RTrPPG: An Ultra Light 3DCNN for Real-Time Remote Photoplethysmography

D. Botina-Monsalve, Y. Benezeth, J. Miteran
Univ. Bourgogne Franche-Comté
ImViA EA7535, France
Deivid-Johan.Botina-Monsalve@u-bourgogne.fr

Abstract

The acquisition of remote photoplethysmography (rPPG) signals is important in multiple applications. Recently, deep-learning-based approaches such as 3D convolutional networks (3DCNNs) have outperformed traditional hand-crafted methods. However, despite their robust modeling ability, it is well known that large 3DCNN models have high computational costs and may be unsuitable for real-time applications. In this paper, we propose a study of the 3DCNN architecture, finding the best compromise between heart rate measurement precision and inference time. The fast inference is obtained decreasing the input size while the precision performance is obtained introducing a new time and frequency-based loss function by adding the signal-to-noise-ratio component to the regular Pearson's correlation loss function. In addition, changing the input color space from RGB to YUV also improved heart rate measurement precision. Using the VIPL-HR database, we retained the HR mean absolute error at 3.99 bpm which is comparable to 3.87 bpm of the state-of-the-art, while the GPU and CPU inference process improved around 88% from 51.77 ms to 2.32 ms in GPU and from 241.57 ms to 28.65 ms in CPU. The resulting network is called Real-Time rPPG (RTrPPG). We release the RTrPPG source code to encourage reproducibility[1].

1. Introduction

Heart rate (HR) and pulse rate variability (PRV) are two physiological parameters that allow the analysis of cardiac behavior. Heart rate monitoring can be conducted by invasive and non-invasive methods classified as contact-based and non-contact-based. Two non-invasive techniques commonly used to measure HR and PRV are electrocardiography (ECG) and photoplethysmography (PPG). ECG measures the electrical field caused by heart activity. On the other hand, PPG measures variations in light absorption in tissues due to the pulsatile nature of the cardiovascular system and the variation in blood volume [14]. PPG and ECG perform contact-based HR measurements, and they may cause hygiene issues, discomfort, or even be unrealizable on fragile skins. Due to these possible drawbacks, in [27], Verkruysse *et al.* proved that PPG signals could be measured remotely from a standard video camera, using ambient light as an illumination source. This technique, known as remote photoplethysmography, offers the advantage of measuring the same parameters as PPG in an entirely remote way. In fact, rPPG is the non-contact equivalent to the reflective mode of PPG using a camera as a receptor and ambient light as a source. Thus, blood volume changes are estimated according to subtle skin color variations, which are captured by the camera when lights are reflected by the skin.

PPG and rPPG signals allow measuring several biomedical parameters, such as heart rate, pulse rate variability, vascular occlusion, peripheral vasomotor activity, blood pressure by pulse transit time, and breathing rate [1]. Therefore, there are also multiple applications, including blood pressure prediction [24], mixed reality [12], physiological measurements of car drivers [37], living skin segmentation [30], face anti-spoofing [35], and control of vital signs in newborns [6].

Like Verkruysse *et al.*, early methods used the green channel to estimate rPPG signals [27]. Then, approaches based on a light tissue interaction model to determine a projection vector were proposed, *e.g.* PbV, POS, and Chrom [8, 9, 28], and others based on blind source separation techniques, *e.g.* PCA, ICA, EVM, PVM, WVM [13, 15, 17–19, 31]. Recently, deep-learning models have started to be used for physiological measurements from video sequences [5, 10, 11, 16, 21, 22, 26, 34]. The main advantages of these methods are that they allow achieving good results without the need for the designer to analyze the problem in-depth [36]. The hand-crafted-based pipeline needs to detect and track the region of interest through the frames, combine color channels, filter them and estimate the physiological parameters such as respiration rate or heart

[1] https://github.com/deividbotina-alv/rtrppg

rate. Alternatively, a pipeline-based framework is no longer necessary in the deep-learning-based measurement. Therefore, deep-learning-based approaches are less prone to error propagation in their pipeline. Nevertheless, recent work has focused on heart rate measurement performance rather than understanding [36].

Ablation studies are helpful because they provide insights into the relative contribution of different architectural and regularization components to machine learning and deep-learning performance [25, 36]. For example, in [25], the authors propose a series of experiments that evaluate the importance of the frame rate. The time and frequency domains evaluation suggests that decreasing the frame rate may lead to better network performance due to the increased length of time that a spatio-temporal kernel covers. In [36], the importance of the spatial context is studied in two-dimensional neural networks (2DCNNs). Results suggest that different resolutions cause minor fluctuations in network performance. However, whether this conclusion is valid in a 3D convolutional network is unclear. Some authors have proposed to use channels other than Red, Green, and Blue (RGB). We can find rPPG methods where authors use color channels such as Lab [4], Luv [4], or YCbCr [21]. Interestingly, in deep-learning-based rPPG measurement, the YUV color space has shown promising results [5, 21, 22].

2DCNNs are of great importance when measuring rPPG signals. They have been used to measure rPPG, HR, BR, and PRV [7, 21, 26, 31]. Nevertheless, it is necessary to perform an additional procedure where the temporal context is taken into account, increasing the computation time and making it harder to implement in an end-to-end fashion. Therefore, 2DCNN-based rPPG measurement approaches may be unsuitable in a real-time context. Note that real-time capability typically refers to when a model runs faster than a webcam at 30 fps (33.3 ms).

Alternatively, three-dimensional convolutional neural networks (3DCNNs) can analyze both spatial and temporal characteristics of a video simultaneously. For this reason, the use of 3DCNNs may be more convenient than 2DCNNs for an end-to-end application. For instance, perhaps one of the most iconic 3DCNN is PhysNet, proposed by Zitong Yu *et al*. in [33], as it has been widely exploited in other studies [10, 25, 32]. The authors make a performance comparison of spatio-temporal networks using 2DCNN+LSTM and 3DCNN. The 3DCNN outperformed the combination of 2DCNN and recurrent networks. With this method, it is possible to acquire rPPG signals directly from video. Fig. 1 depicts the difference between a general rPPG framework based on 2D and 3D CNNs.

In recent years, methods based on 3DCNNs have demonstrated promising results measuring rPPG signals and HR [32–34]. In this article, we build upon previously proposed architectures while focusing on optimizing their inference speed for real-time applications (potentially on low-end devices). Optimizing the inference time can be approached systematically through an ablation study, where various network components such as the size and color space of the input images, as well as the loss function are evaluated. To the best of our knowledge, this is the first work where an ablation study is performed on a 3DCNN in the rPPG task to optimize network response time, signal quality, and heart rate measurement precision.

The main contributions of this work, obtained using an ablation study where we tuned network size, loss function and color space, are :

- A new 3DCNN called Real-Time rPPG (RTrPPG). It achieves results comparable to those found in the literature, acquiring rPPG signals from real-time videos. The inference time is around 2.32 ms on GPU and 28.65 ms on CPU.
- A new temporal-frequency-based loss function that allows the 3DCNN to learn the essential features of the rPPG signal acquisition task. Our loss outperforms the baseline temporal-based loss function.

The remainder of this article is organized as follows: In Sec. 2, we show related works. Sec. 3 presents our spatio-temporal neural network, a new temporal-frequency-based loss function, and the proposed ablation study. In Sec. 4 the metrics used to measure network performance are introduced. In Sec. 5, we present the parameters used to conduct our experiments. Then, results are presented in Sec. 6. Finally, in Sec. 7, we conclude the work done in this article.

2. Related works

As explained previously, 2D and 3D CNNs have been used in rPPG measurement and HR acquisition, in [26] for example, the HR-CNN network is proposed. This 2DCNN has *Extractor* and *HR Estimator* modules. First, a video is taken to detect and resize the faces, and then, the cropped video is passed through the *Extractor* to acquire its rPPG component. Finally, the rPPG is the input in the *HR Estimator*, and its HR is the output. Nevertheless, the model does not include temporal reasoning within the network.

Other 2DCNN-based frameworks have proposed an additional process on the input images to consider the temporal context before using the 2DCNN, and this is the case of DeepPhys [7] and EVM-CNN [31]. DeepPhys is a two-branch model comprising the *Motion model* and *Appearance model*. Since a 2DCNN lacks the ability to process temporal cues, the authors propose to normalize the difference of two consecutive frames as input to the *Motion model*. The *Appearance model* behaves as an attention module. Using the two branches together makes acquiring rPPG

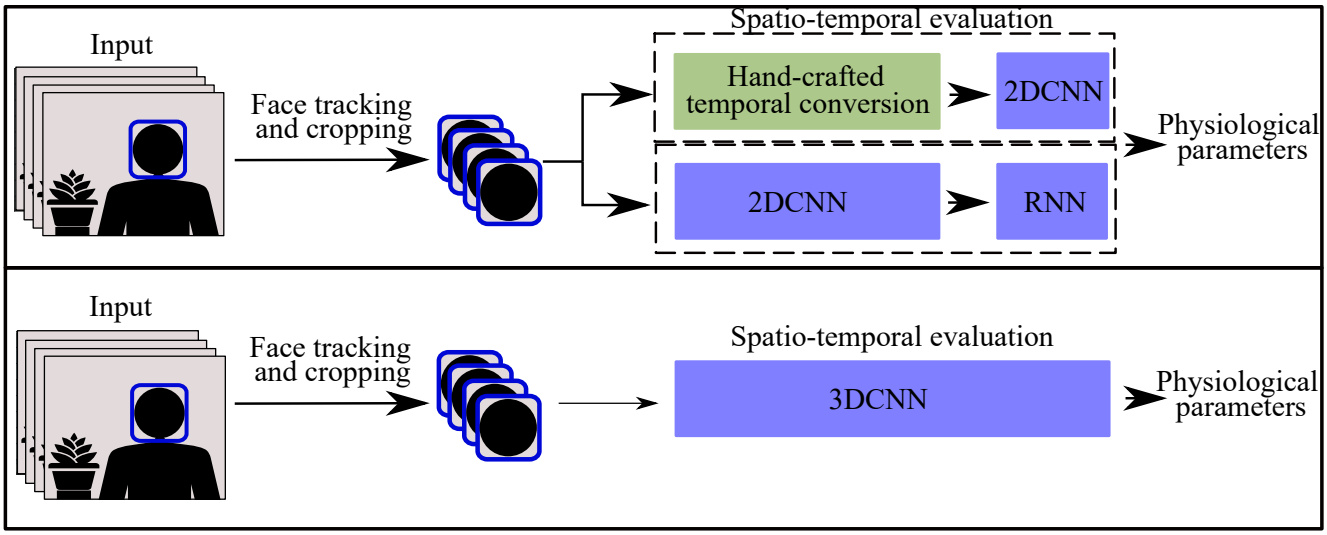

Figure 1. 2DCNN vs 3DCNN rPPG frameworks. 2DCNN frameworks need an additional process to consider the temporal characteristics of the videos. 3DCNNs, on the other hand, allow an end-to-end approach that is easier to implement.

and BR signals possible. In EVM-CNN, the authors take an input video and detect the subject's face on each frame. The face-regions are further processed by spatial decomposition and temporal filtering, generating a *feature image*. This *feature image* is the input to the EVM-CNN network, which gives an output HR value.

To use 2DCNNs while considering the temporal context, it is also possible to use a recurrent neural network at the end of the framework. Bin Huang *et al*. [11], for example, used a network composed of two parts. The first part uses 2D convolutional layers for spatial analysis, and the second one contains a series of stacked long-short-term-memory (LSTM) layers . The resulting LSTM block allows temporal analysis. Thus, this framework can acquire HR from videos.

Finally, it is also possible to combine a spatio-temporal image with 2DCNNs and RNNs, as in the case of RhythmNet [21], where the combined model can measure HR from videos by three main parts. The first part detects the facial landmarks to find the face and divides it into 25 Regions of interest (ROIs). Then, the average value of each color channel in each ROI is calculated. Finally, a sequence is generated from each ROI. The same procedure is repeated for all video frames, resulting in a spatio-temporal map. The second part is a 2DCNN, and the third one is an RNN.

Interestingly, 3DCNNs have also been used to acquire rPPG signals and HR [5,10,32–34]. In [5], a pilot model for measuring pulse rate using 3D convolutional is presented. The CNN acts as an extractor of spatial and temporal features from the input video frames. More specifically, the authors demonstrate the potential of training the network on synthetic videos.

In [34], the authors used two 3DCNNs. The first CNN is a spatio-temporal video enhancement network (STVEN), and the second is called rPPGNet. STVEN is responsible for increasing the resolution of an input video, which is especially useful for highly compressed videos. The rPPGNet is composed of a skin-based attention module that helps to adaptively select skin regions, a partition constraint module that learns a better representation of the rPPG signal features, and a spatio-temporal CNN. The input is the resized face of the subject present in each frame, and the output is an rPPG signal where HR and PRV are measured.

Gideon and Stent in [10] present a contrastive approach where they acquire the cardiac activity of a person from the video of his face. They use a modified version of the 3DCNN-based PhysNet architecture to learn spatio-temporal features over the input video. Interestingly, this is the first approach that allows the acquisition of rPPG signals in a self-supervised way. Moreover, they propose a saliency sampler to obtain an interpretable output to ensure that the system behaves correctly.

3. Methodology

In this work we use an encoder-decoder neural network based on 3DCNNs as a baseline. We propose an ablation study to improve inference speed while maintaining accuracy. We tune image size and color space, and introduce a new temporal-frequency-based loss function.

3.1. Spatio-temporal network

The system input is a series of T-frame images of any three-dimensional color space ($[i_1, i_2, ..., i_T]$). To use only the information related to the skin of the face, we use a neu-

ral network (denoted as Φ) in charge of extracting the face of the subjects found within each frame. Then, we use a resizing procedure denoted as Ω in Eq. (1), in order to have a square image of dimensions bxb. The overall procedure is presented in Eq. (1):

$$[f_1, f_2, ..., f_T] = \Omega(\Phi([i_1, i_2, ..., i_T], \varphi), \omega), \quad (1)$$

where $[f_1, f_2, ..., f_T]$ are the bxb T-frame face images after being resized, φ are the parameters of Φ, and ω is the interpolation process used by Ω.

Inspired by the spatio-temporal network implemented in [33], in this paper, we propose a 3DCNN-Encoder-Decoder denoted as 3DED as a baseline to find the rPPG signals associated with a video. This network is divided in two main parts. The first one is the encoder E where the input data is transformed in a latent-space with more significant spatio-temporal information. The second part, receiving the latent-space feature as an input, is the decoder D that generates the rPPG output $\mathbf{y} = [y_1, y_2, ..., y_T]$. E and D are feed-forward 3DCNNs. The rPPG estimation by the 3DED neural network procedure is presented in Eq. (2):

$$[y_1, y_2, ..., y_T] = \text{3DED}([f_1, f_2, ..., f_T]); \theta), \quad (2)$$

where θ represents the parameters of 3DED.

3.2. Time-frequency based loss function

Pearson's correlation coefficient (ρ) can measure the linear relationship between the temporal characteristics of rPPG and the blood volume pulse ground truth (PPG signal), ignoring the frequency-based characteristics. On the other hand, the frequency domain contains the components related to heart rate and signal quality; therefore, the Signal-to-Noise-Ratio (SNR) can enhance the frequency-based components. Consequently, we use ρ and SNR to optimize the most important characteristics of the rPPG signals. In Eq. (3), we propose the new temporal-frequency-based loss function Negative Pearson's correlation and Signal-to-Noise Ratio (NPSNR) that unites both metrics:

$$\text{NPSNR} = 1 - (\rho + \lambda \text{SNR}), \quad (3)$$

where λ is a constant that balances the frequency component, ρ is the Pearson's correlation between the rPPG signal and its ground truth, $\mathbf{y}$ and $\mathbf{g}$, respectively (Eq. (4)):

$$\rho = \frac{\sum_{j=1}^{T} (\mathrm{y_j} - \overline{y})(\mathrm{g_j} - \overline{g})}{\sqrt{\sum_{j=1}^{T} (\mathrm{y_j} - \overline{y})^2} \sqrt{\sum_{j=1}^{T} (\mathrm{g_j} - \overline{g})^2}}. \quad (4)$$

SNR is the ratio between the rPPG signal power $\bar{P}_{\mathrm{y,signal}}$ and the rPPG signal background $\bar{P}_{\mathrm{y,noise}}$, as described in Eq. (5):

$$\text{SNR} = \frac{\bar{P}_{\mathrm{y,signal}}}{\bar{P}_{\mathrm{y,noise}}}, \quad (5)$$

where the average power is given by $\bar{P}$ [3].

3.3. Ablation study

In this section we propose several experiments to acquire the best compromise between real-time, signal quality and heart rate measurement precision.

In the first approach we gradually decrease the spatial dimensions of the input frames bxb into seven different input sizes d_c where $d_c = \frac{b}{2^c}; c \in [0, 1, ..., 6]$. Then, we propose to replace the temporal-based Negative Pearson's Correlation (NP) loss function with the temporal-frequency-based NPSNR loss function. Finally, we evaluate the performance by changing the RGB color space to Lab, Luv, YUV, and YCbCr.

Fig. 2 depicts the experiments proposed in the ablation study. To cope with decreasing input sizes, we changed the pooling layers while applying the same convolutional operations. These changes only happen in the E encoder. We will refer to the network configurations as 3DEDd_c-ColorChannel-Loss, *e.g.* 3DED8-RGB-NP is the 3DED network with input RGB 8x8 pixels and NP as loss function.

4. Metrics

Template Match Correlation (TMC) and Signal-to-Noise-Ratio (SNR) are used to evaluate the rPPG signal estimation quality. On the other hand, Mean Absolute Error (MAE) and Pearson's correlation coefficient r are used for the evaluation of the heart rate measurement precision. SNR, MAE and r were computed using a 15-second sliding window with a stepping of 0.5 seconds. SNR, TMC, and r are to be maximized, while MAE has to be minimized. MAE results are given in beats per minute (bpm), and decibels (dB) for SNR.

TMC is a coefficient for ECG/PPG signal quality assessment metric [23]. This metric is implemented by detecting the signal peaks and the median beat-to-beat interval of the full-length signals. Then, the pulses are extracted individually centered on their respective peak with a window width equal to the median beat-to-beat interval. A template is calculated as the average of all pulses. Finally, the TMC coefficient is computed as the average correlation of all the pulses with the template. $\mathrm{TMC} = 0$ means that the pulse shape of the signal is non-uniform, while $\mathrm{TMC} = 1$ indicates a perfect uniformity.

The mean absolute error was calculated as the window-wise mean of the heart rate calculated using the contact-based ground truth waveform obtained by pulse oximeter ($\mathbf{hc}$), and the heart rate calculated using the rPPG signal

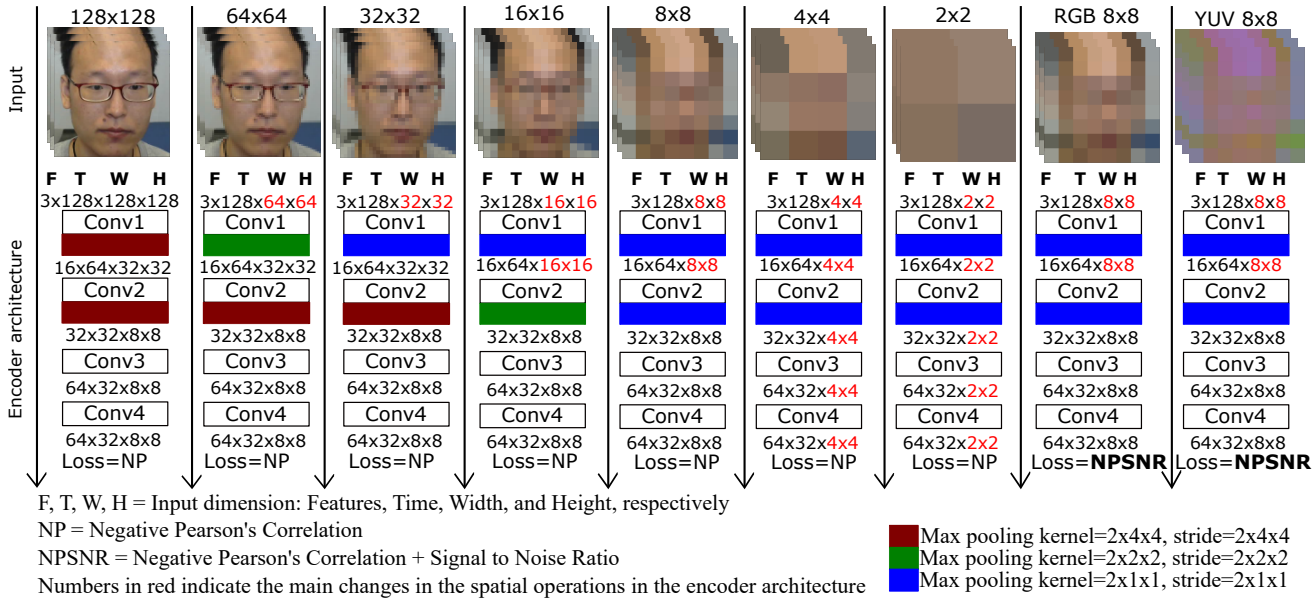

Figure 2. Ablation study. We use a 3DCNN baseline to gradually decrease the input resolution. We also change the loss function and the input color space.

(**hr**). The MAE of the two vectors **hr** and **hc** of size n is presented in Eq. (6):

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|\mathrm{hr_j} - \mathrm{hc_j}\right|, \tag{6}$$

where $\mathrm{hr_j}$ and $\mathrm{hc_j}$ are the value of **hr** and **hc** at position j, respectively.

Pearson's correlation coefficient measures the linear correlation between vectors **hc** and **hr**. $r = -1$ implies a negative linear correlation, while $r = 1$ means a positive total linear correlation, finally, $r = 0$ indicates that there is no linear correlation between the estimations and the reference values. r is given by Eq. (7):

$$r = \frac{\sum_{j=1}^{n}\left(\mathrm{hr_j} - \overline{hr}\right)\left(\mathrm{hc_j} - \overline{hc}\right)}{\sqrt{\sum_{j=1}^{n}\left(\mathrm{hr_j} - \overline{hr}\right)^2}\sqrt{\sum_{j=1}^{n}\left(\mathrm{hc_j} - \overline{hc}\right)^2}}, \tag{7}$$

where $\overline{hr}$ and $\overline{hc}$ are the averages of **hr** and **hc**, respectively.

5. Implementation details

VIPL-HR database: The research in this paper uses the VIPL-HR database collected by the Institute of Computing Technology Chinese Academy of Sciences [20, 21]. The database contains 107 subjects recorded by three different instruments in nine scenarios: stable, motion, talking, dark, bright, long distance, exercise, phone stable, and phone motion. Although this database also contains 752 near-infrared videos, we only consider the 2378 visible light videos. The resolutions of the videos are between 960x720 and 1920x1080, at 25 and 30 fps, respectively. The ground truth photoplethysmography signals were recorded using the CONTEC CMS60C BVP sensor at 60 Hz. Ground truth signals were down-sampled to the respective video sampling rate.

Ground truth adaptation: During the database acquisition, some ground truth signals present anomalies due to the movement of the subjects or failures in the acquisition devices. These inconsistencies (gaps and false peaks) usually happen at the beginning and end of the acquisitions (rarely during the acquisition). However, performing the network training with non-reliable ground truth signals is not ideal. Therefore, a ground truth selection step is necessary. For this purpose, we checked the ground truth signals individually to take only the continuous segment with a reliable ground truth morphology. To ensure reproducible results, we provide one file as supplementary material containing the information of these cropped signals.

The time lag between the blood volume pulse signal measured at the finger and the rPPG signal measured at the face can dramatically reduce heart rate measurement performance during training [36]. To align both signals, we calculated a reference rPPG signal with the POS method [28]. Subsequently, we aligned the ground truth signal to the reference rPPG signal. Finally, we normalized the ground truth between -1 and 1 after using a five-second moving-window average filter.

Model	GPU[ms]	CPU[ms]	MAE[b.p.m]	r	SNR[dB]	TMC	N.T.Param
PhysNet-RGB-NP	51.77	816.47	**3.87**	**0.73**	**5**	**0.91**	768,577
3DED128-RGB-NP	20.38	240.57	6.32	0.53	2.7	0.86	213,633
3DED64-RGB-NP	17.78	79.57	5.13	0.63	3	0.86	213,633
3DED32-RGB-NP	7.5	55.2	5.12	0.65	3	0.86	213,633
3DED16-RGB-NP	3.53	39.75	5.56	0.6	3	0.86	213,633
3DED8-RGB-NP	2.32	28.65	5.12	0.63	3.1	0.86	213,633
3DED4-RGB-NP	1.96	9.91	5.93	0.57	2.7	0.86	213,633
3DED2-RGB-NP	1.96	3.96	7.83	0.45	1.5	0.82	213,633
3DED8-RGB-NPSNR	2.32	28.65	4.37	0.68	4.2	0.88	213,633
3DED8-Lab-NPSNR	2.32	28.65	5.51	0.59	3.5	0.87	213,633
3DED8-Luv-NPSNR	2.32	28.65	4.04	**0.73**	4.4	0.88	213,633
3DED8-YCbCr-NPSNR	2.32	28.65	4.43	0.68	4.0	0.88	213,633
3DED8-YUV-NPSNR (RTrPPG)	**2.32**	**28.65**	3.99	**0.73**	4.6	0.89	213,633

Table 1. Ablation study results. 3DED8-YUV-NPSNR has the best overall configuration.

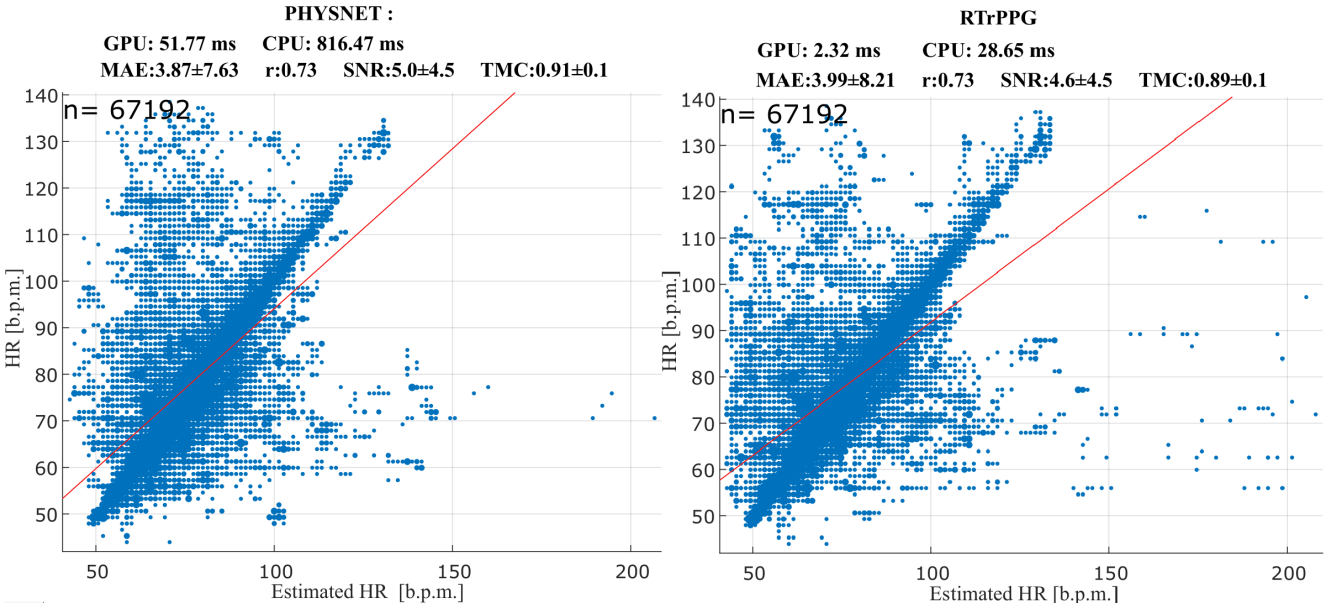

Figure 3. Correlation plot. A comparison between results given by the state-of-the-art network PhysNet and the Real-Time rPPG: RTrPPG.

Hardware and network configuration: For the face detection network Φ we used the MediaPipe implementation[2] based on BlazeFace [2]. The resizing process Ω was done with the *OpenCV* INTER-AREA interpolation method (ω).

We used a personal computer with the following technical specifications: Intel Xeon 2.4 GHz CPU, 16 GB RAM, and an NVIDIA GeForce RTX 2070 GPU. The 3DED network was implemented with *PyTorch* libraries version 1.9.0. We used batch normalization layers after every convolutional layer. The activation functions used were Rectified Linear Unit ReLU for E and Exponential Linear Unit ELU for D. During the training process we adopted a subject-independent 5-fold cross validation evaluation protocol. We used Adam optimizer with learning rate of 0.0001 for NP loss function and 0.00044 for NPSNR. λ was set to 1.32. We set the batch size as 8 and train every model for 15 epochs. The baseline input b was 128, and the number of frames used as input was $T = 128$.

The decoder D used in all experiments has two upsampling layers followed by average pooling, and finally, a channel-wise convolution operation with 1x1x1 kernel. The output of the decoder is a 128-frame rPPG signal (right side of Fig. 4).

[2]https://github.com/google/mediapipe

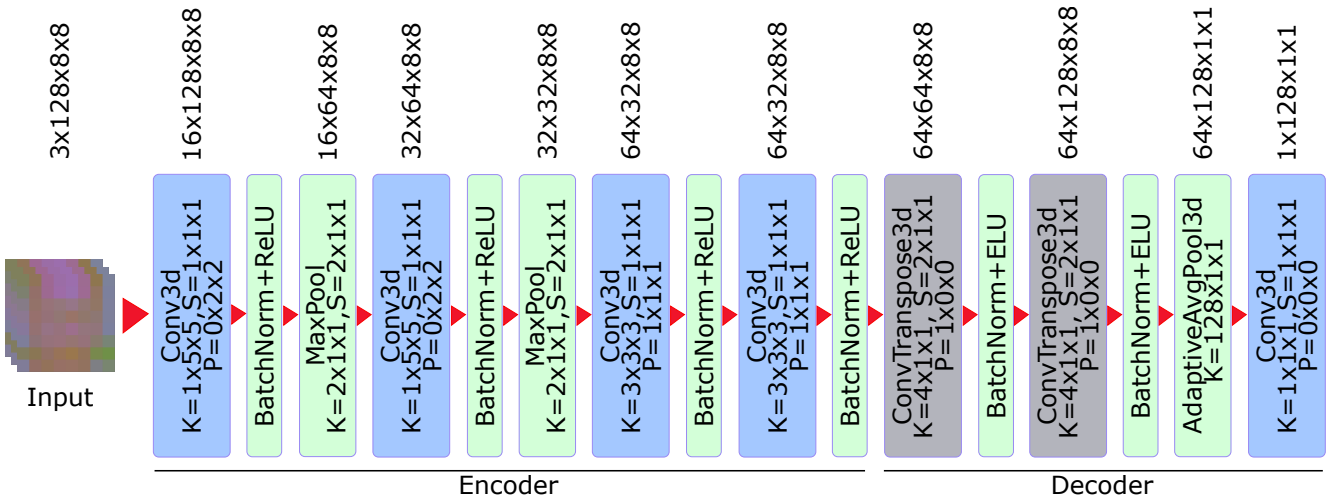

Figure 4. RTrPPG network architecture. Where K=kernel, S=stride, and P=padding.

6. Results and Discussion

In this section we present the results of the experiments proposed in Sec. 3.3. The closest architecture to our reference 3DCNN is PhysNet [33], which we also compare with all 3DED configurations. For each experiment we adopted a subject-independent 5-fold cross validation evaluation protocol. The heart rate is measured from the Fourier transform of the rPPG signal, and the HR value is the frequency corresponding to the peak of greatest magnitude.

Tab. 1 depicts the results of the proposed experiments. The second and third column present the inference time in ms on the GPU and CPU respectively. The next two columns are the metrics related to the heart measurement precision (MAE in b.pm. and r), followed by two more columns with the metrics related to the signal quality (SNR in dB and TMC). Finally, the last column is the number of trainable parameters for each architecture (N.T.Param).

The inference time in 3DED decreases on GPU and CPU when reducing the size of the input images; this is logical since the number of convolutions is also reduced. When the input size is minimum (3DED2-RGB-NP), the inference time is the smallest on CPU and GPU, Even though the MAE increases slightly under this input setting, the low values r=0.45, SNR=1.5, and TMC=0.82 indicate that the signal quality is not reliable. On the other hand, 3DED8-RGB-NP presents balanced metrics between inference time, heart rate measurement precision, and signal quality. However, despite the fact that this network is faster than the baseline and PhysNet, there is still room for improvement in the performance of rPPG signal acquisition. By taking 3DED8-RGB-NP and replacing its temporal-based loss function with the temporal-frequency-based loss function proposed in this paper, it can be seen that all metrics are improved, especially SNR.

By evaluating the RGB, Lab, Luv, YUV, and YCbCr color channels, the best performance is acquired using YUV, which is the empirical space for skin segmentation [29]. Therefore, 3DED8-YUV-NPSNR has the best compromise between real-time, signal quality and heart rate measurement performance, we refer to this architecture as Real-Time rPPG (RTrPPG) and describe its complete architecture in Fig. 4. When comparing the best configuration with the baseline model, all metrics and inference speed are improved. More interestingly, when comparing RTrPPG with the state-of-the-art PhysNet model, very similar metrics are obtained while the inference speed of RTrPPG improves substantially about 88% from 51.77 ms to 2.32 ms in GPU and from 241.57 ms to 28.65 ms in CPU. Fig. 3 shows the HR correlation plots of the best configuration in our ablation study and the PhysNet network. It can be seen that the distribution of the measured HR values using RTrPPG is comparable with PhysNet.

7. Conclusions

3DCNNs are excellent choices for extracting rPPG signals from videos with an end-to-end approach. However, their complex structures may prevent them from real-time applications. In this paper, we proposed a 3DCNN baseline and a series of experiments to find a fast and accurate network for acquiring reliable rPPG signals. The best configuration is referred to as Real-Time rPPG: RTrPPG. We showed that by decreasing the dimension of the input images, the inference speed is improved at the cost of accuracy drop in measuring rPPG signals. We proposed a joint solution showing that a temporal-frequency-based loss function is necessary for the network to learn the fundamental features of the input videos. Likewise, it was also shown that it is better to use the empirical color channel for skin segmen-

tation YUV instead of RGB. Interestingly, when comparing RTrPPG with the state-of-the-art PhysNet, a comparable accuracy to the rPPG signal acquisition is achieved while our model improves the inference speed about 88%, from 51.77 ms to 2.32 ms in GPU and from 241.57 ms to 28.65 ms in CPU. In future works, we will evaluate the performance of the proposed network in near-infrared rPPG signal acquisition applications.

References

[1] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1, 2007. 1

[2] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019. 6

[3] 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. 4

[4] Frédéric Bousefsaf, Choubeila Maaoui, and Alain Pruski. Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. *Biomedical Signal Processing and Control*, 8(6):568–574, 2013. 2

[5] Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, 9(20):4364, 2019. 1, 2, 3

[6] Sitthichok Chaichulee, Mauricio Villarroel, Joao Jorge, Carlos Arteta, Kenny McCormick, Andrew Zisserman, and Lionel Tarassenko. Cardio-respiratory signal extraction from video camera data for continuous non-contact vital sign monitoring using deep learning. *Physiological measurement*, 40(11):115001, 2019. 1

[7] 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

[8] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1

[9] 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. 1

[10] John Gideon and Simon Stent. 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*, pages 3995–4004, 2021. 1, 2, 3

[11] Bin Huang, Che-Min Chang, Chun-Liang Lin, Weihai Chen, Chia-Feng Juang, and Xingming Wu. Visual heart rate estimation from facial video based on cnn. In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1658–1662. IEEE, 2020. 1, 3

[12] Christophe Hurter and Daniel McDuff. Cardiolens: remote physiological monitoring in a mixed reality environment. In *ACM siggraph 2017 emerging technologies*, pages 1–2. 2017. 1

[13] Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999. 1

[14] Alexei A Kamshilin, Ervin Nippolainen, Igor S Sidorov, Petr V Vasilev, Nikolai P Erofeev, Natalia P Podolian, and Roman V Romashko. A new look at the essence of the imaging photoplethysmography. *Scientific reports*, 5(1):1–9, 2015. 1

[15] Magdalena Lewandowska, Jacek Ruminski, Tomasz Kocejko, and Jedrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011. 1

[16] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 1

[17] Duncan Luguern, Richard Macwan, Yannick Benezeth, Virginie Moser, L Andrea Dunbar, Fabian Braun, Alia Lemkaddem, and Julien Dubois. Wavelet variance maximization: a contactless respiration rate estimation method based on remote photoplethysmography. *Biomedical Signal Processing and Control*, 63:102263, 2021. 1

[18] Richard Macwan, Yannick Benezeth, and Alamin Mansouri. Remote photoplethysmography with constrained ica using periodicity and chrominance constraints. *Biomedical engineering online*, 17(1):1–22, 2018. 1

[19] Richard Macwan, Serge Bobbia, Yannick Benezeth, Julien Dubois, and Alamin Mansouri. Periodic variance maximization using generalized eigenvalue decomposition applied to remote photoplethysmography estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1332–1340, 2018. 1

[20] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018. 5

[21] 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. 1, 2, 3, 5

[22] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *European Conference on Computer Vision*, pages 295–310. Springer, 2020. 1, 2

[23] Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE journal of biomedical and health informatics*, 19(3):832–838, 2014. 4

[24] Fabian Schrumpf, Patrick Frenzel, Christoph Aust, Georg Osterhoff, and Mirco Fuchs. Assessment of deep learning based blood pressure prediction from ppg and rppg signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3820–3830, 2021. 1

[25] Jeremy Speth, Nathan Vance, Patrick Flynn, Kevin Bowyer, and Adam Czajka. Unifying frame rate and temporal dilations for improved remote pulse detection. *Computer Vision and Image Understanding*, 210:103246, 2021. 2

[26] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, pages 3–6, 2018. 1, 2

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

[28] 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. 1, 5

[29] Wenjin Wang, Sander Stuijk, and Gerard De Haan. Unsupervised subject detection via remote ppg. *IEEE Transactions on Biomedical Engineering*, 62(11):2629–2637, 2015. 7

[30] Wenjin Wang, Sander Stuijk, and Gerard de Haan. Living-skin classification via remote-ppg. *IEEE Transactions on biomedical engineering*, 64(12):2781–2792, 2017. 1

[31] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Fredo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31(4):1–8, 2012. 1, 2

[32] 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. 2, 3

[33] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 2, 3, 4, 7

[34] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 151–160, 2019. 1, 2, 3

[35] Pong Chi Yuen, Siqi Liu, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography, Aug. 13 2019. US Patent 10,380,444. 1

[36] Qi Zhan, Wenjin Wang, and Gerard de Haan. Analysis of cnn-based remote-ppg to understand limitations and sensitivities. *Biomedical Optics Express*, 11(3):1268–1283, 2020. 1, 2, 5

[37] Qi Zhang, Yimin Zhou, Shuang Song, Guoyuan Liang, and Haiyang Ni. Heart rate extraction based on near-infrared camera: Towards driver state monitoring. *IEEE Access*, 6:33076–33087, 2018. 1