LSTC-rPPG: Long Short-Term Convolutional Network for Remote Photoplethysmography

Jun Seong Lee[1], Gyutae Hwang[2], Moonwook Ryu[1], Sang Jun Lee[2,*]
[1]Electronics and Telecommunications Research Institute, Republic of Korea
[2] Jeonbuk National University, Republic of Korea
eejunseonglee@gmail.com, gyutae741@jbnu.ac.kr,
moonwook@etri.re.kr, sj.lee@jbnu.ac.kr

Abstract

Remote photoplethysmography (rPPG) is a non-contact technique for measuring blood pulse signals associated with cardiac activity. Although rPPG is considered an alternative to traditional contact-based photoplethysmography (PPG) because of its non-contact nature, obtaining reliable measurements remains a challenge owing to the sensitiveness of rPPG. In recent years, deep learning-based methods have improved the reliability of rPPG, but they suffer from certain limitations in utilizing long-term features such as periodic tendencies over long durations. In this paper, we propose a deep learning-based method that models long short-term spatio-temporal features and optimizes the long short-term features, ensuring reliable rPPG. The proposed method is composed of three key components: i) a deep learning architecture, denoted by LSTC-rPPG, which models long short-term spatio-temporal features and combines the features for reliable rPPG, ii) a temporal attention refinement module that mitigates temporal mismatches between the long-term and short-term features, and iii) a frequency scale invariant hybrid loss to guide long-short term features. In experiments on the UBFC-rPPG database, the proposed method demonstrated a mean absolute error of 0.7, root mean square error of 1.0, and Pearson correlation coefficient of 0.99 for heart rate estimation accuracy, outperforming contemporary state-of-the-art methods.

1. Introduction

Remote photoplethysmography (rPPG) is an optical measurement technique that enables non-contact measurement of blood pulse signals related to cardiac activity [6, 35, 41]. In contrast to conventional photoplethysmography (PPG), rPPG does not require direct skin contact for physiological measurements. Instead, rPPG utilizes facial videos and analyzes facial color changes to extract blood pulse signals, referred to as rPPG signals.

The non-contact nature of rPPG makes it an attractive alternative to conventional contact-based PPG, with a wide range of prospective applications in the domains of healthcare and human–computer interaction [6, 21, 25, 41]. For example, by minimizing discomfort or skin irritation to subjects, rPPG can offer a viable solution for patients with skin sensitivities, who may experience discomfort during direct skin contact. However, given the non-contact nature of rPPG, the amplitude of rPPG signals is insufficient and susceptible to contamination, making it difficult to obtain accurate and reliable measurements. Therefore, supplementary measures are necessary to improve the accuracy and reliability of rPPG signal measurement.

Recent advancements in deep learning techniques have made it possible to achieve accurate and reliable measurements of rPPG signals. These achievements are largely based on the spatial and temporal feature extraction capabilities of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The use of these networks facilitates the modeling of subtle changes in facial color and periodic fluctuations, both of which are key information for estimating rPPG signals. Some methods [16, 24] utilize the sequential application of 2D convolutional neural networks (2DCNNs) and RNNs to model spatial and temporal features, respectively, while others [3, 26, 40] employ 3D convolutional neural networks (3DCNNs) to more interactively model the spatial as well as the temporal features. However, the aforementioned approaches have some limitations in modeling and utilizing long-term, i.e., global temporal, features of rPPG signals such as periodic patterns over the entire period. Those 3DCNN-based methods often focus on short-term, i.e., local temporal, features and fail to exploit the long-term features. In methods that combine 2DCNNs and RNNs, although RNNs can theoretically utilize the long-term features, in practice, the use of the long-term features for long input is challenging because of concerns such

as long-term dependencies and gradient vanishing.

To address these limitations, methods [14, 42] based on the transformer architecture [34] have been proposed. The transformer-based methods can refer to long-term features by using the attention mechanism. Although these methods have potential benefits, realizing the full benefits of the transformer in the field of rPPG research remains a challenge. This is because the transformer requires a substantial amount of data to fully utilize its capabilities owing to its low inductive bias, but collecting sufficient data in the field of rPPG research is challenging in practice.

Inspired by the considerations mentioned above, we propose a deep learning-based method that can effectively utilize long-term features to reliably predict rPPG signals, even when limited data is available. The proposed method models long short-term features through a 3DCNN-based hourglass structure and combines the long short-term features to take into account long-term as well as short-term information. Furthermore, to address the issue of information mismatches in the integration of the long short-term features, we introduce the incorporation of a temporal attention refinement module. Finally, we suggest a hybrid loss that consists of time-domain and frequency-domain losses to jointly learn the long short-term characteristics of targeted rPPG signals such as instantaneous signal changes and long-term periodicity. The hybrid loss is characterized by frequency-domain scale invariance, which facilitates the proposed model to easily converge.

The contributions of this study are as follows:

1. A novel deep learning architecture denoted as LSTC-rPPG is proposed. LSTC-rPPG robustly predicts rPPG signals by modeling long short-term features through a 3D hourglass structure and by combining the long short-term features.

2. A temporal attention refinement module (TARM) is proposed. TARM reduces mismatches between long short-term features in terms of temporal receptive fields, resulting in more reliable measurements of rPPG signals.

3. A hybrid loss is proposed. The hybrid loss includes frequency-domain and time-domain losses to guide long-term and short-term characteristics, respectively, of targeted rPPG signals. Furthermore, the frequency-domain loss is constrained by scale invariance to facilitate the proposed model's training.

4. Experimental results demonstrating superior performance compared to state-of-the-art methods are obtained for the UBFC-rPPG dataset, even in the absence of data augmentation.

2. Related work

2.1. Signal Processing-based rPPG

Early attempts to measure rPPG signals involved the detection of facial regions in videos, followed by the spatial averaging of the RGB channels for the facial regions [31, 35]. However, these methods were only applicable in controlled environments such as laboratory settings. Consequently, to facilitate more robust and realistic measurements, methods based on conventional signal processing have been proposed. These include three major methods: region of interest (ROI)-based methods, signal decomposition-based methods, and color transformation-based methods. ROI-based methods aim to select improved ROIs such as ROIs based on facial landmarks [18], dynamic ROIs [12], and unsupervised skin ROIs [1]. Appropriate ROIs allow for the reduction of background noise and head movement artifacts [1, 12, 18, 30]. Signal decomposition-based methods aim to enhance the signal-to-noise ratio, including ICA [20, 27, 28], PCA [37], blind source separation on random patches [15], matrix completion [33], temporal rotation of the spatial subspace of skin pixels [38], and mathematical skin reflection [36]. Color transformation-based methods, such as a linear combination of the chrominance signals [9] and a blood volume pulse signature [10], weight and combine color channels to yield better results. Despite their potential benefits, these methods may not fully utilize spatial and temporal features such as the spatial color changes of facial skin and the periodic tendencies of rPPG signals, thus restricting their effectiveness.

2.2. Deep Learning-based rPPG

Early deep learning-based methods for measuring rPPG signals were based on 2DCNNs that exhibited innovative performance in pattern recognition through their capabilities to model spatial features. These methods used 2DCNNs to capture skin pixel segmentation [4, 32] or subtle color changes in facial skin [5] and outperformed the conventional signal processing-based methods. However, their modeling of temporal features for sequential inputs could potentially be enhanced, given that modeling solely spatial features within a single image is insufficient in capturing sequential dependencies present in data.

Inspired by the potential of modeling temporal features, several methods have been proposed for spatial and temporal modeling. These methods include the integration of 2DCNNs and RNNs [16, 24], the use of hand-crafted preprocessing for generating spatio-temporal maps(STmaps) [22, 23], and the utilization of 3DCNNs [3, 26, 40]. These methods enabled spatial and temporal modeling and demonstrated their performance, but could not model and utilize long-term information such as periodicity in rPPG signals.

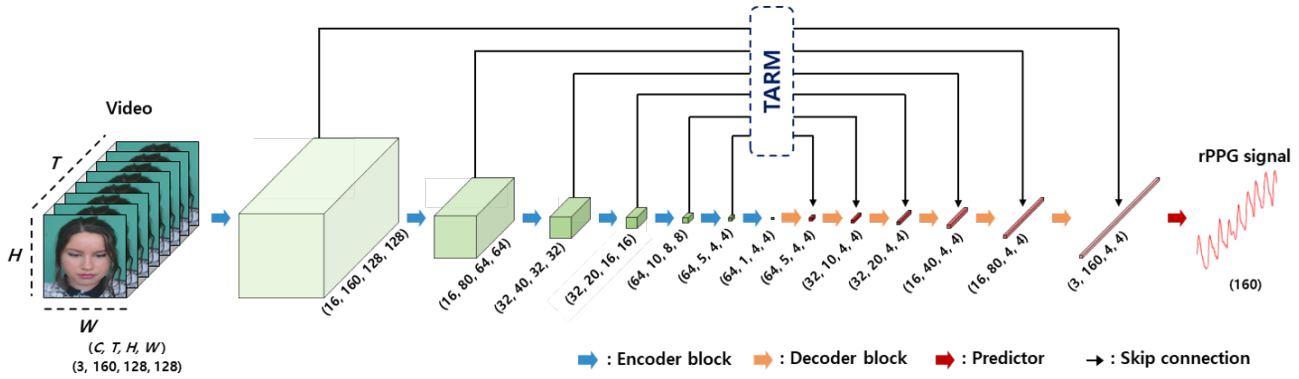

Figure 1. Framework of the proposed architecture.

As alternatives, some studies have proposed the utilization of the principles of the transformer, which enables direct referencing of long-term information, to measure rPPG signals. These studies involve combining the transformer with either 3DCNNs [42] or STMap [14]. By utilizing long-term information, these studies have achieved remarkable performance. However, owing to the low inductive bias, the transformer-based methods require a considerable amount of data to achieve optimal utilization of the transformer. In the field of rPPG research, where data collection can be challenging, the complete potential of the transformer-based methods may sometimes be hindered by this particular aspect.

3. Methodology

In this section, we first introduce LSTC-rPPG as the proposed baseline architecture in Sec. 3.1. Subsequently, the proposed temporal attention refinement module is introduced in Sec. 3.2. Finally, the descriptions of the proposed hybrid loss are presented in Sec. 3.3.

3.1. Long Short-Term Convolutional rPPG

The framework of LSTC-rPPG is illustrated in Fig. 1, where dimensions of height (H), width (W), channel (C), and time (T) are denoted using the (C, T, H, W) format. As an example, in Fig. 1, the input video has dimensions of three channels, a height of 128 pixels, a width of 128 pixels, and a length of 160 frames. LSTC-rPPG performs encoding, decoding, and prediction on an input video of size (3, 160, 128, 128) to estimate a 160-length rPPG signal corresponding to the number of frames. The spatial and temporal information of the input video is represented as compressed features through the encoder blocks. Subsequently, the compressed features are reconstructed along the temporal axis through the decoder blocks. Finally, the predictor estimates an rPPG signal from the decoded features.

Detailed descriptions of LSTC-rPPG are presented in Tab. 1. The notation for height, width, channel, and time used in Fig. 1 is also applied consistently in Tab. 1, and (T, H, W)@N means the use of N filters of size (T, H, W). Furthermore, the notation 3D Conv denotes 3D convolutional operations, 3D Transposed Conv denotes 3D transposed convolutional operations, 3D Avg Pool refers to the application of 3D average pooling, ELU refers to the application of exponential linear unit [7], and BN indicates the execution of batch normalization [13]. In all cases, 3D convolutional operations are executed with a stride of (1, 1, 1), while 3D transposed convolutional operations are executed with a stride of (2, 1, 1) except for Decoder 6, where a stride of (1, 1, 1) is applied for symmetry.

The proposed LSTC-rPPG can be characterized by a 3DCNN-based hourglass structure in the temporal dimension and skip connections (Fig. 1). The hourglass structure enables long short-term spatio-temporal modeling for an input video. During the encoding process, short-term features are progressively compressed into long-term features. The resulting compressed features eventually become one-dimensional in the temporal dimension, fully capturing

Table 1. Details of the proposed architecture.

Block name	Layers	Output shape	Block name	Layers	Output shape
Encoder 1	[3D Conv (3,3,3)@16 ELU BN] × 2	(16,160,128,128)	Decoder 6	3D Transposed Conv (5,1,1)@64 3D Conv (3,3,3)@64 ELU BN	(64,5,4,4)
Encoder 2	3D Avg Pool (2,2,2) [3D Conv (3,3,3)@16 ELU BN] × 2	(16,80,64,64)	Decoder 5	3D Transposed Conv (4,1,1)@64 [3D Conv (3,3,3)@32 ELU BN] × 2	(32,10,4,4)
Encoder 3	3D Avg Pool (2,2,2) [3D Conv (3,3,3)@32 ELU BN] × 2	(32,40,32,32)	Decoder 4	3D Transposed Conv (4,1,1)@64 [3D Conv (3,3,3)@32 ELU BN] × 2	(32,20,4,4)
Encoder 4	3D Avg Pool (2,2,2) [3D Conv (3,3,3)@32 ELU BN] × 2	(32,20,16,16)	Decoder 3	3D Transposed Conv (4,1,1)@32 [3D Conv (3,3,3)@16 ELU BN] × 2	(16,40,4,4)
Encoder 5	3D Avg Pool (2,2,2) [3D Conv (3,3,3)@64 ELU BN] × 2	(64,10,8,8)	Decoder 2	3D Transposed Conv (4,1,1)@16 [3D Conv (3,3,3)@16 ELU BN] × 2	(16,80,4,4)
Encoder 6	3D Avg Pool (2,2,2) 3D Conv (3,3,3)@64 ELU BN	(64,5,4,4)	Decoder 1	3D Transposed Conv (4,1,1)@16 [3D Conv (3,3,3)@3 ELU BN] × 2	(3,160,4,4)
Encoder 7	3D Conv (5,3,3)@64 ELU BN	(64,1,4,4)	Predictor	3D Conv (1,4,4)@1	(1,160,1,1)

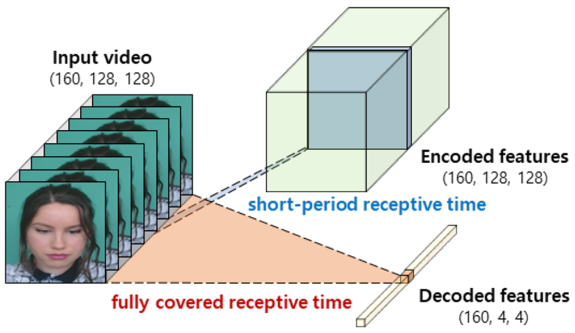

Figure 2. Comparison of receptive fields between encoded and decoded features in terms of a temporal perspective.

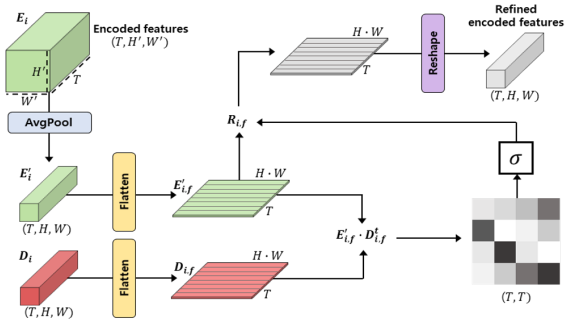

Figure 3. Schematic diagram of temporal attention refinement module.

the receptive time of the input video. The decoding process decompresses the resulting compressed features along the temporal dimension. Decoded features generated through the decompression capture the full receptive time of the input. Consequently, predicting an rPPG signal from the decoded features enables more accurate measurements of the rPPG signal by considering the entire time information, even when estimating individual sample values that constitute the rPPG signal.

However, decoding from the encoded features with significant loss of short-term information due to extreme compression is challenging. Skip connections compensate for this issue. The skip connections are performed by element-wise summation between the same-length encoded and decoded features in the symmetric hourglass structure (Fig. 1). The difference in the spatial size between the encoded and decoded features during the element-wise summation is resolved by applying spatial average pooling, which aligns the spatial dimensions of the encoded features with the decoded features. For instance, the features generated by Encoder 1 (Tab. 1), with a size of (16, 160, 128, 128), are first resized to (16, 160, 4, 4) using spatial average pooling and then combined with the decoded features, with the size of (16, 160, 4, 4), generated by the 3D transposed convolutional operations of Decoder 1 (Tab. 1) through element-wise summation.

3.2. Temporal Attention Refinement Module

The hourglass structure of LSTC-rPPG generates the encoded features that capture short-term information and the decoded features that capture long-term information (Fig. 2), leading to temporal information mismatches between these features. This observation suggests that the proposed skip connections, which perform the summation of individual elements between the encoded and decoded features, can be improved by addressing the temporal information mismatches. To address the mismatches, we propose a temporal attention refinement module (Fig. 3), abbreviated as TARM, which calculates temporal correlations between the encoded and decoded features and utilizes the resulting correlations to refine the encoded features.

The schematic diagram of TARM is illustrated in Fig. 3. Let $E_i \in \mathbb{R}^{(T\times H'\times W')}$ and $D_i \in \mathbb{R}^{(T\times H\times W)}$ denote the i-th channel of encoded and decoded features used as inputs of the skip connections. Here, T denotes the temporal dimension of both E_i and D_i. H' and W' represent the height and width of E_i, respectively, while H and W represent the height and width of D_i, respectively. Regarding E_i and D_i, TARM performs the following steps. First, the difference in spatial dimensions between E_i and D_i is aligned through spatial average pooling as follows:

$$E_i^{'} = \theta(E_i) \in \mathbb{R}^{T\times H\times W}, \tag{1}$$

where $\theta(\cdot)$ means the application of the spatial adaptive average pooling layer. Second, $E_i^{'}$ and D_i are reshaped from 3D feature tensors to 2D matrices as follows:

$$E_{i,f}^{'} = \phi(E_i^{'}) \in \mathbb{R}^{T\times H\cdot W}, \tag{2}$$

$$D_{i,f} = \phi(D_i) \in \mathbb{R}^{T\times H\cdot W}, \tag{3}$$

where $\phi(\cdot)$ is a function that flattens a 3D input into 2D output. Third, $E_{i,f}^{'}$ is temporally refined based on the following formula:

$$R_{i,f} = \sigma(E_{i,f}^{'} \times D_{i,f}^{t}) \times E_{i,f}^{'} \in \mathbb{R}^{T\times H\cdot W}, \tag{4}$$

where $\sigma(\cdot)$ is the SoftMax function and superscript t denotes the transpose of a matrix. Finally, $R_{i,f}$ is reshaped into the original 3D feature tensor, and the i-th channel of temporal attention refined features can be represented as:

$$R_i = \phi^{-1}(R_{i,f}) \in \mathbb{R}^{T\times H\times W}, \tag{5}$$

where $\phi^{-1}(\cdot)$ is the inversion of $\phi(\cdot)$.

TARM addresses the issue of temporal information misalignment that can occur during the skip connections between the encoded and decoded features, from the perspective of considering the temporal receptive field. Furthermore, from the perspective of the attention mechanism wherein the encoded features serve as queries, the decoded features serve as keys, and the encoded features serve again as values, TARM offers another advantage. Specifically, TARM enables the selective concentration of more relevant features at a variety of temporal positions within the input sequence. For example, in an rPPG signal, TARM helps features corresponding to a certain peak timing to intensively refer to features corresponding to other peak timings.

3.3. Frequency scale invariant hybrid loss

Time-domain losses and frequency-domain losses impose distinct constraints when supervising deep neural models for rPPG [39, 42]. The time-domain losses predominantly enforce instantaneous constraints such as the amplitude of target rPPG signals [5, 40], whereas the frequency-domain losses tend to guide periodic features of the signals across their entirety, such as their frequency spectrum [23]. In this regard, LSTC-rPPG, which utilizes long-term as well as short-term features, can be expected to benefit from more effective guidance by jointly utilizing time-domain and frequency-domain losses.

Motivated by this, we propose a novel hybrid loss. The proposed hybrid loss is made up of a time-domain loss and a frequency-domain loss as follows:

$$L_{hybrid} = \alpha \cdot L_{time} + \beta \cdot L_{freq}, \tag{6}$$

where α and β are the weight factors for balancing the losses and are equal to 1.0 and 0.5, respectively. The mean square error between the estimated and ground-truth rPPG signal is adopted for L_{time}. L_{freq} employs the scale invariant error [11] and is defined as follows:

$$L_{freq} = \frac{1}{n}\sum_{i=1}^{n} d_i^2 - \frac{\lambda}{n^2}\left(\sum_{i=1}^{n} d_i\right)^2, \tag{7}$$

$$d_i = \log \hat{p}_i - \log p_i, \tag{8}$$

where the power spectral density (PSD) of the estimated rPPG signal and the PSD of the ground-truth rPPG signal are denoted by $\hat{p}_i$ and p_i, respectively, both having length n indexed by i and the hyperparameter λ is 0.2.

There have been attempts to improve the performance of deep neural models for rPPG by defining losses in the frequency domain [23, 39, 42]. However, the proposed hybrid loss differs in that it aims to utilize the relative magnitude in the frequency domain for learning. The proposed frequency-domain loss helps measure the relationships between powers in PSD, irrespective of the absolute global scale. Inferring the relative magnitude of PSD is a less complex task than predicting the absolute magnitude of PSD, which makes it easier for deep neural models to learn. Furthermore, for the task of inferring HR, which can correspond to the dominant frequency of PSD, the frequency scale invariant loss is suitable for reducing errors in the performance metrics.

4. Experiments

First, a benchmark dataset is introduced in Sec. 4.1, followed by the descriptions of experimental implementation details and performance metrics in Sec. 4.2. Subsequently, in Sec. 4.3, we compare the proposed method with previous methods. Section 4.4 provides the visualizations of experimental results, while the effectiveness of each component consisting of the proposed method is represented in Sec. 4.5 through ablation studies.

4.1. Dataset

UBFC-rPPG [2] is a dataset for rPPG analysis and includes 42 videos along with corresponding ground-truth rPPG signals and HRs from 42 subjects. The UBFC-rPPG dataset was generated through the utilization of customized C++ software for capturing videos, utilizing a cost-effective and uncomplicated webcam (Logitech C920 HD Pro) at a frame rate of 30 frames per second, and capturing at a resolution of 640 $\times$ 480 in uncompressed 8-bit RGB format. To acquire PPG signals and HRs as the ground truth of the captured videos, a CMS50E transmissive pulse oximeter was utilized. During the recording process, there was a varying amount of sunlight and indoor illumination. The subjects were seated approximately 1 meter away from the camera, with their face fully visible. They were instructed to engage in a time-sensitive mathematical game intended to raise their HR, intending to simulate a natural human–computer interaction scenario.

4.2. Implementation Details and Metrics

Following previous studies [17, 42], we used RGB video clips of size 128 $\times$ 128 pixels with a frame length of 160 as the input for the proposed method. In the initial stage of our experiments, we detected a facial region within the first frame of each video in the UBFC-rPPG dataset using MTCNN [43]. Subsequently, we fixed and cropped facial ROIs that were 1.6 times larger than the detected facial region across each video and then resized the facial ROIs into 128 $\times$ 128 pixels. Each resized video was divided into clips of 160 frames. Based on the criteria from previous studies [8, 17], we partitioned the set of 42 videos into two subsets, comprising 28 and 14 videos for training and testing, respectively. For the training data, video clips were created by shifting 30 frames, resulting in an overlap of 130 frames between adjacent clips without any augmentation.

Table 2. Performance comparison with previous methods.

Method	MAE ↓	RMSE ↓	r ↑	Data aug.
CHROM [9]	3.44	4.61	0.97	×
POS [36]	2.44	6.61	0.94	×
DeepPhys [5]	2.90	3.63	0.98	○
PhysNet [40]	2.95	3.67	0.98	○
AND-rPPG [19]	2.67	4.07	0.96	○
TDM [8]	2.32	3.08	0.99	○
Physformer [42]	1.36	2.41	-	○
PhysNet+PFE+TFA [17]	0.76	1.62	-	○
Ours	0.70	1.00	0.99	×

Table 3. Computational complexity analysis.

Method	# Params. ↓	MACs ↓
PhysNet [40]	0.77 M	70.21 G
Physformer [42]	7.03 M	47.01 G
PhysNet+PFE+TFA* [17]	1.34 M	46.34 G
Ours	0.91 M	28.62 G

∗ We attempted to reproduce the number of parameters and MACs using the code available at https://github.com/LJW-GIT/Arbitrary_Resolution_rPPG, which resulted in a parameter count of 2.24 M and MACs count of 169.82 G. Nonetheless, we utilized the values reported by Li et al. [17].

For the testing data, video clips were created without any overlap. The proposed model was trained on PyTorch and two NVIDIA RTX 3090 GPUs. The Adam optimizer was used with a learning rate of 5e-5.

To validate the proposed method, we adopted the most commonly used performance metrics for rPPG evaluation, including the mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (r) between the ground-truth HR and the estimated HR through predicted signals. The estimated HR was computed by finding a dominant frequency using PSD of the estimated rPPG signal. Similar to previous studies [5, 8], a band-pass filter with cutoff frequencies of 0.7–4 Hz was applied to the estimated rPPG signal before HR calculations. In addition, we utilized the number of parameters and multiply-accumulate operations (MACs) using a PyTorch-based counting tool [29] to assess the computational efficiency of the proposed method.

4.3. Comparison with previous methods

Table 2 presents the performance comparison between the proposed method and previous methods [5, 8, 9, 17, 19, 36, 40, 42] on the UBFC-rPPG dataset. The two traditional methods (CHROM, POS) achieved lower overall performance compared to the deep learning-based methods (DeepPhys, PhysNet, AND-rPPG, TDM, Physformer, and PhysNet+PFE+TFA). DeepPhys based on vanilla 2DCNNs and PhyNet based on vanilla 3DCNNs showed improved stability by presenting comparable MAE and lower RMSE compared to the two traditional methods. Physformer, which leverages the transformer architecture for considering long-term features, presented notable advancements in performance relative to the previous methods (CHROM, POS, DeepPhys, PhysNet, AND-rPPG, and TDM), demonstrating that global features are a crucial factor for reliable measurements of rPPG signals. Our proposed method exhibited superior performance across all performance metrics compared to all other methods, achieving an MAE of 0.70, an RMSE of 1.00, and an r of 0.99.

When compared to the latest state-of-the-art method (PhysNet+PFE+TFA), our proposed method showed a slight improvement in MAE by 0.06, but a substantial improvement was observed in RMSE, which decreased to 1.00. These results demonstrated the enhanced stability of the proposed method. In addition, compared to Physformer, our proposed method achieved better performance even without any data augmentation. This demonstrated that our proposed method effectively utilizes long-term as well as short-term features in a more appropriate manner, even with less data. Furthermore, in contrast to all other deep learning-based methods, the performance of our proposed method was achieved without even performing data augmentation. This observation suggests that our proposed method can be more stable and generalizable, leading to more reliable and consistent performance in practical scenarios.

In order to analyze the computational efficiency of the proposed method, the number of parameters and MACs were calculated and compared with the previous methods [17, 40, 42] in Tab. 3. Regarding the number of parameters, PhysNet demonstrated the best performance with 0.77 M, while for MACs, the proposed method showed the best performance with 28.62 G. Our proposed method is based on 3DCNNs, similar to PhysNet. However, compared to PhysNet, our proposed method demonstrated a performance improvement of approximately 59.24% in terms of MACs, while having slightly more parameters. The performance improvement observed in MACs can be attributed to the compression of features in spatial as well as temporal dimensions during the encoding process, as well as the preservation of the compressed feature size for spatial dimensions during the decoding process. Furthermore, our proposed method outperformed both Physformer and PhysNet+PFE+TFA in terms of the number of parameters and MACs. Compared to Physformer, the proposed method exhibited a reduction of approximately 87.06% in the number of parameters and a 39.12% reduction in MACs. Compared to PhysNet+PFE+TFA, the proposed method demonstrated a decrease of approximately 32.09% in the number of pa-

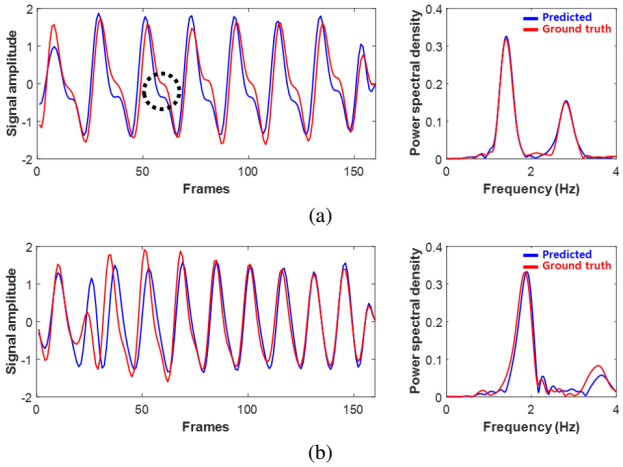

Figure 4. Comparison of predicted rPPG signal and PSD (blue) with ground-truth rPPG signal and PSD (red) for two video clips in testing data: (a) one from Subject 42 and (b) another from Subject 47.

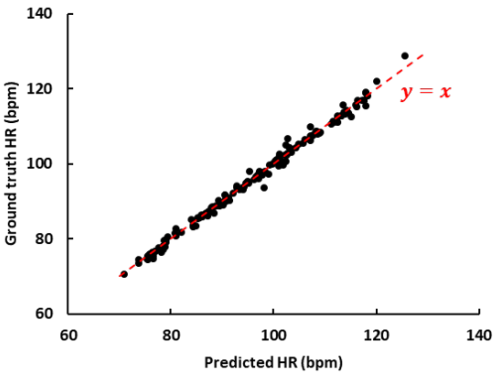

Figure 5. Comparison of predicted HRs with ground-truth HRs for testing data.

rameters and a 38.24% reduction in MACs.

When considering the trade-off between computational complexity and performance metrics (Tab. 2, Tab. 3), the proposed method had slightly more parameters than PhysNet but showed notable performance improvements in MACs and MAE. Furthermore, for Physformer and PhysNet+PFE+TFA the proposed methods presented better performance across all performance metrics, the number of parameters, and MACs.

4.4. Visualization of results

We visualized both the predicted rPPG signals and their PSDs, as well as the ground-truth signals and their PSDs on two video clips in testing data in Fig. 4. The similarity between the predicted and ground-truth rPPG signals and PSDs shown in Fig. 4a demonstrated the effectiveness of our proposed method. Furthermore, detecting dicrotic notch and diastolic peaks, as shown by the black dashed circle (Fig. 4a), further reinforced the effectiveness. In Fig. 4b, sudden changes in HR were observed in the early part of the ground-truth signal. Despite the sudden changes in HR, the proposed method demonstrated a comparable prediction of the rPPG signal. Furthermore, the observation of the similarity between the PSDs of predicted and ground-truth signals in Fig. 4a and Fig. 4b indicated that the frequency scale invariant loss effectively guided the learning in the frequency domain.

Figure 5 displays the scatter plot of the predicted and ground-truth HRs for the testing data, wherein the x-axis represents the predicted HRs and the y-axis represents the ground-truth HRs. The overall alignment of the scatter plot with the $y = x$ line was observed. Such alignment was maintained even at relatively low and high HRs. In a deep learning-based approach, achieving such alignment is not an easy task owing to the data imbalance where the HR distribution of the training data is heavily concentrated in specific beats per minute range, leading to potential overfitting on this range. However, our proposed method could accurately predict the wide distribution of HRs even in the absence of HR distribution data augmentation.

4.5. Ablation studies

To validate the effectiveness of TARM and the use of frequency scale invariant loss, ablation studies were conducted. The results of ablation studies are summarized in Tab. 4, where Siloss denotes the frequency-domain scale invariant loss, TARM denotes the temporal attention refinement module, and the symbol ✓signifies whether to apply Siloss and TARM or not. The employment of Siloss and TARM individually resulted in a minor improvement in MAE compared to the baseline architecture. In contrast, the utilization of both Siloss and TARM together showed a substantial increase in MAE. It can be assumed that the use of TARM for the optimization of long short-term features, combined with the frequency scale invariant hybrid loss to guide their learning, may have resulted in a synergistic effect that led to improved performance.

Table 4. Effectiveness of Siloss and TARM.

Siloss	TARM	MAE ↓	RMSE ↓	r
		0.78	1.1	0.99
✓		0.75	1.0	0.99
	✓	0.74	1.1	0.99
✓	✓	0.70	1.0	0.99

5. Conclusion and Future Work

This paper proposed a deep learning-based method to accurately and reliably measure rPPG signals from facial videos. In Sec. 3.1, a novel baseline architecture, denoted as LSTC-rPPG, was proposed with the objective of modeling long short-term spatio-temporal features and integrating the long short-term features for improved measurement of rPPG signals. In Sec. 3.2, a temporal attention refinement module, referred to as TARM, was proposed to address the temporal mismatches between the long-term and short-term features. Furthermore, in Sec. 3.3, a frequency scale invariant hybrid loss was proposed. The frequency scale invariant hybrid loss helped LSTC-rPPG consider long short-term information of target rPPG signals and also made it easier to learn frequency-domain characteristics through scale invariance. To verify the proposed method, experiments were conducted on the UBFC-rPPG dataset. Our proposed method demonstrated state-of-the-art performance showing an MAE of 0.70, RMSE of 1.00, and r of 0.99.

Future studies will focus on more efficient and improved architecture as well as intra-dataset and cross-dataset experiments on a larger variety of datasets to supplement our evaluation on a limited dataset.

Acknowledgement

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2022 (Project Name: Development of a functional content platform for stress relief, Project Number: R2020060003, Contribution Rate: 100%)

References

[1] Serge Bobbia, Yannick Benezeth, and Julien Dubois. Remote photoplethysmography based on implicit living skin tissue segmentation. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 361–365. IEEE, 2016. 2

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 5

[3] Deivid Botina-Monsalve, Yannick Benezeth, and Johel Miteran. Rtrppg: An ultra light 3dcnn for real-time remote photoplethysmography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2146–2154, 2022. 1, 2

[4] Sitthichok Chaichulee, Mauricio Villarroel, Joao Jorge, Carlos Arteta, Gabrielle Green, Kenny McCormick, Andrew Zisserman, and Lionel Tarassenko. Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 266–272. IEEE, 2017. 2

[5] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 2, 5, 6

[6] Xun Chen, Juan Cheng, Rencheng Song, Yu Liu, Rabab Ward, and Z Jane Wang. Video-based heart rate measurement: Recent advances and future prospects. *IEEE Transactions on Instrumentation and Measurement*, 68(10):3600–3615, 2018. 1

[7] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 3

[8] Joaquim Comas, Adria Ruiz, and Federico Sukno. Efficient remote photoplethysmography with temporal derivative modules and time-shift invariant loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2182–2191, 2022. 5, 6

[9] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 2, 6

[10] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 2

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 5

[12] Litong Feng, Lai-Man Po, Xuyuan Xu, Yuming Li, Chun-Ho Cheung, Kwok-Wai Cheung, and Fang Yuan. Dynamic roi based on k-means for remote photoplethysmography. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1310–1314. IEEE, 2015. 2

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 3

[14] Jiaqi Kang, Su Yang, and Weishan Zhang. Transppg: Two-stream transformer for remote heart rate estimate. *arXiv preprint arXiv:2201.10873*, 2022. 2, 3

[15] Antony Lam and Yoshinori Kuno. Robust heart rate measurement from video using select random patches. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3640–3648, 2015. 2

[16] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 392–409. Springer, 2020. 1, 2

[17] Jianwei Li, Zitong Yu, and Jingang Shi. Learning motion-robust remote photoplethysmography through arbitrary resolution videos. *arXiv preprint arXiv:2211.16922*, 2022. 5, 6

[18] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014. 2

[19] Birla Lokendra and Gupta Puneet. And-rppg: A novel denoising-rppg network for improving remote heart rate estimation. *Computers in biology and medicine*, 141:105146, 2022. 6

[20] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering*, 61(10):2593–2601, 2014. 2

[21] Daniel J McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W Picard. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4000–4004, 2016. 1

[22] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 2

[23] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 295–310. Springer, 2020. 2, 5

[24] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4955–4964, 2021. 1, 2

[25] Byoung-Jun Park, Eun-Hye Jang, Myung-Ae Chung, and Sang-Hyeob Kim. Design of prototype-based emotion recognizer using physiological signals. *ETRI Journal*, 35(5):869–879, 2013. 1

[26] Olga Perepelkina, Mikhail Artemyev, Marina Churikova, and Mikhail Grinenko. Hearttrack: Convolutional neural network for remote video-based heart rate monitoring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 288–289, 2020. 1, 2

[27] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 2

[28] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 2

[29] Vladislav Sovrasov. ptflops: a flops counting tool for neural networks in pytorch framework, 2018. 6

[30] Yu Sun, Sijung Hu, Vicente Azorin-Peris, Stephen Greenwald, Jonathon Chambers, and Yisheng Zhu. Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise. *Journal of biomedical optics*, 16(7):077010–077010, 2011. 2

[31] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007. 2

[32] Chuanxiang Tang, Jiwu Lu, and Jie Liu. Non-contact heart rate monitoring by combining convolutional neural network skin detection and remote photoplethysmography via a low-cost camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1309–1315, 2018. 2

[33] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016. 2

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[35] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 1, 2

[36] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 2, 6

[37] Wenjin Wang, Sander Stuijk, and Gerard De Haan. Exploiting spatial redundancy of image sensor for motion robust rppg. *IEEE transactions on Biomedical Engineering*, 62(2):415–425, 2014. 2

[38] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on biomedical engineering*, 63(9):1974–1984, 2015. 2

[39] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, 27:1245–1249, 2020. 5

[40] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 1, 2, 5, 6

[41] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38(6):50–58, 2021. 1

[42] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4186–4196, 2022. 2, 3, 5, 6

[43] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 5