technologies

Article

Transformative Approach for Heart Rate Prediction from Face Videos Using Local and Global Multi-Head Self-Attention

Smera Premkumar [1], J. Anitha [1], Daniela Danciulescu [2] and D. Jude Hemanth [1,*]

[1] Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore 641114, India; smerapremkumar@karunya.edu.in (S.P.); anithaj@karunya.edu (J.A.)
[2] Department of Computer Science, University of Craiova, 200585 Craiova, Romania; daniela.danciulescu@edu.ucv.ro
* Correspondence: judehemanth@karunya.edu

Abstract: Heart rate estimation from face videos is an emerging technology that offers numerous potential applications in healthcare and human–computer interaction. However, most of the existing approaches often overlook the importance of long-range spatiotemporal dependencies, which is essential for robust measurement of heart rate prediction. Additionally, they involve extensive pre-processing steps to enhance the prediction accuracy, resulting in high computational complexity. In this paper, we propose an innovative solution called LGTransPPG. This end-to-end transformer-based framework eliminates the need for pre-processing steps while achieving improved efficiency and accuracy. LGTransPPG incorporates local and global aggregation techniques to capture fine-grained facial features and contextual information. By leveraging the power of transformers, our framework can effectively model long-range dependencies and temporal dynamics, enhancing the heart rate prediction process. The proposed approach is evaluated on three publicly available datasets, demonstrating its robustness and generalizability. Furthermore, we achieved a high Pearson correlation coefficient (PCC) value of 0.88, indicating its superior efficiency and accuracy between the predicted and actual heart rate values.

Keywords: remote photoplethysmography; transformer; heart rate prediction

Citation: Premkumar, S.; Anitha, J.; Danciulescu, D.; Hemanth, D.J. Transformative Approach for Heart Rate Prediction from Face Videos Using Local and Global Multi-Head Self-Attention. *Technologies* **2024**, *12*, 2. https://doi.org/10.3390/technologies12010002

Academic Editors: Jeffrey W. Jutai, Bang Wang and Pietro Zanuttigh

Received: 6 October 2023
Revised: 21 November 2023
Accepted: 13 December 2023
Published: 22 December 2023

1. Introduction

Heart rate (HR) monitoring is a fundamental part of healthcare, since it offers valuable information about individuals' health. Conventional HR measurements use specialized medical devices like electrocardiograms (ECG) and photoplethysmography (PPG) sensors, which require direct contact with the skin and are suitable in clinical settings. These approaches are not suitable for long-term monitoring, particularly in vulnerable people like infants and the elderly, and it also causes skin irritation. The development of non-invasive camera-based technologies has revolutionized HR monitoring which predicts the heart rate by analyzing subtle changes in skin coloration, resulting from the blood flow variation [1], and this method is referred to as remote photoplethysmography (rPPG). These subtle changes on the skin are imperceptible to the human eye, but camera sensors can capture them with precision. By processing these image pixels over time using specialized signal processing techniques, we can extract the PPG signal and thereby predict the physiological parameters including heart rate, Heart Rate Variability (HRV) and Blood Pressure (BP). With the introduction of digital cameras, remote heart rate monitoring has become accessible across diverse fields, encompassing hospital care [2], telemedicine [3,4], fitness assessment [5,6], motion recognition [7], and the automotive industry [8,9]. This remote method has extended its applications to numerous areas including mental stress detection, cardiovascular function variations, sleep quality assessment, and drowsiness identification [10–12].

Initially, the possibilities of rPPG from face videos were introduced in [13]. Since the rPPG method is a camera-based technique, it is affected by changes in lighting, compression and motion artifacts, all of which can impact the accuracy of predicted output. The main objective is to extract the required signal by addressing these issues, and it needs extensive pre-processing steps including face and region of interest (ROI) detection, signal extraction, and normalization. While these steps were essential for ensuring robust results, they often imposed computational burdens and time constraints. Signal processing methods were the techniques used to extract remote photoplethysmography (rPPG) signals in the beginning.

Later, there was a shift toward learning-based approaches to enhance precision and real-world applicability. These models, particularly, convolutional neural networks (CNNs), have demonstrated their effectiveness in extracting features from facial videos. In the literature, there have been some studies that propose end-to-end methods; even in these cases, several pre-processing steps are required to regenerate the data before inputting it into the network. A comprehensive survey dealing with the developments of this area can be found in [14–16]. For instance, handcrafted methods are employed [17] to regenerate and normalize frames, which are then used as inputs to the network. Similarly, [18,19] utilizes feature maps called MSTmaps, which involve landmark detection and pixel averaging in various color spaces. The accuracy of the output prediction depends on the selection of pre-processing steps and the algorithms employed. Nevertheless, deep learning methods come with their own set of challenges. They often demand substantial amounts of labeled data for training. Moreover, deep learning models may necessitate multiple steps to prepare input data, including face detection and alignment, normalization, and data augmentation, all of which can introduce errors in rPPG extraction. The importance of non-contact methods for HR prediction from facial videos has been further discussed in recent times, especially during the COVID-19 pandemic, where the demand for remote healthcare solutions has increased.

To address the limitations discussed above, this paper proposes an efficient, end-to-end framework that explores the transformer-based network called LGTransPPG. Transformers are known for their capability to understand the long-range dependencies directly from raw data by enabling a multi-head self-attention mechanism. By employing a hybrid of local and global multi-head self-attention techniques, our method is capable of capturing spatiotemporal dependencies and thus enabling a more robust heart rate estimation. Our proposed approach aims to eliminate the need for extensive pre-processing steps and thereby reduce the computational complexity. The local aggregation mechanism focuses on capturing detailed elements within specific facial regions, while the global aggregation mechanism considers the overall spatial context of the face video.

To evaluate the performance of our approach, we conduct experiments on three publicly available datasets—MAHNOB, COHFACE and UCLA–rPPG, which are specifically designed for heart rate estimation from face videos. In addition to intra-dataset validation, we also perform cross-dataset validation to evaluate the generalizability of our approach. This paper aims to develop a more accurate and robust method for heart rate prediction. The contributions of this work include the following:

1. Introducing LGTransPPG, an end-to-end transformer-based framework for heart rate prediction from face videos that eliminates the need for extensive pre-processing steps.
2. Incorporating local and global aggregation techniques to capture both fine-grained facial details and spatial context, improving the accuracy of heart rate estimation.
3. The Frequency-Aware Correlation Loss (FACL) is designed to facilitate accurate heart rate prediction by emphasizing the alignment of frequency components.
4. We evaluate our approach on multiple publicly available datasets, showcasing its efficiency and accuracy compared to existing approaches.

In the subsequent sections, we provide a comprehensive description of the proposed LGTransPPG framework, detailing its architectural design, the experiments and the evaluation metrics employed. Furthermore, we analyze and discuss the experimental results, highlighting the advantages of our method over state-of-the-art methods.

2. Related Works

In this section, we provide an overview of the existing approaches of heart rate prediction from facial videos. We start by discussing state-of-art rPPG methods and then delve into the advancements in the research. We subsequently discuss the role of attention mechanisms and transformers in improving the accuracy of heart rate estimation.

Traditional approaches in remote heart rate estimation initially used Blind Source Separation (BSS) methods [20–22]. The aim of Blind Source Separation (BSS) algorithms is to extract the target PPG signal from noise and artifacts by using the correlation between the signals and thereby improving the signal-to-noise ratio. BSS methods for remote heart rate estimation such as Independent Component Analysis (ICA) [13] and Principal Component Analysis (PCA) [20] are applied to temporal RGB color signal sequences to identify the dominant component associated with heart rate. These are purely based on signal processing and thus difficult to incorporate deeper understandings of physiological and optical processes involved in rPPG measurement.

Taking these challenges into account, model-based methods have been proposed. One such method is the chrominance method (CHROM) [23], which aims to reduce motion challenges by utilizing orthogonal chrominance signals. It eliminates the specular reflection component to build orthogonal chrominance signals from RGB data. Another approach, known as spatial subspace rotation (2SR) [24], utilizes a spatial subspace of skin pixels and measures the temporal rotation within this subspace to extract pulse signals. These methods offer computational efficiency and ease of implementation. They assume an equal contribution of each pixel to the rPPG estimate, which makes them sensitive to noise and limits their applicability in real-world scenarios.

Later, with the development of computer vision, several convolutional neural network (CNN)-based methods have been proposed. DeepPhys [17] was the first learning-based approach, and it shows remarkable improvement on conventional signal processing methods by utilizing deep neural networks. However, these methods are based on two dimensional CNN, and it is primarily designed for image processing. While they are good at capturing spatial features, their effectiveness at capturing temporal features which involve changes over time is limited. Spatial processing involves identifying the skin area within the frame where the pulse signal is located, while temporal processing refers to separating the pulse signal in the time domain at various time instants.

Unlike static images, videos contain dynamic information that captures the variations in physiological signals over time. Therefore, successful rPPG methods need to incorporate not only spatial features but also temporal dynamics to predict the physiological signal accurately. A three-dimensional CNN based PhysNet [25] adopts various spatial–temporal modeling based on convolutional neural networks (CNNs). MTTS–CAN [26] illustrated an efficient on-device architecture that utilizes tensor-shift modules and 2D convolutional operations for spatial–temporal information extraction. Dual-GAN [19] employed an adversarial learning approach using facial landmarks to map video frames to pulse waveforms, while the concept of spatiotemporal maps (STmaps) [27] involved dividing the face into patches and concatenating them into sequences for analysis. The efficacy of three-dimensional architecture with two-dimensional CNN is demonstrated in [28]. However, ST maps serve as a pre-processing module and make a significant computational load to the entire pipeline. Hybrid methods [29–32] also suggested the integration of handcrafted features with deep learning networks for heart rate estimation. Motion and appearance frames are utilized in [31] to predict PPG signals and mitigate noise and illumination artifacts.

Another approach [33] utilized the CHROM method [23] and time-frequency representation for pulse signal extraction. These methods often require pre-processing steps, such as calculating difference frames and image normalization, to enhance the data quality. So, it is evident that these methods still need an improvement in measurement accuracy, as even minor input variations can have a significant impact on output predictions.

Attention Mechanism and Transformers

Transformers, originally popularized in natural language processing, have found their way into various sequence modeling tasks, including rPPG analysis. By incorporating self-attention mechanisms, transformers can effectively capture long-range dependencies, making them well suited for heart rate estimation from facial videos. The attention mechanism enables the model to focus on relevant facial regions and frames, enhancing the accuracy of heart rate prediction while suppressing the influence of noise and other artifacts. This approach eliminates the need for manual feature extraction and predefined filtering techniques. Unlike signal processing methods that require careful parameter tuning and assumptions about signal characteristics, transformer mechanisms automatically learn and adapt to diverse conditions.

The era of transformers began with their introduction in natural language processing [34], which transformed with the concept of attention mechanisms [35]. Attention mechanisms enable transformers to capture dependencies effectively by attending to relevant parts of the input sequence. Subsequently, transformers were successfully applied to image classification [36,37] and video understanding [38], showcasing their capability across different areas. Transformers outperformed traditional models like convolutional neural networks (CNNs), leveraging the self-attention mechanism to capture long-range dependencies [39,40].

In recent research, the Swin Transformer architecture was proposed [41], integrating shifted windows and a self-attention mechanism with non-overlapping windows. It helps to improve the model's ability to capture both local and global information. Later, an inductive bias of locality in the video transformers [42] introduces a better trade-off between speed and accuracy when compared to previous approaches that compute self-attention globally even with spatial–temporal factorization. While the use of transformers in rPPG prediction is hardly explored, there are some relevant works in this area. A temporal difference transformer was introduced [43,44] with the quasi-periodic rPPG features to represent the local spatiotemporal features and evaluated the results on multiple public datasets. Another notable contribution, Efficientphys [45], introduced an end-to-end neural architecture for device-based physiological sensing and conducted a comprehensive comparison with convolutional neural networks. It uses a Swin Transformer layer to extract spatial–temporal features, utilizing tensor shift modules. Despite its less favorable inference time compared to convolution-based methods, it demonstrates promising results. A detailed survey of visual transformers and attention mechanisms can be found in [46–48].

3. Proposed Approach

The proposed framework comprises an embedding module and local–global aggregation based on a multi-head self-attention mechanism followed by a predictor head for rPPG signal prediction. Inspired by [48], we present a novel approach for heart rate prediction from face videos by integrating a token embedding module as the initial step and processing raw video frames as input.

This methodology effectively captures crucial spatiotemporal information by employing a global–local aggregation technique with multi-head self-attention mechanisms, enabling the model to capture both local and global features. Lastly, it incorporates a prediction head to generate accurate heart rate predictions. Through the combination of these techniques, our method achieves both high accuracy in heart rate estimation and computational efficiency on state-of-art methods. An overview of the proposed approach can be seen in Figure 1. We start explaining the token embedding block and then two types of attention modules called Local Multi-Head Self-Attention (L-MHSA) and Global MHSA (G-MHSA).

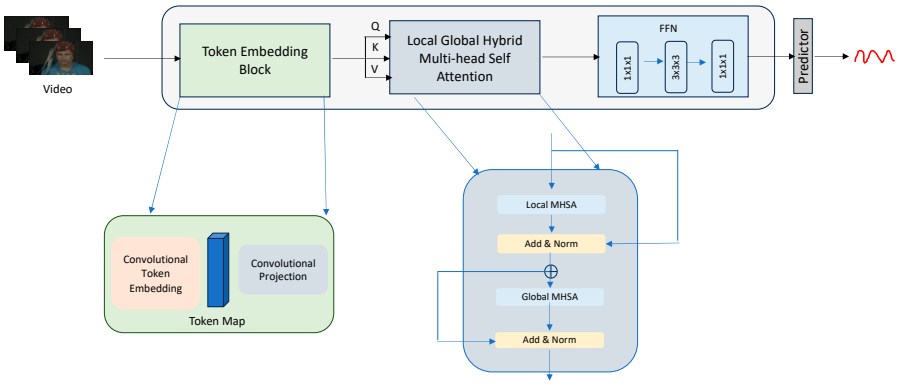

Figure 1. Pipeline of proposed method. A transformer-based rPPG prediction from face videos. It includes an embedding module, a multi-head self-attention mechanism based on local and global aggregation and a prediction head.

3.1. Token-Embedding Module

The token-embedding module consists of a convolution token embedding (CTE) and convolution projection (CP). This module incorporates into this architecture to enhance both performance and robustness all while preserving computational efficiency. CTE aims to model local spatial attribute information from raw frames using a hierarchical multistage approach. Instead of partitioning each frame into patches, the convolutional layer tokenizes the input video [48].

At the start of each stage, a convolutional token-embedding step conducts an overlapping convolution operation on the token map, and then a layer normalization is applied. This approach enables the model to capture local details and achieve spatial downsampling. The convolutional token embedding layer provides the flexibility to modify the token feature dimension and the token count at each stage by adjusting the convolution operation's parameters. Consequently, in each stage, we achieve a gradual reduction in the token sequence length while expanding the token feature dimension. This empowers the tokens to capture more complex visual patterns over larger spatial areas similar to the feature layers in CNNs. An illustration of the token-embedding module can be seen in Figure 2.

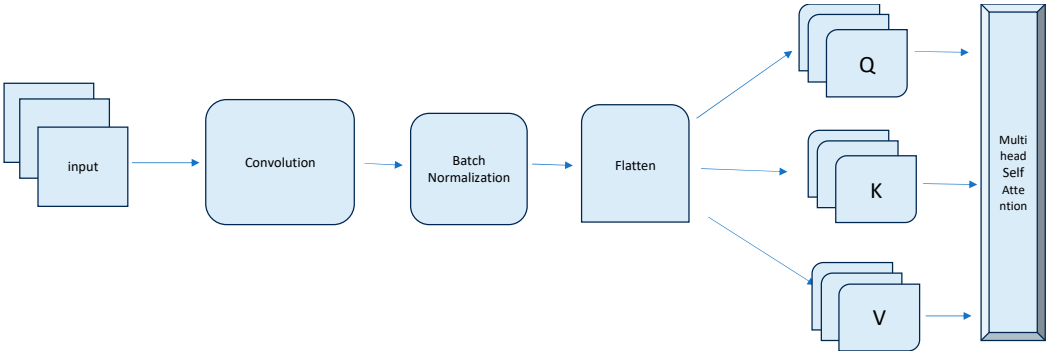

Figure 2. Illustration of token-embedding module [48].

Given an input video X of size T × H × W × C where T, H, W, and C represent the sequence length, width, height and channel, respectively. Given an input video X, the LGTransPPG can be represented as:

$$X_{TM} = CTE(X) \tag{1}$$

$$Q, K, V = CP(X_{TM}) \tag{2}$$

Then, the token map will be $X_{TM} \in R^{T \times H_{tm} \times W_{tm} \times C_{tm}}$ where T is the number of frames. To project a token-embedding tensor into the queries, keys and values, a convolution projection mechanism is used. Convolutional projection (CP) is used to encode query (Q), key (K) and value (V) feature maps. This Q/K/V are then fed into the LGTransPPG module to learn the long-range dependencies over frames. Finally, a feed-forward network (FFN) is employed to fuse the result.

$$X_{TM} \in \mathbb{R}^{T \times H_{tm} \times W_{tm} \times C_{tm}} \tag{3}$$

$$Y = H + X_{TM} \tag{4}$$

$$Z = (FFN(Norm(Y)) + Y) \tag{5}$$

The convolutional projection layer introduced here aims to enhance the local spatial context and, secondly, to increase efficiency by enabling downsampling of the K and V matrices. Essentially, the transformer block incorporating the convolutional projection represents an extension of the original transformer block. We use depth-wise separable convolutions, thereby creating the convolutional projection layer.

3.2. Local–Global Hybrid MHSA Transformer

Multi-head self-attention is an important component of the transformer architecture, which is designed to enhance the model's capacity to capture complex relationships and patterns within sequences of data. Unlike the conventional method that employs coarse image patches with a single scale, the MHSA module enables the extraction of fine-grained representations by partitioning the feature maps of query (Q), key (K), and value (V) into multi-scale patches. A combination method allows short-range attention within a frame to focus on local details and long-range attention over frames to capture dependencies. By incorporating local and global MHSA into a single transformer in a coarse-to-fine manner, our approach enhances the model's ability to learn fine-grained representations. This effectively combines the advantages of short-range and long-range spatiotemporal attention, resulting in improved performance. LGTransPPG consists of L blocks and each block consists of Local Multi-Head Self-Attention (L-MHSA) and Global Multi-Head Self-Attention (G-MHSA) to learn spatiotemporal feature representation [49].

The spatial and temporal modeling of video data are essential for the successful extraction of rPPG signals. Spatial processing involves the identification of the skin area within a frame, where the pulse signal is located, enabling us to find the relevant region for further analysis. On the other hand, temporal processing is related to the pulse signal over time. It is important to note that the spatial distribution of the pulse signal can change significantly at different time instants, and its intensity can fluctuate over time. Therefore, achieving accurate spatial and temporal modeling is critical for improving the reliability and accuracy of rPPG extraction methods. We perform L-MHSA within the non-overlapping window to obtain the local interactions [50,51]. The input is feature map $F \times H \times W$. Then, we flatten the token within the window (i, j) $X_{i,j} \in \mathbb{R}^{(fhw) \times d}$ The multi-head local attention of the k-th block is formulated as

$$Y_k = X_{i,j}^{k-1} + MSA(LN\ (X_{i,j}^{k-1})) \tag{6}$$

While our local attention mechanism is efficient in capturing local information, it may not fully capture global correlations across the frame sequence [50]. To address this limitation and capture long-range dependencies, we incorporate a multi-head global attention mechanism. This additional attention mechanism complements the local attention by enabling the model to capture and learn global information. By leveraging the multi-head global attention, our approach gains the capability to capture the broader context and relationships between frames, enhancing its ability to understand long-term dependencies within the video sequence. In our approach, we utilize each region to propagate global information to every query token. This is achieved through the formulation of the multi-head global attention mechanism [52]. By incorporating the MHSA, the model enables the

exchange in valuable information across different regions, facilitating an understanding of the entire video sequence.

To mitigate computational overhead, we introduce a down sampling approach by applying window-wise pooling to the input feature maps K and V. This involves dividing the feature maps into non-overlapping regions through convolutional operations and pooling. Each region represents a spatiotemporal abstraction of the feature map, and its purpose is to convey global contextual information to each query token. Formally, the multi-head global attention is formulated as an integration of these pooled regions with the query, key, and value inputs, allowing for the fusion of local and global information to facilitate more comprehensive and accurate heart rate prediction.

$$Q^k = W_Q^k LN\left(Y^K\right) \tag{7}$$

$$G - MHSA\left(Y^K\right) = Y^K + softmax\left(\frac{Q^k\left(K^K\right)^T}{\sqrt{D_h}}\right) \tag{8}$$

where $K^K = W_k^k\ LNPool\left(Y^K\right)$. Then, the whole local global attention complexity can be formulated as

$$C = FHWfhw + (FHW)2/fhw \tag{9}$$

In our method, we introduce a novel approach that builds upon existing techniques such as PhysFormer [53,54]. A tube tokenizer-based approach was employed [53,54] for measuring long-distance spatial–temporal rPPGs in facial videos, showcasing appealing outcomes. It highlights the advantages of attention mechanisms in capturing salient features and achieving accurate results. Instead of utilizing a spatiotemporal feed-forward network, we incorporate a spatio feed-forward network. Additionally, inspired by the divided space-time attention-based rPPGTr [55], we introduce a feature fusion module that combines local features and global dependencies to leverage their respective strengths. This is an encoder–decoder architecture that starts with detecting and sampling faces from the original video. The encoder has three encoding layers of different spatial scales, each consisting of a Convolution Block, Divided Space–Time Attention Block, and Feature Fusion Block. The fusion process involves concatenating the local and global features, which is followed by max pooling and average pooling operations. The resulting features are then activated using the sigmoid function and combined through addition. Finally, a 3D convolution with a kernel size of 1 is applied to adjust the channel dimension of the feature map.

4. Experiments

All experiments were conducted in Python 3.8 and Pytorch. We used Pearson loss [56] and FACL for training and also reproduced several existing models, namely PhysNet [25], TS-CAN [26], and DeepPhys [17] based on their open-source code. The Physnet and TS CAN were trained on Tensorflow 2.6. In the case of TS-CAN and DeepPhys, we used a lower resolution of 36 × 36 instead of 72 × 72 from the open-source code to trade-offs between computational efficiency and model performance. Furthermore, we made certain adjustments to the training parameters of PhysNet to achieve improved performance. It involved adjusting the learning rate, and experimenting with batch sizes, to enhance the performance of PhysNet. These adjustments were made to optimize the model's training process, improve its generalizability and align effectively with our dataset. To test their mobile CPU inference performance, we used Raspberry Pi CCPU: Cortex–A72 4 cores.

The videos in the MAHNOB-HCI dataset were downsampled to 30 fps for efficiency. In the training stage, we randomly sampled RGB face clips with size 160 × 128 × 128 as model inputs. For the task of heart rate estimation on the MAHNOB-HCI dataset, the low illumination and high compression videos in MAHNOB-HCI raised convergence difficulties. We fine-tuned the pre-trained model on the dataset for an additional 30 epochs.

Additionally, the frequency-aware correlation loss (FACL) is designed to achieve accurate heart rate prediction from the inspiration of the work [57] by emphasizing the alignment of frequency components. A successful approach to model training in previous works is the Mean Squared Error (MSE) loss function. However, MSE loss trains the model to capture both amplitude and frequency. In the context of heart rate estimation, we focus more on the frequency of the pulsatile signal, and we opt for a frequency-aware correlation loss in the frequency domain instead of the time domain. To assess the robustness of the loss function, we compute the selected loss on the data after each epoch and choose the model with the lowest loss during testing. Our evaluation of RMSE performance of different loss functions shows that Pearson's correlation coefficient (PCC) and frequency-aware correlation loss (FACL) consistently perform well. Based on these findings, we use FACL as the preferred loss function.

Unlike traditional correlation computations in the time domain, FACL performs the correlation computation in the frequency domain. The FACL loss is defined as follows:

$$L(y, \hat{y}) = \frac{-c \times Max|F(y) \times F(\hat{y})|}{\sigma(y) \times \sigma(\hat{y})} \tag{10}$$

where y represents the ground-truth heart rate signal and $\hat{y}$ is the predicted heart rate signal. F denotes the Fourier-transform operator, and σ(y) and σ($\hat{y}$) are the standard deviations of y and $\hat{y}$, respectively. In FACL, the Fourier-transformed representations of y and $\hat{y}$ is element-wise multiplied, highlighting the alignment of frequency components. By dividing this product by the respective standard deviations, we ensure that the loss is appropriately scaled.

The coefficient c represents the ratio of power within the frequency range associated with heart rate to the total power. It allows us to weigh the loss based on the importance of frequencies for heart rate prediction. By employing FACL, we encourage the model to align the frequency characteristics of the predicted heart rate with the ground truth. This approach effectively captures the underlying pulsatile nature of the signal and enhances the accuracy of heart rate prediction.

Datasets and Evaluation Metrics

We conducted experiments on three public datasets, COHFACE [56], MAHNOB-HCI [58], and UCLA rPPG [59]. To assess the accuracy of the rPPG pulse extraction algorithms, we employed evaluation metrics derived from recent publications [33,34,60]. These evaluation metrics mean absolute HR error (HR_{mae}), root mean squared HR error (HR_{rmse}), and Pearson's correlation coefficients (ρ). These metrics serve as benchmarks for evaluating the performance of the algorithms.

$$\text{Mean absolute Error } HR_{MAE} = \frac{1}{n}\sum_{i=1}^{n} HR_{est}^{i} - HR_{gnd}^{i} \tag{11}$$

$$\text{Root Mean Square Error } HR_{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(HR_{est}^{i} - HR_{gnd)}^{i}\right)^{2}} \tag{12}$$

$$\text{Pearson Correlation Coefficient } \rho = \frac{\sum_{i=1}^{n} \left|\left(X^{i} - \overline{X}\right)\left(Y^{i} - \overline{Y}\right)\right|}{\sqrt{\sum_{i=1}^{n}\left(X^{i} - \overline{X}\right)^{2}}\sqrt{\sum_{i=1}^{n}\left(Y^{i} - \overline{Y}\right)^{2}}} \tag{13}$$

The COHFACE [56] dataset consisted of a total of 160 videos including 40 healthy individuals, 28 males and 12 females. The recordings were made using a Logitech C525 camera with a frame rate of 20 and a resolution of 640 × 480. The dataset was recorded in various experimental scenarios, including both studio and natural lighting conditions. In the studio setting, natural light was avoided, and additional light from a spotlight was used to ensure proper illumination of the subject's face. On the other hand, the natural lighting scenario involved turning off all the lights in the room. To capture the face area

in its entirety, all participants in the dataset were instructed to remain still in front of the webcam for four sessions, each lasting approximately 1 min.

MAHNOB-HCI [58] was made available in 2011, comprising 527 videos. These videos featured 27 subjects, consisting of 15 males and 12 females. The frame rate for these recordings was 61, and the resolution was set at 780 × 580. During the recording process, the ECG signal values were simultaneously recorded to ensure synchronization

UCLA rPPG [59] is a real dataset comprising 104 subjects. However, due to faulty settings, we excluded the samples of two subjects, resulting in a final dataset of 102 subjects encompassing a diverse range of skin tones, ages, genders, ethnicities, and races. Each subject was recorded in five videos, each lasting approximately one minute, capturing a total of 1790 frames at a frame rate of 30 fps. Some of the example frames from these datasets can be seen in Figure 3.

Figure 3. Example frames from public datasets MAHNOB, COHFACE and UCLA rPPG.

5. Results and Discussion

A comparative analysis of ground truth HR and predicted HR have been performed. From Figure 4, we can see the strong correlation between the HR values that shows the accuracy and robustness of the suggested method. It exhibits minimal bias and low variance concerning the regression line. Figure 4 shows the graphical representation of the comparison between the ground truth rPPG and predicted rPPG, which are both normalized to keep the amplitudes in the same range.

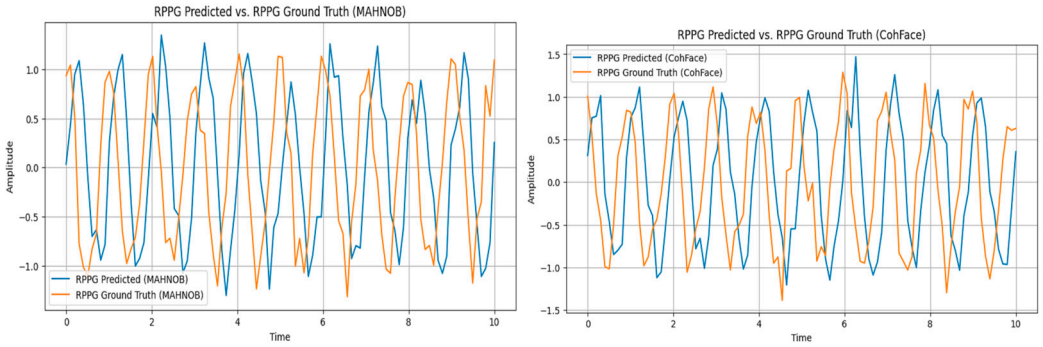

Figure 4. Graphical representation of rPPG predicted vs. ground truth on MAHNOB and COHFACE.

To further investigate the relationship between the predicted HR and the ground truth HR, we selected 300 samples from our test dataset for comparative analysis. The results of this analysis are shown in Figure 5, which includes Bland–Altman (BA) plots and regression plots. The solid line in the BA plot denotes the mean value, while the two dashed lines represent the interval. Meanwhile, a dashed line in the regression plot represents the standard line where the predicted HR matches precisely with the ground truth HR. From the figure, we can indicate that the LGTransPPG exhibits a smaller standard deviation, and its predicted HR aligns more closely with the ground truth HR. This outcome underscores the enhanced prediction accuracy and robustness of the recommended method.

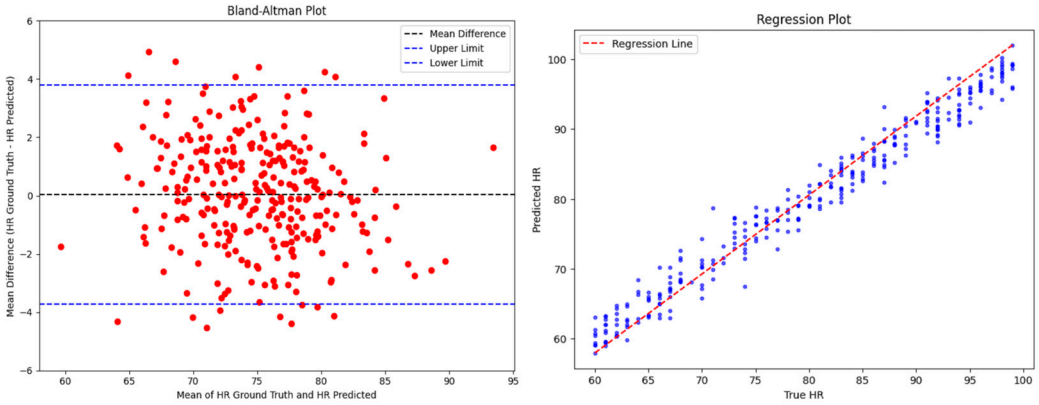

Figure 5. Bland–Altman analysis and regression plot between ground truth HR and predicted HR.

To compare our method with state-of-the-art methods, we perform an intra-dataset validation. We chose both traditional methods (ICA [13], CHROM [23], POS [61]) and CNN-based methods (TS-CAN [26], HR-CNN [62], PhysNet [25], DeepPhys [17]) and tested on three datasets MAHNOB, COHFACE and UCLA rPPG. The results are shown in Table 1. We applied a bandpass filter with a cut-off frequency of 0.7 Hz and 4 Hz to calculate the performance metrics. We used the average heart rate and calculated performance metrics for each video.

Table 1. Quantitative analysis results of the proposed method in comparison with three publicly available datasets, COHFACE, MAHNOB-HCI and UCLA-rPPG.

Methods		**MAHNOB**				**COHFACE**				**UCLA-rPPG**			
		SNR (dB)	MAE	RMSE	ρ	SNR (dB)	MAE	RMSE	ρ	SNR (dB)	MAE	RMSE	ρ
Signal Processing methods	ICA [13]	2.47	5.14	7.84	0.15	1.23	18.99	14.67	0.21		13.54	24.68	0.43
	CHROM [23]	1.74	4.32	9.64	0.23		16.05	23.54	0.17	1.25	19.44	11.23	0.22
	POS [61]	3.43	8.33	10.24	0.45	0.76	17.67	18.76	0.13	0.61	15.78	13.33	0.11
Learning Based methods	TS-CAN [26]	3.22	1.06	4.55	0.76	2.32	8.32	9.74	0.54	1.98	8.04	12.48	0.43
	HR-CNN [62]	2.91	2.42	6.26	0.44		10.04	19.38	0.37	5.51	9.81	14.98	0.58
	PhysNet [25]	8.67	1.43	2.45	0.67	**5.01**	2.46	9.61	0.87	3.79	10.27	2.45	0.43
	DeepPhys [17]	9.54	**0.87**	1.67	0.89	4.03	1.96	6.43	0.94	**6.38**	1.96	1.98	**0.97**
	Ours	**10.01**	**0.94**	**1.36**	**0.94**	**4.16**	**1.33**	**4.43**	**0.97**	**6.22**	**1.36**	**1.97**	**0.97**

Performance on MAHNOB: We split the 527 videos of the MAHNOB dataset and used 422 videos for training and 105 for testing. The performance was compared with those of other state-of-the-art methods and reported in Table 1. Starting with SNR, the suggested approach ("Ours") achieved the highest SNR value of 10.01 dB, indicating superior noise control and signal preservation. DeepPhys is closely followed with a competitive SNR of 9.54 dB, emphasizing its capability to minimize noise. ICA, HR-CNN, and TS-CAN showed lower SNR values, which denotes less effective noise reduction. On Moving to MAE, "Ours" maintains its lead by achieving the lowest MAE of 0.94, showcasing exceptional accuracy

in heart rate prediction. DeepPhys, with an MAE of 0.87, also demonstrates remarkable predictive precision. Conversely, ICA, HR-CNN, and TS-CAN still exhibit comparatively higher MAE values, indicating larger prediction errors.

Regarding RMSE, "Ours" remains the best, delivering the smallest RMSE value of 1.36, highlighting its effectiveness in predicting errors. DeepPhys closely follows with an RMSE of 1.67, showcasing robust performance. ICA, HR-CNN, and TS-CAN continue to display higher RMSE values, suggesting less accurate predictions. Taking into account the provided standard deviation values (ρ), it is evident that DeepPhys maintains its strong performance with a ρ of 0.89, indicating a relatively stable and consistent prediction. "Ours" also demonstrates noteworthy stability with a ρ of 0.94. In contrast, ICA, HR-CNN, and TS-CAN show higher ρ values, suggesting greater variability and less consistency in their predictions. In summary, while "Ours" excels in SNR, MAE, and RMSE metrics, showcasing impressive accuracy, precision, and error minimization, DeepPhys closely follows with remarkable stability and consistent performance. This comprehensive analysis highlights the strengths of these two methods in heart rate prediction from facial video data, considering both accuracy and stability.

Performance on COHFACE: We use five-fold subject independent protocol [27] and the results are reported in Table 1. Our method ("Ours") achieved an SNR of 4.16 dB, indicating a commendable level of noise control, which was closely followed by DeepPhys at 4.03 dB. While both methods displayed competitive noise reduction capabilities, others such as TS-CAN and HR-CNN exhibited comparatively lower SNR values. Comparing this dataset to MAHNOB, the COHFACE dataset appears to present a more challenging noise environment. Moving to MAE, "Ours" achieved an MAE of 1.33, showing accuracy on output predictions. However, TS-CAN and HR-CNN displayed higher MAE values, indicating high prediction errors. The COHFACE dataset introduced increased prediction challenges in comparison to MAHNOB.

In terms of RMSE, "Ours" maintained an RMSE of 4.43, and DeepPhys closely followed with an RMSE of 6.43, indicating robust performance. On the other side, TS-CAN and HR-CNN exhibited higher RMSE values, implying less accuracy in predictions. Considering the standard deviation (ρ) values, DeepPhys remained stable with a ρ value of 0.94, closely followed by "Ours" at 0.97, indicating consistent performance. However, the absence of standard deviation values in the initial analysis limited our ability to evaluate stability. These ρ values now provide critical insights into the reliability and consistency of each method's performance. In conclusion, our method ("Ours") and DeepPhys showcased strong performance in SNR, MAE, and RMSE metrics on the challenging COHFACE dataset, indicating their capacity for accurate and precise heart rate predictions. These findings underscore the adaptability of these methods across diverse datasets.

Performance on UCLA rPPG: To compare more objectively, we randomly split all the subjects; 80% of them were selected for training and 20% were used for testing. We report the results in Table 1. Starting with SNR, our method ("Ours") achieved an SNR of 6.22 dB, while DeepPhys closely followed with an SNR of 6.38 dB, signifying strong noise reduction capabilities. HR-CNN also exhibited a relatively high SNR value at 5.51 dB. These results highlight the effectiveness of these methods in mitigating noise in the UCLA rPPG dataset, enhancing the accuracy of heart rate predictions. Compared to previous datasets (MAHNOB and COHFACE), the UCLA rPPG dataset presents a different noise profile, necessitating tailored approaches.

Moving to MAE, "Ours" had an MAE of 1.36, and DeepPhys also demonstrated exceptional accuracy with an MAE of 1.96. However, HR-CNN and PhysNet exhibited higher MAE values. Regarding RMSE, "Ours" retained its lead with an RMSE of 1.97, and HR-CNN and PhysNet exhibited relatively higher RMSE values, suggesting less accurate predictions. Considering the standard deviation (ρ) values, both "Ours" and DeepPhys maintained stability with ρ values of 0.97, emphasizing consistent performance. This stability is a critical aspect of reliability in real-world applications. We also present error statistics for all videos in Figure 6, showing that the majority (80%) have predicted errors of

less than 5 bpm. These results underscore the effectiveness of our method in estimating heart rates.

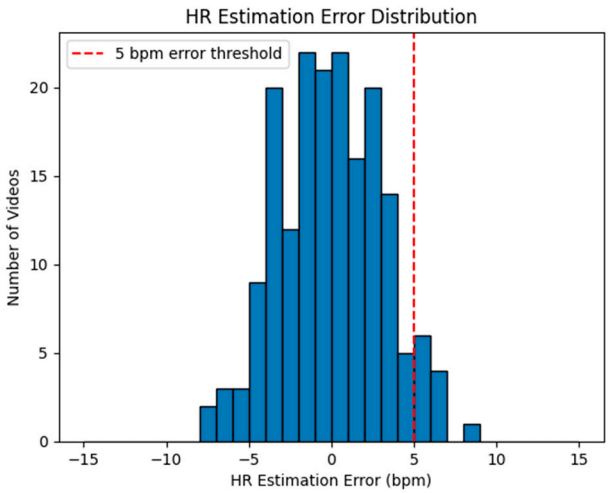

Figure 6. Heart rate estimation error distributions of the proposed approach.

From the obtained results, it is evident that when compared to signal processing methods, our proposed approach outperforms significantly across all three datasets. Learning-based approaches exhibit larger root mean square error (RMSE) and mean absolute error (MAE). This could be due to the overfitting of extracted features in learning-based methods. Our method achieves better results without following any additional steps. It is worth highlighting that even when compared to PhysNet, a benchmarked learning-based method, our approach performs comparably, showcasing its ability to learn directly from scratch. These results are achieved across three different publicly available datasets, indicating the generalizability of our method and its capability to learn subtle features. In summary, our method ("Ours") consistently showcased performance across all three datasets, and it shows their adaptability to the diverse dataset and their ability to provide accurate, precise, and stable heart rate predictions. These results show the efficacy of our method even in the absence of pre-processing steps.

We also performed a comparative assessment of the computational costs between the proposed method and PhysNet [25]. The average training time per sample was calculated. Our approach, LGTransPPG, demands 0.45 s per sample and outperformed PhysNet, which requires 0.63 s per sample. This demonstrates that our proposed approach is 28.57% faster, emphasizing its pathway to practical applications.

5.1. Cross Dataset Validation

For generalizability, we have conducted cross-dataset evaluation. Table 2a shows the results trained on COHFACE and tested on MAHNOB. We conducted cross-dataset validation following the five-fold cross-validation protocol [27]. The evaluation metrics include the mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient (ρ). The results highlight the effectiveness of our method for heart rate estimation across datasets and its potential for generalizability.

Table 2. Cross-dataset results in terms of MAE, RMSE and ρ. Datasets were trained on COHFACE and tested on MAHNOB, and cross-dataset results were tested on UCLA-rPPG.

Method	(a) MAHNOB			(b) UCLA rPPG		
	MAE	**RMSE**	ρ	**MAE**	**RMSE**	ρ
ICA [13]	7.92	5.98	0.72	8.28	9.28	0.55
POS [61]	3.42	6.72	0.85	5.43	3.4	0.75
CHROM [23]	5.54	6.01	0.87	3.69	5.41	0.77
DeepPhys [17]	3.82	2.78	0.98	1.42	1.8	0.84
Ours	**2.98**	**3.12**	**0.99**	**0.94**	**1.76**	**0.87**

Notably, the MAE values show the precision of our method's heart rate predictions, with a low value of 2.98. The low RMSE value of 3.12 further emphasizes the effectiveness of our approach in minimizing prediction errors. The high Pearson correlation coefficient (ρ) of 0.99 indicates a strong linear relationship between the predicted and ground truth heart rates across datasets. Comparatively, our method outperformed other heart rate prediction methods, including ICA, POS, CHROM, DeepPhys, and HR-CNN, in this cross-dataset scenario. These results highlight the superior generalizability of our method. In addition, we evaluate the model training on COHFACE, and we test on all subjects in UCLA-rPPG. All the results are reported in Table 2b. Our model achieved the lowest MAE and RMSE (0.94 and 1.76, respectively). In conclusion, our cross-dataset evaluation demonstrates that LGTransPPG excels in heart rate estimation even without pre-processing steps and shows its generalizability.

5.2. Ablation Study

In this section, we analyze the factors influencing the performance of our proposed heart rate prediction model, considering feature fusion. To validate the efficacy of feature fusion, we use two networks: LGTransPPG-local to capture local features and LGTransPPG-global to capture global features.

The ablation study results show the impact of feature fusion and model architecture on heart rate prediction performance. LGTransPPG-local focuses on local features, whereas LGTransPPG-global emphasizes global features. Comparing these models with our methodology, we observe significant differences in performance. LGTransPPG-global consistently outperforms LGTransPPG-local across all metrics. Our approach, which integrates both local and global features, is highlighted as the most effective approach. It achieves the lowest MAE, RMSE, and HRMSE and shows high accuracy. This result validates the combination of both local and global sources, enhancing the model's accuracy. The Pearson correlation coefficient (*PCC*) highlights the reliability of the suggested method. With the highest *PCC* value of 0.88, it shows a strong linear relationship between predicted and ground truth heart rates. To assess how well the model fits the data, we calculated the R-squared value, which is a measure of the proportion of the variance in one variable that is predictable from another variable, and it ranges from 0 to 1. A higher R-squared value indicates a stronger relationship between the variables. From our experiments, we achieved 0.77, and it shows a stronger linear relationship between the variables. This highlights the model's consistency. This study draws motivation and inspiration from the rPPGTr paper authored by [43,55].

This suggests that the integration of both local and global features in our proposed approach effectively captures relevant information for heart rate estimation. The superior performance of the proposed model highlights the importance of feature fusion in capturing comprehensive spatiotemporal features and represents its efficacy in improving heart rate estimation accuracy and robustness across the UCLA RPPG and COHFACE datasets. Results can be seen in Figure 7.

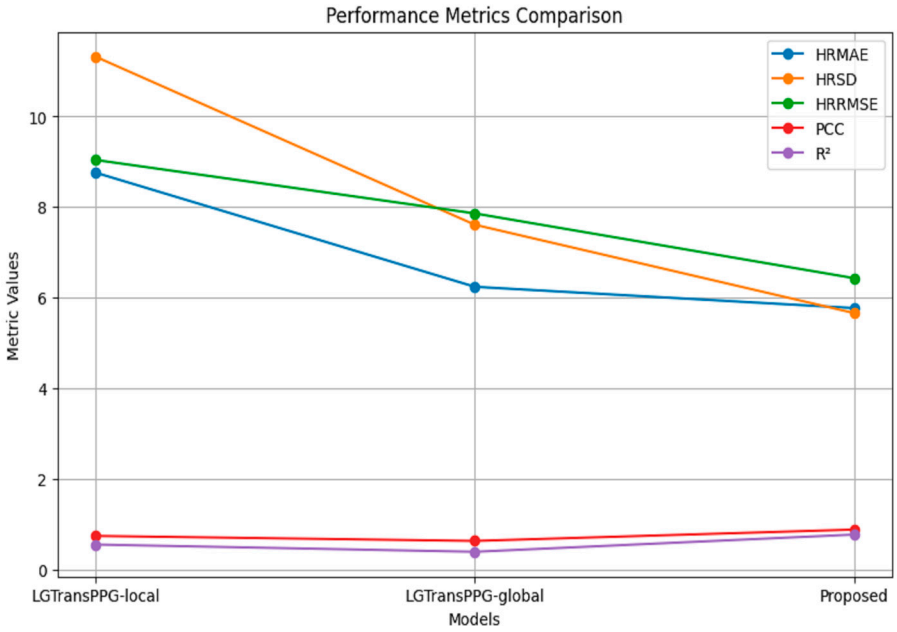

Figure 7. Ablation study on different feature fusion methods.

The ablation study conducted on the MAHNOB and COHFACE datasets explores the impact of various spatial scales on heart rate estimation performance. The first set of experiments focuses on spatial size invariance while keeping the channel size invariant. The results show the model's ability to handle different spatial scales, as shown in Figure 6 by HRMAE, HRSD, HRRMSE, and Pearson correlation coefficient (PCC) as an evaluation indicator in the rPPG signal.

When the channel size is increased, there is a performance improvement. To find the influence of the temporal dependency, we conducted an ablation that includes a comparison with rPPGTr [55], and the results are reported in Figure 8. Our method exhibits the smallest parameter when the spatial size decreases and the channel size increases. Experimental results affirm that the utilization of the local–global hybrid MHSA in this study contributes to enhanced prediction accuracy. The second set of experiments investigates the effect of decreasing spatial size while keeping the channel size either invariant or increased. The proposed approach with a smaller spatial size and increased channel size achieves superior performance compared to the spatial size invariant. The lower HRMAE, HRSD and HRRMSE values highlight the effectiveness of incorporating fine-grained information from smaller spatial scales. The higher PCC values indicate a stronger correlation between the predicted and ground truth rPPG signals.

These results show the significance of considering both spatial scales and channel sizes for accurate heart rate estimation. The study suggests that combining information from smaller spatial scales, with an increased channel size, helps the model capture finer details and increases its ability to estimate heart rates more accurately. This indicates the importance of multi-scale processing in rPPG-based heart rate estimation algorithms. By effectively adapting to different spatial scales and increased channel sizes, the proposed method demonstrates its efficacy in improving heart rate estimation performance on the MAHNOB and COHFACE datasets.

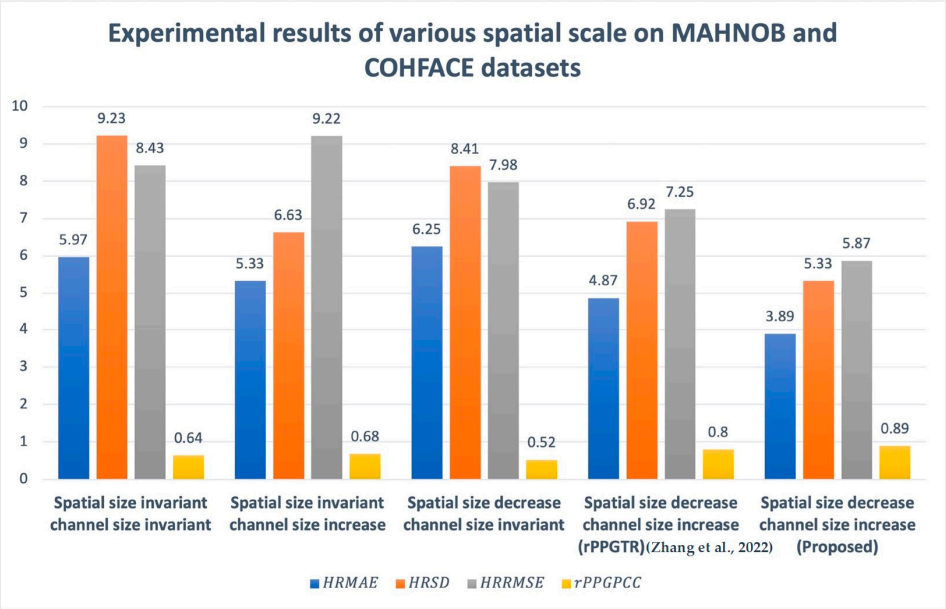

Figure 8. Ablation study spatial scale on MAHNOB and COHFACE [55].

6. Conclusions

In conclusion, this paper presents a comprehensive investigation into heart rate prediction from face videos using a novel transformer-based framework. The proposed approach incorporates token embedding, convolution, and convolution projection as initial processing steps, which is followed by a hybrid attention mechanism that combines local and global information. The results obtained from extensive experiments on multiple publicly available datasets demonstrate the efficacy of the recommended technique. With a high Pearson correlation coefficient (PCC) value of 0.88, our method demonstrates a robust and accurate prediction, which is further validated by Bland–Altman plots and regression plots. We conduct comprehensive intra and cross-dataset validation experiments to evaluate the performance of the proposed approach. LGTransPPG delivers superior performance across multiple datasets, achieving remarkable signal-to-noise ratios (SNRs) of 10.01 dB on MAHNOB and 6.22 dB on UCLA rPPG, demonstrating its robustness in noise control capabilities.

Looking at the results, we can see that our method performs well compared to the best existing methods. What is notable is that our method does not require any preprocessing steps; it learns directly from the raw data while considering both spatial and temporal aspects. Despite this, it maintains robustness even when compared to methods that use preprocessing steps. This is a significant advantage for real-world applications. However, we observed a notable decrease in accuracy when applied to individuals with darker skin tones. We conducted additional testing using self-collected images of individuals with darker skin, which again highlighted this challenge. To ensure diverse populations' applicability, we realize the need for improvement in this specific area. We are actively addressing this limitation to enhance generalizability, considering the lighter skin tone bias in existing datasets.

The accuracy, computational efficiency, and adaptability of the recommended method make it a promising solution for real-time heart rate estimation tasks. Future work could focus on exploring different datasets and further optimizing the proposed framework to enhance its performance.

Author Contributions: Conceptualization, S.P.; Methodology, J.A.; Formal Analysis, D.D.; Supervision, D.J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset available for public use is used in this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Poh, M.-Z.; McDuff, D.J.; Picard, R.W. Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Trans. Biomed. Eng.* **2010**, *58*, 7–11. [CrossRef] [PubMed]
2. Yu, X.; Laurentius, T.; Bollheimer, C.; Leonhardt, S.; Antink, C.H. Noncontact Monitoring of Heart Rate and Heart Rate Variability in Geriatric Patients Using Photoplethysmography Imaging. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 1781–1792. [CrossRef] [PubMed]
3. Sasangohar, F.; Davis, E.; Kash, B.A.; Shah, S.R. Remote patient monitoring and telemedicine in neonatal and pediatric settings: Scoping literature review. *J. Med. Internet Res.* **2018**, *20*, e295. [CrossRef] [PubMed]
4. Hebbar, S.; Sato, T. Motion Robust Remote Photoplethysmography via Frequency Domain Motion Artifact Reduction. In Proceedings of the 2021 IEEE Biomedical Circuits and Systems Conference (BioCAS), Berlin, Germany, 7–9 October 2021; pp. 1–4. [CrossRef]
5. Sinhal, R.; Singh, K.; Raghuwanshi, M.M. An Overview of Remote Photoplethysmography Methods for Vital Sign Monitoring. *Adv. Intell. Syst. Comput.* **2020**, *992*, 21–31. [CrossRef]
6. Chang, M.; Hung, C.-C.; Zhao, C.; Lin, C.-L.; Hsu, B.-Y. Learning based Remote Photoplethysmography for Physiological Signal Feedback Control in Fitness Training. In Proceedings of the 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), Kristiansand, Norway, 9–13 November 2020; pp. 1663–1668. [CrossRef]
7. Zaunseder, S.; Trumpp, A.; Wedekind, D.; Malberg, H. Cardiovascular assessment by imaging photoplethysmography-a review. *Biomed. Tech* **2018**, *63*, 529–535. [CrossRef]
8. Huang, P.W.; Wu, B.J.; Wu, B.F. A Heart Rate Monitoring Framework for Real-World Drivers Using Remote Photoplethysmography. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 1397–1408. [CrossRef]
9. Wu, B.F.; Chu, Y.W.; Huang, P.W.; Chung, M.L. Neural Network Based Luminance Variation Resistant Remote-Photoplethysmography for Driver's Heart Rate Monitoring. *IEEE Access* **2019**, *7*, 57210–57225. [CrossRef]
10. Kuncoro, C.B.D.; Luo, W.-J.; Kuan, Y.-D. Wireless Photoplethysmography Sensor for Continuous Blood Pressure Bio signal Shape Acquisition. *J. Sens.* **2020**, *2020*, 7192015. [CrossRef]
11. Hilmisson, H.; Berman, S.; Magnusdottir, S. Sleep apnea diagnosis in children using software-generated apnea-hypopnea index (AHI) derived from data recorded with a single photoplethysmogram sensor (PPG): Results from the Childhood Adenotonsillectomy Study (CHAT) based on cardiopulmonary coupling analysis. *Sleep Breath.* **2020**, *24*, 1739–1749. [CrossRef]
12. Wilson, N.; Guragain, B.; Verma, A.; Archer, L.; Tavakolian, K. Blending Human and Machine: Feasibility of Measuring Fatigue through the Aviation Headset. *Hum. Factors* **2020**, *62*, 553–564. [CrossRef]
13. Verkruysse, W.; Svaasand, L.O.; Nelson, J.S. Remote plethysmographic imaging using ambient light. *Opt. Express* **2008**, *16*, 21434–21445. [CrossRef] [PubMed]
14. McDuff, D. Camera Measurement of Physiological Vital Signs. *ACM Comput. Surv.* **2023**, *55*, 176. [CrossRef]
15. Premkumar, S.; Hemanth, D.J. Intelligent Remote Photoplethysmography-Based Methods for Heart Rate Estimation from Face Videos: A Survey. *Informatics* **2022**, *9*, 57. [CrossRef]
16. Malasinghe, L.; Katsigiannis, S.; Dahal, K.; Ramzan, N. A comparative study of common steps in video-based remote heart rate detection methods. *Expert Syst. Appl.* **2022**, *207*, 117867. [CrossRef]
17. Chen, W.; McDuff, D. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In Proceedings of the 15th European Conference, Munich, Germany, 8–14 September 2018. Proceedings, Part II. [CrossRef]
18. Niu, X.; Yu, Z.; Han, H.; Li, X.; Shan, S.; Zhao, G. Video-based remote physiological measurement via cross-verified feature disentangling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part II 16*; Springer International Publishing: Cham, Switzerland, 2020; pp. 295–310.
19. Lu, H.; Han, H.; Zhou, S.K. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12404–12413.
20. Lewandowska, M.; Rumiński, J.; Kocejko, T.; Nowak, J. Measuring pulse rate with a webcam—A non-contact method for evaluating cardiac activity. In Proceedings of the 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), Szczecin, Poland, 18–21 September 2011; pp. 405–410.
21. Zhang, B.; Li, H.; Xu, L.; Qi, L.; Yao, Y.; Greenwald, S.E. Noncontact heart rate measurement using a webcam, based on joint blind source separation and a skin reflection model: For a wide range of imaging conditions. *J. Sens.* **2021**, *2021*, 9995871. [CrossRef]
22. Poh, M.-Z.; McDuff, D.J.; Picard, R.W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **2010**, *18*, 10762–10774. [CrossRef]
23. de Haan, G.; Jeanne, V. Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2878–2886. [CrossRef]

24. Wang, W.; Stuijk, S.; de Haan, G. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE Trans. Biomed. Eng.* **2016**, *3*, 1974–1984. [CrossRef]
25. Yu, Z.; Li, X.; Zhao, G. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv* **2019**, arXiv:1905.02419.
26. Liu, X.; Fromm, J.; Patel, S.; McDuff, D. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19400–19411.
27. Niu, X.; Shan, S.; Han, H.; Chen, X. RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. *IEEE Trans. Image Process.* **2019**, *29*, 2409–2423. [CrossRef] [PubMed]
28. Yu, Z.; Peng, W.; Li, X.; Hong, X.; Zhao, G. 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, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 151–160.
29. Qiu, Y.; Liu, Y.; Arteaga-Falconi, J.; Dong, H.; El Saddik, A. EVM-CNN: Real-time contactless heart rate estimation from facial video. *IEEE Trans. Multimed.* **2018**, *21*, 1778–1787. [CrossRef]
30. Hu, M.; Qian, F.; Guo, D.; Wang, X.; He, L.; Ren, F. ETA-rPPGNet: Effective time-domain attention network for remote heart rate measurement. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12. [CrossRef]
31. Niu, X.; Han, H.; Shan, S.; Chen, X. Synrhythm: Learning a deep heart rate estimator from general to specific. In *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018*; IEEE: New York, NY, USA; pp. 3580–3585.
32. Song, R.; Chen, H.; Cheng, J.; Li, C.; Liu, Y.; Chen, X. PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 1373–1384. [CrossRef]
33. Hsu, G.S.; Ambikapathi, A.; Chen, M.S. Deep learning with time-frequency representation for pulse estimation from facial videos. In *Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, 1–4 October 2017*; IEEE: New York, NY, USA; pp. 383–389.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2017; pp. 5998–6008.
35. Minissi, M.E.; Chicchi Giglioli, I.A.; Mantovani, F.; Alcaniz Raya, M. Assessment of the autism spectrum disorder based on machine learning and social visual attention: A systematic review. *J. Autism Dev. Disord.* **2022**, *52*, 2187–2202. [CrossRef]
36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
37. Liu, L.; Hamilton, W.; Long, G.; Jiang, J.; Larochelle, H. A universal representation transformer layer for few-shot image classification. *arXiv* **2020**, arXiv:2006.11702.
38. Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; Xia, H. End-to-end video instance segmentation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 8741–8750.
39. Gao, H.; Wu, X.; Shi, C.; Gao, Q.; Geng, J. A LSTM-based realtime signal quality assessment for photoplethysmogram and remote photoplethysmogram. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 3831–3840.
40. Lee, E.; Chen, E.; Lee, C.Y. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Part XXVII 16*; Springer International Publishing: Cham, Switzerland, 2020; pp. 392–409.
41. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
42. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3202–3211.
43. Shi, C.; Zhao, S.; Zhang, K.; Wang, Y.; Liang, L. Face-based age estimation using improved Swin Transformer with attention-based convolution. *Front. Neurosci.* **2023**, *17*, 1136934. [CrossRef]
44. Li, L.; Lu, Z.; Watzel, T.; Kürzinger, L.; Rigoll, G. Light-weight self-attention augmented generative adversarial networks for speech enhancement. *Electronics* **2021**, *10*, 1586. [CrossRef]
45. McDuff, D.J.; Wander, M.; Liu, X.; Hill, B.L.; Hernández, J.; Lester, J.; Baltrušaitis, T. SCAMPS: Synthetics for Camera Measurement of Physiological Signals. *arXiv* **2022**, arXiv:2206.04197.
46. Selva, J.; Johansen, A.S.; Escalera, S.; Nasrollahi, K.; Moeslund, T.B.; Clap'es, A. Video Transformers: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12922–12943. [CrossRef] [PubMed]
47. Hassanin, M.; Anwar, S.; Radwan, I.; Khan, F.S.; Mian, A.S. Visual Attention Methods in Deep Learning: An In-Depth Survey. *arXiv* **2022**, arXiv:2204.07756.
48. Wu, H.; Xiao, B.; Codella, N.C.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. CvT: Introducing Convolutions to Vision Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.

49. Liang, Y.; Zhou, P.; Zimmermann, R.; Yan, S. DualFormer: Local-Global Stratified Transformer for Efficient Video Recognition. *arXiv* **2021**, arXiv:2112.04674.
50. Ma, F.; Sun, B.; Li, S. Logo-Former: Local-Global Spatio-Temporal Transformer for Dynamic Facial Expression Recognition. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
51. Ming, Z.; Yu, Z.; Al-Ghadi, M.; Visani, M.; Luqman, M.M.; Burie, J.-C. Vitranspad: Video Transformer Using Convolution and Self-Attention for Face Presentation Attack Detection. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 4248–4252. [CrossRef]
52. Aksan, E.; Kaufmann, M.; Cao, P.; Hilliges, O. A spatio-temporal transformer for 3d human motion prediction. In *Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021*; IEEE: New York, NY, USA; pp. 565–574.
53. Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Torr, P.; Zhao, G. PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 4176–4186. [CrossRef]
54. Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Cui, Y.; Zhang, J.; Torr, P.; Zhao, G. PhysFormer++: Facial Video-Based Physiological Measurement with SlowFast Temporal Difference Transformer. *Int. J. Comput. Vis.* **2023**, *131*, 1307–1330. [CrossRef]
55. Zhang, X.; Yang, C.; Yin, R.; Meng, L. An End-to-End Heart Rate Estimation Scheme Using Divided Space-Time Attention. *Neural Process. Lett.* **2022**, *55*, 2661–2685. [CrossRef]
56. Heusch, G.; Anjos, A.; Marcel, S. A reproducible study on remote heart rate measurement. *arXiv* **2017**, arXiv:1709.00962.
57. Revanur, A.; Dasari, A.; Tucker, C.S.; Jeni, L.A. Instantaneous physiological estimation using video transformers. In *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*; Springer International Publishing: Cham, Switzerland, 2022; pp. 307–319.
58. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Trans. Affect. Comput.* **2012**, *3*, 42–55. [CrossRef]
59. Wang, Z.; Ba, Y.; Chari, P.; Bozkurt, O.D.; Brown, G.; Patwa, P.; Vaddi, N.; Jalilian, L.; Kadambi, A. Synthetic generation of face videos with plethysmograph physiology. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20587–20596.
60. Zheng, K.; Ci, K.; Li, H.; Shao, L.; Sun, G.; Liu, J.; Cui, J. Heart rate prediction from facial video with masks using eye location and corrected by convolutional neural networks. *Biomed. Signal Process. Control.* **2022**, *75*, 103609. [CrossRef]
61. Wang, W.; Den Brinker, A.C.; Stuijk, S.; De Haan, G. Algorithmic principles of remote PPG. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 1479–1491. [CrossRef]
62. Wang, Z.-K.; Kao, Y.; Hsu, C.-T. Vision-Based Heart Rate Estimation via a Two-Stream CNN. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3327–3331. [CrossRef]