electronics

MDPI

Article

Research on Heart Rate Detection from Facial Videos Based on an Attention Mechanism 3D Convolutional Neural Network

Xiujuan Sun, Ying Su, Xiankai Hou, Xiaolan Yuan, Hongxue Li and Chuanjiang Wang *

College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China; cxjsun@sdust.edu.cn (X.S.); 202283080053@sdust.edu.cn (Y.S.); 202283080040@sdust.edu.cn (X.H.); 202283080028@sdust.edu.cn (X.Y.); 202383080128@sdust.edu.cn (H.L.)
* Correspondence: cxjwang@sdust.edu.cn; Tel.: +86-135-5303-9802

Abstract: Remote photoplethysmography (rPPG) has attracted growing attention due to its non-contact nature. However, existing non-contact heart rate detection methods are often affected by noise from motion artifacts and changes in lighting, which can lead to a decrease in detection accuracy. To solve this problem, this paper initially employs manual extraction to precisely define the facial Region of Interest (ROI), expanding the facial area while avoiding rigid regions such as the eyes and mouth to minimize the impact of motion artifacts. Additionally, during the training phase, illumination normalization is employed on video frames with uneven lighting to mitigate noise caused by lighting fluctuations. Finally, this paper introduces a 3D convolutional neural network (CNN) method incorporating an attention mechanism for heart rate detection from facial videos. We optimize the traditional 3D-CNN to capture global features in spatiotemporal data more effectively. The SimAM attention mechanism is introduced to enable the model to precisely focus on and enhance facial ROI feature representations. Following the extraction of rPPG signals, a heart rate estimation network using a bidirectional long short-term memory (BiLSTM) model is employed to derive the heart rate from the signals. The method introduced here is experimentally validated on two publicly available datasets, UBFC-rPPG and PURE. The mean absolute errors were 0.24 bpm and 0.65 bpm, the root mean square errors were 0.63 bpm and 1.30 bpm, and the Pearson correlation coefficients reached 0.99, confirming the method's reliability. Comparisons of predicted signals with ground truth signals further validated its accuracy.

Keywords: BiLSTM; attention mechanism; convolutional neural network; facial video; rPPG

Received: 4 December 2024
Revised: 8 January 2025
Accepted: 9 January 2025
Published: 10 January 2025

Citation: Sun, X.; Su, Y.; Hou, X.; Yuan, X.; Li, H.; Wang, C. Research on Heart Rate Detection from Facial Videos Based on an Attention Mechanism 3D Convolutional Neural Network. *Electronics* **2025**, *14*, 269. https://doi.org/10.3390/electronics14020269

1. Introduction

With the improvement of people's living standards, various problems caused by unhealthy lifestyles have become increasingly prominent, and cardiovascular disease has become one of the main factors endangering human health and life [1]. The accurate real-time monitoring of heart rate plays a crucial role in assessing personal health status. Traditional contact heart rate detection equipment needs to use specific sensors to directly contact the human body to obtain physiological signals. Long-term contact measurement may cause discomfort to the human body, especially for infants and patients with skin allergies. In contrast, non-contact heart rate measurement [2] has gained widespread attention from academia and industry because of its low cost and lack of direct contact with the human body.

As shown in Figure 1, the skin reflection model describes how the light source illuminates the skin, and the reflected light captured by the camera consists of both specular

and diffuse reflections. The diffuse reflection, caused by the capillaries, carries pulse information. The rPPG technology aims to extract the diffuse reflection component containing the pulse signal. The resulting signal, extracted by rPPG, is known as the rPPG signal. The face video heart rate detection based on rPPG [3–5] refers to the facial video captured by the camera and analyzing the slight color variations on the skin surface, which result from changes in blood volume within the blood vessels, thereby extracting the rPPG signal that synchronizes with the cardiac cycle [6]. This method does not require the participant to wear additional sensors, effectively reducing sensory discomfort. Furthermore, due to its simple equipment, low cost, wide adaptability, and high accuracy, only a signal camera can easily detect heart rate, making it stand out among various non-contact heart rate detection methods. It has become an important method favored by many researchers.

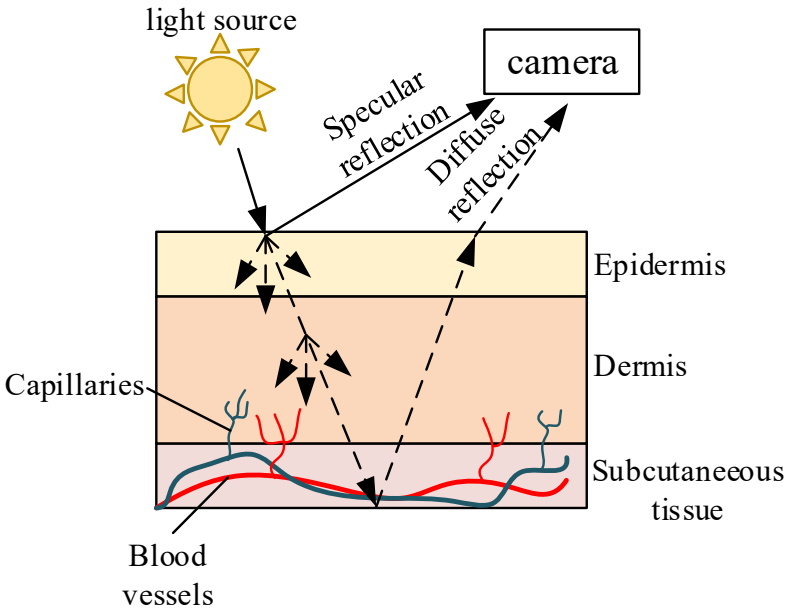


Figure 1. Skin reflection model.

The rPPG method was first introduced by Verkruysse [7] in 2008, who selected the forehead region of the face as the ROI and averaged the RGB color channels within this region to obtain the rPPG signal. The heart rate was then derived through Fourier transformation. This study provided the basis for subsequent research on non-contact heart rate detection. Poh et al. [8] later incorporated independent component analysis (ICA) [9] into facial video-based heart rate detection, using it to extract the rPPG signal from the mixed signal collected by the camera. In recent years, based on the powerful feature learning and representation capabilities of deep learning, a growing number of researchers have applied deep learning technology to non-contact heart rate detection. Hsu [10] proposed converting the rPPG signal extracted from facial videos into a time-frequency representation and using the VGG15 [11] network to create a mapping between the time-frequency representation and the heart rate. Špetlík et al. [12] developed a two-step HR-CNN model, in which the Extractor extracts the rPPG signal from the video, and the Estimator predicts the heart rate from the extracted signal. McDuff designed an end-to-end DeepPhys [13] model and introduced an attention mechanism to improve the accuracy of the extracted signals. Song et al. [14] introduced a new model architecture, PulseGAN, which takes rPPG signals detected by the traditional CHROM method as input. The model uses a generative adversarial network (GAN) to produce precise rPPG signals, thereby achieving denoising.

Currently, methods for heart rate detection from facial videos are not yet fully matured. Most studies are still affected by noise from motion artifacts and lighting fluctuations, which

lead to the low accuracy in heart rate measurements. The quality of rPPG signal obtained from different facial ROI is also different, so it is still worth exploring how to select an appropriate ROI for the face and remove the interference of noise [15].

To resolve these problems, this study introduces a facial video heart rate detection algorithm based on a 3D-CNN with an attention mechanism. Inspired by CVD [16], we manually select the ROI on the face, aiming to maximize the facial area while avoiding rigid regions such as the eyes, mouth, and eyebrows, which tend to undergo significant displacement during facial expression changes and are prone to motion artifacts. By excluding these areas, we can effectively reduce the interference of motion artifacts. Due to the effect of lighting variations, differences in the emission characteristics of different facial regions under different lighting conditions may cause signal distortion and substantial errors. To mitigate this, we perform lighting normalization on videos with uneven illumination during training. Specifically, we apply adaptive histogram equalization to compensate for the brightness channel in the image while keeping other chrominance channels unchanged. This approach allows for the adaptive adjustment of the lighting level in the image while preserving facial details more effectively. Finally, based on the improvement of the existing 3D-CNN, the SimAM attention mechanism is introduced. It is a lightweight attention mechanism that can be combined with 3D-CNN to improve the feature representation of models in complex backgrounds, without significantly increasing the model's parameter count and computational complexity. After the initial rPPG signal is extracted, it is sent into the BiLSTM model, and the heart rate value is further extracted from the rPPG signal. The BiLSTM model analyzes the signal bidirection in time series. More comprehensive temporal correlations can be captured to improve the accuracy of heart rate predictions.

The contributions of this work are as follows:

1. Manually selecting the facial ROI to enlarge the facial area while minimizing the regions prone to motion artifacts, thus avoiding their interference.
2. Performing normalization on videos with uneven illumination, allowing for the adaptive adjustment of lighting variations in the video and preserving facial details to reduce the impact of lighting changes.
3. Introducing the lightweight attention mechanism SimAM, based on the 3D-CNN, aiming to reduce computational complexity while accurately extracting rPPG signals and minimizing the influence of noise on signal extraction.
4. Incorporating BiLSTM to extract heart rate information from rPPG signals through bidirectional processing, long-term and short-term dependency modeling, and temporal feature learning, improving the accuracy and generalization ability of heart rate estimation.

2. Algorithm Description

2.1. General Block Diagram

Figure 2 depicts the overall block diagram of the proposed model. It is primarily composed of three stages: face detection and ROI extraction, rPPG signal extraction and denoising, and heart rate estimation. In the first stage, Multi-Task Cascaded Convolutional Networks (MTCNN) [17] are used to quickly screen possible face regions from face videos, and then Dlib [18] is used to further accurately locate faces and obtain key information. The appropriate facial ROI region based on the obtained key points is manually cropped out. In the second stage, the cropped facial ROI was sent into the 3D-SimAM network model to extract the initial rPPG signal. Since the rPPG signal extracted initially contains significant noise, in order to enhance the signal quality, the initial signal is denoised by a bandpass filter, which cuts off at a frequency of [0.6, 4] Hz. In the third stage, the optimized rPPG signal is input into the heart rate estimation network to predict the heart rate value.

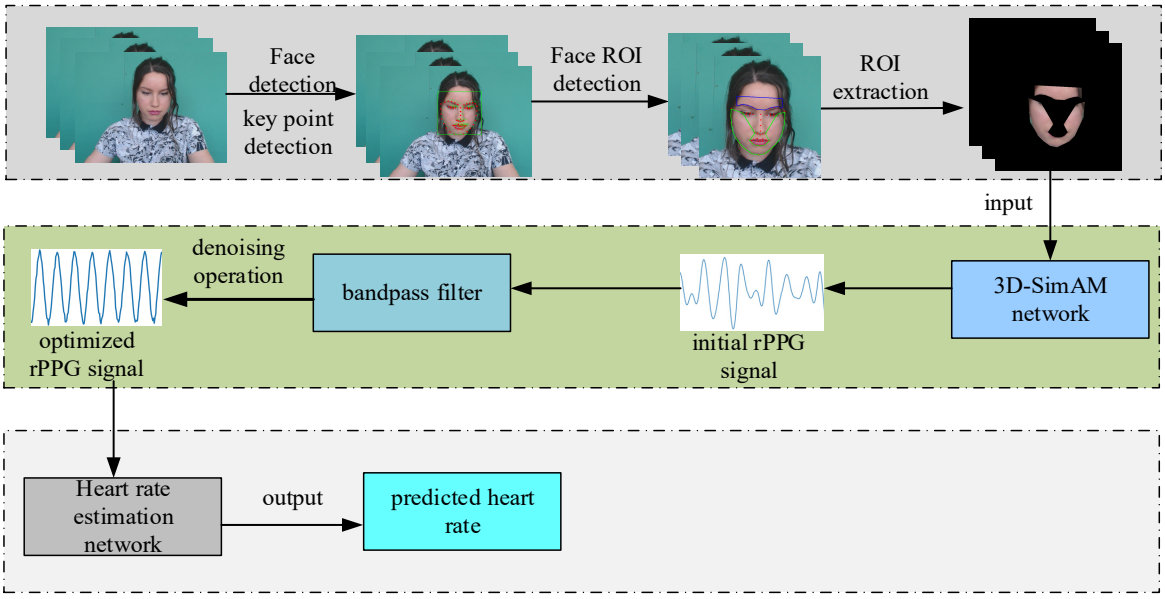


Figure 2. General block diagram.

2.2. SimAM Branches

SimAM, a lightweight attention mechanism [19], aims to enhance the performance of convolution neural networks. Unlike the existing channel dimension and spatial dimension, SimAM is able to infer 3D attention weights for each neuron in the feature map without introducing additional parameters, the basic idea being to use a simple module to capture important information in the feature map. By introducing an adaptive weighting mechanism, efficient feature enhancement is realized.

The SimAM architecture is depicted in Figure 3c. It modifies the existing attention mechanism, which the channel attention mechanism (Figure 3a) and the spatial attention mechanism (Figure 3b) cannot be applied to simultaneously. Some weights will be generated from the input feature map, and each neuron will generate its own attention weights according to its similarity or difference with the surrounding neurons, so as to identify the important neurons. The generated attention weights are then dimensionally matched with the features of the input feature map in an extended way, which enables each neuron to obtain a corresponding importance weight. Then, the generated weights are used on the original feature map and the values of each neuron are readjusted to highlight key features and suppress irrelevant information. The operation of the attention mechanism mainly depends on the choice of the defined energy function, which reduces the excessive adjustment of the network structure.

For the efficient application of the attention mechanism, evaluating the importance of each neuron is crucial. Drawing on neuroscience theories, neurons that carry more information usually demonstrate firing patterns distinct from neighboring neurons. Neurons exhibiting spatial inhibition effects should be given higher weight. Based on this characteristic, a corresponding energy function is defined:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{1}$$

where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$, $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2$. Formula (1) shows that the difference between neuron t and its surrounding neurons is related to e_t^*. Therefore, the importance of each neuron can be obtained by $\frac{1}{e_t^*}$. Unlike other neural networks, the SimAM attention mecha-

nism operates on individual neurons, using scaling operators instead of addition for feature refinement. The stages of the entire refinement of the module are:

$$\tilde{X} = sigmoid(\frac{1}{E}) \odot X \tag{2}$$

where E groups all e_t^* in channel and spatial dimensions, the sigmoid function [20] is applied to limit values that are too large in E.

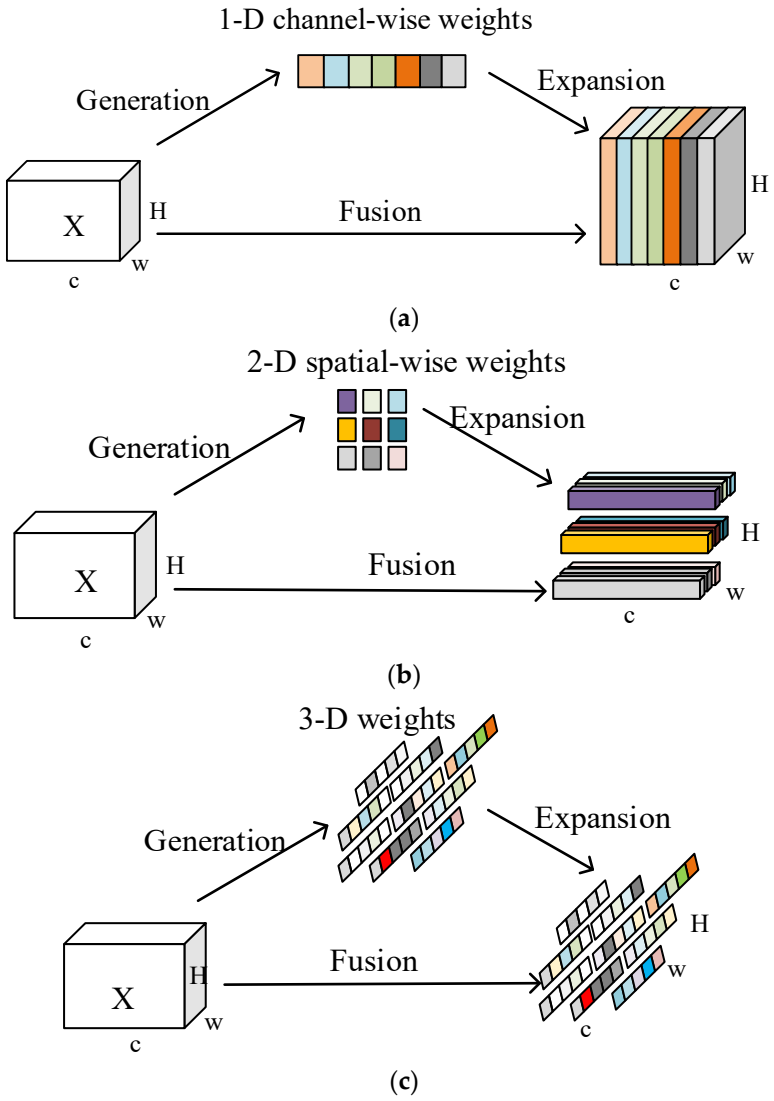


Figure 3. Comparison of different attention mechanisms. (**a**) Channel attention structure diagram; (**b**) spatial attention structure diagram; (**c**) SimAM attention mechanism structure diagram.

2.3. 3D-SimAM Structure

The traditional 2D-CNN takes the image as the input, and can only perform the convolution calculation on a single frame image, which cannot fully utilize the inter-frame information in the time dimension. In an effort to address this issue, this paper selects 3D-CNN as the backbone network, which is good at capturing global features from spatiotemporal data and can extract features in both the temporal and spatial dimensions simultaneously. To enhance the model's performance in complex backgrounds while focusing on areas directly related to heart rate changes, the SimAM attention mechanism was combined with 3D-CNN, which allows the model to better process global and local features, improving the accuracy of heart rate detection.

Figure 4 illustrates the structure of 3D-SimAM. The input is a video segment $V \in R^{3\times T\times H\times W}$ containing T frames of the facial ROI, where T is the number of frames. The initial layer uses 3D convolution kernels of size $1 \times 5 \times 5$ with a stride of 1, primarily responsible for extracting low-level spatiotemporal features from the video frames. This layer extracts features in both spatial and temporal dimensions to ensure that subsequent layers can extract deeper features based on this foundation. Next, the feature map passes through the first recurrent layer, where average pooling is applied to reduce the spatial dimensions of the feature map. In the recurrent layer, two 3D convolution layers, both with kernel sizes of $3 \times 3 \times 3$, are used to reduce the computational complexity while retaining the average feature information. These layers further extract more complex features, followed by Batch Normalization (BN) and ELU activation functions to alleviate the vanishing gradient problem, accelerate training and enhance the network's representational ability.

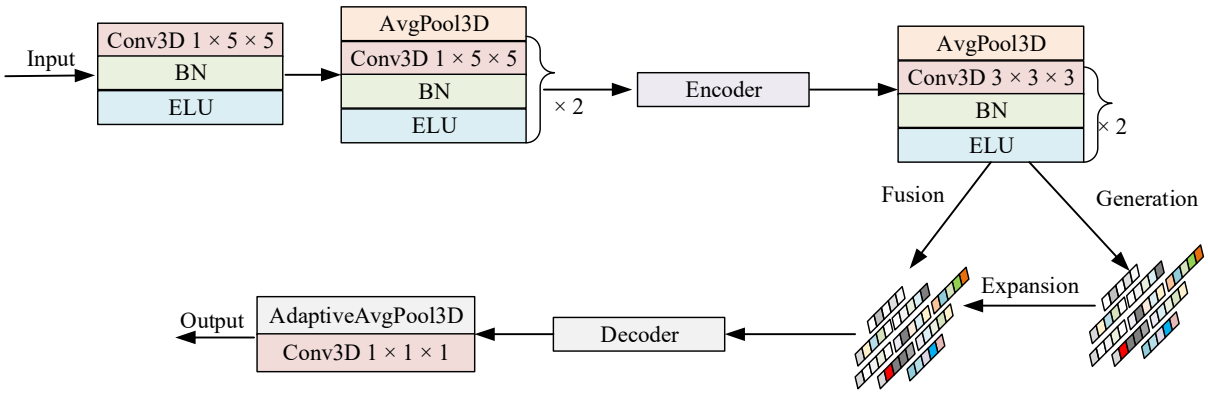


Figure 4. 3D-SimAM structure.

2.4. Heart Rate Estimation Module

Common Recurrent Neural Networks (RNN) encounter problems such as gradient vanishing and gradient explosion when dealing with long time series. To solve this problem, the BiLSTM model [21] is introduced, which is a variant of a Long Short-Term Memory Network (LSTM). BiLSTM is unique in its ability to capture time-dependent information from both the front and back of the data simultaneously, significantly improving the ability to extract sequence features. With this bidirectional approach, BiLSTM can more fully understand the dynamics of sequence data when analyzing them. The specific heart rate estimation network is shown in the Figure 5. The optimized rPPG signal is input into the BiLSTM model, and the output of the BiLSTM is mapped to the heart rate estimate through a fully connected layer.

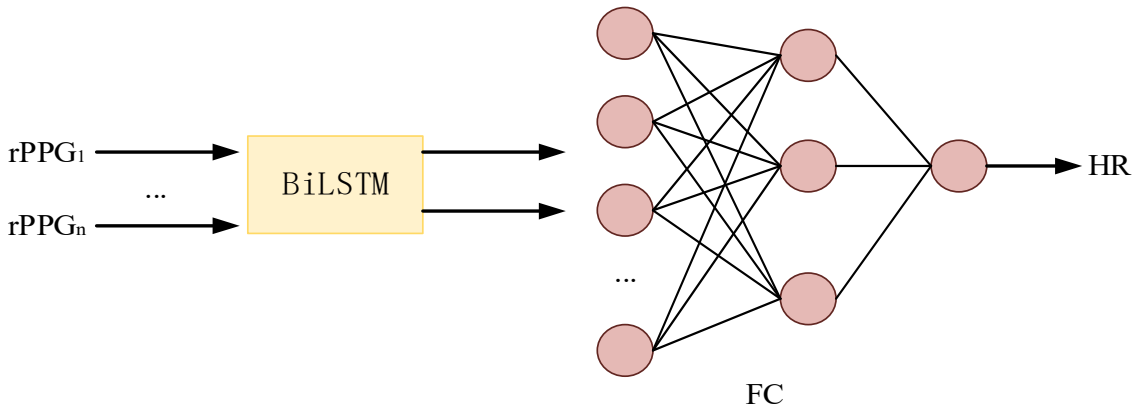


Figure 5. Heart rate estimation module.

The structure of the BiLSTM is shown in Figure 6. The forward LSTM is responsible for processing the input sequence chronologically from the past to the future, while the reverse LSTM processes it in reverse order from the future to the past, and finally combines the output of both to form a global feature representation. BiLSTM has excellent long- and short-term memory, making it more suitable for processing rPPG signals over long periods of time. In the process, BiLSTM is able to filter out some short-term noise effects and focus on extracting more critical timing features. In addition, it can flexibly process input sequences of different time lengths and adapt rPPG signals of different sampling frequencies. Therefore, the BiLSTM model contributes to the enhancement of the accuracy of the heart rate signal prediction and provides more reliable results.

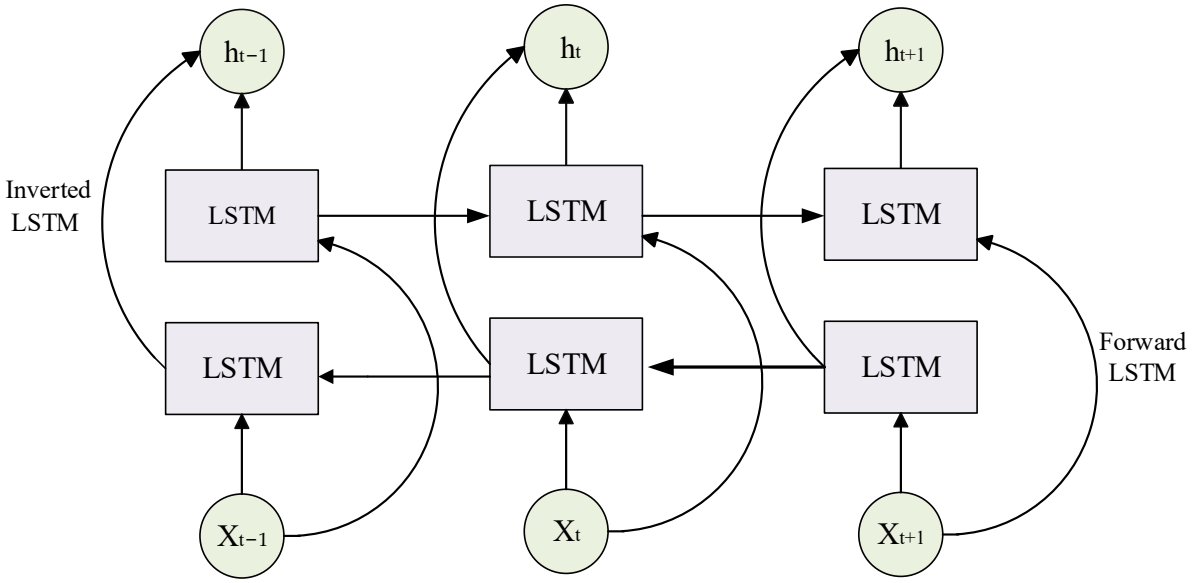


Figure 6. BiLSTM module network structure diagram.

The structural calculation process of BiLSTM is akin to a single LSTM, where the forward LSTM state and backward LSTM state are combined to obtain the state of the BiLSTM network. Its calculation formula is as follows:

$$\overrightarrow{h_t} = LSTM(h_{t-1}, x_t) \tag{3}$$

$$\overleftarrow{h_t} = LSTM(h_{t+1}, x_t) \tag{4}$$

$$h_t = \alpha\overrightarrow{h_t} + \beta\overleftarrow{h_t} \tag{5}$$

where $x_t, \overrightarrow{h_t}, \overleftarrow{h_t}$ represent the input data at time t, the output of the forward LSTM hidden layer, and the output of the reverse LSTM hidden layer, respectively. α, β are constant coefficients, respectively, of the weight of $\overrightarrow{h_t}, \overleftarrow{h_t}$.

2.5. Evaluation Indicators

The loss functions used in this paper mainly include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson correlation coefficient R. Each indicator is calculated as follows. Where n represents the number of samples, HR_{pred} represents the predicted heart rate corresponding to each video, HR_{label} represents the true heart rate value of each video, $\overline{HR}_{pred}$ represents the average value of the entire predicted sample, and $\overline{HR}_{label}$ represents the average value of all true heart rate values.

2.5.1. Mean Absolute Error (MAE)

MAE is the average of the absolute values of the difference between the predicted and true values. Its calculation formula is shown in the Equation (6). $|\cdot|$ indicates that the absolute value is taken, and the average absolute error HR_{MAE} can avoid the problem of positive and negative cancelling each other in the estimation error, so it can accurately reflect the actual estimation error.

$$HR_{MAE} = \frac{1}{n}\sum_{i=1}^{n} |HR_{pred}{}^{(i)} - HR_{label}{}^{(i)}| \tag{6}$$

2.5.2. Root Mean Squared Error (RMSE)

RMSE is the mean square root of the squared error between all the estimated heart rates and the true heart rate. Its calculation formula is shown in Equation (7). HR_{RMSE} can describe the deviation degree of the error between the estimated value of the algorithm and the true value.

$$HR_{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(HR_{pred}{}^{(i)} - HR_{label}{}^{(i)}\right)^2} \tag{7}$$

2.5.3. Pearson Correlation Coefficient (R)

R is a statistical indicator used to assess the linear relationship between two variables. Its value ranges from -1 to 1. The Pearson correlation coefficient is determined by calculating the covariance between two variables divided by their respective standard deviations. The greater the absolute value of R, the stronger the correlation, that is, the closer R is to 1 (positive correlation) or -1 (negative correlation), the stronger the correlation between the predicted value and the true value. Conversely, the closer R is to 0, the weaker the correlation. The calculation formula is shown in Equation (8).

$$R = \frac{\sum_{i=1}^{n}\left(HR_{pred}{}^{(i)} - \overline{HR}_{pred}\right)\left(HR_{label}{}^{(i)} - \overline{HR}_{label}\right)}{\sqrt{\sum_{i=1}^{n}\left(HR_{pred}{}^{(i)} - \overline{HR}_{pred}\right)^2}\sqrt{\sum_{i=1}^{n}\left(HR_{label}{}^{(i)} - \overline{HR}_{label}\right)^2}} \tag{8}$$

The smaller the MAE and the RMSE are, the closer the predicted values are to the true values, indicating more accurate measurement results. The R is used to reflect the degree of linear correlation between two random variables. The closer to 0, the weaker the correlation is; the closer to 1, the stronger the correlation is.

3. Experiment

3.1. Dataset

3.1.1. UBFC-rPPG Dataset

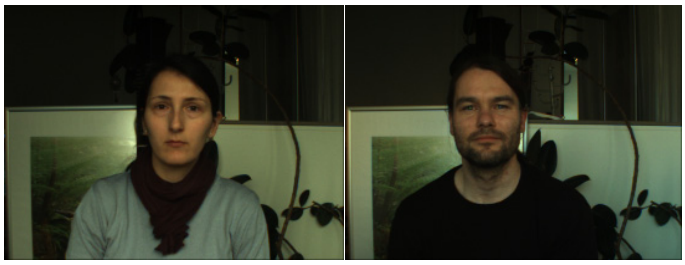
The UBFC-rPPG dataset [22] recorded videos of 42 individuals in real environments, using a low-cost webcam (Logitech C920 HD Pro), each of which was about 2 min long, with a sampling rate of 30 fps and a resolution of 640 × 480. The subjects sat about 1 m away from the imaging device and were required to play digital games to induce changes in heart rate. Physiological data were collected by a CMS50E fingertip pulse oximeter while recording videos, with the collected signals including pulse waves and blood oxygen saturation. All experiments were performed indoors under varying lighting conditions, such as sunlight and different intensities of indoor lighting. A portion of the dataset is illustrated in Figure 7.

(**a**) (**b**)

Figure 7. The part of UBFC-rPPG datasets are shown in (**a**,**b**).

3.1.2. PURE Dataset

The PURE dataset [23] consisted of ten participants, with eight males and two females, at a distance of approximately 1.1 m from the imaging device. Six types of head movement were determined, including steady, speaking, slow translation, fast translation, small rotation, and moderate rotation. The dataset contains a total of 60 video clips, each recording a duration of 1 min. Reference data were synchronized using a pulse oximeter (pulox CMS50E), offering pulse waveforms and blood oxygen saturation (SpO_2) measurements at a sampling rate of 60 Hz. The lighting conditions consisted of sunlight passing through a large window and illuminating the subject's face, with slight variations in illumination due to changes in cloud cover. All recordings were carried out while the subject was in a resting state. A portion of the dataset is illustrated in Figure 8.

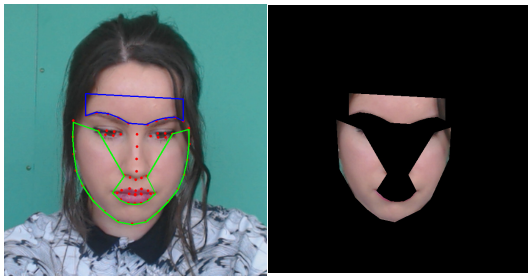


(**a**) (**b**)

Figure 8. The part of PURE datasets are shown in (**a**,**b**).

3.2. ROI Selection

The facial ROI is used to identify key areas in facial images or videos, and is a core step in extracting heart rate signals. By precisely selecting and extracting the facial ROI, noise interference can be effectively reduced, improving the accuracy and stability of the heart rate signal. Proper ROI selection and extraction are crucial steps for the success of non-contact heart rate detection. Drawing inspiration from the ROI used in CVD, this paper employs a manual method to select facial ROIs that contain rich information. Specifically, the face is divided into two regions using the Dlib 68-point facial landmark detection technique. Region 1 is the forehead, defined by landmarks 18–27 and extended upwards by 50 pixels. Region 2 is defined by sequentially connecting landmarks 1–17, 46, 36, 55–60, 49, 32, 37, and landmark 1, covering the lower half of the face. The combination of these two regions forms the final ROI used for rPPG signal extraction, covering multiple areas of the face that reflect blood perfusion fluctuations while avoiding areas like the eyes, mouth, and eyebrows that are prone to noise. This significantly enhances the robustness and accuracy of signal extraction. The specific ROI regions are shown in Figure 9a below. After manually selecting the facial ROI, a closing operation is applied to extract the selected facial ROI, as illustrated in Figure 9b below.



(**a**) (**b**)

Figure 9. (**a**) ROI selection; (**b**) ROI extraction.

3.3. Illumination Normalization

Illumination normalization is accomplished by adjusting the brightness and contrast of an image to reduce or eliminate the effects of varying lighting conditions, ensuring consistent brightness across different environments. In non-contact heart rate detection, signals like rPPG are extracted by analyzing subtle changes in the color of the facial skin, which are related to periodic fluctuations in blood volume. These color changes are influenced by both the heartbeat and external factors like variations in ambient lighting. Changes in lighting conditions can affect the brightness and color distribution of facial images, potentially causing errors or distortions in rPPG signal extraction. Illumination normalization can effectively reduce noise caused by lighting changes, ensuring that the rPPG signals extracted from the face are more stable and accurate, thus enhancing the accuracy and robustness of heart rate estimation. The image with illumination normalization is shown in Figure 10.

(**a**) (**b**)

Figure 10. (**a**) Original image; (**b**) illumination normalized image.

3.4. Data Augmentation

Given that the UBFC and PURE datasets are relatively small, data augmentation is required to enhance the model's robustness by generating new data samples. Common augmentation methods include rotation, flipping, cropping, and scaling. Convolutional neural networks treat horizontally flipped images as different images, which can increase the number of training samples, reduce overfitting, and improve the model's generalization ability.

In order to help the model better capture temporal information and context, we introduced the Sliding Window Strategy for data augmentation. The sliding window strategy is widely used in time-series data, especially for video or audio. This strategy extracts multiple subsequences from the original sequence by sliding a window along the time axis with a fixed or variable step size, thus generating more training samples. Specifically, for the UBFC and PURE datasets, the first 50 s of each video are taken, and each video is downsampled to 30 frames per second to be suitable for the model in this paper. The sliding window strategy is then applied to crop the facial video (with a window length of 5 s and a time step of 3 s) to generate a new augmented video sample dataset. For

example, for a 50-s video in the first dataset, the first sample dataset will be from 0–5 s, the second sample dataset will be from 3–8 s, and so on. Each adjacent sample dataset overlaps by 2 s, which helps the model better understand the relationships between adjacent time windows. In this way, 16 sub-videos can be obtained from one video, greatly expanding the dataset. The specific cropping method is shown in Figure 11.

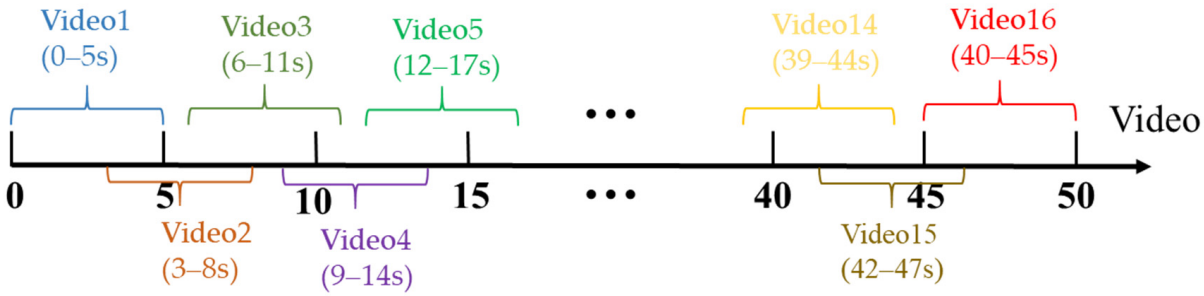


Figure 11. Data augmentation using sliding window.

3.5. Implementation Details

This model adopts Python language and the PyTorch deep learning framework, and all training and testing are carried out using a server with an NVIDIA Tesla V100-PCIE. The version is python 3.8 and PyTorch 1.11. The PURE dataset and the UBFC-rPPG dataset divide the training set and the test set according to 8:2 to ensure that the test sample does not appear in the training sample. Specifically, 538 videos from the UBFC dataset are used for training, and 135 videos for testing; 768 videos from the PURE dataset are used for training, and 192 videos for testing. In the training process, the initial learning rate was set to 1×10^{-5}, and the Adam_W optimizer was used for optimization. Since the videos were all RGB videos, in_ch was set to 3, and the model training was 60 rounds in total.

3.6. Experimental Results and Comparison

3.6.1. Analysis of Experimental Results

In order to judge the fitting degree of the proposed model to the rPPG signal, intra-dataset tests were carried out on the UBFC and PURE datasets, respectively. Figures 12 and 13, respectively, show the comparison graphs between the estimated signals on the UBFC and PURE datasets, and their corresponding ground truth signals. It can be intuitively seen from the graphs that the rPPG signal curve obtained by the model in this paper presents a high consistency compared with the real signal curve on the ground. It is proved that the model has high accuracy in predicting heart rate signal.

In addition, to better visualize the effect of the model used in this study, Figure 14 shows the scatter plots of predicted and true heart rates generated in the UBFC and PURE datasets. The horizontal axis represents the true heart rate value, the vertical axis represents the predicted heart rate value, and the red line y = x represents the position where the predicted value is equal to the true value. The closer the sample point is to y = x, the better the prediction effect is.

As can be seen from the two figures (a) and (b) in Figure 14, about 5% of the points deviate from the straight line y = x. On the whole, the data points are concentrated and located near y = x. This shows that the predicted heart rate and the real heart rate in the two datasets have strong consistency.

The Bland–Altman plot is a two-dimensional scatter plot where each point represents a test result. The horizontal coordinate represents the average value of the true heart rate values and the predicted heart rate values, while the vertical coordinate represents the interpolation of the true heart rate values and the predicted heart rate values. The

two dotted lines represent the 95% confidence interval $[\mu - 1.96\sigma, \mu + 1.96\sigma]$, and the solid line in the middle represents the average. If this point falls within the confidence interval, it indicates that the predicted heart rate is consistent with the true heart rate.

Figure 15 shows the Bland–Altman plots of predicted and true heart rate generated in the UBFC dataset and PURE dataset. For the UBFC dataset, the confidence interval shown in the Bland–Altman plot is $[-4.8, 5.1]$, and the confidence interval shown in the PURE dataset is $[-7, 6.8]$, both in a small range, and most of the points are within the confidence interval. Therefore, it is proved that the proposed model can retain good predictive performance in the face of different datasets, and the error is controlled at a low level, ensuring the reliability and practicability of the prediction results.

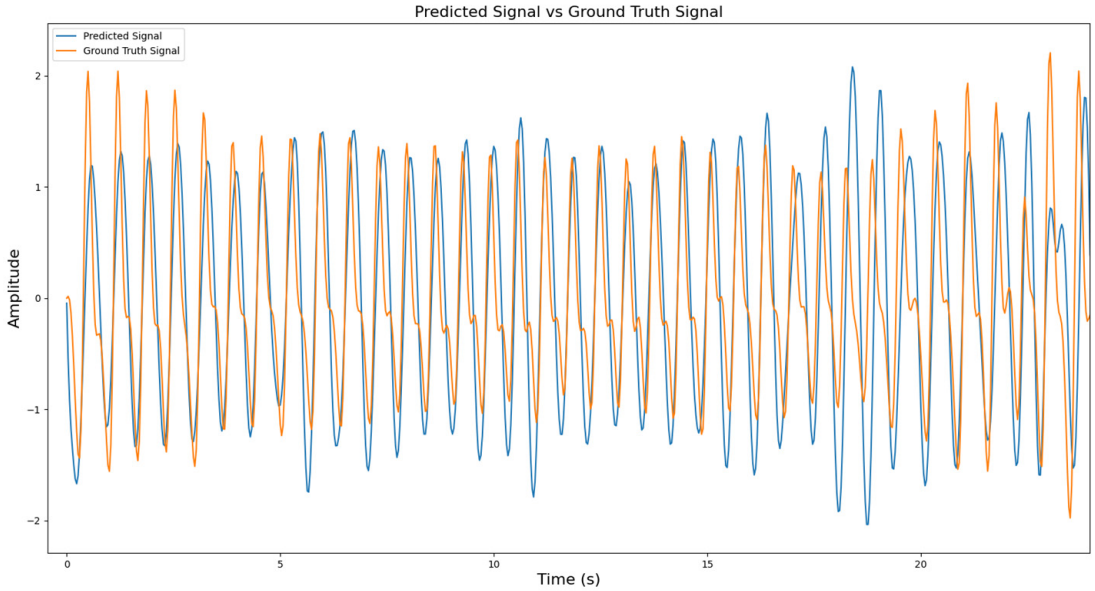


Figure 12. Comparison of prediction signal and ground truth signal in UBFC-rPPG dataset.

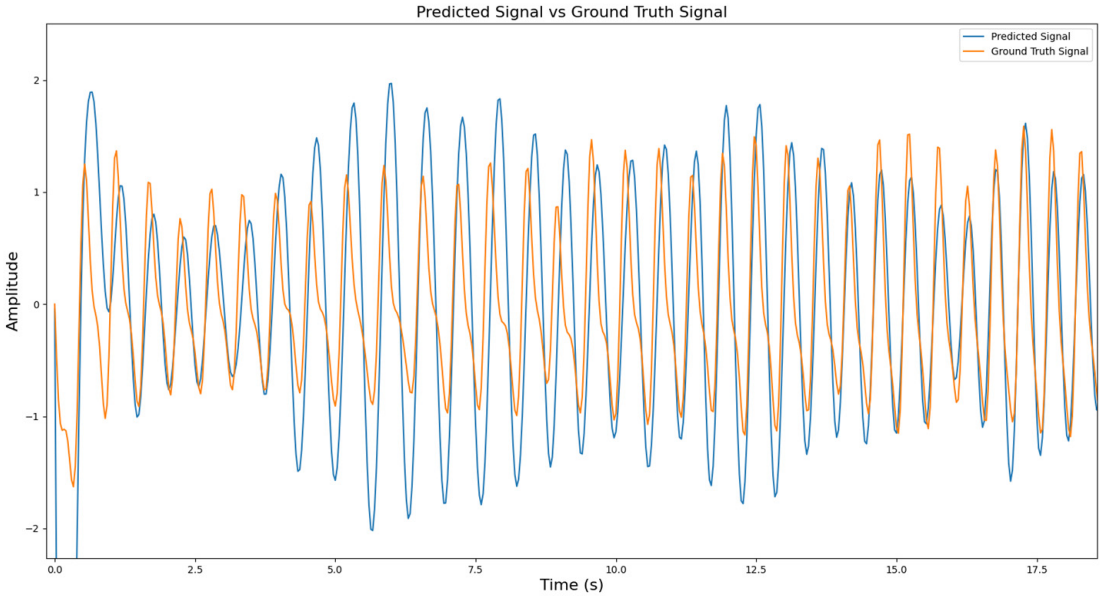
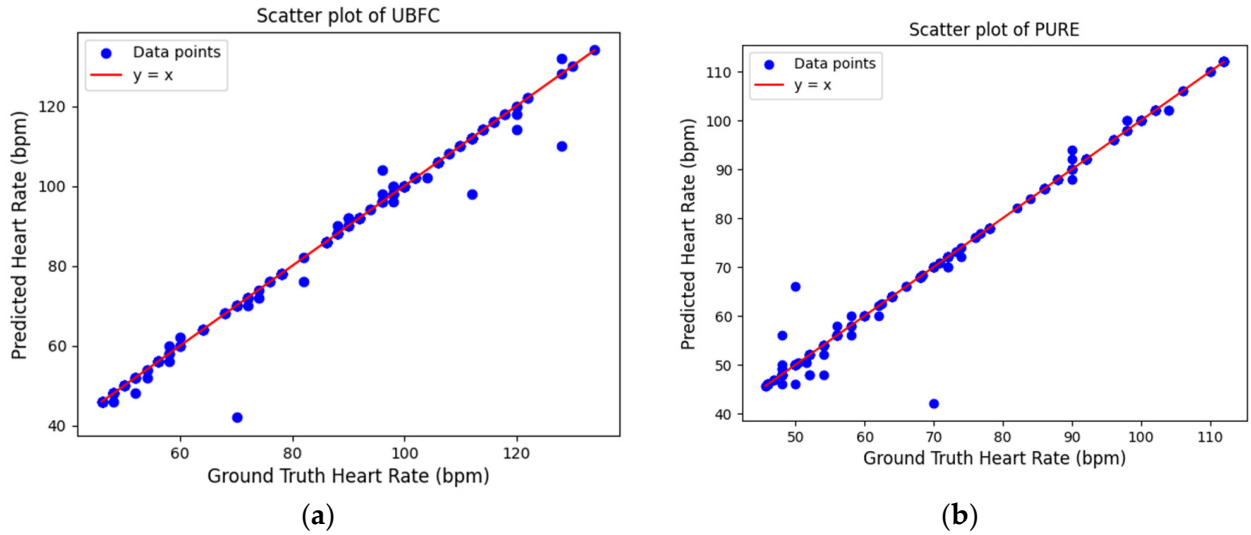
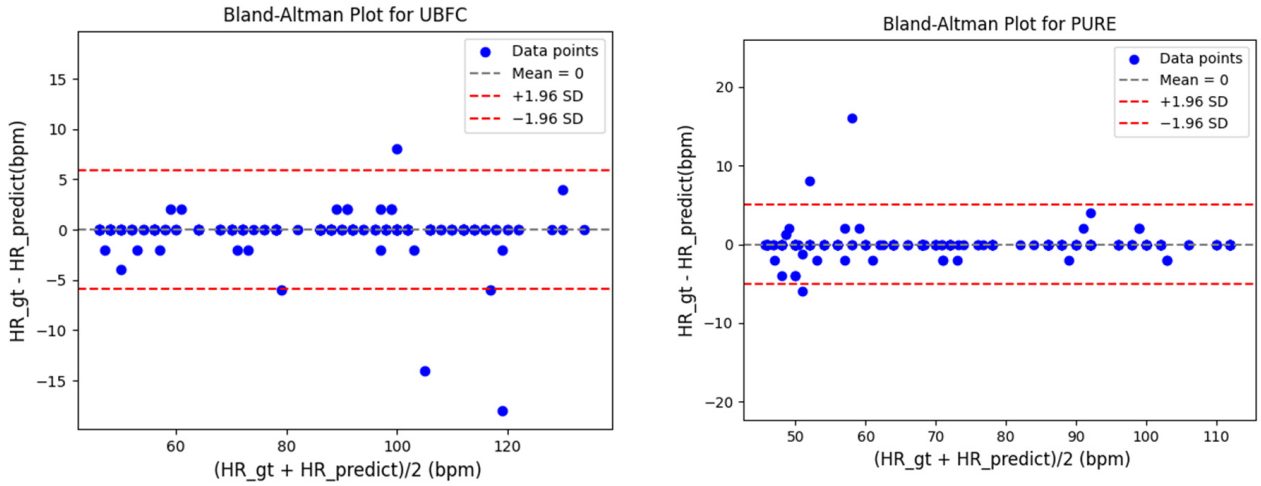


Figure 13. Comparison of prediction signal and ground truth signal in PURE dataset.



(**a**) (**b**)

Figure 14. Scatter plot of predicted heart rate and true heart rate. (**a**) UBFC-rPPG data distribution point diagram; (**b**) PURE data distribution point diagram.



(**a**) (**b**)

Figure 15. Scatter plot of predicted heart rate and true heart rate. (**a**) Bland-Altman plot for the UBFC-rPPG dataset; (**b**) Bland-Altman plot for the PURE dataset.

3.6.2. Comparative Experiment

The evaluation indexes of the model mainly used MAE, RMSE, and Pearson correlation coefficient R. The method used in this study is compared with previous methods, including some traditional methods like ICA, CHROM, POS, and Deep leaning methods such as Contrast-phys, SiNC, Dual-GAN, etc. The comparison results are shown in Tables 1 and 2.

Table 1 presents the comparison results of our model with other methods on the UBFC dataset. The results of MAE and RMSE are both excellent, reaching 0.24 bpm and 0.65 bpm, respectively, which is better than most methods. The results of the tests on the PURE dataset are listed in Table 2. The MAE obtained by using the 3D-SimAM model is 0.63 bpm, which is very close to the SiNC method (0.61 bpm). RMSE was 1.30 bpm and performed best among all the comparison models. The Pearson correlation coefficient reached 0.99, indicating a very high correlation. Overall, the 3D-SimAM model shows satisfactory results on both the UBFC and PURE datasets, which proves the validity and reliability of the proposed method in heart rate estimation tasks.

Table 1. Comparison of test results of the UBFC-rPPG dataset.

Method	MAE (bpm) ↓	RMSE (bpm) ↓	R ↑
ICA [9]	5.17	11.76	0.65
CHROM [24]	2.37	4.91	0.89
POS [25]	4.05	8.75	0.78
SynRhythm [26]	5.59	6.82	0.75
PulseGAN [14]	1.19	2.10	0.98
Gideon2021 [27]	1.85	4.28	0.93
Contrast-Phys [28]	0.64	1.00	0.99
SiNC [29]	0.59	1.83	0.99
Dual-GAN [30]	0.44	0.67	0.99
Contrast-Phys + (100%) [31]	0.21	0.80	0.99
3D-SimAM(Ours)	0.24	0.65	0.99

In this case, "↑" ("↓") signifies that a higher (lower) value is better.

Table 2. Comparison of test results of the PURE dataset.

Method	MAE (bpm) ↓	RMSE (bpm) ↓	R ↑
CHROM [24]	2.07	9.92	0.99
2SR [32]	2.44	3.06	0.98
HR-CNN [12]	1.84	2.37	0.98
PhysNet [33]	2.10	2.60	0.99
Dual-GAN [30]	0.82	1.31	0.99
SiNC [29]	0.61	1.84	0.99
Gideon2021 [27]	2.3	2.90	0.99
Contrast-Phys [28]	1.00	1.40	0.99
3D- SimAM(Ours)	0.63	1.30	0.99

In this case, "↑" ("↓") signifies that a higher (lower) value is better.

3.6.3. Testing Across Datasets

To further access the robustness of the proposed model, cross-dataset tests are carried out. Specifically, the model is trained on the UBFC dataset and tested on the PURE dataset. The results are shown in Table 3. The MAE obtained using the 3D-SimAM model was 2.05 bpm, and the Pearson correlation coefficient also increased to 0.87 bpm. From the experimental data, there is a high correlation between the heart rate predicted by the model and the real heart rate.

Table 3. Comparison of test results across datasets.

Method	MAE (bpm) ↓	RMSE (bpm) ↓	R ↑
CHROM [24]	-	13.97	0.55
PhysNet [33]	2.20	6.85	0.86
Physformer [34]	2.68	7.01	0.86
Contrast-Phys [28]	2.43	7.43	0.86
3D- SimAM(Ours)	2.15	7.08	0.88

In this case, "↑" ("↓") signifies that a higher (lower) value is better.

3.7. Ablation Experiment

To evaluate the contribution of the SimAM attention mechanism in improving facial heart rate detection accuracy, ablation experiments were designed on the UBFC-rPPG and PURE datasets, respectively. The ablation experiment was designed as follows:

(1) Without incorporating any attention mechanism module;
(2) Add SimAM attention mechanism to the existing model;
(3) Add CBAM attention mechanism to the existing model [35];

(4) Add SKAttention attention mechanism to the existing model.

The experimental results are shown in Table 4:

Table 4. Experimental results of the UBFC-rPPG dataset.

SimAM Attention	CBAM Attention	SK Attention	MAE (bpm) ↓	RMSE (bpm) ↓	R ↑
			0.64	1.00	0.99
√			0.24	0.65	0.99
	√		0.49	1.73	0.986
		√	0.20	0.57	0.98

In this case, "↑" ("↓") signifies that a higher (lower) value is better. "√" indicates that the corresponding attention mechanism has been added.

As shown in Tables 4 and 5, the model performs worst when no attention mechanism module is added. After introducing the attention mechanism into the model, the evaluation indexes MAE, RMSE, and R are improved to some extent compared with the original network model. According to the experimental results, the effect of introducing the SimAM attention mechanism is better than that of introducing the other two attention mechanisms. On the UBFC dataset, MAE and RMSE increased by 40% and 35%, respectively, compared with the model without any attention mechanism. On the PURE dataset, MAE improved by 37% and RMSE improved by 10%. This result shows that by introducing the SimAM attention mechanism into the network, the model can better focus on the region of interest, eliminate irrelevant noise interference, improve the network's robustness, and enhance its performance.

Table 5. Experimental results of the PURE dataset.

SimAM Attention	CBAM Attention	SK Attention	MAE (bpm) ↓	RMSE (bpm) ↓	R ↑
			1.00	1.40	0.99
√			0.63	1.30	0.99
	√		1.08	0.97	0.99
		√	2.43	7.29	0.98

In this case, "↑" ("↓") signifies that a higher (lower) value is better. "√" indicates that the corresponding attention mechanism has been added.

4. Conclusions and Prospects

Facial video-based heart rate detection offers advantages such as low cost and no contact, with promising applications in remote health monitoring, monitoring patients with skin injuries, and evaluating driver status. In this paper, we proposed a 3D facial video heart rate detection algorithm based on an attention mechanism to address the challenge of noise interference in existing non-contact methods. Compared with traditional 3D-CNN, the model with the attention mechanism can better focus on regions and spatiotemporal features related to heart rate, thereby improving sensitivity to facial dynamic changes and helping to accurately extract the rPPG signal. At the same time, by manually selecting facial ROI regions and normalizing video images with uneven lighting, the influence of motion artifacts and lighting changes is effectively reduced, enhancing the accuracy of heart rate detection and the system's robustness. The results are tested on both the UBFC-rPPG dataset and the PURE dataset, showing significant performance in both within-dataset and cross-dataset evaluations.

Despite achieving good results on both datasets, there are still certain limitations in practical applications that need to be addressed in future studies. This algorithm performs well when a single face appears in the video, but the detection accuracy decreases when

multiple faces are present. Future work will focus on heart rate detection in multi-person scenarios. We also plan to incorporate an automated adjustment mechanism in the ROI selection process, which will dynamically adjust the position and size of the ROI based on real-time facial movements, further improving performance in complex environments.

Author Contributions: Conceptualization, Y.S. and X.S.; methodology, Y.S.; software, Y.S. and X.H.; validation, Y.S., X.Y. and H.L.; formal analysis and investigation, Y.S., X.S. and X.H.; resources and data curation, Y.S. and C.W.; writing—original draft preparation, Y.S.; writing—review and editing, X.S. and C.W.; project administration, Y.S. and C.W.; funding acquisition, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Key Science and Technology Innovation Programs in Shandong Province (No.2017CXGC0919).

Data Availability Statement: The datasets used in this study are available upon request. Please contact the corresponding author via email to obtain access to the datasets. UBFC-rPPG: https://sites.google.com/view/ybenezeth/ubfcrppg. PURE: https://www.tu-ilmenau.de/neurob/data-sets-code/pulse-rate-detection-dataset-pure, (accessed on 8 January 2025).

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

References

1. Cook, M.S.; Togni, M.C.; Schaub, M.; Wenaweser, P.; Hess, O.M. High heart rate: A cardiovascular risk factor? *Eur. Heart J.* **2006**, *27*, 2387–2393. [CrossRef] [PubMed]
2. Cao, J.; Li, Y.; Zhang, K.; Gool, L.V. Video Super-Resolution Transformer. *arXiv* **2021**, arXiv:2106.06847.
3. Xun, C.; Cheng, J.; Song, R.; Liu, Y.; Rabab, W.; Wang, Z.J. Video-based heart rate measurement: Recent advances and future prospects. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 3600–3615.
4. Xin, L.; Patel, S.; McDuff, D. RGB Camera-based physiological sensing: Challenges and future directions. *arXiv* **2021**, arXiv:2110.13362.
5. Yu, Z.; Li, X.; Zhao, G. Facial-Video-Based Physiological Signal Measurement: Recent advances and affective applications. *IEEE Signal Process. Mag.* **2021**, *38*, 50–58. [CrossRef]
6. Xiao, H.; Liu, T.; Sun, Y.; Sun, Y.; Li, Y.; Zhao, S.; Avolio, A. Remote photoplethysmography for heart rate measurement: A review. *Biomed. Signal Process. Control* **2024**, *88*, 105608. [CrossRef]
7. Verkruysse, W.; Svaasand, L.; Nelson, J. Remote plethymographic imaging using ambient light. *Opt. Express* **2008**, *16*, 21434–21445. [CrossRef] [PubMed]
8. Poh, M.-Z.; MCDuff, D.; Picard, R. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **2010**, *18*, 10762–10774. [CrossRef]
9. Comon, P. Independent component analysis, A new concept? *Signal Process.* **1994**, *36*, 287–314. [CrossRef]
10. Hsu, G.; Ambikapathi, A.; Chen, M. Deep learning with time-frequency representation for pulse estimation from facial videos. In Proceedings of the IEEE International Joint Conference on Biometrics(IJCB), Denver, CO, USA, 1–4 October 2017; pp. 383–389.
11. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
12. Špetlík, K.; Franc, V.; Matas, J. Visual heart rate estimation with convolutional neural network. In Proceedings of the British Machine Vision Conference, Newcastle, UK, 3–6 September 2018; pp. 3–6.
13. Chen, W.; McDuff, D. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 349–365.
14. 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] [PubMed]
15. Gupta, A.; Ravelo-García, A.G.; Dias, F.M. Availability and performance of face based non-contact methods for heart rate and oxygen saturation estimations: A systematic review. *Comput. Methods Programs Biomed.* **2022**, *219*, 106771. [CrossRef] [PubMed]
16. Niu, X.; Yu, Z.; Hu, H.; Li, X.; Shan, S.; Zhao, G. Video-Based Remote Physiological Measurement via Cross-Verified Feature Disentangling. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Volume 12347, pp. 295–310.
17. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
18. King, D.E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.

19. Yang, L.; Zhang, R.; Li, L.; Xie, X. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR: New York, NY, USA, 2021; pp. 11863–11874.
20. Mazhar, N.; Malik, F.M.; Raza, A.; Khan, R. Predefined-time control of nonlinear systems: A sigmoid function based sliding manifold design approach. *Alex. Eng. J.* **2022**, *61*, 6831–6841. [CrossRef]
21. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef] [PubMed]
22. Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; Dubois, J. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.* **2019**, *124*, 82–90. [CrossRef]
23. Stricker, R.; Miller, S.; Gross, H.M. Non-contact video-based pulse rate measurement on a mobile service robot. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25–29 August 2014; pp. 1056–1062.
24. De Haan, G.; Jeanne, V. Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2878–2886. [CrossRef] [PubMed]
25. 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]
26. 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; pp. 3580–3585.
27. Gideon, J.; Stent, S. The Way to My Heart Is Through Contrastive Learning: Remote Photoplethysmography From Unlabelled Video. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 3995–4004.
28. Sun, Z.; Li, X. Contrast-Phys: Unsupervised Video-Based Remote Physiological Measurement via Spatiotemporal Contrast. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Volume 13672, pp. 492–510.
29. Speth, J.; Vance, N.; Flynn, P.; Czajka, A. Non-Contrastive Unsupervised Learning of Physiological Signals From Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 14464–14474.
30. 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(CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12404–12413.
31. Sun, Z.; Li, X. Contrast-Phys+: Unsupervised and Weakly-supervised Video-based Remote Physiological Measurement via Spatiotemporal Contrast. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; Volume 14, pp. 5835–5851.
32. Wang, W.; Stuijk, S.; Haan, G. A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1974–1984. [CrossRef]
33. Yu, Z.; Li, X.; Zhao, G. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. *arXiv* **2019**, arXiv:1905.02419.
34. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4186–4196.
35. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.