CNN) with an attention mechanism to establish a mapping between video frames and the required physiological information. In the same year, HR-CNN proposed a two-step CNN method consisting of feature extraction and a heart rate estimator to estimate heart rate from a series of face images. Yu and his colleagues introduced the PhysNet spatiotemporal network, which used 3D convolutions to extract both the temporal and spatial information from videos [13]. Niu and others introduced spatiotemporal features of heart rate information and designed a transfer learning strategy from general to specific features for heart rate estimation. Later, they proposed a cross-validation decoupling feature separation method [14], inputting multiscale spatiotemporal maps (STMaps) into two autoencoders to predict physiological signals by cross-validation decoupling of encoded features. They also applied channel and spatiotemporal attention mechanisms to further improve heart rate estimation from facial videos [15]. Liu and colleagues [16] introduced a temporal shift convolution attention architecture (MTTS-CAN) that promotes information exchange across multiple frames by moving convolutional kernels along the time dimension. Yu and others [17] proposed time-difference convolution operations to capture inter-frame dependencies. Song and colleagues mapped a new spatiotemporal feature to the corresponding heart rate value using CNN and designed a feature decoder framework that utilizes transfer learning to reduce the need for training data while speeding up model convergence [18].

PhysFormer introduced an end-to-end transformer architecture for remote physiological measurements [19], which adaptively aggregates local and global spatiotemporal features to enhance rPPG representations. Later, the authors made improvements based on PhysFormer and proposed PhysFormer++ with a dual-stream SlowFast architecture [20]. Unlike the need for pretraining from large-scale datasets, PhysFormer++ can be easily trained from scratch using rPPG datasets.

III. Methodology

A. Decompose rPPG Signal with EEMD

Ensemble Empirical Mode Decomposition (EEMD) [22] is an adaptive time-frequency analysis algorithm used for signal processing, aiming to address the problem of mode mixing that may occur when decomposing non-stationary signals with EMD [21]. By introducing white noise into the original signal, signal decomposed by EEMD can distribute evenly across various time scales, thus reducing mode mixing. The uniform distribution characteristic of the white noise spectrum allows signals of different scales to separate effectively. Since the mean value of white noise is 0, the influence of noise can be offset by averaging multiple decompositions, resulting in more accurate Intrinsic Mode Function (IMF) components. The implementation of the EEMD algorithm is as follows:

1) Add noise $\omega_n(t)$ to the original time series data $X(t)$, where $n = 1, 2, ..., N$, to obtain the time series data $Xn(t)$:
$$X_n(t) = X(t) + \omega_n(t) \tag{1}$$
2) $X_n(t)$ is decomposed by EMD and gains in the corresponding IMF components $c_n(t)$, where $i = 1, 2, ..., m$, and $r_{m,n}(t)$ is the residual series.
$$X_n(t) = \sum_{i=1}^{m} c_{i,n}(t) + r_{m,n}(t) \tag{2}$$
3) Repeat the above step (2), adding new white noise each time and obtain N sets of IMF components and the corresponding residual sequences.
4) The final decomposition result is obtained by averaging the N sets of IMF components and the residual sequence sets:
$$c_n(t) = \frac{1}{n} \sum_{i=1}^{N} c_{i,n}(t) \tag{3}$$
$$r_m(t) = \frac{1}{N} \sum_{n=1}^{N} r_{m,n}(t) \tag{4}$$

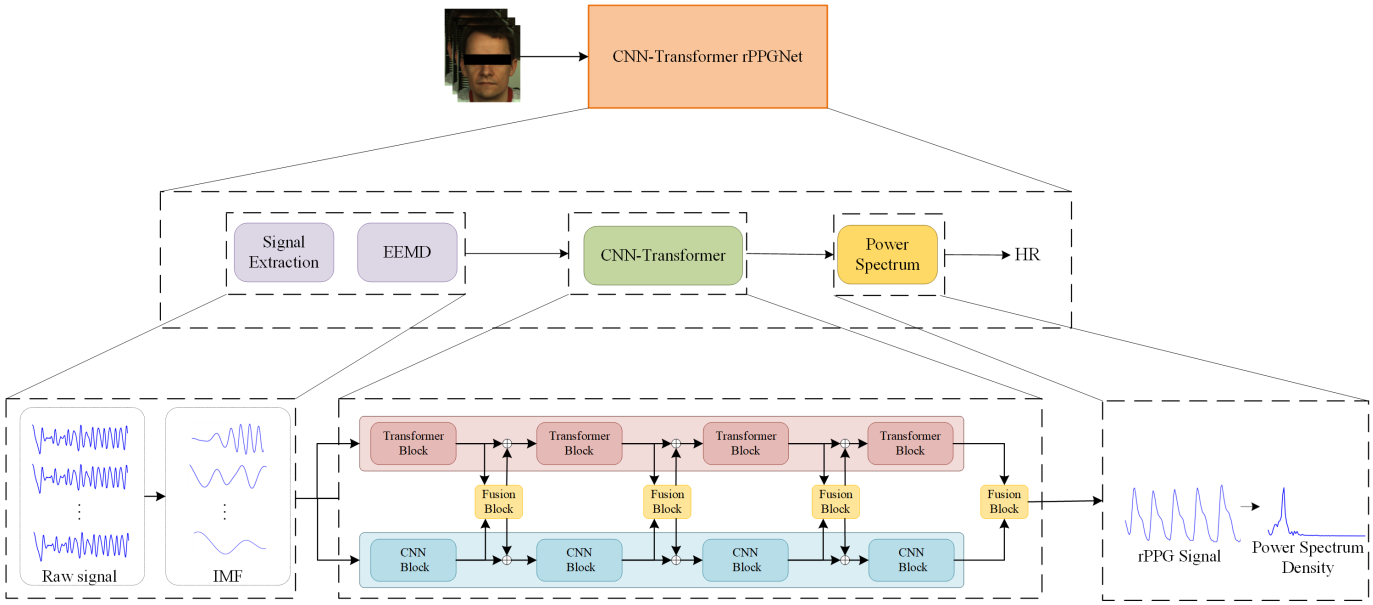


Fig. 1: Overview of the proposed approach for rPPG based remote HR measurement

B. CNN-Transformer rPPGNet

Convolutional Neural Networks (CNNs) have the ability to extract local features and retain the positional information of features during the feature learning process. And the Transformer structure has significant advantage in capturing global features due to its self-attention structure. Therefore, we design a parallel network structure that combines CNN to extract local features and Transformer to extract global feature expressions. The entire network is shown in Fig. 1.

First, the IMFs are input into CNN to extract local features, which are then passed into the Transformer branch, enabling the Transformer branch to process local features. Then, the global features encoded by the Transformer are fed back to the CNN branch, allowing CNN to integrate global features when extracting local features, thereby enhancing global perception. These two processes interact with each other.

The convolutional layer of CNN takes the input feature map and computes a weighted sum with a given convolution kernel to obtain the output feature map. The calculation method is shown as follows:

$$F^l = \phi(Conv1d(\omega, F^{l-1}, b^l)) = LeakyReLu(\omega_f^l * F^{l-1} + b^l) \quad (5)$$

In the formula, $Conv1d$ represents the one-dimensional convolutional neural network. F^{l-1} is the feature of the $l-1$ th layer, ω_f^l represents the shared convolution kernel weight coefficients of length f for the l th layer, F^l is the feature of the l th layer, b^l is the bias coefficient of the l th layer, and ϕ denotes the activation function within the network. We uses the asterisk $*$ to represent the convolution operation. The features at each position in the output feature map are only related to a local region of length f in the input feature map. Therefore, the local features of the signal can be extracted after passing through the CNN.

The Transformer encoder mainly consists of Multi-Head Attention and fully connected Feed-Forward layers. The data volume in heart rate estimation task is not large, therefore we only utilizes the encoder structure of the Transformer to extract global high-dimensional contextual feature vectors using the Self-Attention mechanism.

The multi-head attention mechanism takes the input vectors and feeds them into multiple parallel attention computations. It then concatenates the output vectors and maps them back to the space of the original input vectors to obtain the final attention vector. The specific calculation is as shown as follows:

$$MultiHead(Q, K, V) = Concat(head_i, ..., head_h)W^0 \quad (6)$$

$$head_i(Q, K, V) = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (7)$$

Here, $Concat$ denotes the operation of concatenating multiple matrices along a certain dimension, h represents the number of parallel Attention operations performed, and $head_i$ indicates the computation of the Attention for the i th head. QW_i, KW_i, and VW_i are the mapping matrices for the i th head. We also use Self-Attention to extract the inner correlations within the sequence and obtain the global context features.

To ensure that the output dimension is consistent with the dimension of the previous layer, a one-dimensional deconvolutional layer is required for dimensionality reduction before the output. Subsequently, the $SoftMax$ function is applied to achieve linear layering and normalization, resulting in the final output.

C. PostProcessing

After gaining the output rPPG signal of the CNN-Transformer network, we analyze these rPPG signals in frequency domain thereby obtaining the power spectrum and peak frequency chart. Identify the frequency corresponding to the maximum peak value and the corresponding heart rate can be calculated based on this frequency value.

$$HR = 60 \times F_{max}(\omega) \tag{8}$$

In the above formula, $F_{max}(\omega)$ represents the frequency matched at the peak maximum.

D. Loss Function

In order to accurately reconstruct the rPPG signal, we consider both time-domain and frequency-domain loss. To ensure that the predicted rPPG signal exhibits a similar trend to the ground truth label, we use negative Pearson correlation coefficient to calculate the time-domain loss, which is shown as follows:

$$L_r = 1 - \frac{\sum_{t=1}^{T}(x_t - \overline{x_t})(y_t - \overline{y_t})}{\sqrt{\sum_{t=1}^{T} x_t - \overline{x_t}^2}\sqrt{\sum_{t=1}^{T} y_t - \overline{y_t}^2}} \tag{9}$$

In Equation(9), T represents the length of the video, x_t represents the tth predicted rPPG signal, and y_t represents the tth ground truth PPG signal value.

Additionally, we also introduce cross-entropy loss to calculate frequency-domain loss, which is shown as follows:

$$L_f = CE(PSD(X), HR_{gt}) \tag{10}$$

In Equation(10), $PSD(X)$ represents the predicted power spectral density of the rPPG signal, HR_{gt} denotes the average heart rate value of the ground truth label, and $L_f = CE(X, Y)$ is used to calculate the cross-entropy loss between the predicted values and the true values.

Finally, the overall loss L will be calculated using $L = L_r + L_f$

IV. Experimental Results

In this section, we first introduce the datasets and then describe the experimental setting and implementation details. Next, we present our testing results under different experimental settings and show comparison with existing methods to evaluate its effectiveness.

A. Datasets and Evaluation Metrics

We have provided evaluation results for the public UBFC-rPPG [23] and PURE [24] datasets. The UBFC-rPPG dataset consists of facial videos of 2 minutes from 42 subjects. Videos are recorded in a resolution of 640 $\times$ 480 in 8-bit RGB format at the frame rate of 30 frames per second. The PURE dataset consisits of facial videos of 1 minute from 10 subjects. with each video lasting approximately 1 minute, a frame rate of 30Hz, and a resolution of 640 $\times$ 480 pixels. We report the most commonly used performance metrics for evaluation, including the mean absolute error (MAE), standard deviation (SD), root mean square error (RMSE) and Pearson Correlation Coefficient (r).

(a) UBFC-rPPG (b) PURE

Fig. 2: Examples of UBFC-rPPG and PURE datasets.

B. Implementation Details

The method we proposed is based on the PyTorch framework and is implemented using an NVIDIA TITAN GPU for training. All videos and ground truth PPG signals are downsampled to 30 Hz. The total epoch is 30 epochs, using the Adam optimizer, with learning rate of 0.001 and a batch size of 8. The ratio of the training set and the test set is 7:3.

C. IMF Extraction

After seperate the green channel signals from two facial ROIs, we perform the EEMD decomposition on the preprocessed signals to extract the IMFs that contain heart rate information to determine the heart rate values.

The EEMD decomposition result is shown in Fig. 3 and we obtain a total of 5 IMFs. Then we use fast Fourier transform to transform the IMFs to frequency domain. The component with the highest peak in the frequency spectrum within the heart rate range is identified as the target signal. It can be observed that IMF2 has the highest peak in the heart rate range, indicating that this signal has the strongest periodicity within the heart rate range.

D. Experimental Results

We trained and tested our proposed algorithm on the UBFC-rPPG and PURE datasets and compared it with 6 heart rate prediction algorithms, the results are shown in Table II and Table I.

From Table II and I, it can be concluded that compared to the deep learning-based methods (EfficientPhys [25], PhysNet, TS-CAN, and PhysFormer), the overall performance of these two traditional methods (CHROM [26], POS [27]) is relatively lower. Compared with the two traditional methods, PhysNet and EfficientPhys, which are based on 3DCNNs, have significantly lower Mean Absolute Error (MAE) and lower Root Mean Square Error (RMSE), thereby improving stability. PhysFormer utilizes the Transformer architecture to consider long-term features and has made significant progress in performance compared to the previous methods (CHROM, POS, EfficientPhys, PhysNet, and TS-CAN), indicating that global features are key factors in measuring rPPG signals.

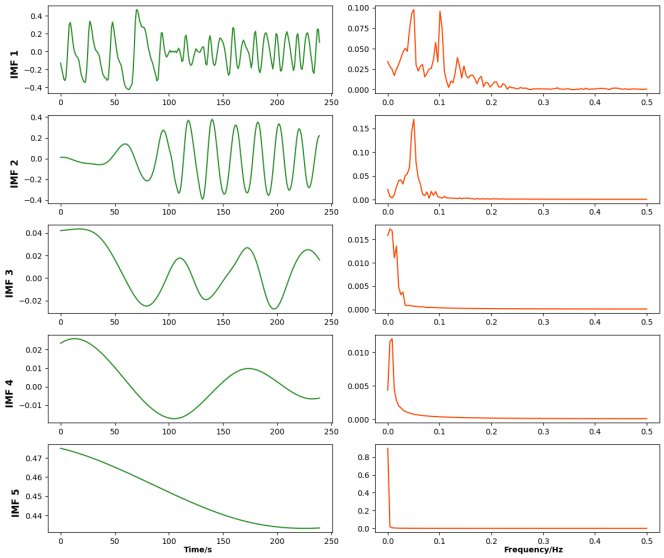


Fig. 3: Example of an EEMD-based rPPG decomposition

Compared to all other methods, our proposed method shows excellent performance in all performance metrics, with an MAE of 0.87 bpm on PURE dataset and 0.81 bpm on UBFC-rPPG dataset, an RMSE of 1.72 bpm on PURE dataset and 1.65 bpm on UBFC-rPPG dataset, and a correlation coefficient (r) of 0.99 on both dataset.

TABLE I: Performance Comparsion Trained on PURE Dataset and Test on UBFC-rPPG and PURE Datasets.

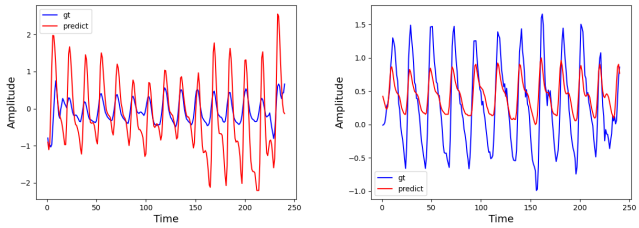
Method	UBFC-rPPG			PURE		
	MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑
CHROM	3.97	8.72	0.88	3.98	3.78	0.91
POS	3.99	7.58	0.92	4.00	3.86	0.92
EfficientPhys	1.59	2.70	0.95	2.07	2.10	0.93
PhysNet	3.33	4.05	0.75	1.63	1.68	0.93
TS-CAN	2.73	3.45	0.96	1.29	1.50	0.92
PhysFormer	4.33	4.56	0.71	2.11	2.98	0.88
Ours	**0.87**	**1.72**	**0.99**	**0.92**	**1.34**	**0.99**

TABLE II: Performance Comparsion Trained on UBFC-rPPG Dataset and Test on UBFC-rPPG and PURE Datasets.

Method	UBFC-rPPG			PURE		
	MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑
CHROM	3.98	3.78	0.89	5.77	11.52	0.88
POS	4.00	3.86	0.92	3.67	7.25	0.92
EfficientPhy	1.77	2.21	0.92	2.13	2.38	0.93
PhysNet	1.75	2.43	0.76	5.54	5.32	0.89
TS-CAN	2.69	3.71	0.95	5.36	5.84	0.92
PhysFormer	3.18	4.66	0.68	4.48	5.56	0.88
Ours	**0.81**	**1.65**	**0.99**	**1.13**	**1.98**	**0.99**

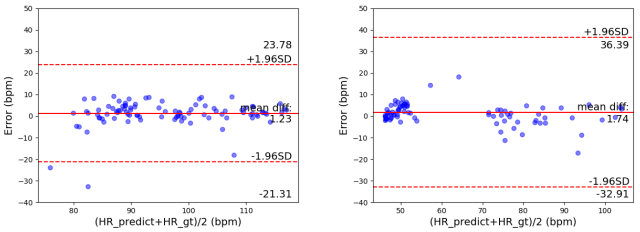
We visualized both the predicted rPPG signals and the ground-truth signals on two video clips from the testing data, as shown in Fig. 4. The similarity between the predicted rPPG signals and the ground-truth PPG signals is depicted in Fig. 4. It can be observed from Fig. 4 that the predicted rPPG signals fit well with the ground-truth PPG signals, indicating that the method proposed in this paper can construct high-quality rPPG signals from facial videos.

In addition, as shown in Fig. 5 and Fig. 6, we use the Bland-Altman plots and Scatter plots to further evaluate the correspondence between the ground truth HR_{gt} and the predicted HR_{pred}. From the plot, it can be seen that the average difference is approximately 0.88, and 97% of the error values in the detection data fall within the 95% consistency interval. This indicates a high level of consistency between the HR_{gt} and the HR_{pred}. The method we proposed contributes to better consistency and a smaller standard deviation.



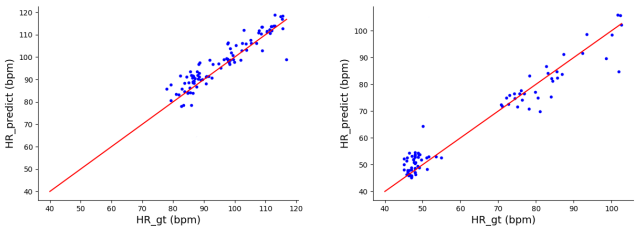
(a) UBFC-rPPG (b) PURE

Fig. 4: Visualization of rPPG signal reconstruction in the UBFC-rPPG and PURE dataset.



(a) UBFC-rPPG (b) PURE

Fig. 5: Bland–Altman plots between the ground truth HR_{gt} and the predicted HR_{pred} on UBFC-rPPG and PURE dataset.



(a) UBFC-rPPG (b) PURE

Fig. 6: Scatter plot comparing the ground truth HR_{gt} and the predicted HR_{pred} on UBFC-rPPG and PURE dataset.

E. Ablation Study

We also conducted ablation experiments for the method proposed in this paper. In Table III, we conduct 3 experiments in

HR detection to prove the effectiveness of each branch within the network. The results show that our method provides better performance than pure CNN and Transformer frameworks.

CNN framework excels at extracting local features but has limited capability in handling sequential data because it is not adept at capturing dependencies in time sequences, thus fails at capturing global features. While the Transformer structure is good at capturing global features but has higher computational costs when dealing with long sequences. The CNN+Transformer structure proposed in this paper achieves more powerful feature extraction capabilities while saving computational resources.

Additionally, we also evaluated the complexity of the network, the results indicate that the method we proposed has a lower complexity in terms of the number of parameters and FLOPs.

TABLE III: Ablation Study of the proposed method on UBFC-rPPG and PURE dataset

CNN branch	Transformer branch	UBFC-rPPG MAE	UBFC-rPPG RMSE	PURE MAE	PURE RMSE
✓		1.75	2.43	3.33	4.05
	✓	3.18	4.66	4.33	4.56
✓	✓	**0.81**	**1.65**	**0.87**	**1.72**

V. Conclusion

To overcome the issue that non-contact heart rate detection is easily affected by environment and the movement of the subjects, we propose a method that employs Ensemble Empirical Mode Decomposition (EEMD) to decompose the rPPG signal, and then put the extracting Intrinsic Mode Functions (IMFs) into a feature fusion model based on Transformer and CNN networks, which is better at dealing with different feature types and interference noise. Comprehensive experiments based on real-world datasets show the excellent performance of the method in both intra-dataset and cross-dataset testing. In our subsequent work, we will further extend the model to provide an end-to-end HR estimation network and adapt the model to more complex real-world scenarios.

References

[1] G. O. Young, "Synthetic Structure of Industrial Plastics," *Plastics*, New York, pp. 15-64, 1964.

[2] W. Chen, "Linear Networks and Systems," *Brooks/Cole Engineering Division*, Belmont, CA, USA: Wadsworth, pp. 123-135, 1993.

[3] W. Verkruysse, L. O. Svaasand, J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[4] M. Z. Poh, D. J. McDuff, R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics express*, vol. 18, no. 10, pp. 10762–10774, 2010.

[5] M. Lewandowska, J. Rumiński, T. Kocejko and J. Nowak, "Measuring pulse rate with a webcam—A non-contact method for evaluating cardiac activity," in *Proc. of FedCSIS*, pp. 405–410, 2011.

[6] G.S. Hsu, A. Ambikapathi, and M. S. Chen, "Deep learning with timefrequency representation for pulse estimation from facial videos," in *Proc. of IJCB*, pp. 383–389, 2017.

[7] X. Niu, H. Han, S. Shan, and X. Chen, "Vipl-hr: A multi-modal database or pulse estimation from less-constrained face video," in *Proc. of ACCV*, pp. 562–576, 2018.

[8] W. Chen, D, McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proc. of ECCV*, pp. 349-365, 2018.

[9] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, X. Chen, "PulseGAN: Learning to Generate Realistic Pulse Waveforms in Remote Photoplethysmography," *IEEE JBHI*, vol. 25, no. 5, pp. 1373-1384, 2021.

[10] X. Niu, H. Han, S. Shan and X. Chen, "SynRhythm: Learning a Deep Heart Rate Estimator from General to Specific," in *Proc. of ICPR*, pp. 3580-3585, 2018.

[11] W. Chen, D. Mcduff, "DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks," in *Proc. of ECCV*, pp. 356-373, 2018.

[12] R. Spetlík, V. Franc, J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proc. of BMVC* pp. 3-6, 2018.

[13] Z. Yu, X. Li, G. Zhao, "Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks," in *Proc. of BMVC*, 2019.

[14] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, G. Zhao, "Video-Based Remote Physiological Measurement via Cross-Verified Feature Disentangling," in *Proc. of ECCV*, 2020.

[15] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, X. Chen, "Robust remote heart rate estimation from face utilizing spatial-temporal attention," in *Proc. of FG*, 2019.

[16] X. Liu, J. Fromm, S. Patel, D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Proc. of NIPS*, 2020.

[17] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, "Autohr: a strong end-to-end baseline for remote heart rate measurement with neural searching," *IEEE Signal Processing Letters*, vol. 27, pp. 1245–1249, 2020.

[18] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, X. Chen, "Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks," *IEE TIM*, vol. 69, no. 10, pp. 7411-7421, 2020.

[19] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. Torr and G. Zhao, "PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer," in *Proc. of CVPR*, 2022.

[20] Z. Yu, y. Shen, J. Shi, H. Zhao, y. Cui, J. Zhang, P. Torr, G. Zhao, "PhysFormer++: Facial Video-based Physiological Measurement with SlowFast Temporal Difference Transformer," in *Proc. of CVPR*, 2023.

[21] E. Huang, Z. Shen, R. Long, C. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proc. R. Soc. Lond. A*, pp. 454903–995, 1998.

[22] H. Wu, E.Huang, "Ensemble empirical mode decomposition: A noise assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, pp. 1-41, 2009.

[23] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, pp. 82–90, 2019.

[24] R. Stricker, S. Müller, H. M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. of 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1056–1062, 2014.

[25] X. Liu, B. L. Hill, Z. Jiang, S. Patel, D. McDuff, "EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Vitals Measurement, ", in *in Proc. WACV*, 2021.

[26] G. de Haan, V. J, "Robust Pulse Rate From Chrominance-BasedrPPG," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878-2886, 2013.

[27] R. Song, s. Zhang, J. Cheng, C. Li, X. Chen, "New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method," *Computers in Biology and Medicine*, vol. 116, 2019.