Multimodal Video-Based Heart Rate Estimation with Temporal Difference Transformer

Zhiqin Zhou[1], Songping Wang[1] and Caifeng Shan[1,*]

[1]State Key Laboratory for Novel Software Technology and School of Intelligence Science and Technology, Nanjing University, Suzhou, 215123, China

Abstract

With the advancement of remote photoplethysmography (rPPG), video-based physiological signal measurement has emerged as a convenient and contactless method for heart rate (HR) estimation. Most existing video-based rPPG methods rely on either RGB or NIR imaging, each with distinct advantages and limitations. RGB-based methods demonstrate higher accuracy but are sensitive to lighting and skin tone, while NIR-based methods are more robust to such variations but less responsive to blood volume changes, leading to lower accuracy. Therefore, we propose a temporal difference transformer (TDT)-based multimodal fusion framework for robust HR estimation. We employ separate 3D convolutional encoders to independently extract modality-specific spatio-temporal representations from RGB and NIR input streams. Subsequently, the TDT module enhances quasi-periodic features while adaptively aligning and fusing RGB and NIR representations. Additionally, a composite loss function is introduced to provide simultaneous supervision across temporal and spectral domains. In the 4th RePSS Challenge, the proposed method achieved second place with the root mean square error (RMSE) of 12.32 bpm on the official test set, demonstrating strong performance in multimodal HR estimation.

Keywords

Remote photoplethysmography, heart rate estimation, temporal difference transformer, multimodal fusion

1. Introduction

Heart rate (HR) is a crucial physiological indicator for assessing health and emotional states [1]. While traditional contact-based methods can cause discomfort, remote photoplethysmography (rPPG) has emerged as a promising non-contact alternative. By analyzing subtle color variations in facial videos caused by blood volume changes with each cardiac pulse, rPPG enables HR estimation without physical contact. With the rapid advancement of deep learning techniques [2], video-based rPPG methods have made significant progress [3, 4]. Numerous studies have adopted convolutional neural networks (CNNs) [5] and transformer-based architectures [6, 7] to model spatio-temporal features of facial blood flow. These learning frameworks have demonstrated promising performance across various public datasets.

Despite the success of RGB-based deep learning methods for HR estimation, they still face considerable challenges in real-world scenarios. In low-light conditions, limited visible light reduces the signal-to-noise ratio of rPPG signals [8]. Illumination changes introduce color shifts that can obscure true physiological rhythms. Skin tone differences also affect reflectance: darker skin absorbs more light, leading to weaker signals and lower rPPG amplitude. In contrast, NIR cameras capture only the infrared portion of the incident light, which is less sensitive to ambient illumination changes and penetrates deeper into the skin. This enables more stable signal acquisition in low-light conditions and reduces reflectance disparities across different skin tones. HR estimation methods based on NIR cameras and infrared illumination exhibit advantages in such settings [9]. However, due to the limited sensitivity of NIR light to blood volume changes, NIR-based methods generally result in less accurate HR estimation than RGB-based methods [10].

IJCAI 2025: International Joint Conference on Artificial Intelligence, August 28, 2025, Guangzhou, China

*Corresponding author.

zin@smail.nju.edu.cn (Z. Zhou); theone@buaa.edu.cn (S. Wang); cfshan@nju.edu.cn (C. Shan)

https://caifeng-shan.github.io/ (C. Shan)

0009-0001-1229-4400 (Z. Zhou); 0009-0001-4513-7284 (S. Wang); 0000-0002-2131-1671 (C. Shan)

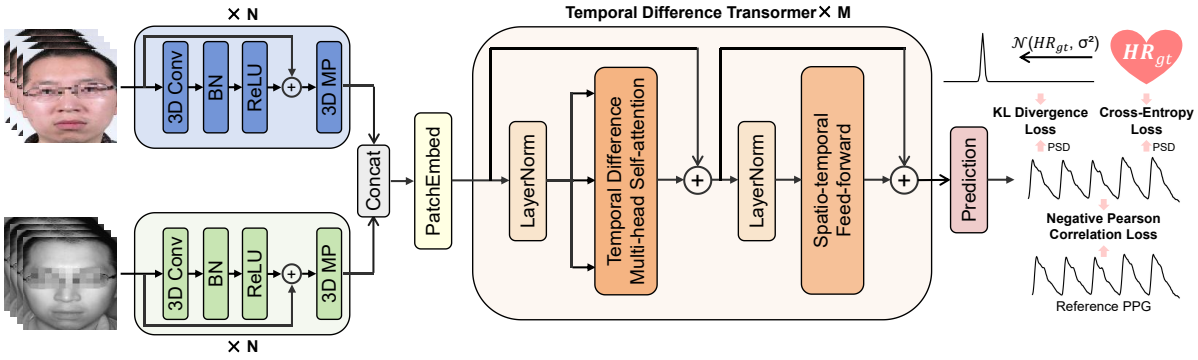

Figure 1: The whole framework of our multimodal video-based HR estimation network.

To address these limitations, integrating RGB and NIR modalities shows promise for achieving a balance between illumination robustness and sensitivity to blood flow signals. By leveraging the complementary strengths of both modalities, multimodal fusion has the potential to improve the accuracy and generalizability of rPPG estimation in diverse and complex environments. To explore the potential of multimodal data fusion in rPPG, the 4th Remote Physiological Signal Sensing (RePSS) Challenge was launched in conjunction with the International Joint Conference on Artificial Intelligence (IJCAI 2025).

In this challenge, we propose an end-to-end multimodal fusion method based on the temporal difference transformer (TDT) [11]. First, separate 3D convolutional encoders extract modality-specific spatio-temporal representations from the RGB and NIR input streams. The TDT module then enhances quasi-periodic features while adaptively aligning and fusing cross-modal RGB and NIR representations. Furthermore, a composite loss function enforces simultaneous supervision in both temporal and spectral domains to optimize feature learning. By effectively integrating features from both modalities and domains, the proposed method enhances the robustness of HR estimation under challenging scenarios. On the official challenge test dataset, the proposed method achieved the root mean square error (RMSE) of 12.32, ranking second among all participating teams. These results demonstrate the effectiveness and competitiveness of the proposed approach in multimodal HR estimation tasks.

2. Methodology

As shown in Fig. 1, our method consists of three main components: modality-specific spatio-temporal encoding, multimodal feature fusion, and rPPG prediction. Furthermore, a composite loss function is designed to provide supervision in both the time and frequency domains. In the following sections, we provide a detailed description of each component.

2.1. Modality-specific Spatio-temporal Encoding

Given a sequence of facial video frames from two modalities (RGB denoted as $\mathbf{X}_{\text{RGB}}$ and NIR as $\mathbf{X}_{\text{NIR}}$), each modality is independently processed through a 3D convolutional encoder to extract spatio-temporal features specific to that modality. The RGB input has three channels, while the NIR input has one channel, with both sharing the same temporal and spatial dimensions.

Each encoder is composed of multiple stacked blocks, where each block includes a 3D convolutional layer followed by batch normalization, ReLU activation, and 3D max pooling. To improve training stability and maintain feature quality, residual connections are introduced within each block by adding the input features to the output of the transformation.

Through this encoding process, we obtain modality-specific feature representations: $\mathbf{F}_{\text{RGB}} \in \mathbb{R}^{D\times T_1\times H_1\times W_1}$ and $\mathbf{F}_{\text{NIR}} \in \mathbb{R}^{D\times T_1\times H_1\times W_1}$, both sharing the same shape in terms of temporal and spatial resolution after downsampling, but carrying distinct semantic information from their respective inputs.

2.2. Multimodal Feature Fusion

In temporal difference multi-head self-attention (TD-MHSA) [11], we utilize a temporal difference convolution (TDC) operator to explicitly encode temporal dynamics into attention computations. Given an input sequence $\boldsymbol{x} \in \mathbb{R}^{2D \times T_2 \times H_2 \times W_2}$, the TDC operation is formally defined as:

$$\text{TDC}(\boldsymbol{x}) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot \boldsymbol{x}(p_0 + p_n) + \theta \cdot \big(- \boldsymbol{x}(p_0) \cdot \sum_{p_n \in \mathcal{R}'} w(p_n) \big) \tag{1}$$

where p_0 denotes the current spatio-temporal location, $\mathcal{R}$ is the local $3 \times 3 \times 3$ receptive field, and $\mathcal{R}'$ represents the adjacent temporal neighborhood.

For each attention head i, we compute the query/key/value matrices by:

$$Q_i = \mathcal{S}\big(\text{TDC}\left(\text{LN}(\boldsymbol{x})\right)\big), K_i = \mathcal{S}\big(\text{TDC}\left(\text{LN}(\boldsymbol{x})\right)\big), V_i = \mathcal{S}\big(\text{Conv3D}\left(\text{LN}(\boldsymbol{x})\right)\big), \tag{2}$$

where $\text{LN}(\cdot)$ is layer normalization, $\mathcal{S}(\cdot)$ denotes the reshaping operation from video to sequence format, and $Q_i, K_i, V_i \in \mathbb{R}^{(T_2 \cdot H_2 \cdot W_2) \times 2D}$. This design explicitly injects temporal change patterns into the attention mechanism through differential operators.

The self-attention for each head is then computed using the scaled dot-product:

$$\text{SA}_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{D_h}}\right) V_i, \tag{3}$$

where $D_h = \frac{2D}{h}$ is the dimensionality of each attention head, h represents the number of attention heads.

The outputs from all h heads are concatenated and linearly projected, then reshaped back to the original video format via $\mathcal{V}(\cdot)$, yielding the final TD-MHSA output:

$$\text{TD-MHSA} = \mathcal{V}\big(\text{Concat}(\text{SA}_1, \text{SA}_2, \ldots, \text{SA}_h) \cdot U\big), \tag{4}$$

where U is a learnable output projection matrix, and $\mathcal{V}(\cdot)$ is the inverse reshape operator (sequence to video).

To enhance local feature consistency and positional awareness, a spatio-temporal feed-forward (ST-FF) module is employed in place of standard linear layers. The module consists of a $1 \times 1 \times 1$ pointwise convolution for channel expansion, a $3 \times 3 \times 3$ depthwise-separable 3D convolution to model spatio-temporal interactions, and a second $1 \times 1 \times 1$ pointwise convolution for dimensionality reduction. Each stage is followed by batch normalization and ELU activation to promote training stability and non-linearity.

2.3. rPPG Prediction

After the TDT module completes modality fusion and temporal modeling, the model employs a two-stage upsampling process to progressively restore temporal resolution while reducing channel dimensionality to extract key sequential features. Spatial average pooling is then applied to integrate local information and produce a more compact temporal representation. Finally, a regression module maps the temporal features into a rPPG signal sequence.

To obtain the final HR, the fast Fourier transform (FFT) is applied to the predicted rPPG signal and identify the dominant frequency component within the physiological range (0.7–4 Hz). The frequency with the highest amplitude is selected as the pulse frequency f_{peak}.

2.4. Loss Function

To enable the model to effectively reconstruct the rPPG waveform and accurately estimate HR, we design a composite loss function consisting of three components, providing supervision from both the time and frequency domains.

KL Divergence Loss [12]: This loss provides soft supervision in the frequency domain. Given the ground truth $\mathrm{HR}_{\mathrm{gt}}$, a Gaussian distribution centered at $\mathrm{HR}_{\mathrm{gt}}$ with standard deviation σ is used as the target distribution p_k over the discretized frequency bins [11]:

$$p_k = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(k - (\mathrm{HR}_{\mathrm{gt}} - \mathrm{HR}_{\min}))^2}{2\sigma^2}\right), \tag{5}$$

where $\mathrm{HR}_{\min}$ denotes the theoretical minimum HR. Let $\hat{p}$ represent the softmax-normalized result of the power spectral density (PSD) of the predicted rPPG signal. The KL divergence loss is defined as:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{k=0}^{L-1} p_k \cdot \log\left(\frac{p_k}{\hat{p}_k}\right). \tag{6}$$

Cross-Entropy Loss [13]: Let $\hat{z}_k$ represent the spectral energy at frequency bin k in the PSD of the predicted rPPG, where y is the index of the bin closest to the $\mathrm{HR}_{\mathrm{gt}}$. The cross-entropy loss is then:

$$\mathcal{L}_{\mathrm{CE}} = -\log\left(\frac{\exp(\hat{z}_y)}{\sum_{k=0}^{L-1}\exp(\hat{z}_k)}\right). \tag{7}$$

Negative Pearson Correlation Loss [14]: This loss optimizes temporal similarity between the predicted signal and ground truth. Given predicted signal x and ground truth y, the loss is defined as:

$$\mathcal{L}_{\mathrm{Pearson}} = 1 - \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}, \tag{8}$$

where $\mathrm{Cov}(x, y)$ represents the covariance between x and y, and σ_x, σ_y are the standard deviations of the predicted signal and ground truth, respectively.

Overall Loss: The final training objective combines the three losses with weighting coefficients:

$$\mathcal{L}_{\mathrm{total}} = \alpha \cdot \mathcal{L}_{\mathrm{Pearson}} + \beta \cdot (\mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{CE}}), \tag{9}$$

where α and β control the relative importance of the temporal and frequency-domain losses.

3. Experiments

3.1. Datasets

The training data comes from the **VIPL-HR** [15, 16] dataset, which includes paired RGB and NIR videos from 107 subjects, along with synchronized ground truth PPG signals. It covers diverse real-world scenarios such as talking, body movement, and varying lighting conditions.

The test set consists of paired RGB-NIR clips from 100 subjects in the **VIPL-V2** [15] dataset and 100 subjects from the **OBF** [17] dataset. All videos are divided into 10 s segments for evaluation. **VIPL-V2** provides diverse lighting conditions, while **OBF** includes subjects with varied skin tones, enabling a comprehensive assessment of the model's generalization ability.

3.2. Implement and evaluation metric

The model proposed in this study is implemented using the PyTorch framework and trained on a high-performance computing system equipped with eight NVIDIA GeForce RTX 4090 GPUs. Each input sample consists of a facial video segment containing 160 frames, with each frame resized to a resolution of 128×128 pixels. During training, the batch size is set to 8, and the total number of training epochs is 50. The model is optimized using the Adam optimizer, with a weight decay coefficient of 5×10^{-5}. The initial learning rate is set to 1×10^{-4}, and a StepLR learning rate scheduler is employed, which halves the learning rate every 25 epochs.

Table 1
The final leaderboard of the 4th RePSS Challenge's top five teams.

Ranking	Team Name	Captain Affiliation	RMSE (bpm)
1	HFUT-VUT	Hefei University of Technology	11.89505
2	**IST** (Ours)	Nanjing University	**12.31846**
3	xjgroupscu	Sichuan University	12.70790
4	NJU_TEAM	Nanjing University	14.51449
5	Sgt.Pepper's	Hefei University of Technology	14.69105

Table 2
Comparison of different models and input modalities for HR estimation. PhysFormer (RGB-NIR) denotes a variant where the corresponding RGB and NIR frames are directly concatenated and used as the input to the PhysFormer model. The best result is highlighted in bold.

Method	RMSE (bpm)
PhysFormer (RGB-only)	12.95383
PhysFormer (RGB-NIR)	12.92008
Ours (RGB-NIR)	**12.31846**

The evaluation metric used to evaluate the prediction accuracy is the RMSE, which is defined as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\text{HR}_{\text{pred}_i} - \text{HR}_{\text{gt}_i})^2}{N}}, \tag{10}$$

where N denotes the number of video segments, HR_{gt_i} represents the ground truth HR of the i-th video segment, and $\text{HR}_{\text{pred}_i}$ denotes the predicted HR of the i-th video segment.

3.3. Results

As shown in Table 1, our team, IST (Nanjing University), achieved the RMSE of 12.31846 bpm in the HR estimation task, ranking second in the final leaderboard of the 4th RePSS Challenge. Compared to the winning team, HFUT-VUT, which achieved the RMSE of 11.89505 bpm, our result is only approximately 0.42 bpm higher, demonstrating the strong performance and competitiveness of our proposed method.

Table 2 compares our method with PhysFormer under different input settings. PhysFormer achieved the RMSE of 12.95383 bpm with RGB input, and a slightly improved 12.92008 bpm when using concatenated RGB and NIR frames. In contrast, the proposed method, with a more effective RGB-NIR fusion strategy, achieved a significantly lower RMSE of 12.31846 bpm, reflecting a relative improvement of about 4.9% over the RGB-only PhysFormer. This demonstrates the importance of both modality integration and effective fusion design to fully leverage the complementary information from RGB and NIR inputs.

4. Conclusion

This paper presents a multimodal HR estimation method that fuses RGB and NIR video information using a TDT architecture. Modality-specific spatio-temporal encoders are first applied to extract robust spatio-temporal features, providing a strong foundation for subsequent fusion. The TDT enables the model to capture subtle physiological variations in facial regions while effectively fusing RGB and NIR features. To further improve prediction accuracy, the model is trained with a composite loss function combining time-domain and frequency-domain components, guiding it to focus on both dominant frequency localization and waveform shape consistency. In the 4th RePSS Challenge, the proposed method achieved second place on the official test set, demonstrating its effectiveness and competitiveness in multimodal HR estimation. We believe that future improvements could lead to even better outcomes.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62350068).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] D. McDuff, S. Gontarek, R. Picard, Remote measurement of cognitive stress via heart rate variability, in: 2014 36th annual international conference of the IEEE engineering in medicine and biology society, IEEE, 2014, pp. 2957–2960.

[2] X. Wei, S. Wang, H. Yan, Efficient robustness assessment via adversarial spatial-temporal focus on videos, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 10898–10912.

[3] D.-Y. Kim, S.-Y. Cho, K. Lee, C.-B. Sohn, A study of projection-based attentive spatial–temporal map for remote photoplethysmography measurement, Bioengineering 9 (2022) 638.

[4] J. Wang, C. Shan, L. Liu, Z. Hou, Camera-based physiological measurement: Recent advances and future prospects, Neurocomputing 575 (2024) 127282.

[5] J. Comas, A. Ruiz, F. Sukno, Efficient remote photoplethysmography with temporal derivative modules and time-shift invariant loss, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 2182–2191.

[6] L. Pang, X. Li, Z. Wang, X. Lei, Y. Pei, Self-supervised augmented vision transformers for remote physiological measurement, in: 2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), IEEE, 2023, pp. 623–627.

[7] R.-X. Wang, H.-M. Sun, R.-R. Hao, A. Pan, R.-S. Jia, Transphys: Transformer-based unsupervised contrastive learning for remote heart rate measurement, Biomedical Signal Processing and Control 86 (2023) 105058.

[8] K. Kurihara, D. Sugimura, T. Hamamoto, Non-contact heart rate estimation via adaptive rgb/nir signal fusion, IEEE Transactions on Image Processing 30 (2021) 6528–6543.

[9] E. Magdalena Nowara, T. K. Marks, H. Mansour, A. Veeraraghavan, Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 1272–1281.

[10] W. Wang, A. C. den Brinker, G. De Haan, Discriminative signatures for remote-ppg, IEEE Transactions on Biomedical Engineering 67 (2019) 1462–1473.

[11] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, G. Zhao, Physformer: Facial video-based physiological measurement with temporal difference transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4186–4196.

[12] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Transactions on Image Processing 26 (2017) 2825–2838.

[13] 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 27 (2020) 1245–1249.

[14] Z. Yu, W. Peng, X. Li, X. Hong, G. 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, 2019, pp. 151–160.

[15] X. Niu, H. Han, S. Shan, X. Chen, Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video, in: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14, Springer, 2019, pp. 562–576.

[16] X. Niu, S. Shan, H. Han, X. Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, IEEE Transactions on Image Processing 29 (2019) 2409–2423.

[17] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, G. 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), IEEE, 2018, pp. 242–249.