electronics

MDPI

Article

Improving Remote Photoplethysmography Performance through Deep-Learning-Based Real-Time Skin Segmentation Network

Kunyoung Lee [1,†], Jaemu Oh [2,†], Hojoon You [2] and Eui Chul Lee [3,*]

[1] Department of Computer Science, Graduate School, Sangmyung University, Seoul 03016, Republic of Korea; 201933048@sangmyung.kr

[2] Department of AI & Informatics, Graduate School, Sangmyung University, Seoul 03016, Republic of Korea; 202231059@sangmyung.kr (J.O.); 202132041@sangmyung.kr (H.Y.)

[3] Department of Human Centered Artificial Intelligence, Graduate School, Sangmyung University, Seoul 03016, Republic of Korea

[*] Correspondence: eclee@smu.ac.kr; Tel.: +82-2-781-7553

[†] These authors contributed equally to this work.

Abstract: In recent years, health-monitoring systems have become increasingly important in the medical and safety fields, including patient and driver monitoring. Remote photoplethysmography is an approach that captures blood flow changes due to cardiac activity by utilizing a camera to measure transmitted or reflected light through the skin, but it has limitations in its sensitivity to changes in illumination and motion. Moreover, remote photoplethysmography signals measured from nonskin regions are unreliable, leading to inaccurate remote photoplethysmography estimation. In this study, we propose Skin-SegNet, a network that minimizes noise factors and improves pulse signal quality through precise skin segmentation. Skin-SegNet separates skin pixels and nonskin pixels, as well as accessories such as glasses and hair, through training on facial structural elements and skin textures. Additionally, Skin-SegNet reduces model parameters using an information blocking decoder and spatial squeeze module, achieving a fast inference time of 15 ms on an Intel i9 CPU. For verification, we evaluated Skin-SegNet using the PURE dataset, which consists of heart rate measurements from various environments. When compared to other skin segmentation methods with similar inference speeds, Skin-SegNet demonstrated a mean absolute percentage error of 1.18%, showing an improvement of approximately 60% compared to the 4.48% error rate of the other methods. The result even exhibits better performance, with only 0.019 million parameters, in comparison to DeepLabV3+, which has 5.22 million model parameters. Consequently, Skin-SegNet is expected to be employed as an effective preprocessing technique for facilitating efficient remote photoplethysmography on low-spec computing devices.

Keywords: deep learning; skin segmentation; remote photoplethysmography; vital sign monitoring

check for updates

Citation: Lee, K.; Oh, J.; You, H.; Lee, E.C. Improving Remote Photoplethysmography Performance through Deep-Learning-Based Real-Time Skin Segmentation Network. *Electronics* **2023**, *12*, 3729. https://doi.org/10.3390/electronics12173729

Academic Editor: Jiann-Shiun Yuan

Received: 17 July 2023
Revised: 23 August 2023
Accepted: 2 September 2023
Published: 4 September 2023

1. Introduction

The field of health monitoring has grown significantly in recent years owing to the increasing interest in personal healthcare such as telemedicine and wearable healthcare devices. One of the most important health-monitoring approaches is the measurement of cardiac activity. There are several methods for the contact measurement of the heart rate, e.g., electrocardiography (ECG). However, ECG devices, which record the electrical potentials generated by the heart, are not suitable for home healthcare because they are typically bulky or inconvenient owing to the use of contact sensors. Photoplethysmography (PPG) is an alternative to ECG that measures changes in blood flow through light transmitted through the skin or reflected light after passing through the skin. In addition, ECG features are highly correlated with PPG features [1]. Compared to ECG, PPG is easier to perform and uses inexpensive equipment.

Also, PPG can be performed using remote cameras such as a webcam or a mobile camera. This is called remote PPG (rPPG) [2,3]. rPPG has the advantage that it can be used by anyone, anywhere, as it does not require a high-performance camera and can be performed using a consumer-level camera, such as a smartphone camera. However, to measure cardiac activity with high reliability based on rPPG, there is also a limitation that noise factors caused by motion and illumination must be sufficiently controlled. There are various approaches for noise reduction, and noise can be reduced through accurate skin segmentation, as shown in Figure 1. In this paper, we propose a region-of-interest (ROI) selection method for rPPG improvement that enables real-time inference even on edge devices through deep learning-based fast skin segmentation.

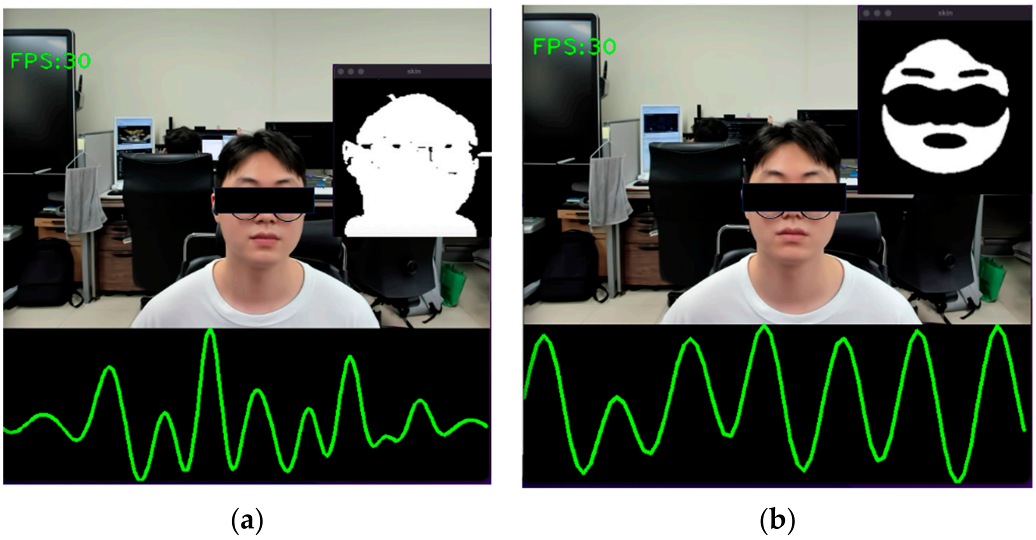

Figure 1. Examples of rPPG measurements according to the skin segmentation method. (**a**) When the skin is not properly segmented, and (**b**) when the skin is correctly segmented.

The contributions of this study are as follows:

- Existing studies on rPPG have emphasized the importance of skin segmentation. However, few attempts have been made to improve rPPG by skin segmentation. This study confirmed a noise reduction and rPPG improvement with Skin-SegNet and confirmed that the rPPG improvement is higher in noisy environments caused by talking or motion artifacts. Skin-SegNet shows an average improvement of 20% in terms of the MAPE compared with existing threshold-based skin segmentation methods (YCbCr [4], HSV [5]). Additionally, the average success rate of heart rate estimation within 5 bpm in a talking environment is 9.5%.
- Due to the nature of image processing, there is a trade-off between accuracy and processing speed. However, Skin-SegNet achieves state-of-the-art (SOTA) performance in terms of accuracy and speed. Skin-SegNet-based rPPG measurement shows 10 times faster processing speed than existing deep learning methods of ROI selection, and there is no significant difference in performance. The inference time of Skin-SegNet was 15 ms, which is a level capable of real-time processing (>30 frames per second (FPS)).

2. Related Works

In recent years, studies have explored the challenges of rPPG, such as changes in ambient light and the movement of subjects. Representative rPPG extraction methods include principal component analysis (PCA) [6], CHROM [7], POS [8], and OMIT [9].

PCA [6] is similar to independent component analysis, and it is used to extract rPPG signals. The average and variance of the RGB signals in each area are calculated by dividing the ROI into the entire face, forehead, and cheeks. Results have shown that the forehead is the most uniform area. Furthermore, it has been experimentally shown that noise increases as the size of the ROI decreases.

CHROM [7] is an rPPG algorithm that linearly combines color-difference signals by assuming a normalized skin color. A simple face detector and skin mask are used to select the ROI. A clean rPPG signal is obtained by removing the pixels that contaminate the rPPG signal, such as a mustache.

POS [8] is based on CHROM, and it is robust to movement. It extracts a pulse signal through a projection plane orthogonal to skin. A support-vector-machine-based discriminator and learning-based object tracker are used for ROI selection.

OMIT [9] is a powerful rPPG algorithm for compression artifacts that is based on QR decomposition. Two methods are used for ROI selection: U-Net-based skin segmentation and patch selection based on facial landmark coordinates.

Generally, the rPPG process is categorized into ROI selection, pulse signal extraction, and signal processing. ROI selection requires skin segmentation because only the skin pixel region is relevant for rPPG signals [10]. As PPG signals are extracted from skin pixels, preprocessing is crucial for extracting accurate skin regions, regardless of the rPPG extraction method. The exclusion of a few skin pixels negligibly affects the rPPG signal extraction performance. However, the inclusion of eyebrows, hair, reflected light, and shadow areas other than the skin can have a significant adverse effect on the performance. There are several advantages of performing skin segmentation during rPPG.

First, rPPG is sensitive to motion artifacts. It is difficult to remain still without moving the face and to not blink, which can result in noise when rPPG is performed using a camera [6]. This motion noise can be reduced by excluding the relatively active areas of the eyelids and lips.

Second, glasses, hair, and beards can contaminate rPPG signals. To reduce the effect of these factors, studies have simply narrowed the face area or divided it into smaller areas without dividing the skin [9,11]. However, this approach reduces the ROI, thereby making rPPG more sensitive to noise [7].

Finally, convex surfaces, such as the nose and lips, can cause specular reflection and negatively affect the rPPG performance when these areas are included as skin pixels. These three problems can be solved by accurately segmenting the largest possible skin area to obtain a reliable signal.

PCA and CHROM perform rPPG signal extraction by using threshold-based skin segmentation methods such as YCbCr [4] and HSV [5] to select skin regions as ROIs to obtain pulse signals. Bobia's work [12] emphasizes that ROI selection during rPPG signal extraction is an important step for obtaining reliable signals. In addition, another study by Bobia [13] states that the quality of the ROI directly affects the quality of rPPG signals.

The simplest skin segmentation method is threshold-based, which uses a limited range of colors of human skin. However, this method is not suitable for the reliable measurement of rPPG signals because it is based on colors.

In recent years, methods that use superpixels and deep learning have been developed as alternatives to threshold-based skin segmentation. Superpixel-based methods are frequently used for image segmentation and object tracking. An image is segmented into small, uniform regions with similar characteristics, and image processing is performed using these regions as basic units.

Recently, a superpixel method that uses the simple linear iterative clustering (SLIC) algorithm was developed [14]. This method, which is relatively faster than the superpixel method, uses a superpixel extracted via an energy-driven sampling algorithm. Bobia's work [13] applied the SLIC-based superpixel method to skin segmentation. This work reports that accurate skin segmentation is important when extracting rPPG signals and that the SLIC-based superpixel method provides satisfactory accuracy at 25 fps for a 640 × 480 pixel image. However, we aim to use skin segmentation for rPPG signal extraction. This is a part of the preprocessing stage, and thus, it should not require a long time. Although a speed of 25 fps is not low, we require a higher speed. Therefore, we do not use a superpixel method in comparative study because it is difficult to find the correct value for the superpixel hyperparameter, i.e., the speed.

Deep learning is being applied to various fields owing to recent developments, and its performance is significantly higher than that of existing methods.

Tran et al. [15] attempted to obtain a clean rPPG signal via skin segmentation using DeepLabV3+ [16]. A heart rate estimation accuracy of over 90% was achieved using DeepLabV3+ and the adaptive pulsatile plane (APP) method. The highest accuracies were similar to those obtained using CHROM and POS.

It could not be confirmed whether skin segmentation had a significant effect on rPPG signals because signals were obtained using the APP method. In addition, owing to the processing speed problem, face detection was performed only at the beginning of the process. Signals were extracted by performing skin segmentation only once in the region where the face was found. Even small movements of a subject can cause signals to be extracted from the incorrect skin area.

SINet [17] shows good performance in the field of portrait segmentation. SINet reduces the number of parameters by more than 90% while maintaining the accuracy of several existing models. Furthermore, it is suitable for our purpose because it successfully runs at over 100 fps on mobile devices. Therefore, the structure of our proposed network inherits the information blocking decoder and spatial squeeze module, which are lightweight and performance-enhancing techniques of SINet [17].

Lee et al. [18] proposed the extreme lightweight skin segmentation networks (ELSNet) for measuring rPPG. Despite the fast inference speed of 167 frames per second (FPS), the overall performance of the rPPG measurement was improved, and the improvement was greater in environments with frequent motion.

In this study, we propose a real-time skin segmentation network (Skin-SegNet) that measures reliable signals. Skin-SegNet is specialized for the improvement of rPPG signal extraction by fast and accurate ROI selection. It learns facial structural elements to remove the eyebrows, hair, eyes, nose, and mouth, which may interfere with rPPG signal extraction. Skin-SegNet improves the success rate of heart rate estimation by approximately 6% within 5 bpm. In addition, it was confirmed that these rPPG enhancements are generalized enhancement methods that can be applied to the PCA, CHROM, POS, and OMIT rPPG extraction algorithms. An average performance improvement of 9.5% and 20% is confirmed in a talking environment and in terms of the mean absolute percentage error (MAPE), respectively. Skin-SegNet shows an inference time of 15 ms using only an Intel i9 CPU. This implies that real-time processing is possible even if the model runs during the preprocessing of rPPG.

3. Method

This section describes the structure of the proposed skin segmentation model and the experimental setup used to confirm the improvement in the performance of the rPPG signal extraction algorithm by the skin segmentation model. A network design of Skin-SegNet consists of the information blocking decoder and spatial squeeze module proposed by SINet [17].

In addition, the rPPG performance comparison shown in Figure 2 is performed to compare the performance between Skin-SegNet and the existing methods. The first step in the process is to acquire face images. The performance of each method is evaluated using the PURE dataset. In the third process, the existing methods and the Skin-SegNet-based image processing method are performed to confirm the improvement of rPPG performance through skin segmentation improvement. In addition, an rPPG performance comparison was performed according to fair-skin segmentation by applying all four types of rPPG extraction methods.

During the rPPG measurement process, the same method of the traditional rPPG measurement process was applied except for the skin segmentation process, which is the second step in Figure 2. After removing noise through signal processing, rPPG extraction is performed. After that, the heart rate is calculated using frequency analysis through FFT, and performance is compared using rPPG metrics such as MAPE, coverage, SNR, and Pearson correlation coefficient. In the fifth step of Figure 2, four types of rPPG extractions are performed. This is also a process of verifying performance improvement through skin segmentation, and performance comparison experiments were conducted only based on the type of skin segmentation method.

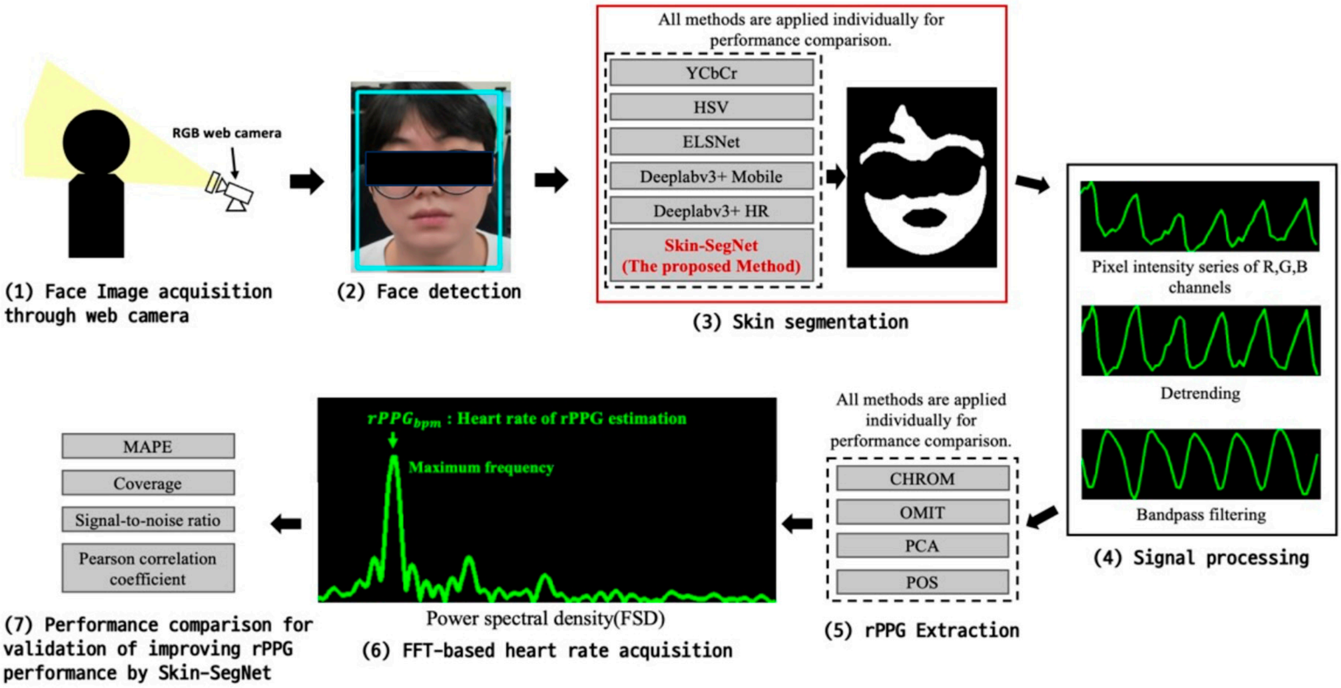
RGB web camera
(1) Face Image acquisition through web camera
(2) Face detection
All methods are applied individually for performance comparison.
YCbCr
HSV
ELSNet
Deeplabv3+ Mobile
Deeplabv3+ HR
Skin-SegNet (The proposed Method)
(3) Skin segmentation
Pixel intensity series of R,G,B channels
Detrending
Bandpass filtering
(4) Signal processing
All methods are applied individually for performance comparison.
CHROM
OMIT
PCA
POS
(5) rPPG Extraction
$rPPG_{bpm}$: Heart rate of rPPG estimation
Maximum frequency
Power spectral density(FSD)
(6) FFT-based heart rate acquisition
MAPE
Coverage
Signal-to-noise ratio
Pearson correlation coefficient
(7) Performance comparison for validation of improving rPPG performance by Skin-SegNet

Figure 2. Overview of rPPG performance comparison to validate improvement through skin segmentation using Skin-SegNet.

3.1. Information Blocking Decoder

Information blocking can reduce the ambiguity of the model during object boundary segmentation to decrease the errors between the background and object, which are commonly encountered in segmentation. A typical case is when the model inaccurately judges the background or an edge. Figure 3 shows an example in which the clothes and head parts are not correctly segmented.

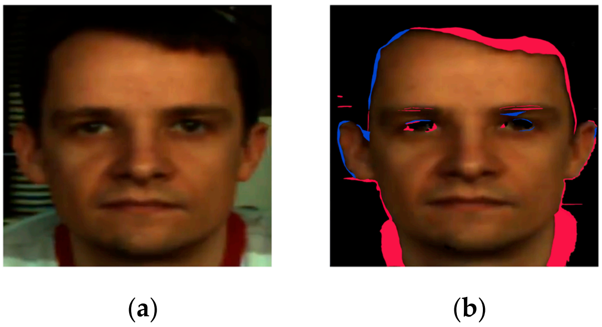
(**a**) (**b**)

Figure 3. Example of segmentation errors that occur mainly around object boundaries. (**a**) Original image, and (**b**) typical example of segmentation errors. Blue and red indicate false negatives and false positives, respectively.

Recent studies [16,19–21] addressed this problem by reusing high-resolution feature maps in encoders. However, unlike other studies, SINet imports only the parts about which the model is uncertain without importing the information of the high-resolution feature map. This is referred to as information blocking. The descriptions and related expressions for information blocking are as follows:

$$c = \max\left(softmax(X)\right) \tag{1}$$

In Equation (1), is the confidence score, is the feature map, and is the probability value of each class.

$$Information\ Blocking\ Map = 1 - \mathrm{c} \tag{2}$$

The information blocking map expressed by Equation (2) is obtained by subtracting the reliability score obtained from Equation (1) from 1. Elementwise convolution is performed on the information blocking map, and a high-resolution feature map is obtained in this manner. This ensures that only the low-confidence parts (the high-value parts of the information blocking map) can obtain information from the high-resolution feature map.

3.2. Spatial Squeeze Module

Unlike the multipath structure [22], the S2 block uses a multireceptive field scheme that processes global information while reducing latency by applying average pooling. The S2 module used in SINet [17] consists of a split transform–merge structure [19,20,23] to cover the multireceptive field using two S2 blocks [17]. The S2 module uses pointwise convolution, channel shuffle, and group pointwise convolution to reduce computation. The size of feature maps is reduced by half. Then, they are passed through S2 blocks and merged. The feature map merged with the input feature map is added as a residual connection. The PReLU activation function is used. Figure 4 shows the S2 module and S2 block.

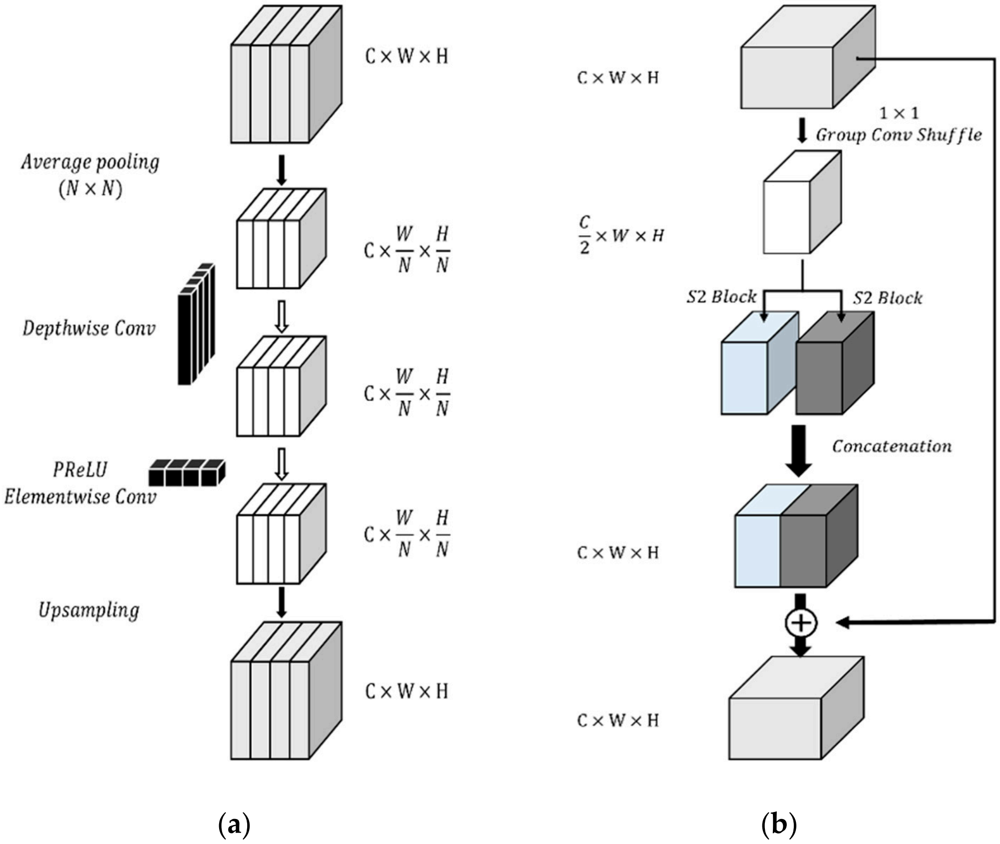

Figure 4. Architecture of Skin-SegNet. The S2 module consists of the bottleneck layer. Information blocking is a method for fine skin segmentation. (**a**) Depthwise separable convolution, and (**b**) squeeze and excitation.

3.3. Real-Time Skin Segmentation Network (Skin-SegNet)

In SINet, a structure optimized for portrait segmentation is created using a network. In this study, we design a network optimized for skin segmentation in detected face regions using the S2 block, group convolution, and information blocking. However, we use a different model architecture and network depth. The overall model structure is shown in Figure 5.

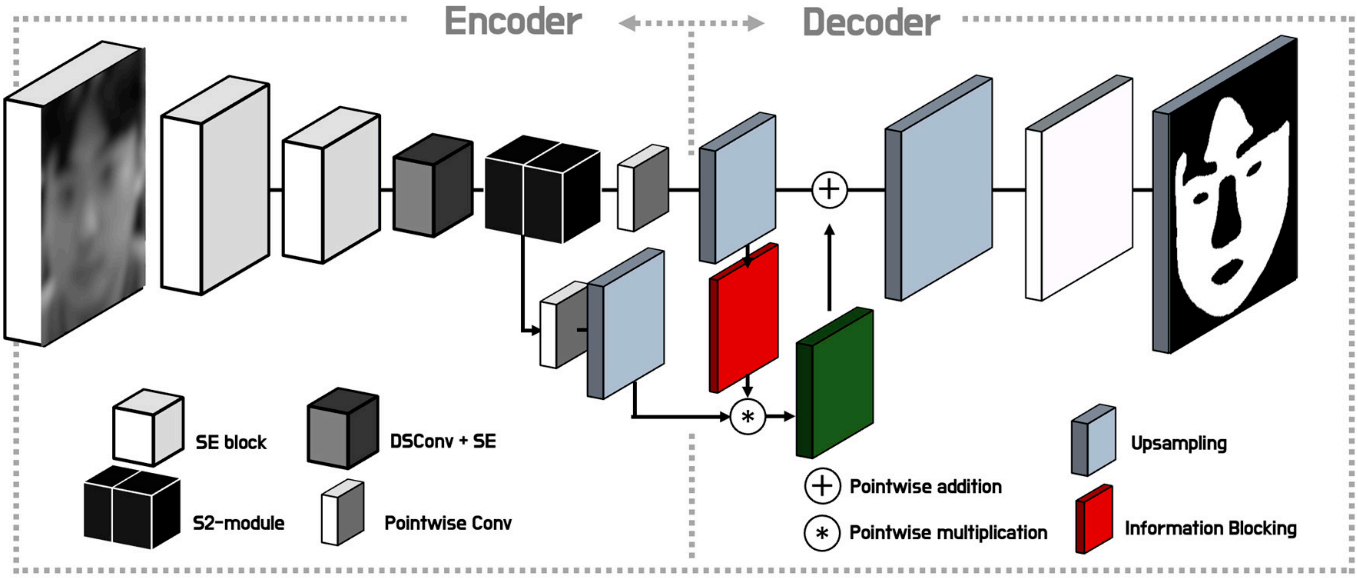

Figure 5. Architecture of Skin-SegNet. The S2 module consists of the bottleneck layer. Information blocking is a method for fine skin segmentation. DS: depthwise separable convolution, SE: squeeze and excitation.

The input image is passed through two layers, i.e., a convolutional layer with two strides and a batch normalization layer, and PReLU before the encoder. Thereafter, the feature map passes through depthwise separable (DS) convolution and a DS + squeeze-and-excitation (SE) lock composed of SE blocks. At the back of the encoder, the feature map is passed through the S2 block six times, and the final encoder output is computed using point-by-point convolution. The setup of Skin-SegNet is described in Table 1. The performances of Skin-SegNet and existing skin segmentation methods [4,5] are compared for the four rPPG methods introduced in Section 2.

Table 1. Architecture of Skin-SegNet. The S2 module consists of the bottleneck layer. Information blocking is a method for fine skin segmentation. DS: depthwise separable convolution, SE: squeeze and excitation.

#	Input	Operation	Output	*k*, *p*
1	3 × 244 × 244	SE block	12 × 112 × 112	Downsampling
2	12 × 112 × 112	SE block	16 × 56 × 56	Downsampling
3	16 × 56 × 56	DS + SE block	16 × 28 × 28	Downsampling
4	16 × 28 × 28	S2 module	32 × 28 × 28	[k = 3, p = 1], [k = 5, p = 1]
5	32 × 28 × 28	S2 module	32 × 28 × 28	[k = 5, p = 1], [k = 3, p = 2]
6	32 × 28 × 28	S2 module	32 × 28 × 28	[k = 5, p = 2], [k = 3, p = 4]
7	32 × 28 × 28	S2 module	32 × 28 × 28	[k = 5, p = 1], [k = 5, p = 1]
8	32 × 28 × 28	S2 module	32 × 28 × 28	[k = 3, p = 2], [k = 3, p = 4]
9	32 × 28 × 28	S2 module	32 × 28 × 28	[k = 3, p = 1], [k = 5, p = 2]

Table 1. *Cont.*

#	Input	Operation	Output	*k*, *p*
10	48 × 28 × 28	Concatenation, Conv2d	2 × 28 × 28	Encoder output
11	2 × 28 × 28	Bilinear2d	2 × 56 × 56	Upsampling
12	16 × 56 × 56	Conv2d	2 × 56 × 56	Shortcut
13	2 × 56 × 56	Gate function	2 × 56 × 56	Information blocking operation
14	2 × 56 × 56	Bilinear2d	2 × 112 × 112	Upsampling
15	2 × 112 × 112	Bilinear2d, Conv2d	2 × 224 × 224	Upsampling

Skin-SegNet removes the areas that may interfere with rPPG signal extraction and detects the maximum skin area. It is experimentally confirmed that accurate and fast skin segmentation using Skin-SegNet significantly improves the rPPG measurement performance in various movement and conversation environments.

4. Result

4.1. Datasets

Training dataset: The CelebAMask-HQ dataset [24] contains 30,000 high-resolution facial images selected from the CelebA dataset. It consists of 19 facial components, e.g., skin, nose, eyes, eyebrows, lips, hair, and accessories. The CelebAMask-HQ dataset allows for the editing of facial components; therefore, we trained Skin-SegNet using this dataset to detect only skin regions.

Evaluation dataset: The PURE dataset [25] validates the robustness of Skin-SegNet against noise artifacts caused by illumination and motion. The PURE dataset contains rPPG data in six environments, which include resting, talking, slow and fast movement, and small and medium rotation.

4.2. Evaluation

Table 2 presents the rPPG performance obtained when Skin-SegNet-based skin segmentation is applied to the PURE dataset. Three rPPG performance metrics are used, i.e., the MAPE and two values of coverage, as given by Equations (3)–(6).

$$MAPE = \frac{100}{k}\sum_{i=1}^{k}\left|\frac{y_i - \hat{f}(x_i)}{y_i}\right| \tag{3}$$

Table 2. Average evaluation result of four rPPG methods in all experimental environments.

Methods	MAPE	Coverage5	Coverage3	SNR	$r(\sqrt{r^2})$
YCbCr [4]	5.58	80.25%	66.0%	0.9290	0.83
HSV [5]	9.53	73.25%	60.50%	0.8782	0.75
ELSNet [18]	4.48	85.25%	71.50%	1.0004	0.88
Deeplabv3+ Mobile [16]	2.01	96.78%	95.17%	1.5479	0.88
Deeplabv3+ HR [16]	1.99	96.80%	95.11%	1.5647	0.88
Skin-SegNet (ours)	1.81	96.78%	95.17%	1.6077	0.89

In Equation (3), k is the index of the 1 s sliding window. y_i is the exact heart rate corresponding to the index. $\hat{f}(x_i)$ is the estimated heart rate corresponding to the exponent.

$$b_k(T) = \begin{cases} 0, \ if\ d(k) > T \\ 1, \ if\ d(k) < T \end{cases} \tag{4}$$

In Equation (4), k is the index of the 1 s sliding window. $d(k)$ is the heart rate error function corresponding to the index k. T is the bpm threshold of the binary function $b_k(T)$.

$$Coverage\ T = \frac{\sum_k b_k(T)}{K} \tag{5}$$

The coverage given by Equation (5) represents the success rate of the time series obtained using the binary function. Therefore, the performance improves as the coverage increases.

$$SNR = \frac{\sum_{n=A}^{B} spectrum(n)}{\sum_{n=A}^{B} spectrum(n) - \sum_{n=f_{min}}^{f_{max}} spectrum(n)} \tag{6}$$

In Equation (6), the minimum and maximum human heart rates are 42 bpm and 210 bpm, respectively, which correspond to 0.7 Hz (f_{min}) and 3.5 Hz (f_{max}) in the frequency domain, respectively. The peak frequency is defined as the most dominant frequency in the bpm range defined above. A and B denote the peak frequency with a margin of 0.7 Hz subtracted from or added to it, respectively. Figure 6 shows a visual representation of f_{min}, f_{max}, A, and B.

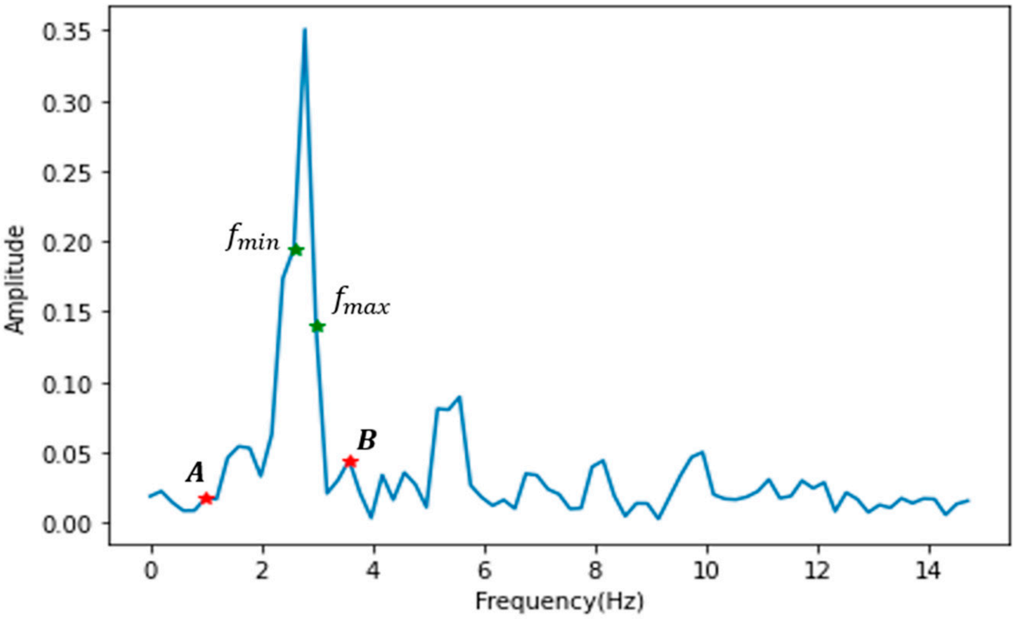

Figure 6. Examples of f_{min}, f_{max}, A, and B in Equation (6) (f_{max} and f_{min} are given by peak frequency ±0.7 Hz; A and B are frequencies corresponding to 42 bpm and 210 bpm, respectively).

4.3. Evaluation Result

Table 2 shows the average performance of all rPPG methods. The performance of deep learning-based methods outperforms the color domain-based methods YCbCr and HSV in all evaluation metrics. This result shows that rPPG extraction performance can be improved only by improving ROI selection based on image processing. In Table 3, Skin-SegNet has the fastest inference time and fewest model parameters among the deep learning-based methods. Table 3 shows that Skin-SegNet can perform more than 30 FPS with only CPU operation. Also, Table 2 shows that the heart rate-based rPPG evaluation results are better or there is no significant difference, even though the speed is more than 10 times faster and the model parameters are less than 1/100 the weight.

Table 3. Inference times of skin segmentation methods using only CPU. (CPU: Intel(R) Core(TM) i9-9900K, ms: milliseconds, G: giga, M: million).

Methods	Mean (ms)	Max. (ms)	MACs (G)	Parameters (M)
Deeplabv3+ Mobile [16]	124	130	35.66	5.22
Deeplabv3+ HR [16]	324	359	6.03	71.71
ELSNet [18] *	5	7	0.023	0.01
Skin-SegNet (ours) *	12	15	0.047	0.019

* This symbol indicates that the model is capable of real-time inference.

The improvement of rPPG performance through deep learning-based ROI selection is noticeably increased in a talking environment. Table 4 shows the performance improvement in a talking environment on the PURE dataset. Compared to conventional methods, rPPG measurement using deep learning-based ROI selection shows performance improvements ranging from a minimum of 5% to a maximum of 20% compared to YCbCr and HSV. Coverage T in Table 2 is the probability that the error between the estimated bpm and ground-truth bpm is less than T bpm. A detailed explanation is shown in Equations (4) and (5).

Table 4. Evaluation result of four rPPG methods in talking environment.

rPPG Method	Skin Segmentation Method	MAPE	Coverage5	Coverage3
CHROM [7]	YCbCr [4] *	9	63.41%	49.57%
	HSV [5] *	11.7	54.45%	39.49%
	ELSNet [18] *	6.7	71.22%	51.0%
	Deeplabv3+ Mobile [16]	1.9	93.83%	90.51%
	Deeplabv3+ HR [16]	1.9	94.47%	91.48%
	Skin-SegNet (ours) *	2.3	93.65%	91.18%
OMIT [9]	YCbCr [4] *	10.2	60.45%	48.67%
	HSV [5] *	11.2	58.76%	42.33%
	ELSNet [18] *	6.9	67.09%	55.39%
	Deeplabv3+ Mobile [16]	1.7	95.52%	94.29%
	Deeplabv3+ HR [16]	1.7	95.46%	94.03%
	Skin-SegNet (ours) *	1.8	95.19%	93.31%
PCA [6]	YCbCr [4] *	8.7	65.49%	49.18%
	HSV [5] *	20.9	42.32%	32.43%
	ELSNet [18] *	6.7	68.87%	54.11%
	Deeplabv3+ Mobile [16]	1.6	95.92%	94.60%
	Deeplabv3+ HR [16]	1.5	96.67%	94.58%
	Skin-SegNet (ours) *	2.1	94.81%	93.35%
POS [8]	YCbCr [4] *	14.7	44.38%	34.47%
	HSV [5] *	16.0	43.80%	30.75%
	ELSNet [18] *	14	46.66%	35.12%
	Deeplabv3+ Mobile [16]	1.9	96.72%	92.30%
	Deeplabv3+ HR [16]	1.8	96.47%	91.74%
	Skin-SegNet (ours) *	2.4	95.81%	93.05%

* This symbol indicates that the model is capable of real-time inference.

Table 5 shows the rPPG evaluation results in the fast translation environment that includes motion noise caused by head movement. Skin-SegNet shows SOTA-level rPPG measurement in motion and illumination noise reduction despite having the fastest inference time. Therefore, in Tables 4 and 5, the improvement due to Skin-SegNet was confirmed more clearly in an environment with frequent motion noise. This means that rPPG pulse signal quality can be improved only by improving skin segmentation.

Table 5. Evaluation result of four rPPG methods in fast translation environment.

rPPG Method	Skin Segmentation Method	MAPE	Coverage5	Coverage3
CHROM [7]	YCbCr [4] *	9.5	62.84%	48.58%
	HSV [5] *	12.9	60.48%	45.11%
	ELSNet [18] *	7.4	69.17%	54.41%
	Deeplabv3+ Mobile [16]	1.1	98.03%	96.36%
	Deeplabv3+ HR [16]	1.1	98.61%	97.60%
	Skin-SegNet (ours) *	1.2	98.63%	97.76%
OMIT [9]	YCbCr [4] *	10.6	62.14%	47.23%
	HSV [5] *	12.2	62.42%	49.41%
	ELSNet [18] *	7.6	67.97%	55.08%
	Deeplabv3+ Mobile [16]	1.0	98.56%	97.39%
	Deeplabv3+ HR [16]	1.2	98.18%	97.69%
	Skin-SegNet (ours) *	1.2	98.59%	97.39%
PCA [6]	YCbCr [4] *	9.8	63.32%	49.70%
	HSV [5] *	26.4	60.29%	45.40%
	ELSNet [18] *	7.4	68.31%	55.37%
	Deeplabv3+ Mobile [16]	1.0	98.39%	97.85%
	Deeplabv3+ HR [16]	1.2	98.13%	97.60%
	Skin-SegNet (ours) *	1.4	98.19%	97.10%
POS [8]	YCbCr [4] *	14.6	53.46%	39.70%
	HSV [5] *	18.2	55.47%	42.08%
	ELSNet [18] *	13.4	56.27%	41.83%
	Deeplabv3+ Mobile [16]	1.4	98.68%	96.36%
	Deeplabv3+ HR [16]	1.4	98.34%	96.91%
	Skin-SegNet (ours) *	1.0	98.42%	97.80%

* This symbol indicates that the model is capable of real-time inference.

Figure 7 shows that real-time processing is possible at an average speed of 30 fps on an Intel i7 CPU when it is operated simultaneously with the face detection model using OpenCV in the Python environment.

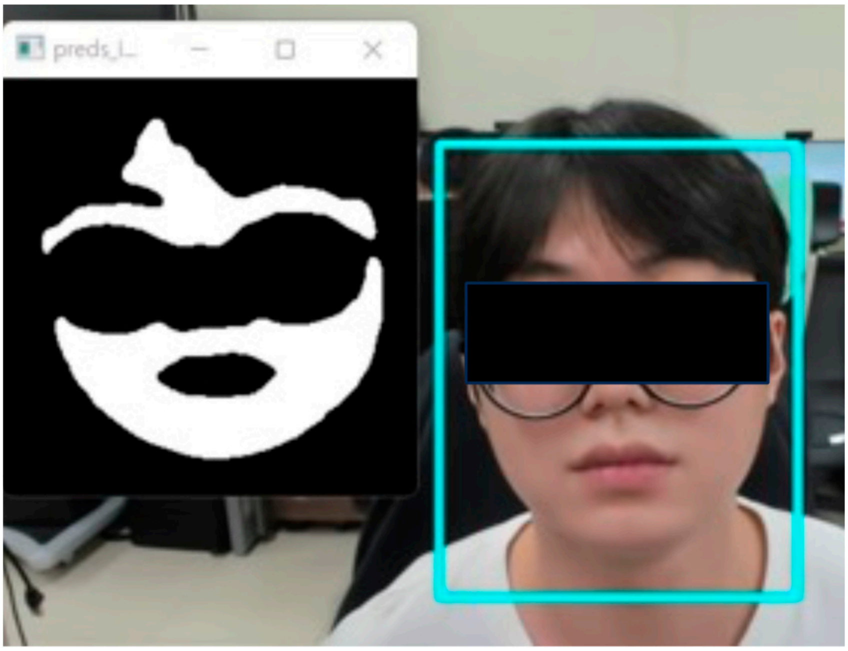

Figure 7. Real-time processing at an average speed of 30 fps on the i7 CPU when it is operated simultaneously with the face detection model using OpenCV in Python.

5. Discussion

rPPG has been extensively investigated in recent years. However, the improvement in the performance of rPPG through image processing has not been examined in detail. Furthermore, most studies have focused on pulse signal extraction and signal-processing-based noise reduction rather than ROI selection. In this study, it was experimentally verified that rPPG signal extraction could be improved by utilizing skin segmentation. The results demonstrated the importance of ROI selection in the rPPG process.

To the best of our knowledge, most existing ROI selection methods use face-landmark-based regions [26–28] and threshold-based skin segmentation [4,5]. Superpixel-based [13,14] and deep learning-based methods [15,16] have been used to improve rPPG signal extraction performance through ROI selection. Although studies have examined deep learning-based ROI selection using DeepLabV3+ [16], they have not investigated performance improvement by optimizing rPPG skin segmentation. Therefore, this study proposes Skin-SegNet to improve the rPPG performance based on the fastest skin segmentation network. In addition, it is shown that the improvement in the rPPG performance through skin segmentation is better in an environment with motion artifacts, such as a talking and fast translation environments.

Experimental results show that ROI selection and pulse signal extraction are essential for applying the rPPG technology to fitness, mobile, or driving environments with considerable noise. Moreover, for general applications, rPPG must be robust against the facial expressions, head movements, and facial movements caused by talking, and lighting changes must be considered. Therefore, future studies can examine the mitigation or removal of the motion noise generated in a wild environment through ROI selection to improve the rPPG performance. This work uses the PURE dataset, which consists of a talking environment, head movement, and rotation. However, these occur concurrently in a wild environment. Therefore, in the future, experiments can be performed using datasets from real-world environments.

Additionally, as shown in Table 6, we performed additional experiments to confirm the accuracy according to heart rate on the PURE dataset with a heart rate range of 40~150 bpm. For heart rates lower than 50 beats per minute, the error is higher than the overall average MAPE of 1.81. However, in the case of data over 100 bpm, it was confirmed that the error was lower than the average. In the PURE dataset, there were 10 subjects, and 2 subjects corresponded to high heart rate data of more than 100 bpm, and even this was about 10% when compared with the ratio. In addition, the cases of 80–90 and 90–100 bpm, which have the highest MAPE, had few data, and this was unreliable because the result was for one subject. This is a limitation of this study, considering the balance of heart rate ranges, and acquiring a dataset with a sufficient number of at least 10 subjects per heart range to confirm rPPG characteristics will be pursued in future work. In addition, similar studies include studies on rPPG characteristics according to skin type and transparency.

Table 6. MAPE and time length according to heart rate in PURE dataset [25].

$HeartRate(HR)Range$	MAPE (%)	Number of Frames	Time (s)
$40 \leq HR < 50$	2.59	14,737	491
$50 \leq HR < 60$	1.54	24,886	830
$60 \leq HR < 70$	1.21	20,433	681
$70 \leq HR < 80$	0.51	26,184	873
$80 \leq HR < 90$	26.9	4596	153
$90 \leq HR < 100$	5.24	663	22
$100 \leq HR < 110$	0.75	800	27
$110 \leq HR < 120$	0.3	512	17
$120 \leq HR < 130$	0.5	5483	183
$130 \leq HR < 140$	1.59	3979	133
$140 \leq HR < 150$	0.52	263	9

The attention mechanism can be applied to skin segmentation in deep learning-based rPPG signal measurement. Attention-based performance improvements in facial expression and motion recognition have been obtained in recent years [29,30]. In addition, an ROI mask can be used for training deep learning models. Hwang et al. [31] used a torso ROI mask to train a breathing signal output for a convolutional neural network model. The value of the ROI mask contained information about the direction and amplitude of respiration. Therefore, this information was used as training data. This ROI-mask-based data preprocessing helped train a noise-tolerant model, which is a training method that considers noise and can be applied to deep learning-based rPPG model training.

6. Conclusions

It is experimentally confirmed that rPPG performance can be improved through skin segmentation during ROI selection. We propose Skin-SegNet, which rapidly and accurately finds only the pixels in the skin region. Skin-SegNet prevents motion noise by excluding the lip region and eyelids from the ROI. In addition, it helps obtain a reliable rPPG signal by removing the parts that may contaminate the signal, such as glasses and hair, to find only actual skin pixels. Skin-SegNet performs feature extraction using a small amount of computation by applying the S2 block and group convolution to SINet for real-time processing. Skin-SegNet is a lightweight model that is more than 10 times faster than the Deeplabv3+ structure but shows SOTA-level rPPG measurement. Also, Skin-SegNet shows a performance improvement of up to 20% compared to conventional color domain approaches such as YCbCr and HSV. Skin-SegNet shows the fastest skin segmentation inference time at 15 ms on an Intel i9 CPU and has SOTA-level rPPG performance. Therefore, Skin-SegNet can be used in the preprocessing part of various rPPG application fields, and it can be expected to be used in low-end devices such as mobile devices.

Author Contributions: Conceptualization, E.C.L.; data curation, K.L. and H.Y.; formal analysis, K.L. and J.O.; investigation, K.L. and J.O.; methodology, K.L. and J.O.; project administration, K.L. and E.C.L.; software, K.L. and J.O.; supervision, K.L. and E.C.L.; validation, K.L. and E.C.L.; visualization, K.L., J.O. and H.Y.; writing—original draft, K.L. and J.O.; writing—review and editing, E.C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors have no relevant financial or nonfinancial interest to disclose.

References

1. Gil, E.; Orini, M.; Bailon, R.; Vergara, J.M.; Mainardi, L.; Laguna, P. Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. *Physiol. Meas.* **2010**, *31*, 1271. [CrossRef] [PubMed]
2. Wieringa, F.P.; Mastik, F.; Steen, A.V.D. Contactless multiple wavelength photoplethysmographic imaging: A first step toward “SpO_2 camera” technology. *Ann. Biomed. Eng.* **2005**, *33*, 1034–1041. [CrossRef] [PubMed]
3. Humphreys, K.; Ward, T.; Markham, C. Noncontact simultaneous dual wavelength photoplethysmography: A further step toward noncontact pulse oximetry. *Rev. Sci. Instrum.* **2007**, *78*, 044304. [CrossRef]
4. Phung, S.L.; Bouzerdoum, A.; Chai, D. A novel skin color model in ycbcr color space and its application to human face detection. In Proceedings of the International Conference on Image Processing, Rochester, NY, USA, 22–25 September 2002.
5. Dahmani, D.; Cheref, M.; Larabi, S. Zero-sum game theory model for segmenting skin regions. *Image Vis. Comput.* **2020**, *99*, 103925. [CrossRef]
6. Lewandowska, M.; Rumiński, J.; Kocejko, T.; Nowak, J. Measuring pulse rate with a webcam—A non-contact method for evaluating cardiac activity. In Proceedings of the 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), Szczecin, Poland, 18–21 September 2011; pp. 405–410.
7. De Haan, G.; Jeanne, V. Robust pulse rate from chrominance-based rPPG. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2878–2886. [CrossRef]
8. Wang, W.; Den Brinker, A.C.; Stuijk, S.; De Haan, G. Algorithmic principles of remote PPG. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 1479–1491. [CrossRef] [PubMed]

9. Casado, C.A.; López, M.B. Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces. *arXiv* **2022**, arXiv:2202.04101. [CrossRef] [PubMed]
10. Scherpf, M.; Ernst, H.; Misera, L.; Malberg, H.; Schmidt, M. Skin Segmentation for Imaging Photoplethysmography Using a Specialized Deep Learning Approach. In Proceedings of the 2021 Computing in Cardiology (CinC), Brno, Czech Republic, 13–15 September 2021; pp. 1–4.
11. Verkruysse, W.; Svaasand, L.O.; Nelson, J.S. Stuart. Remote plethysmographic imaging using ambient light. *Opt. Express* **2008**, *16*, 21434–21445. [CrossRef] [PubMed]
12. Bobbia, S.; Benezeth, Y.; Dubois, J. Remote photoplethysmography based on implicit living skin tissue segmentation. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 361–365.
13. Bobbia, S.; Luguern, D.; Benezeth, Y.; Nakamura, K.; Gomez, R.; Dubois, J. Real-time temporal superpixels for unsupervised remote photoplethysmography. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1341–1348.
14. Nikolskaia, K.; Ezhova, N.; Sinkov, A.; Medvedev, M. Skin detection technique based on HSV color model and SLIC segmentation method. In Proceedings of the 4th Ural Workshop on Parallel, Distributed, and Cloud Computing for Young Scientists, Ural-PDC 2018, CEUR Workshop Proceedings, Yekaterinburg, Russia, 15 November 2018; pp. 123–135.
15. Tran, Q.V.; Su, S.F.; Sun, W.; Tran, M.Q. Adaptive pulsatile plane for robust noncontact heart rate monitoring. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *51*, 5587–5599. [CrossRef]
16. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
17. Park, H.; Sjosund, L.; Yoo, Y.; Monet, N.; Bang, J.; Kwak, N. Sinet: Extreme lightweight portrait segmentation networks with spatial squeeze module and information blocking decoder. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2066–2074.
18. Lee, K.; You, H.; Oh, J.; Lee, E.C. Extremely Lightweight Skin Segmentation Networks to Improve Remote Photoplethysmography Measurement. In Proceedings of the International Conference on Intelligent Human Computer Interaction, Tashkent, Uzbekistan, 20–22 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 454–459.
19. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 552–568.
20. Park, H.; Yoo, Y.; Seo, G.; Han, D.; Yun, S.; Kwak, N. C3: Concentrated-comprehensive convolution and its application to semantic segmentation. *arXiv* **2018**, arXiv:1812.04920.
21. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
22. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
23. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9190–9200.
24. Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5549–5558.
25. Stricker, R.; Müller, S.; Gross, H.M. Non-contact video-based pulse rate measurement on a mobile service robot. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25–29 August 2014; pp. 1056–1062.
26. Li, X.; Chen, J.; Zhao, G.; Pietikainen, M. Remote heart rate measurement from face videos under realistic situations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
27. Tulyakov, S.; Alameda-Pineda, X.; Ricci, E.; Yin, L.; Cohn, J.F.; Sebe, N. 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, Las Vegas, NV, USA, 27–30 June 2016.
28. Poh, M.Z.; McDuff, D.J.; Picard, R.W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **2010**, *18*, 10762–10774. [CrossRef] [PubMed]
29. Toisoul, A.; Kossaifi, J.; Bulat, A.; Tzimiropoulos, G.; Pantic, M. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.* **2021**, *3*, 42–50. [CrossRef]

30. Du, W.; Wang, Y.; Qiao, Y. Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Trans. Image Process.* **2017**, *27*, 1347–1360. [CrossRef] [PubMed]
31. Hwang, H.; Lee, K.; Lee, E.C. A Real-time Remote Respiration Measurement Method with Improved Robustness based on a CNN Model. *Appl. Sci.* **2022**, *12*, 11603. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.