X-iPPGNet: A novel one stage deep learning architecture based on depthwise separable convolutions for video-based pulse rate estimation

Yassine Ouzar, Djamaleddine Djeldjli, Frédéric Bousefsaf, Choubeila Maaoui

▶ **To cite this version:**

Yassine Ouzar, Djamaleddine Djeldjli, Frédéric Bousefsaf, Choubeila Maaoui. X-iPPGNet: A novel one stage deep learning architecture based on depthwise separable convolutions for video-based pulse rate estimation. Computers in Biology and Medicine, 2023, 154, pp.106592. 10.1016/j.compbiomed.2023.106592 . hal-04144444

HAL Id: hal-04144444
https://hal.science/hal-04144444v1

Submitted on 28 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

X-iPPGNet: a Novel One Stage Deep Learning Architecture based on Depthwise Separable Convolutions for Video-based Pulse Rate Estimation

Yassine Ouzar, Djamaleddine Djeldjli, Frédéric Bousefsaf* and Choubeila Maaoui

Université de Lorraine, LCOMS, F-57000 Metz, France

Abstract

Pulse rate (PR) is one of the most important markers for assessing a person's health. With the increasing demand for long-term health monitoring, much attention is being paid to contactless PR estimation using imaging photoplethysmography (iPPG). This non-invasive technique is based on the analysis of subtle changes in skin color. Despite efforts to improve iPPG, the existing algorithms are vulnerable to less-constrained scenarios (i.e., head movements, facial expressions, and environmental conditions). In this article, we propose a novel end-to-end spatio-temporal network, namely **X-iPPGNet**, for instantaneous PR estimation directly from facial video recordings. Unlike most existing systems, our model learns the iPPG concept from scratch without incorporating any prior knowledge or going through the extraction of blood volume pulse signals. Inspired by the Xception network architecture, color channel decoupling is used to learn additional photoplethysmographic information and to effectively reduce the computational cost and memory requirements. Moreover, X-iPPGNet predicts the pulse rate from a short time window (2 seconds), which has advantages with high and sharply fluctuating pulse rates. The experimental results revealed high performance under all conditions including head motions, facial expressions, and skin tone. Our approach significantly outperforms all current state-of-the-art methods on three benchmark datasets: MMSE-HR ($MAE = 4.10$; $RMSE = 5.32$; $r = 0.85$), UBFC-rPPG ($MAE = 4.99$; $RMSE = 6.26$; $r = 0.67$), MAHNOB-HCI ($MAE = 3.17$; $RMSE = 3.93$; $r = 0.88$).

Keywords: Pulse rate estimation, convolutional neural networks, end-to-end learning, imaging photoplethysmography, Xception network

Preprint submitted to Computers in Biology and Medicine

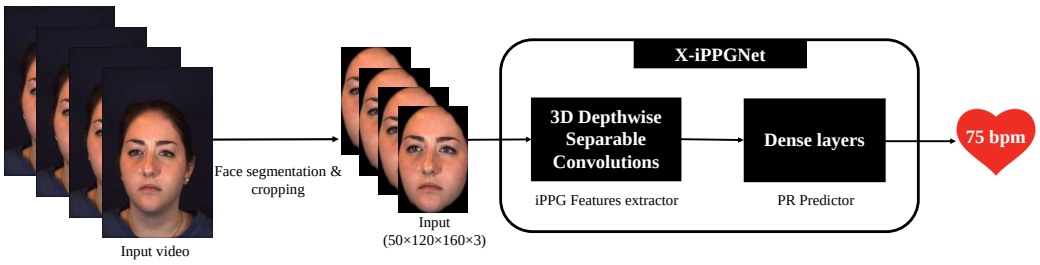


Figure 1: Overview of the proposed framework for visual pulse rate estimation. Face segmentation and cropping are performed first on the input video to get rid of non-skin areas. Then the facial image sequences are fed to a deep neural network (X-iPPGNet) consisting of 3D Depthwise Separable Convolutions for spatial and temporal features extraction, and Dense layers for pulse rate prediction.

1. Introduction

Pulse rate (PR) is one of the important indicators of a person's health that needs to be monitored routinely to identify a range of health issues. Electrocardiography and Photoplethysmography (PPG) are the main ways of measuring heart rate activity. Both techniques use contact sensors that need to be attached to body parts. Despite the high accuracy and robustness provided by these devices, specific conditions are required to acquire accurate measurements. Moreover, contact with skin can be inconvenient or even infeasible in some critical cases such as burns, skin ulcers, or contagious diseases [1]. These constraints limit their use in realistic scenarios. Over the last decade, great progress has been made in non-contact pulse rate estimation using imaging photoplethysmography, due to its wide application domains [2, 3, 4, 5, 6, 7, 8]. iPPG is an optical technique allowing a remote assessment of the pulse rate by observing the blood-volume variations on a person's face using a simple camera.

Conventional iPPG algorithms are based on hand-crafted features approaches, which generally involve multi-stage pipelines and require multiple image and signal processing steps [9, 2, 3, 6, 4, 5]. Most of these methods have been carried out under constrained environments and rely on certain assumptions regarding light-skin interaction and head motions. Therefore, they perform reasonably well under controlled conditions. However, their

performance degrades significantly under challenging scenarios such as large head movement, poor lighting conditions, and very dark skin [8, 10].

Inspired by the recent breakthroughs in computer vision tasks [11, 12, 13, 14], current state-of-the-art algorithms incorporate deep learning architectures in different stages of the conventional imaging photoplethysmography pipeline. Deep neural networks have been used to accurately extract the iPPG signal [7, 8, 15, 16]. However, several limitations remain to be resolved. These systems are not end-to-end, so they still require pre-processing or post-processing steps as well as a larger time-span window to estimate pulse rate. Furthermore, heart rate activity should be measured even in unconstrained scenarios. Many factors can affect the measurement: the person may move his head or express emotions, his face can be partially occluded or light conditions may be changing continuously. These situations affect the quality of the extracted iPPG signal, thus degrading the accuracy of the predicted PR values.

To address these drawbacks, we developed an end-to-end deep learning model (X-iPPGNet) for instantaneous pulse rate estimation directly from raw facial videos. The architecture is fully automatic and does not require any prior knowledge or special pre-processing or post-processing. This work is an extension and improvement of the method proposed as part of the Vision for Vitals Challenge [17]. We propose a new efficient architecture and evaluate its effectiveness using public databases. We also examine the impact of challenging conditions on performance.

The main contributions of this study are summarized as follows:

1. We propose a novel one-stage approach based on an end-to-end trainable neural network.

2. X-iPPGNet predicts pulse rate from short-time video excerpts (2 seconds), which is particularly relevant in the case of high and sharply fluctuating pulse rates.

3. Color channels decoupling is used to extract additional photoplethysmographic information.

4. The first use of the BP4D+ database in conjunction with data augmentation.

5. Extensive evaluations on multiple public databases to analyze the effectiveness and generalizability of the proposed method against a range of challenging factors.

The remainder of the article is organized as follows: related works are briefly exposed in Section II. Section III presents the materials and methods. Experimental results are presented and discussed in Section IV and V respectively. Finally, conclusions and future works are given in Section VI.

2. Related works

By surveying existing research on contactless pulse rate estimation using iPPG, we can identify the existence of two major approaches according to the way of iPPG signal extraction, either manually using conventional methods [9, 2, 3, 6, 4, 5], or automatically using deep learning models [7, 8, 15, 16]. Earlier works on iPPG relied on hand-crafted features approaches that generally include image and signal processing operations. The image processing techniques are first applied to locate the skin regions containing relevant information about the subtle color changes associated with blood flow. Different color spaces and different regions of interest (ROI) were exploited to constitute raw iPPG signals using a spatial averaging operation. Verkruysse et al. [9] have initially computed raw iPPG signals from the green channel using a set of predefined ROI. Several face detectors and trackers have been used to extract the entire face or sub-regions from the face such as the forehead or cheeks [26, 30, 34, 35, 36]. Bousefsaf et al. [27] proposed to select only the pixels of interest using a custom skin segmentation, while Tulyakov et al. [22] developed an approach to choose dynamically the ROI using self-adaptive matrix completion. Furthermore, different color spaces have been studied besides the standard RGB. For example, the u* component from the CIE L*u*v* color space [27] and V from YUV have been exploited [28].

In the second step, signal processing algorithms are performed to increase the signal-to-noise ratio and remove the noise from iPPG signal. Some of the popular studies include blind source separation methods, such as independent component analysis [26] and principal components analysis [30]. On the other hand, Haan and his group achieved further improvements by proposing model-based approaches [3, 4, 5]. They developed different color subspace transformations to overcome motion artifacts and improve the quality of iPPG signal.

Table 1: A brief summary of existing iPPG-based PR estimation approaches and their pros & cons.

	Multiple stage		**One stage**
	Conventional	**Deep learning**	
Input	Thermal [18]	Thermal [19]	R.G.B [20]
	Monochromatic [21]		
	R.G.B [22, 23]	R.G.B [7, 16]	Synthetic Data [24]
	Five band [25]		
Preprocessing	Face ROI detection & tracking [22, 26]	Face ROI detection & tracking [8]	Face ROI detection & tracking [20]
	Color space transformation [27, 28]	Spatial temporal maps [8]	
	Signal decomposition [2]	Video magnification [29]	
Postprocessing	Filtering [26, 30, 3]	FFT [31]	-
	FFT [32, 2]	Peaks detection [15]	
	Peaks detection [26, 25]	Deep learning model [8, 33]	
iPPG signal extraction	Spatial average [6, 26, 1]	Deep regression model [7, 15]	-
Pros	Allows pulse wave features extraction	Good generezability	Good generezability
		Allows pulse wave features extraction	Easy to deploy
			Short time span window
Cons	Hard to deploy	Hard to deploy	Does not allow pulse wave features extraction
	Require pre-processing or post-processing steps	Require pre-processing or post-processing steps	
	Large time span window	Large time span window	
	Poor generezability		

With the great success of deep learning and more specifically convolutional neural networks for medical imaging and computer vision tasks [37, 38, 12], several groups developed deep learning-based methods for iPPG estimation. According to the recent review of Ni et al. [39], existing methods are built using VGG-style CNN [7, 33, 40], or combine CNN and LSTM to take into account the temporal information [41, 8, 20], or use 3D-CNN directly to simultaneously learn spatial and temporal features [15, 42, 24, 43, 44]. To name some of the promising works, Chen and McDuff [7] proposed a convolutional attention network named DeepPhys, which consists of two-stream CNN to extract blood volume pulse waveform from facial video under varying lighting and significant head motions. They used an appearance model based on an attention mechanism to find the appropriate regions of interest (ROI) and to guide the motion representation model. Radim et al. [33] proposed a two-stage convolutional neural network method composed of 2D CNN and 1D CNN respectively. The first one extracts the iPPG signal while the second regresses pulse rate values. Niu et al. [8] generated spatial-temporal maps from multiple ROI over the face and then trained a CNN-RNN network to regress the average PR value. Yu et al. [15] introduced a spatial-temporal deep neural network (PhysNet) to extract iPPG signals from raw facial videos, and then measure the averaged PR and HRV features. AutoHR is a recent contribution proposed by Yu et al. [16]. The authors used temporal difference convolution beside a strong backbone discovered via neural architecture search to estimate accurately the iPPG signal from image sequences.

All the methods mentioned above are based on several processing stages. They mainly use deep learning to recover iPPG signals from facial videos. However, some works have adopted deep neural networks to pulse rate estimation in an end-to-end manner without passing by iPPG signal extraction. Bousefsaf et al. [24] were the first to demonstrate the possibility of pulse rate estimation from a face video without any additional processing. They put forward a 3D CNN trained purely on synthetic data. Huang et al. [20] developed a one-stage spatio-temporal network that combines 3D convolutional and LSTM modules to extract spatial and temporal features and a Dense layer to pulse rate value estimation. Ouzar et al. [45] proposed an efficient model built on a linear stack of depthwise separable convolution layers concatenated with residual connections. This method has advantages in terms of speed and simplicity and can run in real-time both on CPUs and GPUs. Existing iPPG-based PR measurement approaches are summarized in Table 1.

3. Materials and Methods

3.1. Datasets

The availability of huge databases and advanced neural architectures have underpinned the great success of deep learning approaches in computer vision tasks. In the field of remote PR estimation, the lack of large-scale heart rate (HR) datasets has limited the use of deep learning models [8]. Existing public domain HR databases are quite limited not only in data size but also in diversity. Head motion, facial expressions, occlusion, and skin tone correspond to the main challenging conditions that affect the performance of contactless pulse rate measurement from facial videos. However, previous works had not addressed all of these problems due to the quality and scale of the aforementioned databases.

For this study, we used four public datasets for pulse rate estimation to evaluate the performance of the proposed method. We trained X-iPPGNet on BP4D+ [46], a public large-scale database, while MAHNOB-HCI [47], UBFC-rPPG [48], and MMSE-HR [46] were used for testing. We briefly describe each of these three datasets in the subsequent paragraphs while we present in detail the BP4D+ database as we are the first to use it for training deep neural networks. Table 2 gives detailed comparisons between the different databases used in our experiments.

Table 2: Summary of the public-domain databases used in our experiments.

Database	**Nb of participants**	**Nb videos**	**FPS**	**Ethnicity**	**Task/Condition**
MMSE-HR [46]	40	102	25	Latino/Hispanic, White, African American, Asian, and Others	Emotion elicitation
MAHNOB-HCI [47]	27	527	61	Caucasian and Asian	Emotion elicitation
UBFC-rPPG [48]	42	42	30	-	Interaction
BP4D+ [46]	140	1400	25	Latino/Hispanic, White, African American, Asian, and Others	Emotion elicitation

3.1.1. MMSE-HR

MMSE-HR [46] was collected for contactless pulse rate estimation under challenging conditions. It consists of 102 RGB facial videos recorded at 25

frames-per-second (fps) from 40 subjects (17 males and 23 females) with various ethnic/racial ancestries. The corresponding average pulse rates were gathered using a contact BVP sensor (sampling frequency: 1K HZ).

3.1.2. MAHNOB-HCI

MAHNOB-HCI [47] is a commonly used benchmark to assess the effectiveness and generalizability of non-contact pulse rate estimation methods. It includes 527 videos from 27 subjects (12 males and 15 females) along with their corresponding physiological signals. All videos are recorded at 61 fps with a resolution of 780×580 pixels. ECG signal has been used to calculate the ground truth pulse rate values.

3.1.3. UBFC-rPPG

UBFC-rPPG [48] consists of 42 videos from 42 subjects. The videos were recorded using a low-cost webcam at 30 fps and a resolution of 640×480 pixels. The duration of each recording varies between 50 and 90 seconds. A Contec Medical CMS50E finger pulse oximeter is synchronized with the video recordings to establish the ground truth PPG signal.

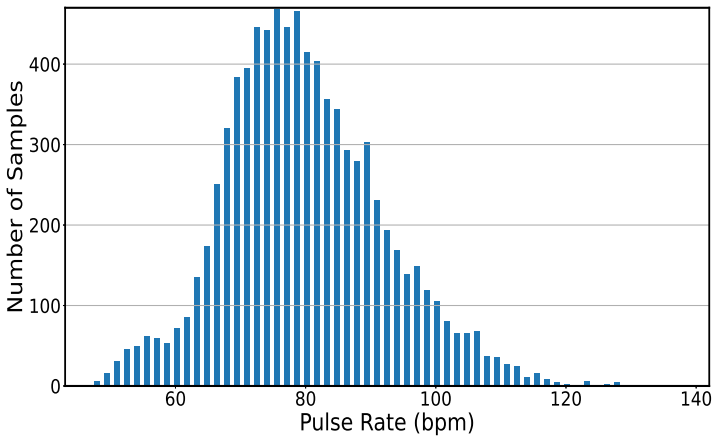


Figure 2: Distribution of the ground truth pulse rates in BP4D+.

3.1.4. BP4D+ [46]

is a large-scale public database mainly dedicated to multimodal spontaneous emotion recognition based on facial expressions and physiological

parameters. It includes several physiological signals such as heart rate, respiratory rate, and blood pressure. Compared to existing pulse rate databases, BP4D+ is significantly larger in terms of data amount and ethnic diversity (including Black, White, Asian, and Hispanic/Latino). Additionally, it was collected under challenging scenarios such as significant head motions, wild pulse rate range, facial expressions, and occlusions. 140 subjects (82 females and 58 males) participated in ten sessions set up to elicit different emotions. 1400 RGB videos lasting 30 seconds to 1 minute were recorded at 25 fps. The resolution of each video is 1040×1392 pixels. Pulse rate and other physiological signals were collected with contact sensors at 1K Hz. Figure 2 shows the histogram of ground truth pulse rate distribution in BP4D+. Pulse rate values vary from 47 to 139 beats per minute (bpm), which almost covers the typical pulse rate range. The histogram forms an inverse Gaussian distribution because most healthy and relaxed adults have a resting heart rate comprised between 70 and 90 beats per minute (see Figure 2). On the other hand, due to a large amount of corrupted ground truth signals (see a typical example in Figure 3), we recalculated the pulse rates from the blood pressure signals available in the database. We also removed segments where facial regions are outside the image.

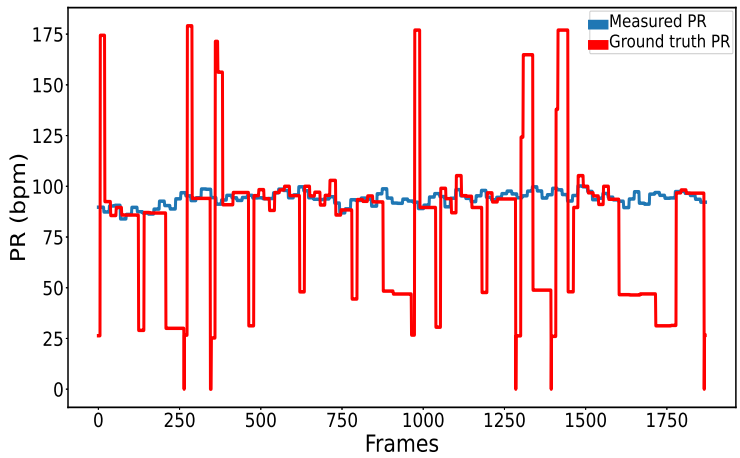


Figure 3: Example of ground truth pulse rates (participant F005) showing strong inconsistencies. Red curve: ground truth pulse rate provided by the database; Blue curve: pulse rate computed from the raw blood pressure signal.

3.2. Proposed framework

The general framework for pulse rate estimation from facial videos is illustrated in Figure 1. We treat this task as a one-stage regression problem that takes batches of 50 frames (corresponding to 2 seconds) as input and regresses the pulse rate value as output. First, face segmentation is performed to eliminate the background and non-skin areas [49]. Then the face region is cropped from the segmented face image according to the coordinates of the first non-zero pixel on each side of the image. Finally, the face image sequences are scaled and fed to a 3D fully convolutional neural network. We assume that the proposed architecture can automatically focus on the most vascularized areas of the face. It then learns the spatio-temporal features associated with iPPG.

3.2.1. Face segmentation

The extraction of regions of interest (ROI) is the first step of almost all video-based pulse rate estimation [26, 8, 50, 15, 20]. It aims to maximize the signal-to-noise ratio by only keeping the skin pixels that carry the iPPG information. Several face and facial landmarks detectors have been employed to locate ROI. However, these techniques often fail in situations involving head movement, occlusion, or facial expressions. Many other factors can also affect ROI extraction, such as lighting and background. We compared the performance of the three most popular face detectors used for iPPG extraction in terms of efficiency, i.e., Viola&Jones [51], Dlib [52], and MTCNN [53]. Table 3 illustrates the number of missed images on the MMSE-HR dataset [46] presented in section 3.1. MMSE-HR has been widely used as a test set in several works and contains about 108117 images. The results show that the three face detectors mentioned above fail to perform well in unconstrained scenes.

To overcome the limitations of face detectors, especially in unconstrained scenarios, we performed face segmentation using one of the state-of-the-art algorithms [49] (see Table 3). This method, originally proposed for face-swapping ideally works in all conditions without missing any frames. Faces are properly segmented from backgrounds and occlusions with high accuracy. Some processed images extracted from the MMSE-HR database are shown in Figure 4.

Table 3: Number of missed images according to the most popular face detection algorithms.

Face detector	Number of missed frames
Viola-Jones [51]	1375
Dlib [52]	227
MTCNN [53]	48
Face segmentation [49]	**0**

Figure 4: Examples showing the ability of the face segmentation model to work in difficult scenarios. Top figures: raw images, bottom figures: corresponding segmentations.

3.2.2. Pulse rate estimation neural network

Most of the existing video-based PR estimation approaches that integrate a deep learning model rely on a VGG-style CNN. Temporal information is processed using recurrent networks [20, 8], spatio-temporal convolutions [24, 15], or by incorporating another temporal branch in parallel [7]. The VGG-style CNN is a basic architecture that uses a standard convolution stack with no residual blocks [54]. Despite its simplicity, it is more prone to overfitting. It also performs worse than other deep learning architectures on many computer vision tasks [55]. In addition, standard convolution considers all spatial and color channel information together. However, previous studies

showed that color channels have different physiological properties and that pulsatile activity varies from one color to another [56]. Although the green channel featuring the strongest plethysmographic signal and carries more PPG information compared to the other channels, the red and blue channels also contained useful and complementary plethysmographic information that should not be neglected [26]. Nevertheless, and to the best of our knowledge, all deep learning-based approaches have combined RGB channels. This can lead to loss of useful features across channels, affecting measurement accuracy.

In this study, we designed an end-to-end deep regression framework based on a modified Xception network [57]. This architecture outperforms other deep learning models in several computer vision tasks [55, 58]. Furthermore, it relies on depthwise separable convolution instead of standard convolution operations that require larger amounts of memory and computational cost. A depthwise separable convolution extension for 3D volumes is used [1] to learn the relevant features associated with the cardiac rhythm of each color channel separately.

The idea behind the depthwise separable convolution is that the depth and spatial dimension of a filter can be decoupled within a convolutional layer. First, the video embedding dimensions are separated and an independent spatio-temporal convolution is performed for each color channel. This operation is called depthwise convolution. It aims to extract local features from each color channel of the input image sequences separately and to capture the temporal relationships among the spatial feature sequences. Then, a pointwise convolution is performed on the convoluted tensor to merge the feature maps across channels in the embedding dimension. This effectively reduces computational costs and memory requirements.

Figure 5 presents the overall architecture of the proposed X-iPPGNet, which consists of three blocks (entry, middle, and exit). It includes 36 convolutional layers structured in 14 modules, all linked with shortcuts as in the ResNet architecture, except for the first and last modules. Since the network is very deep, these residual connections allow reducing the impact of gradient vanishing. Each convolutional layer is followed by a batch-normalization to stabilize the training process and accelerate the convergence. ReLU activation functions are also used to perform nonlinear mapping. The features

[1] https://github.com/alexandrosstergiou/keras-DepthwiseConv3D

extraction output is flattened and fed into two dense layers of 1024 and 1 neurons, respectively, to estimate the pulse rate value.

In summary, the proposed non-contact pulse rate estimation framework is a one-stage pipeline that predicts the average pulse rate in only 2 seconds video fragments. The input is represented as a 5-dimensional tensor ($Nbacth \times Nbframes \times ImHeight \times ImWeight \times Channel$) (where $Nbatch$ is the batch-size; $Nbframes$ is the length of face video clip; $ImHeight$, $ImWeight$, and $Channel$ are the size of each frame) and the output is the estimated pulse rate in beats per minute.

We consider pulse rate prediction as a one-step regression problem. Training is fully supervised where each 2-seconds video fragment takes a ground truth pulse rate obtained with a contact device as a training label. In the training phase, the network learns to associate the ground truth pulse rate value with each facial video sequence by constructing a mapping relationship between inputs and outputs, i.e., mapping of a three-dimensional tensor (video data) to a single scalar (pulse rate). After the training phase, the network would be able to estimate pulse rate within the trained pulse rate range.

3.2.3. Implementation Details

3.1. Training

The proposed architecture is implemented with Keras and Tensorflow frameworks and trained with two Nvidia Quadro P6000s. The videos have been cut into sequences of 50 frames (corresponding to 2 seconds). The size of each frame is $160 \times 120 \times 3$ ($ImHeight \times ImWeight \times Channel$). The total number of sequences is 39762. Inspired by the SWATS optimization procedure [59], we started training with a Rectified Adam (RAdam) optimizer [60] before switching to Stochastic Gradient Descent (SGD) [61] when the validation accuracy stops improving. The learning rate was initially set to 10^{-4}, and then decreased to 10^{-6}. We train the network for about 25 epochs with a batch size of 64 ($Nbacth = 64$) and using the mean-squared-error loss function. In addition, a dropout technique [62] is applied before the final dense layer of the network (the dropout rate is set to 0.4). L1 and L2 regularization strategies are employed as well, which help to overcome overfitting issues and improve the model generalizability to new data.

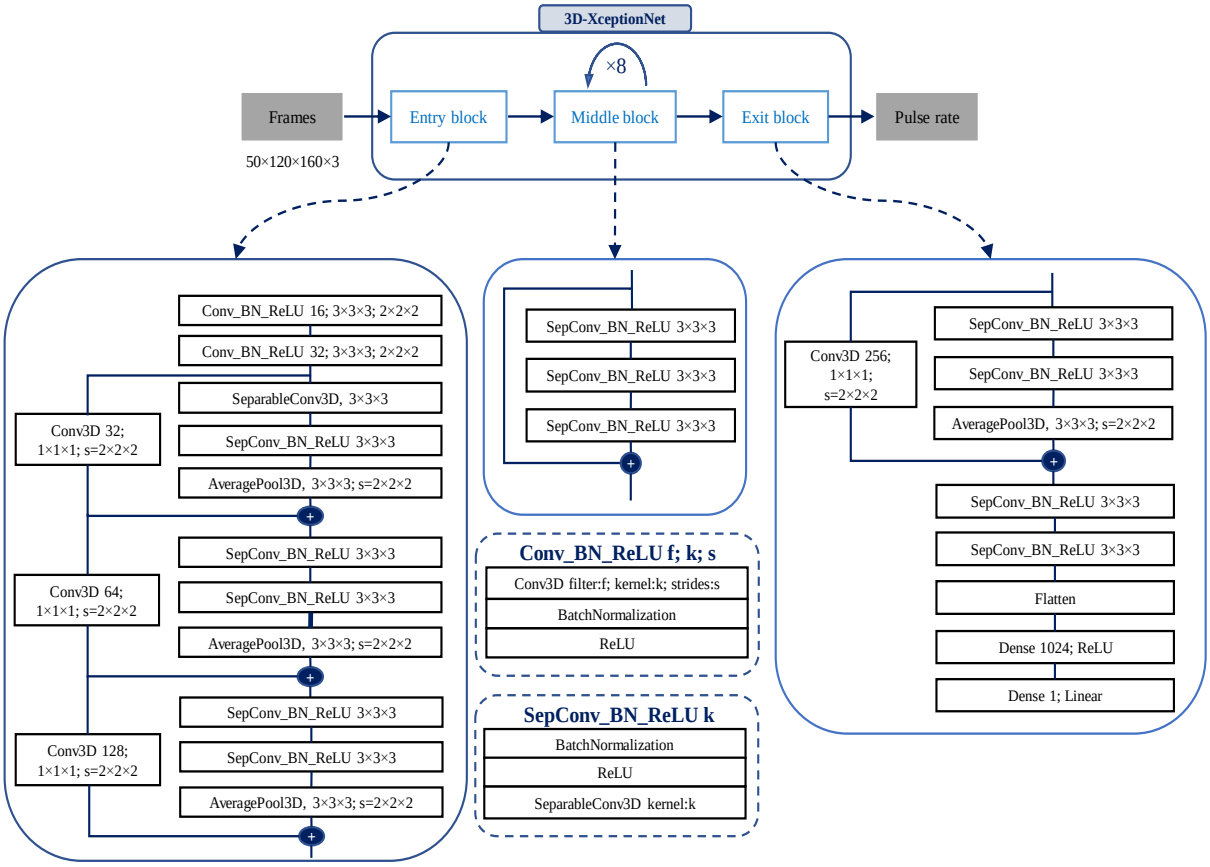


Figure 5: X-iPPGNet architecture proposed in this work. It corresponds to a modified version of the Xception network. 2D depthwise separable convolution layers are replaced by 3D depthwise separable convolution to capture both spatial and temporal features across video frames. A Dense layer is used instead of a Global Average Pooling layer. The input video fragment first passes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow which ends with a dense layer of 1 neuron, to estimate the corresponding pulse rate.

3.2. Training set augmentation

A common problem with limited and imbalanced datasets when training a neural network is overfitting and poor predictive performance, specifically for minority label samples.

X-iPPGNet was first trained without data augmentation. However, sev-

eral problems that hinder the accuracy of pulse rate predictions have caught our attention. They are mainly caused by the highly imbalanced pulse rate samples in the BP4D+ database and also by the subjects skin tone [46]. Therefore, high and low pulse rate values and the skin color type with fewer samples are more difficult to predict. It is very challenging for a deep model to learn relevant features on poorly represented data. Neural networks tend to focus on targets with large numbers of samples. To address this issue, a data augmentation technique was applied to increase the size of the training set. Since more samples are available in the mid-pulse rates range (70, 90) bpm and less outside this range (see Figure 2), we performed threefold offline data augmentation on the video sequences associated with pulse rate values greater than 90 bpm or lower than 70 bpm. Following the same strategy presented in [63], we performed standard geometric augmentation and video magnification to increase the training set size and improve the robustness of the model. The geometric augmentation involves image transformations such as random clockwise and counterclockwise rotations by up to 20 degrees, scaling (in and out) of up to 20%, and horizontal and vertical video image shifting by 10% of the frame's width and height. The Eulerian video magnification (EVM) technique [64] was used to amplify the subtle colorimetric fluctuations due to iPPG in the videos. The intensity of these fluctuations can be weak for pixels that cover dark skin. The EVM method has been proven effective for PR estimation [29, 65, 64]. This technique takes a cropped ROI video sequence as input and applies spatial decomposition followed by temporal filtering to the frames. Laplacian pyramid is used for spatial decomposition, while temporal filtering is performed by applying the Fourier transform for each pixel. The amplification factor is fixed to 60 while Frequencies outside the cutoff (45-240 bpm) are set to zero. Finally, the inverse Fourier transform is applied to reconstruct the frames. The resulting video is then amplified and reveals hidden subtle changes in the skin color instigated by blood flow in facial vessels.

4. Experiments

We aim to achieve several goals in the conducted experiments. First, we prove the possibility of measuring pulse rate with high accuracy without going through the commonly used iPPG signal extraction step. Secondly, we provide a performance comparison with various developed baseline systems as well as other deep learning approaches recently proposed for contactless

pulse rate estimation using iPPG. Thirdly, we demonstrate the generalization ability of our method under challenging conditions to illustrate the proposed framework's efficiency.

In order to study the generalizability and the effectiveness of the proposed X-iPPGNet presented in Section 3.2, three widely used public-domain databases are employed namely MMSE-HR [46], MAHNOB-HCI [47], and UBFC-rPPG [48]. MMSE-HR is directly used for testing without any additional processing since it was collected under the same conditions as BP4D+ (the training dataset). UBFC-rPPG and MAHNOB-HCI are downsampled from 30 fps and 61 fps to 25 fps in order to harmonize the fps of training and testing videos. For each experiment, we do not use videos of the same subject in both training and testing. We evaluate and compare the performance with other state-of-the-art techniques using different metrics: the standard deviation (SD), the mean absolute error (MAE, see Equation 1), the root mean square error (RMSE, see Equation 2), and the Pearson's correlation coefficient (r, see Equation 3). PR_i and $\widehat{PR_i}$ represent the ground truth and estimated pulse rate, respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|PR_i - \widehat{PR_i}| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(PR_i - \widehat{PR_i}\right)^2} \tag{2}$$

$$r = \frac{\sum_{i=1}^{n}(PR_i - \overline{PR_i})(\widehat{PR_i} - \overline{\widehat{PR_i}})}{\sqrt{\sum_{i=1}^{n}(PR_i - \overline{PR_i})^2(\widehat{PR_i} - \overline{\widehat{PR_i}})^2}} \tag{3}$$

4.1. Results

4.1.1. Evaluation on MMSE-HR

We first evaluate the generalization ability of X-iPPGNet by training the network on BP4D+ and testing it on MMSE-HR (see section 3.1).

Table 4 gives detailed comparisons with several state-of-the-art approaches including hand-crafted methods (Li2014 [66], CHROM [32], SAMC [22]) and deep learning-based methods (EVM-CNN [29], PhysNet [15], RhythmNet [8] and Auto-HR [16]). The X-iPPGNet proposed in this study achieves the best performance (SD = 5.34 bpm; MAE = 4.10 bpm; RMSE = 5.32 bpm and r

= 0.85), outperforming all competing methods. Comparison with the other state-of-the-art methods are taken from [16].

Table 4: PR estimation results by the proposed approach and several state-of-the-art methods on MMSE-HR.

Approach	Method	SD (bpm)	RMSE (bpm)	r
Multiple stage Hand-crafted	Li2014	20.02	19.95	0.37
	CHROM	14.08	13.97	0.55
	SAMC	12.24	11.37	0.71
Multiple stage Deep learnings	RhthmNet	6.98	12.76	0.78
	PhysNet	12.76	13.25	0.44
	AutoHR	5.71	5.87	**0.89**
One stage	**X-iPPGNet (Ours)**	**5.34**	**5.32**	0.85

4.1.2. Evaluation on UBFC-rPPG

In this experiment, we followed the same strategy presented in [20]. 25 videos were randomly selected to fine-tune the model pre-trained on BP4D+. The remaining videos were reserved for testing. Since the UBFC-rPPG dataset contains very limited facial videos (only one video is recorded for each subject), we used a three-fold subject-independent cross-validation strategy. Performance comparison results with other state-of-the-art techniques are taken from [67] and presented in Table 5. The proposed X-iPPGNet achieves good results and generalizes well in unseen domains. It should be noted that we achieved the best SD (6.25 bpm) and RMSE (6.26 bpm) among the existing methods.

4.1.3. Evaluation on MAHNOB-HCI

We further verify the efficiency and generalizability of X-iPPGNet on MAHNOB-HCI [47], which is the most commonly used dataset for non-contact PR estimation. The high compression rate and spontaneous movements caused by emotional stimulation make PR estimation challenging. We used the same three-fold subject-independent cross-validation protocol as for UBFC-rPPG (see Section 4.1.2). We randomized 66% of the videos to fine-tune the model pre-trained on BP4D+ and used the remaining videos

Table 5: PR estimation results by the proposed approach and several state-of-the-art methods on UBFC-RPPG.

Approach	Method	SD (bpm)	MAE (bpm)	RMSE (bpm)	r
Multiple stage Hand-crafted	Green	20.2	10.2	20.6	-
	ICA	18.6	8.43	18.8	-
	CHROM	19.1	10.6	20.3	-
	POS	10.4	**4.12**	10.5	-
Multiple stage Deep learning	Meta-rPPG	7.12	5.97	7.42	0.53
One stage	3DCNN	8.55	5.45	8.64	-
	PRNet	6.45	5.29	7.24	-
	X-iPPGNet (Ours)	**6.25**	4.99	**6.26**	**0.67**

for testing. Table 6 compares the performance of X-iPPGNet with state-of-the-art techniques, including hand-crafted and deep learning-based methods. From the results, we can observe that the X-iPPGNet ranks first on all metrics (SD = 3.93; MAE = 3.17; RMSE = 3.93 and $r = 0.88$). It is clear that our model performs very well under various image acquisition conditions and highly compressed videos.

4.2. Key Components Analysis

We also provide additional analysis to examine the impact of challenging factors, i.e., pulse rate distribution values, skin tone, gender, and head movements. All experiments have been conducted on the MMSE-HR dataset.

4.2.1. Impact of pulse rate distribution values

To further analyze the impact of PR distribution values on the performance of X-iPPGNet, we plot the differences between estimated and ground-truth pulse rate versus ground-truth estimation. This Bland–Altman plot (see Figure 6) shows that the distribution is concentrated inside the 95% limits of agreement (1.96 SD) for low (< 70) and mid (70, 90) pulse rates range. However, predictions of high pulse rates exhibit some outliers (> 90). We suppose that this observation is connected to the imbalanced training set

Table 6: PR estimation results by the proposed approach and several state-of-the-art methods on MAHNOB-HCI.

Approach	Method	SD (bpm)	MAE (bpm)	RMSE (bpm)	r
Multiple stage Hand-crafted	Poh 2011	13.5	-	13.6	0.36
	CHROM	-	13.49	22.36	0.21
	Li 2014	6.88	-	7.62	0.81
	SAMC	5.81	4.96	6.23	0.83
Multiple stage Deep learning	SynRhythm	10.88	-	11.08	-
	DeepPhys	-	4.57	-	-
	HR-CNN	-	7.25	9.24	0.51
	rPPGNet	7.82	5.51	7.82	0.78
	RhythmNet	3.99	-	3.99	0.87
	PhysNet	7.84	5.96	7.88	0.76
	AutoHR	4.73	3.78	5.10	0.86
	PulseGAN	-	4.15	6.53	0.71
One stage	**X-iPPGNet (Ours)**	**3.93**	**3.17**	**3.93**	**0.88**

(see figure 2). Furthermore, the error rate increases significantly for higher pulse rates than for mid and low pulse rates due to their fluctuations over the time window [20].

Moreover, the Bland–Altman exhibits a marked negative trend. The model tends to over-estimate low PR and under-estimate high PR because low and high pulse rates are under-represented in the training dataset. We suppose that this observation is a direct consequence of the dataset imbalance. The model tends to produce predictions oriented towards mid-PR values. The PR difference is therefore positive for low PR and negative for high PR.

4.2.2. Impact of skin tone and gender

MMSE-HR was selected to assess the generalizability of our method to different skin tones. This dataset is more diverse in terms of ethnicity (including black, white, Asian, and Hispanic / Latino) compared to UBFC-rPPG

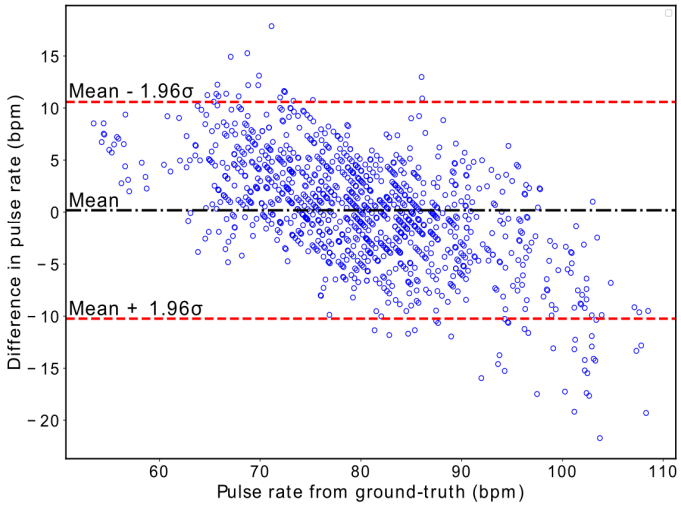


Figure 6: Bland–Altman plot showing the differences in pulse rate between ground-truth and estimated values plotted against the ground-truth measurements for the MMSE-HR dataset (see section 3.1). Mean values are represented by black dash-dot lines and 95% limits of agreement (1.96 SD) by red dashed lines.

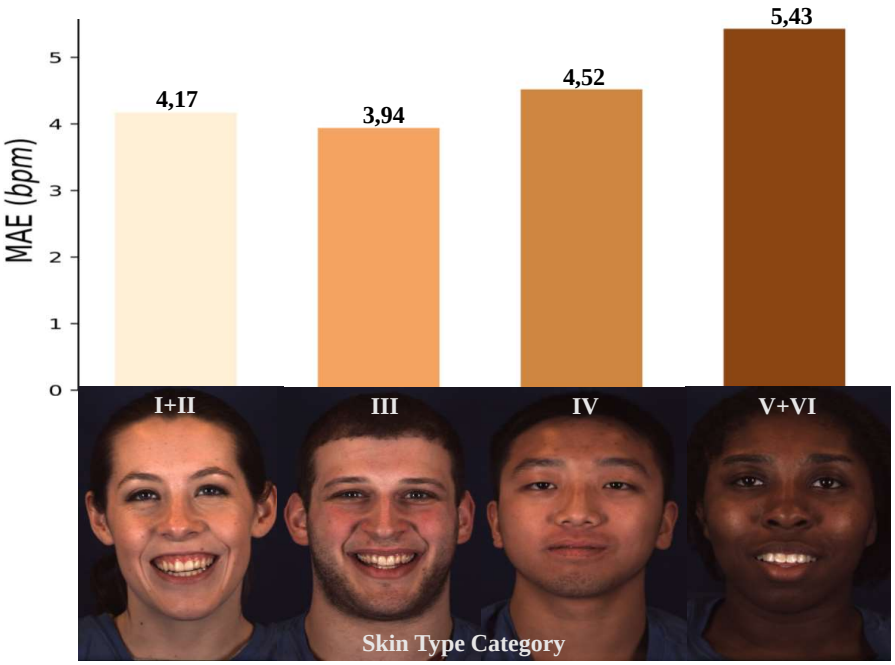
[48] and MAHNOB-HCI [47], which are highly biased towards lighter skin. Following the protocol employed by the authors of [68], which is based on the Fitzpatrick scale [69], we divided the database into 4 categories according to skin tone type. In addition to types III and IV, we grouped skin types I + II and V + VI together as there were relatively few subjects in these categories. The predictions of X-iPPGNet for different skin tones are reported in Table 7. The proposed technique exhibits great performance for all skin types and relatively less for dark skin, considering that participants with darker skin tones are underrepresented in the training set.

We further evaluated the impact of gender on pulse rate estimation. The results obtained show differences in performance between males and females (see Table 8). This confirms the results of previous study showing a slightly lower error rate for males than for females [8].

4.2.3. Impact of head movement

Visual pulse rate estimation in unconstrained environments remains a challenging task. Besides skin color and environmental conditions, head movements and facial expressions should be considered to build a robust

Table 7: PR MAE, RMSE and r for our method by skin type on MMSE-HR.



Fitzpatrick Skin Types	I+II	III	IV	V+VI
MAE (bpm)	4.17	3.94	4.52	5.43
RMSE (bpm)	5.31	5.18	5.76	6.82
r	0.87	0.81	0.84	0.40

pulse rate measurement system. Pulse rate estimation error for videos with stable subjects and those that include facial expressions and head movements has been computed in order to assess how rigid movements (e.g., head tilt and posture changes) and non-rigid movements (e.g., facial expressions) affect the performance of X-iPPGNet. The results are presented in Table 9. We observe a performance degradation for large movements compared to stable videos but the error remains acceptable.

Table 8: Performance of our method on MMSE-HR by gender.

Gender	**Male**	**Female**
MAE (bpm)	3.74	4.53
RMSE (bpm)	4.76	5.84
r	0.79	0.85

Table 9: Performance of our method on MMSE-HR under different head movement conditions.

Head movement conditions	**Stable**	**Large movement**
MAE (bpm)	3.88	4.44
RMSE (bpm)	4.91	5.74
r	0.86	0.82

4.2.4. Time window size

The time window size is an important parameter for video-based pulse rate estimation. Previous studies have reported that a longer window size leads to better performance, especially when using bandpass filter operation or power spectral density [15, 29]. However, this increases the computational cost which is not suitable for real-time applications. Indeed, there is a trade-off in the size of the time window. If the time window is too large, the predicted pulse rate loses instantaneous information as we average pulse rates in the concerned video fragment. Conversely, the input video fragment may not contain a full cycle of two consecutive beats, resulting in an inaccurate pulse rate estimate. Table 10 presents the window size selected in this work in addition with state-of-the-art methods. All previous studies present much longer time windows than our method, except PRNet [20], 3DCNN [24], and rPPGNet [42]. These methods used a 2-seconds video fragment to estimate pulse rate, but with a higher number of frames.

Table 11 presents computation time and accuracy by window size. It is clear that increasing the window size implies more input images and more

Table 10: The time window size of the input video fragment in state-of-the-art methods

Method	**Time window size**
DeepPhys [7]	30 s
Siamese-rPPG [50]	20 s
CHROM [52]	10 s
POS [5]	10 s
SynRhythm [70]	10 s
RhythmNet [8]	10 s
2SR [4]	6 s
EVM-CNN [29]	4/6/8 s
PhysNet [15]	2/4(best)/8 s
rPPGNet [42]	2 s (64 frames)
PRNet [20]	2 s (60 frames)
3DCNN [24]	2 s (60 frames)
X-iPPGNet (Ours)	**2 s (50 frames)**

trainable parameters, thus increasing computation time. The same applies to accuracy where MAE and RMSE raise with increasing time windows, except for 1-second window which does not cover the low-frequency interval. For this reason, the 2-seconds window has been carefully selected to have a complete cardiac cycle and to cover the entire pulse rate range. Computation times of the methods that use a 2-seconds window is reported in Table 12. X-iPPGNet achieves 140 ms inference time behind PRNet [20], which runs the fastest among the six methods. X-iPPGNet is however deeper and outperforms PRNet in terms of accuracy.

Table 11: Performance and computation time of our method on MMSE-HR using different time window sizes.

Window size	**1s**	**2s**	**3s**	**4s**	**6s**
MAE (bpm)	10.21	4.10	6.41	7.75	8.13
RMSE (bpm)	12.89	5.32	7.98	9.77	10.02
Computation time (ms)	120	140	160	180	220

Table 12: Computation time of our approach compared to state-of-the-art methods that use a 2-second input window size.

Method	**Computation time (ms)**
rPPGNet [42]	230
PhysNet [15]	200
3DCNN [24]	155
LCOMS [45]	150
PRNet [20]	**130**
X-iPPGNet (Ours)	140

5. Discussion

This work has been undertaken to optimize and improve iPPG-based systems for pulse rate estimation. Most existing studies extract the iPPG signal using either conventional approaches [2, 32, 6, 66, 4, 5] or deep learning-based methods [7, 15, 8, 16]. Pulse rate is usually computed as the inverse of the average time difference between consecutive beats in the time domain, or as the frequency with the highest power spectrum energy in the frequency domain. Therefore, additional processing steps such as peak detection, Fast Fourier Transform, or Power Spectral Density are required. Moreover, the accuracy depends on the quality of the iPPG waveform and on the accuracy of the main peaks detection. Since publicly available databases are challenging and provide a large number of corrupted and poor-quality PPG signals

[46, 47, 71], this directly affects the main peak location and consequently decreases the accuracy.

The proposed approach corresponds to an end-to-end trainable neural network where pulse rate is directly predicted from facial video recordings without separate iPPG signal recovery and with no prior knowledge. X-iPPGNet merges iPPG signal extraction and pulse rate prediction in one step. We rely on the ability of deep learning models to implicitly learn useful information directly from raw data. The training is fully supervised where each 2-seconds video fragment takes a ground truth pulse rate obtained with a contact device as a training label.

The main advantages of the proposed approach lie in its simplicity and low processing latency. A short time window is used to estimate pulse rate (2 s, 50 video frames). The size of the time window has a direct impact on performances. The larger it is, the higher the error, especially when dealing with higher and sharply fluctuating pulse rates (see Table 11). This is due to the loss of instantaneous information since the pulse rate is estimated by the averaging operation over the time window (As shown in Table 11). Moreover, our approach is more suitable for real-time measurement. The architecture is based on the Xception backbone that significantly reduces the number of parameters and computational costs without any performance degradation.

Since the most important factor when dealing with deep learning-based approaches is data, X-iPPGNet has been trained on BP4D+ to operate accurately in challenging scenarios and enable more robust training. BP4D+ provides a large amount of data and ethnic diversity, as well as challenging conditions. Furthermore, data augmentation is applied to increase the amount of under-represented samples at high and low frequencies. Using such a database in conjunction with data augmentation allows automatic learning of iPPG without hand-crafted features. Additionally, advanced deep learning optimization techniques as well as regularization strategies used in our work help to overcome overfitting issues and improve the model generalizability to new data.

The above experimental results verify the effectiveness of the proposed method and prove the possibility of measuring pulse rate directly from facial videos without going through iPPG signal recovery. Test results on three benchmark databases outperform existing methods and reveal the generalization ability to new data. We also examined the impact of various factors on prediction errors. The evaluation shows good performance in less-constrained scenarios such as head movement, illumination, video compression, and for

different skin tones.

5.1. Limitations

The main limitation of our method concerns the way the pulse rate is measured. Although the framework is end-to-end trainable and superior in terms of speed and simplicity, pulse rate prediction without going through iPPG signal extraction does not allow pulse wave features extraction which is useful in medical applications [1] or for affective state recognition [72]. Furthermore, we have identified several issues that can be improved in future studies. First, most publicly available databases are very limited in terms of amount of data [48, 73, 74]. This lack of data makes training deep learning models more difficult and therefore increases the probability of overfitting and decreases the ability to generalize to new data. Although a few large-scale databases are available [46, 71, 47], they are not very diverse and are highly skewed towards light skin tones and mid-pulse rates. This leads to a lack of generalization and poor performance for under-represented samples. Using synthetic data [75, 76, 24, 70] or combining multiple datasets [77] can solve the problem of the limited amount of data while applying advanced data augmentation strategies can improve performances for under-represented samples by creating additional and different training instances. Secondly, we noticed a high rate of corruption and poor quality ground truth PPG signals in the databases we used [46, 48, 47]. Data preparation and cleaning are essential to properly train the network and avoid overfitting problems. Finally, existing networks often consist of a large number of parameters and require high computational costs, which greatly hampers their application on resource-limited devices such as mobile phones. Therefore, investigating lightweight network models can considerably improve the speed and accuracy while maintaining similar performance.

6. Conclusion and Future Works

In this paper, we proposed a novel one-stage approach (X-iPPGNet) for contactless pulse rate estimation from facial video recordings using a deep spatio-temporal network. This approach is an efficient and elegant way to predict pulse rate without separate iPPG signal extraction and with no prior knowledge. X-iPPGNet is inspired by the Xception network architecture, which has proven to be efficient for general-purpose 2D image tasks in terms

of accuracy, fast convergence speed, and low computational cost. Our extensive experiments showed the effectiveness of the proposed architecture, which achieves higher accuracy and outperforms existing methods on three popular benchmark datasets such as MMSE-HR, UBFC-rPPG, and MAHNOB-HCI. The results of this study demonstrated that pulse rate can be estimated remotely from facial videos without the need for complicated hand-crafted features or iPPG signal extraction.

Looking forward to our future work, we intend to compare the performance between our one-stage-based approach and two-stage-based methods. We will further analyze the effect of combining real and synthetic data on performance. Furthermore, we envisage investigating lightweight networks to develop a faster and more suitable model for real-time applications. We would also like to investigate the effectiveness of the proposed approach for measuring other physiological parameters, such as blood pressure, respiratory rate, and oxygen saturation.

ACKNOWLEDGMENT

This work has been partly funded by the Contrat Plan État Région (CPER) Innovations Technologiques, Modélisation et Médecine Personnalisée (IT2MP) and Fonds Européen de Développement Régional (FEDER).

References

[1] D. Djeldjli, F. Bousefsaf, C. Maaoui, F. Bereksi-Reguig, A. Pruski, Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera, Biomedical Signal Processing and Control 64 (2021) 102242.

[2] M.-Z. Poh, D. J. McDuff, R. W. Picard, Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam, IEEE Transactions on Biomedical Engineering 58 (2011) 7–11. URL: http://ieeexplore.ieee.org/document/5599853/. doi:10.1109/TBME.2010.2086456.

[3] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE Transactions on Biomedical Engineering 60 (2013) 2878–2886.

[4] W. Wang, S. Stuijk, G. De Haan, A novel algorithm for remote photoplethysmography: Spatial subspace rotation, IEEE transactions on biomedical engineering 63 (2015) 1974–1984.

[5] W. Wang, A. C. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote ppg, IEEE Transactions on Biomedical Engineering 64 (2016) 1479–1491.

[6] F. Bousefsaf, C. Maaoui, A. Pruski, Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate, Biomedical Signal Processing and Control 8 (2013) 568–574.

[7] W. Chen, D. McDuff, Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 349–365.

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

[9] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light., Optics express 16 (2008) 21434–21445.

[10] E. Nowara, D. McDuff, A. Veeraraghavan, A meta-analysis of the impact of skin type and gender on non-contact photoplethysmography measurements, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 1148–1155.

[11] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.

[12] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: A brief review, Computational intelligence and neuroscience 2018 (2018).

[13] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in: International workshop on human behavior understanding, Springer, 2011, pp. 29–39.

[14] K. Suzuki, Overview of deep learning in medical imaging, Radiological physics and technology 10 (2017) 257–273.

[15] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, arXiv preprint arXiv:1905.02419 (2019).

[16] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching, IEEE Signal Processing Letters 27 (2020) 1245–1249.

[17] A. Revanur, Z. Li, U. A. Cifti, L. Yin, L. A. Jeni, The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation., in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2021.

[18] S. Bennett, T. N. El Harake, R. Goubran, F. Knoefel, Adaptive eulerian video processing of thermal video: An experimental analysis, IEEE Transactions on Instrumentation and Measurement 66 (2017) 2516–2524. doi:10.1109/TIM.2017.2684518.

[19] D.-Y. Chen, H.-S. Zou, A.-T. Hsieh, Thermal image based remote heart rate measurement on dynamic subjects using deep learning, in: 2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), 2020, pp. 1–2. doi:10.1109/ICCE-Taiwan49838.2020.9258129.

[20] B. Huang, C.-L. Lin, W. Chen, C.-F. Juang, X. Wu, A novel one-stage framework for visual pulse rate estimation using deep neural networks, Biomedical Signal Processing and Control 66 (2021) 102387. URL: https://www.sciencedirect.com/science/article/pii/S1746809420304936. doi:https://doi.org/10.1016/j.bspc.2020.102387.

[21] K. Humphreys, T. Ward, C. Markham, Noncontact simultaneous dual wavelength photoplethysmography: a further step toward noncontact pulse oximetry, Review of scientific instruments 78 (2007) 044304.

[22] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, N. 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, 2016, pp. 2396–2404.

[23] W. Wang, A. D. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote ppg, IEEE Transactions on Biomedical Engineering 64 (2017) 1479–1491.

[24] F. Bousefsaf, A. Pruski, C. Maaoui, 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video, Applied Sciences 9 (2019) 4364. doi:10.3390/app9204364.

[25] D. J. McDuff, S. Gontarek, R. W. Picard, Improvements in remote cardiopulmonary measurement using a five band digital camera, IEEE Transactions on Biomedical Engineering 61 (2014) 2593–2601.

[26] M.-Z. Poh, D. J. McDuff, R. W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation., Optics express 18 (2010) 10762–10774.

[27] F. Bousefsaf, C. Maaoui, A. Pruski, Automatic Selection of Webcam Photoplethysmographic Pixels Based on Lightness Criteria, Journal of Medical and Biological Engineering 37 (2017) 374–385. URL: http://link.springer.com/10.1007/s40846-017-0229-1. doi:10.1007/s40846-017-0229-1.

[28] J. Rumiński, Reliability of pulse measurements in videoplethysmography, Metrology and Measurement Systems 23 (2016).

[29] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, A. E. Saddik, Evm-cnn: Real-time contactless heart rate estimation from facial video, IEEE Transactions on Multimedia 21 (2019) 1778–1787.

[30] M. Lewandowska, J. Rumiński, T. Kocejko, J. Nowak, Measuring pulse rate with a webcam — a non-contact method for evaluating cardiac activity, 2011 Federated Conference on Computer Science and Information Systems (FedCSIS) (2011) 405–410.

[31] X. Liu, J. Fromm, S. Patel, D. McDuff, Multi-task temporal shift attention networks for on-device contactless vitals measurement, Advances in Neural Information Processing Systems 33 (2020) 19400–19411.

[32] G. Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE transactions on bio-medical engineering 60 (2013). doi:10.1109/TBME.2013.2266196.

[33] R. Špetlík, V. Franc, J. Matas, Visual heart rate estimation with convolutional neural network, in: Proceedings of the british machine vision conference, Newcastle, UK, 2018, pp. 3–6.

[34] T. Blöcher, J. Schneider, M. Schinle, W. Stork, An online ppgi approach for camera based heart rate monitoring using beat-to-beat detection, in: 2017 IEEE Sensors Applications Symposium (SAS), 2017, pp. 1–6. doi:10.1109/SAS.2017.7894052.

[35] M. Kumar, A. Veeraraghavan, A. Sabharwal, Distanceppg: Robust non-contact vital signs monitoring using a camera, Biomedical optics express 6 (2015) 1565–1588.

[36] S. Kwon, J. Kim, D. Lee, K. Park, Roi analysis for remote photoplethysmography on facial video, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2015, pp. 4938–4941.

[37] A. S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on mri, Zeitschrift für Medizinische Physik 29 (2019) 102–127. URL: https://www.sciencedirect.com/science/article/pii/S0939388918301181. doi:https://doi.org/10.1016/j.zemedi.2018.11.002, special Issue: Deep Learning in Medical Physics.

[38] E. Goceri, N. Goceri, Deep learning in medical image analysis: Recent advances and future trends, 2017, pp. 305–310.

[39] A. Ni, A. Azarang, N. Kehtarnavaz, A review of deep learning-based contactless heart rate measurement methods, Sensors 21 (2021) 3719. doi:10.3390/s21113719.

[40] A. Reiss, I. Indlekofer, P. Schmidt, K. Van Laerhoven, Deep ppg: Large-scale heart rate estimation with convolutional neural networks, Sensors 19 (2019). URL: https://www.mdpi.com/1424-8220/19/14/3079. doi:10.3390/s19143079.

[41] E. Lee, E. Chen, C.-Y. Lee, Meta-rppg: Remote heart rate estimation using a transductive meta-learner, in: ECCV, 2020.

[42] Z. Yu, W. Peng, X. Li, X. Hong, G. Zhao, Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement, 2019. arXiv:1907.11921.

[43] T. Luguev, D. Seuß, J.-U. Garbas, Deep learning based affective sensing with remote photoplethysmography, in: 2020 54th Annual Conference on Information Sciences and Systems (CISS), 2020, pp. 1–4. doi:10.1109/CISS48834.2020.1570617362.

[44] O. Perepelkina, M. Artemyev, M. Churikova, M. Grinenko, Heartrack: Convolutional neural network for remote video-based heart rate monitoring, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 1163–1171. doi:10.1109/CVPRW50498.2020.00152.

[45] Y. Ouzar, D. Djeldjli, F. Bousefsaf, C. Maaoui, Lcoms lab's approach to the vision for vitals (v4v) challenge, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 2750–2754.

[46] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U. A. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. Cohn, Q. Ji, L. Yin, Multimodal spontaneous emotion corpus for human behavior analysis, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3438–3446.

[47] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Transactions on Affective Computing 3 (2012) 42–55. doi:10.1109/T-AFFC.2011.25.

[48] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters 124 (2019) 82–90. URL: https://www.sciencedirect.com/science/article/pii/S0167865517303860. doi:https://doi.org/10.1016/j.patrec.2017.10.017, award Winning Papers from the 23rd International Conference on Pattern Recognition (ICPR).

[49] Y. Nirkin, I. Masi, A. T. Tran, T. Hassner, G. G. Medioni, On face segmentation, face swapping, and face perception, CoRR abs/1704.06729 (2017). URL: http://arxiv.org/abs/1704.06729. arXiv:1704.06729.

[50] Y.-Y. Tsou, Y.-A. Lee, C.-T. Hsu, S.-H. Chang, Siamese-rppg network: Remote photoplethysmography signal estimation from face videos, in: Proceedings of the 35th Annual ACM Symposium on Applied Computing, SAC '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 2066–2073. URL: https://doi.org/10.1145/3341105.3373905. doi:10.1145/3341105.3373905.

[51] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, 2001, pp. I–I. doi:10.1109/CVPR.2001.990517.

[52] D. E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.

[53] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (2016) 1499–1503.

[54] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[55] P. Zhao, C. Li, M. M. Rahaman, H. Yang, T. Jiang, M. Grzegorzek, A comparison of deep learning classification methods on small-scale image data set: from convolutional neural networks to visual transformers, arXiv preprint arXiv:2107.07699 (2021).

[56] A. V. Moço, S. Stuijk, G. de Haan, Motion robust ppg-imaging through color channel mapping, Biomedical optics express 7 (2016) 1737–1754.

[57] F. Chollet, Xception: Deep learning with depthwise separable convolutions, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 1800–1807.

[58] K. Shaheed, A. Mao, I. Qureshi, M. Kumar, S. Hussain, I. Ullah, X. Zhang, Ds-cnn: A pre-trained xception model based on depth-wise

separable convolutional neural network for finger vein recognition, Expert Syst. Appl. 191 (2022). URL: https://doi.org/10.1016/j.eswa.2021.116288. doi:10.1016/j.eswa.2021.116288.

[59] N. Keskar, R. Socher, Improving generalization performance by switching from adam to sgd, ArXiv abs/1712.07628 (2017).

[60] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the variance of the adaptive learning rate and beyond, in: Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020), 2020.

[61] S. Ruder, An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747 (2016).

[62] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[63] E. M. Nowara, D. McDuff, A. Veeraraghavan, Combining magnification and measurement for non-contact cardiac monitoring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 3810–3819.

[64] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM transactions on graphics (TOG) 31 (2012) 1–8.

[65] N. Miljković, D. Trifunović, Pulse rate assessment: Eulerian video magnification vs. electrocardiography recordings, in: 12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL), IEEE, 2014, pp. 17–20.

[66] X. Li, J. Chen, G. Zhao, M. Pietikäinen, Remote heart rate measurement from face videos under realistic situations, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4264–4271. doi:10.1109/CVPR.2014.543.

[67] E. Lee, E. Chen, C.-Y. Lee, Meta-rppg: Remote heart rate estimation using a transductive meta-learner, in: European Conference on Computer Vision, Springer, 2020, pp. 392–409.

[68] X. Liu, Z. Jiang, J. Fromm, X. Xu, S. N. Patel, D. McDuff, Metaphys: Unsupervised few-shot adaptation for non-contact physiological measurement, ArXiv abs/2010.01773 (2020).

[69] T. Fitzpatrick, The validity and practicality of sun-reactive skin types i through vi., Archives of dermatology 124 6 (1988) 869–71.

[70] X. Niu, H. Han, S. Shan, X. Chen, Synrhythm: Learning a deep heart rate estimator from general to specific, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3580–3585. doi:10.1109/ICPR.2018.8546321.

[71] X. Niu, H. Han, S. Shan, X. Chen, Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video, in: ACCV, 2018.

[72] Y. Ouzar, F. Bousefsaf, D. Djeldjli, C. Maaoui, Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2460–2469.

[73] G. Heusch, A. Anjos, S. Marcel, A reproducible study on remote heart rate measurement, arXiv preprint arXiv:1709.00962 (2017).

[74] R. Stricker, S. Müller, H.-M. Groß, Non-contact video-based pulse rate measurement on a mobile service robot, The 23rd IEEE International Symposium on Robot and Human Interactive Communication (2014) 1056–1062.

[75] D. McDuff, X. Liu, J. Hernandez, E. Wood, T. Baltrusaitis, Synthetic data for multi-parameter camera-based physiological sensing, arXiv preprint arXiv:2110.04902 (2021).

[76] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, X. Chen, Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography, IEEE Journal of Biomedical and Health Informatics 25 (2021) 1373–1384.

[77] B. L. Hill, X. Liu, D. McDuff, Beat-to-beat cardiac pulse rate measurement from video, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 2739–2742.