bioengineering

MDPI

Article

Contactless Blood Oxygen Saturation Estimation from Facial Videos Using Deep Learning

Chun-Hong Cheng [1,*], Zhikun Yuen [2], Shutao Chen [3], Kwan-Long Wong [3], Jing-Wei Chin [3], Tsz-Tai Chan [3] and Richard H. Y. So [3,4]

[1] Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK
[2] Department of Computer Science, University of Ottawa, Ottawa, ON K1H 8M5, Canada; zyuen077@uottawa.ca
[3] PanopticAI, Hong Kong Science and Technology Parks, New Territories, Hong Kong, China; shutaochen@panoptic.ai (S.C.); kylewong@panoptic.ai (K.-L.W.); nickchin@panoptic.ai (J.-W.C.); tericchan@panoptic.ai (T.-T.C.); rhyso@ust.hk (R.H.Y.S.)
[4] Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China
* Correspondence: cc1722@ic.ac.uk

Abstract: Blood oxygen saturation (SpO_2) is an essential physiological parameter for evaluating a person's health. While conventional SpO_2 measurement devices like pulse oximeters require skin contact, advanced computer vision technology can enable remote SpO_2 monitoring through a regular camera without skin contact. In this paper, we propose novel deep learning models to measure SpO_2 remotely from facial videos and evaluate them using a public benchmark database, VIPL-HR. We utilize a spatial–temporal representation to encode SpO_2 information recorded by conventional RGB cameras and directly pass it into selected convolutional neural networks to predict SpO_2. The best deep learning model achieves 1.274% in mean absolute error and 1.71% in root mean squared error, which exceed the international standard of 4% for an approved pulse oximeter. Our results significantly outperform the conventional analytical Ratio of Ratios model for contactless SpO_2 measurement. Results of sensitivity analyses of the influence of spatial–temporal representation color spaces, subject scenarios, acquisition devices, and SpO_2 ranges on the model performance are reported with explainability analyses to provide more insights for this emerging research field.

Keywords: blood oxygen saturation measurement; deep learning; facial videos; non-contact monitoring; remote health monitoring

Citation: Cheng, C.-H.; Yuen, Z.; Chen, S.; Wong, K.-L.; Chin, J.W.; Chan, T.-T.; So, R.H.Y. Contactless Blood Oxygen Saturation Estimation from Facial Videos Using Deep Learning. *Bioengineering* **2023**, *11*, 251. https://doi.org/10.3390/bioengineering11030251

Academic Editors: Fernando Vaquerizo-Villar and Verónica Barroso-García

Received: 5 February 2024
Revised: 26 February 2024
Accepted: 2 March 2024
Published: 4 March 2024

1. Introduction

Human vital signs, such as blood oxygen saturation (SpO_2), heart rate (HR), respiration rate (RR), blood pressure, and body temperature, are standard parameters used to evaluate a person's health status [1,2]. Specifically, SpO_2 readings indicate whether a person has enough oxygen to operate efficiently. SpO_2 readings are a common metric for trauma management and early detection of diseases like hypoxemia, sleep apnea, and heart diseases [3–5].

The COVID-19 pandemic has critically affected many across the globe. According to [6,7], monitoring only an individual's body temperature is insufficient for detecting COVID-19. Given this limitation, researchers have investigated the feasibility of other vital signs for pandemic control. SpO_2 is a logical candidate for such monitoring. It has been observed that COVID-infected individuals displayed low SpO_2 readings before the occurrence of other respiratory symptoms [8,9]. Additionally, some patients have experienced silent hypoxemia, where they exhibit dangerously low SpO_2 readings without signs of respiratory distress [10]. Wide deployment of an accurate tool that can conveniently

and rapidly monitor SpO_2 would greatly enhance a global ability to control inflammatory infectious diseases such as COVID-19.

Currently, SpO_2 is generally measured non-invasively using pulse oximeters and other wearable devices [11,12]. However, contact-based devices have usability limitations and are impractical for long-term monitoring. Usage for extended periods can cause discomfort and are unsuitable for those with skin sensitivity [13]. Moreover, using contact-based devices for health monitoring may facilitate the spread of infectious diseases. Therefore, contactless approaches for SpO_2 measurement have emerged as highly desirable.

Over the last decade, several contactless SpO_2 measurement approaches have been proposed. Researchers have used a variety of cameras, from infrared cameras [14] and high-quality monochrome cameras equipped with special filters [15–18] to off-the-shelf webcams [19–23], to estimate SpO_2 by capturing subtle light intensity changes on the face. Deep learning techniques have achieved state-of-the-art performance for the remote measurement of physiological signs such as HR [24–39] and RR [36,39–47]. However, remote SpO_2 measurement is still in its infancy, with only a few papers using convolutional neural networks (CNNs) to predict SpO_2 from RGB facial videos [48–50]. Additionally, most existing methods are evaluated on private self-collected datasets, preventing a fair comparison of algorithmic performance [51].

In this paper, we utilize a spatial–temporal representation—that is, a spatial–temporal map (STMap), as proposed in [52]—to encode SpO_2-related physiological information from videos recorded by several consumer-grade RGB cameras. Each STMap is fed into various 2D CNNs for predicting SpO_2 in an end-to-end manner. In addition, We explore the explainability of the model and visualize feature maps of each hidden layer to uncover the process of how it addresses input data. This illustrates the advantage of using an STMap instead of taking the spatial average as input. Moreover, we make use of a public benchmark dataset, VIPL HR [52,53], to conduct our experiments and analysis. This research investigates the feasibility of utilizing a spatial–temporal map for remote SpO_2 measurement and evaluates the proposed method on a public dataset for fair comparison. Our deep learning approach offers these contributions to ongoing research:

- It is trained and evaluated on a large-scale multi-modal public benchmark dataset of facial videos.
- It outperforms conventional contactless SpO_2 measurement approaches, showing potential for applications in real-world scenarios.
- It provides a deep learning baseline for contactless SpO_2 measurement. With this baseline, future research can be benchmarked fairly, facilitating progress in this important emerging field.

2. Literature Review

2.1. Contact-Based SpO_2 Measurement

Today, pulse oximeters are being widely utilized to monitor SpO_2 in a non-invasive manner. The principle underlying SpO_2 measurement through pulse oximetry is known as the Ratio of Ratios [54,55]. Pulse oximeters contain Light Emitter Diodes (LEDs) that generate two different light wavelengths, 660 nm (red) and 940 nm (infrared), to measure the different absorption coefficients of oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb) [56]. The photodetector inside the pulse oximeter analyzes the light absorption of these two wavelengths and produces an absorption ratio from which the SpO_2, as a percentage, can be determined from the table in [57]. Healthy SpO_2 values generally range from 95% to 100% [58]. Equation (1) illustrates how pulse oximeters measure SpO_2.

$$SpO_2 = \frac{C_{HbO_2}}{C_{Hb} + C_{HbO_2}} \times 100\% \quad (1)$$

where C_{HbO2} is the concentration of HbO_2 and C_{Hb} is the concentration of Hb.

2.2. SpO_2 Measurement with RGB Cameras

Since smartphones have become ubiquitous in our daily lives, researchers have explored the possibility of SpO_2 measurements through a smartphone camera [11,12]. Using these methods, subjects place their fingertips on top of the smartphone camera, and SpO_2 is estimated based on the reflected light captured by the camera. However, since most smartphone cameras are visible imaging sensors—that is, they only capture light in the visible portion of the spectrum—they cannot capture infrared wavelengths. To overcome this deficiency, Scully et al. [11] proposed to replace the infrared component of the Ratio of Ratios principle with the blue wavelength, since the difference between the absorption coefficients of HbO_2 and Hb are very similar at the two wavelengths [12,59–61]. Equation (2) illustrates the Ratio of Ratios principle for SpO_2 with an RGB camera.

$$SpO_2 = A - B\frac{(AC_{RED})/(DC_{RED})}{(AC_{BLUE})/(DC_{BLUE})} \quad (2)$$

where AC_{BLUE} and AC_{RED} represent the standard deviations of the blue and red color channels, respectively while DC_{BLUE} and DC_{RED} represent the mean of the blue and red color channels, respectively. A and B are experimentally evaluated coefficients that are determined by identifying the line of best fit between the ratios of the red and blue channels and the SpO_2 estimated by a ground truth device. Following Equation (2), remote SpO_2 measurement with an RGB camera was further validated in [21–23,48,50]. However, only two methods used deep learning and were tested on a public benchmark dataset [48,49].

2.3. Deep Learning-Based Remote Vital Sign Monitoring

Over the last decade, many deep learning-based methods have been developed for remote vital sign monitoring, with many studies focusing on HR [24–39], followed by RR [36,39–47]. In general, the underlying principle behind these methods is remote photoplethysmography (rPPG). When body tissues are illuminated by surrounding light, tiny fluctuations in reflected light intensities due to variation in the concentration of hemoglobin can be captured by conventional cameras, producing the so-called rPPG signal [62,63]. After extracting the rPPG signal, subsequent vital signs such as HR or RR can be obtained by further signal processing.

Among the deep learning-based methods for remote SpO_2 measurement based on RGB facial videos [48–50], Hu et al. [48] utilised a multi-model fusion approach and took advantage of the Ratio of Ratios principle. Hamoud et al. [49] used an XGBoost Regressor [64] to measure SpO_2 with the features extracted by a pre-trained CNN. Akamatus et al. [50] made use of spatial–temporal input that is based on the AC and DC components of the Ratio of Ratios principle.

2.4. Spatial–Temporal Representation for Vital Sign Estimation

For remote physiological measurement from facial videos, the crucial information is extracted from the changes in pixel intensity of the subject's face. Since contactless methods are inherently susceptible to noise such as illumination changes and head movements [24], a spatial-averaging operation is generally performed on the region of interest (face) to enhance the quality of the extracted signal. Niu et al. [52] proposed an rPPG-based spatial-temporal representation, spatial–temporal map (STMap), that is widely used for HR estimation as well as face anti-spoofing [39,52,65–68]. The STMap, a low-dimensional spatial–temporal representation in which physiological information of the original video is embedded, can be directly fed into a CNN, which learns and develops a function for mapping a connection between the STMap and the output vital sign. To the best of our knowledge, there is no existing work that has applied rPPG-based STMaps to predict SpO_2. Given the success of spatial–temporal representations for estimating HR, this motivates us to utilize a similar approach for remote SpO_2 measurement.

3. Materials and Methods

3.1. Spatial–Temporal Map Generation

As shown in Figure 1, we followed an approach similar to that proposed in [52] to generate spatial–temporal maps (STMaps). For each video, we randomly sampled 225 consecutive frames and used a face detector (OpenFace [69]) to obtain the subject's face location. The facial frames were down-sampled to 128 × 128 using an average pooling filter (kernel size = 16 and stride = 16) to reduce noise and image dimension. Each frame was then split into 64 patches (8 × 8, from R_1 to R_{64}), and average pooling was applied to each patch for noise removal. Let $P(x,y,t,c)$ be the intensity value of the pixel with the coordinate(x,y) of the tth frame of the video at c color space, and the average pooling of these patches can be denoted as

$$V_{c,i}(t) = \frac{\sum_{x,y \in R_i} P(x,y,t,c)}{A_{R_i}} \tag{3}$$

where A_{R_i} represents the area of the patch R_i. Then, for each patch, we have a sequential signal with length of 225 for each color space c, which is $V_{c,i} = \{V_{c,i}(1), V_{c,i}(2), \ldots, V_{c,i}(225)\}$. For the case of combining RGB and YUV color space, the value of c should be 6. Lastly, these sequential signals are concatenated to form an STMap, a 2D map generated from a video with embedded SpO_2-related information.

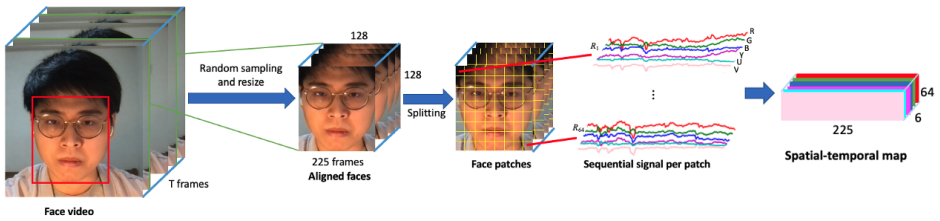

Figure 1. Process of generating a spatial–temporal map in RGB + YUV color spaces.

Other than the traditional RGB color space, an STMap can also be generated from different or a combination of multiple color spaces [65]. In this paper, we transformed the RGB color space to YUV and YCrCb color spaces through Equations (4) and (5), respectively:

$$\begin{aligned} Y &= 0.299 \times R + 0.587 \times G + 0.114 \times B \\ U &= -0.169 \times R - 0.331 \times G + 0.5 \times B + 128 \\ V &= 0.5 \times R - 0.149 \times G - 0.081 \times B + 128 \end{aligned} \tag{4}$$

$$\begin{aligned} Y &= 0.299 \times R + 0.587 \times G + 0.114 \times B \\ Cr &= (R - Y) \times 0.713 + 128 \\ Cb &= (B - Y) \times 0.564 + 128 \end{aligned} \tag{5}$$

The c color dimensions for each face patch were concatenated to produce the final spatial–temporal representation of size 225 × 64 × c. Figure 2 shows a visual example of the STMaps generated from the different color spaces.

3.2. SpO_2 Estimation Using CNNs

We framed SpO_2 estimation as a regression problem and utilized 2D CNNs to predict a single SpO_2 value from an STMap. The STMaps were resized to 225 × 225 to match the input size of the CNNs. We selected and compared three state-of-the-art CNN architectures that are commonly utilized in computer vision tasks, namely ResNet-50 [70], DenseNet-121 [71], and EfficientNet-B3 [72], which were pre-trained with the ImageNet [73] dataset. The last layer of each model was replaced with a regression layer. Table 1 shows their model complexities.

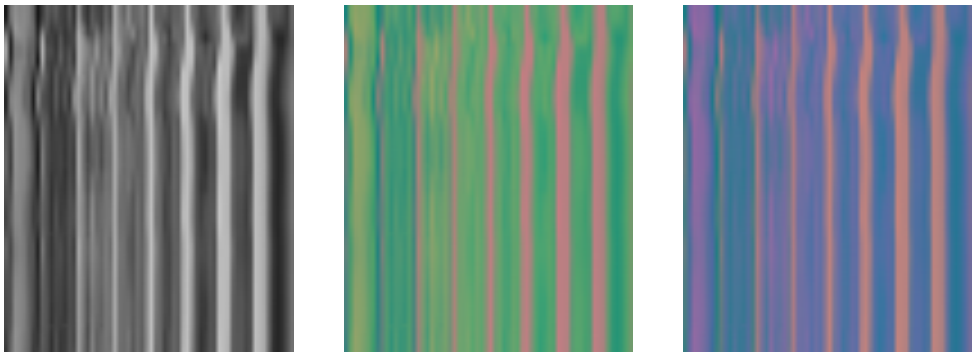

Figure 2. Examples of the spatial–temporal maps (STMaps) in RGB (**left**), YUV (**middle**), and YCrCb (**right**) color spaces generated from the VIPL-HR dataset.

Table 1. Number of parameters (Params) and floating point operations per second (FLOPs) of the selected CNN architectures.

Model	Params	FLOPs
EfficientNet-B3 [72]	9.2 M	1.0 B
ResNet-50 [70]	26 M	4.1 B
DenseNet-121 [71]	8 M	5.7 B

3.3. Dataset

We trained and tested our models on STMaps generated from the VIPL-HR. The VIPL-HR dataset (https://vipl.ict.ac.cn/resources/databases/201811/t20181129_32716.html) (accessed on 20 June 2023) dataset [52,53], is a public-domain dataset originally proposed for remote HR estimation. Since SpO_2 readings were also recorded during the data collection, VIPL-HR can also be used for bench-marking contactless SpO_2 measurement methods. The dataset contains 2378 RGB and 752 near-infrared (NIR) facial videos of 107 subjects (79 males and 28 females, mostly Asians) recorded by four acquisition devices (web camera, smartphone frontal camera, RGB-D camera, and NIR camera). The length of each video is around 30 s, with a frame rate of around 30 frames per second.

For our experiments, we utilized RGB videos of subjects sitting naturally in nine scenarios as follows: (1) at 1 m, (2) while performing large head movements, (3) while reading a text aloud, (4) in a dark environment, (5) in a bright environment, (6) at a long distance (1.5 m instead of 1 m), (7) after doing exercise for 2 min, (8) while holding the smartphone, and (9) while holding the smartphone and performing large head movements. Specific details of the data collection process are listed in [53]. The large variety in the scenarios contributes to the generalizability of the proposed method for different applications. Figure 3 illustrates the distribution of ground truth SpO_2 values for STMaps generated from the VIPL-HR dataset.

3.4. Evaluation Metrics

We utilized the following performance metrics to evaluate the performance of SpO_2 prediction:

- Mean absolute error (MAE) = $\frac{\sum_{i=1}^{N}|x_i - y_i|}{N}$;
- Root mean square error (RMSE) = $\sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}}$.

where x_i is the predicted SpO_2 and y_i is the ground truth SpO_2 in unit of percentage (%).

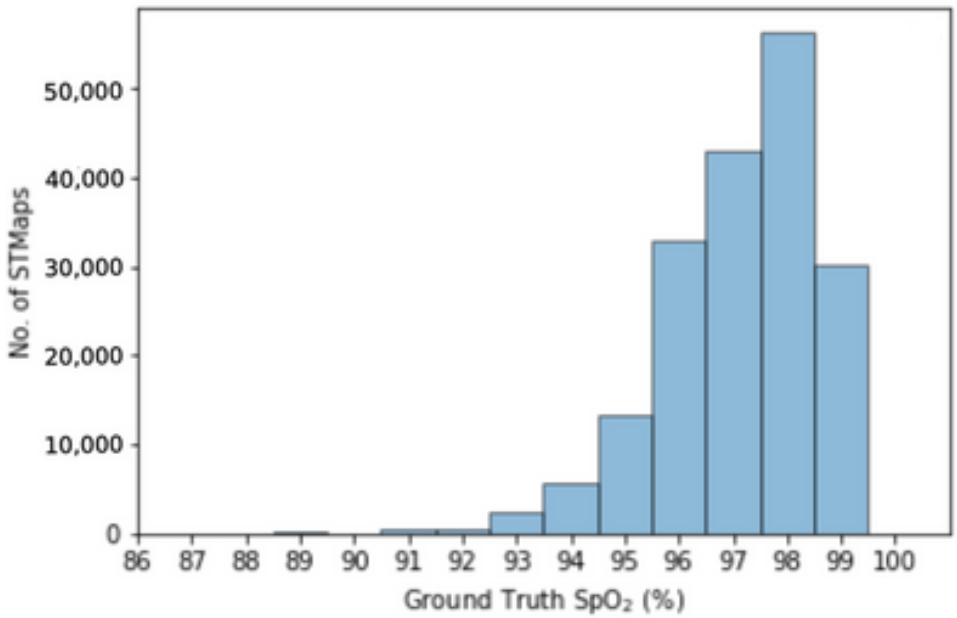

Figure 3. Ground truth SpO2 (%) distribution of STMaps generated from the VIPL-HR dataset.

3.5. Training Settings

To ensure fair evaluation, we performed five-fold subject cross-validation, during which we first separated the subjects into small bins according to the distribution of the SpO2 values of each subject. Each small bin contained at least 5 subjects. Within each bin, the subjects were randomly split into 5 groups. This process guaranteed that the SpO2 values of each fold were equitably distributed. We conducted a Friedman chi-squared test among different folds and the p value was recorded as 0.273, which meant we could not refuse the H_0 hypothesis that the samples were drawn from the same distribution. The final MAE and RMSE results were obtained by averaging over the five folds.

For the training data, we randomly sampled 225 consecutive frames 70 times for each video in the training set to generate STMaps. There are at least 113,068 STMaps for training in each fold. For model training, we used the AdamW optimizer [74] and batch size of 32 on a NVIDIA RTX 3080 GPU. The initial learning rate was set to 0.0001 with a weight decay of 0.001. The RMSE loss function was also utilized for all models. It takes around 12 h to train a single model.

3.6. Feature Map Visualization

While deep learning-based approaches have shown remarkable performance in different vital sign estimation tasks, it is of great interest to uncover what the neural network has learned. A video stream from the dataset was presented to the network and forward-propagated to predict SpO2, during which the responses of hidden convolutional layers on different levels were recorded. The extracted feature maps were averaged over all channels within each layer. As the response map of each layer was a 2D STMap, each row of the feature map that corresponded to a timestamp was detached and separately transformed back to an 8×8 image for better visualization. This process is illustrated in Figure 4.

We applied this process to all convolutional layers to transform the 2D STMap back to interpretable 2D squared image sequences. The results are shown in the next section.

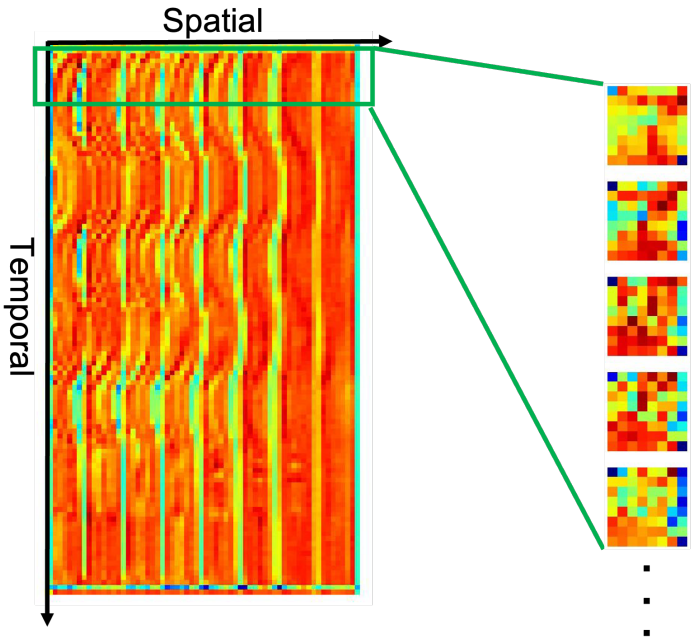

Figure 4. Example of a recorded feature map from the first hidden convolutional layer in blocks. During forward propagation, different color channels were fused; therefore, we average the feature maps over different channels within the layer. Each column corresponds to a patch along the temporal axis and each row corresponds to one frame. For visualization, each row was transformed back to an 8×8 square sequence. The subject's face can still be recognized from the reshaped squares.

4. Results and Discussion

4.1. Performance on STMaps Generated from Different Color Spaces

As mentioned in [52,75], during the generation of the spatial–temporal representation, selecting an appropriate color space can reduce head motion artifacts and improve the overall signal quality of STMaps. To investigate the impact of color space on SpO_2 estimation, we compared the performance of STMaps generated from RGB, YUV, concatenated RGB and YUV, and YCrCb color spaces.

Among the trained models, EfficientNet-B3 trained on concatenated YCrCb STMaps (EfficientNet-B3 + YCrCb) achieved the lowest MAE and RMSE (Table 2) but the combination of YCrCb color space with the other two models resulted in unsatisfactory performance. Moreover, all deep learning models achieved a relatively satisfactory performance when trained on RGB STMaps. This indicates that the introduction of additional color spaces during STMap generation will not improve the deep learning model's performance for SpO_2 estimation, but the selection of appropriate color space will affect the performance. Specifically, RGB color space seems to achieve the most stable performance.

Table 2. Performance of selected deep learning models trained on STMaps generated from different color spaces for SpO_2 estimation.

Model	RGB		YUV		RGB + YUV		YCrCb	
	MAE (%)	RMSE (%)	MAE (%)	RMSE (%)	MAE (%)	RMSE (%)	MAE (%)	RMSE (%)
EfficientNet-B3 [72]	1.274	1.710	1.304	1.756	1.279	1.707	1.273	1.680
ResNet-50 [70]	1.309	1.741	1.307	1.750	1.321	1.781	1.423	1.939
DenseNet-121 [71]	1.284	1.722	1.357	1.783	1.296	1.713	1.421	1.860

4.2. Performance on Different Subject Scenarios and Acquisition Devices

As EfficientNet-B3 + RGB achieved a relatively stable and good performance in the previous experiment, we used EfficientNet-B3 + RGB as our deep learning benchmark for subsequent analysis. We evaluated the performance of our deep learning method against the conventional Ratio of Ratios algorithm for contactless SpO_2 estimation (Equation (2)) with coefficients A and B from previous works [21,22]. We further investigated the performance of these methods in different subject scenarios and acquisition devices in the VIPL-HR dataset. We also included the performance of other deep learning methods [48,49] that have been tested on the VIPL-HR dataset. Additionally, the deep learning method proposed by Hu et al. [48] was first trained on another public dataset, PURE [76], and then fine-tuned on VIPL-HR.

Table 3 highlights that all deep learning methods significantly outperform the conventional Ratio of Ratios algorithm on the VIPL-HR dataset by at least 30% [21] with an up to 66.7% [22] reduction in RMSE. Moreover, the results are within the error range (4%) according to the international standard for a pulse oximeter that can be used for clinical purposes [77], showing the capability of deep learning-based approaches for real-world applications. Notwithstanding, due to the variance in the model's performance between subjects, the historical trends of SpO_2 measurements are often a better indication of the subject's health status than a single measurement at one point in time.

Table 3. Performance of deep learning methods and past analytic methods (Ratio of Ratios) for SpO_2 estimation.

Method	MAE (%)	RMSE (%)
Deep Learning with STMap (EfficientNet-B3 + RGB)	1.274	1.710
Deep Learning [48]	1.000	1.430
Deep Learning [49]	1.170	-
Past Analytic (Ratio of Ratios) [22]	3.334	5.137
Past Analytic (Ratio of Ratios) [21]	1.838	2.489

Figures 5 and 6 show the performance of the tested methods in different subject scenarios in the VIPL-HR dataset (Section 3.3). The deep learning method consistently achieved the lowest MAE (Figure 5) and RMSE (Figure 6) in all cases. Moreover, it is worth noting the significant performance difference between methods in Scenarios 4 and 5, indicating the deep learning method's potential to address illumination variations.

Figures 7 and 8 illustrate the performance of the tested methods on different acquisition devices, including: (1) Logitech C310 web camera (960×720, 25fps), (2) HUAWEI P9 frontal camera (1920×1080, 30fps), and (3) RealSense F200 RGB-D camera (1920×1080, 30fps) in the VIPL-HR dataset. Consistent with the results of subjects in different scenarios, the deep learning method achieved the lowest MAE (Figure 7) and RMSE (Figure 8) for all acquisition devices.

4.3. Performance over Different SpO_2 Ranges

Inspired by Li et al. [78], we analyzed the performance of remote SpO_2 estimation methods over different SpO_2 ranges. The SpO_2 value of a healthy person is usually between 95% and 100% [58]. Based on this classification, we separated the data into two groups: normal ($SpO_2 \geq 95\%$) and abnormal ($SpO_2 < 95\%$).

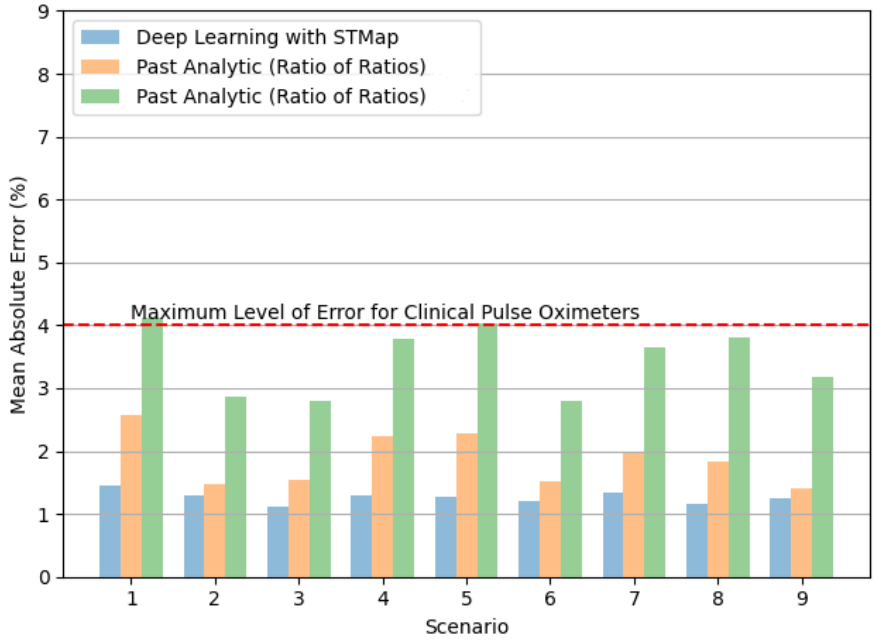

Figure 5. Comparison of mean absolute error (MAE) in remote SpO_2 estimation by deep learning with STMap and past analytic methods (Green refers to [22], Orange refers to [21]) for different subject scenarios of the VIPL-HR dataset.

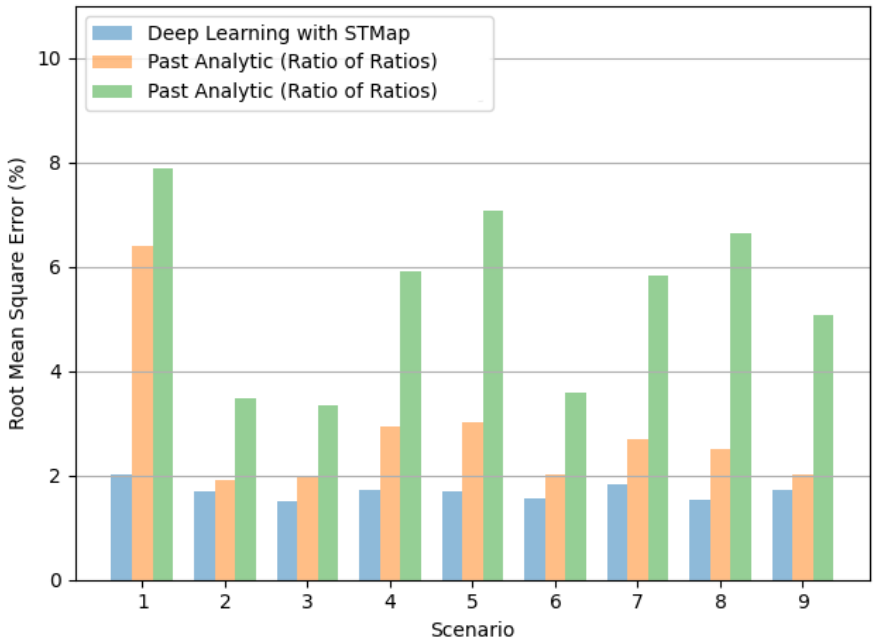

Figure 6. Comparison of root mean square error (RMSE) in remote SpO_2 estimation by deep learning with STMap and past analytic methods (Green refers to [22], Orange refers to [21]) for different subject scenarios of the VIPL-HR dataset.

From Table 4, we observe that the deep learning method outperforms the Ratio of Ratios algorithm in both normal and abnormal SpO_2 ranges. However, the model's MAE and RMSE in the normal range (0.978 and 1.288, respectively) are significantly lower than those in the abnormal range (3.077 and 3.563, respectively). The model's increase in prediction error in the abnormal range may be because the distribution of the training dataset contains fewer low SpO_2 values. Similar to the conclusion drawn in [78] for

predicting HR values in the higher and lower ranges, the challenge of predicting abnormal SpO_2 measurements should be a focus of future works.

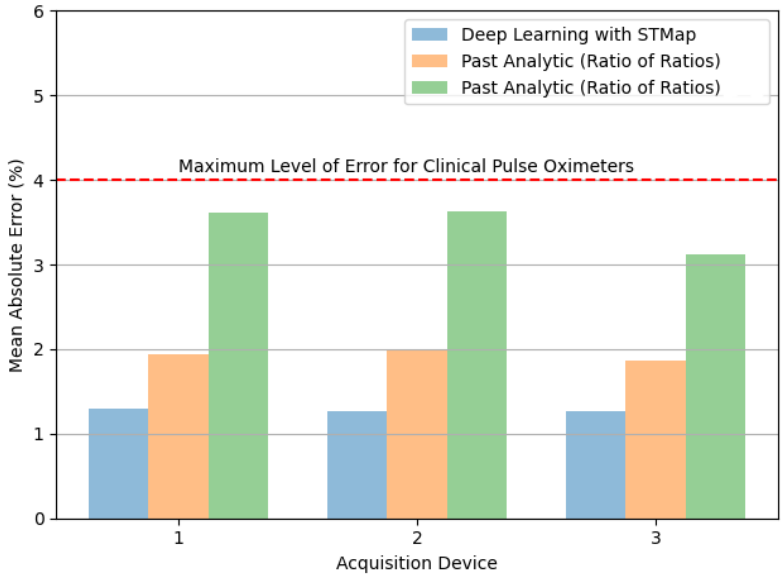

Figure 7. Comparison of mean absolute error (MAE) in remote SpO_2 estimation by deep learning with STMap and past analytic methods (Green refers to [22], Orange refers to [21]) for different acquisition devices (1 = Web Camera, 2 = Smartphone Frontal Camera, 3 = RGB-D Camera) of the VIPL-HR dataset.

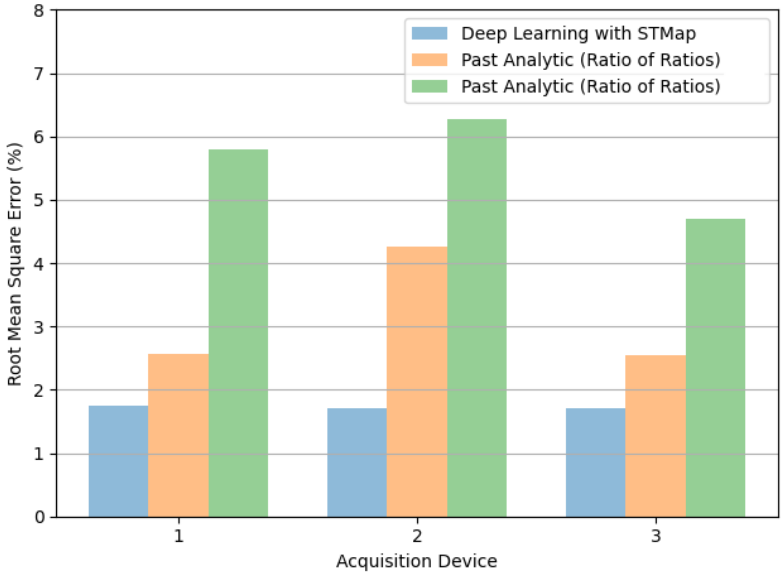

Figure 8. Comparison of root mean square error (RMSE) in remote SpO_2 estimation by deep learning with STMap and past analytic methods (Green refers to [22], Orange refers to [21]) for different acquisition devices (1 = Web Camera, 2 = Smartphone Frontal Camera, 3 = RGB-D Camera) of the VIPL-HR dataset.

Table 4. Performance of deep learning with STMap (EfficientNet-B3 + RGB) and past analytic methods (Ratio of Ratios) for SpO_2 estimation in normal (≥95%) and abnormal (<95%) ranges.

Method	Normal		Abnormal	
	MAE (%)	RMSE (%)	MAE (%)	RMSE (%)
Deep Learning with STMap (EfficientNet-B3 + RGB)	0.978	1.288	3.077	3.563
Past Analytic (Ratio of Ratios) [22]	3.140	4.972	6.798	7.496
Past Analytic (Ratio of Ratios) [21]	1.690	2.264	4.482	5.034

4.4. Feature Maps Learned by CNN Model

In Figure 9, the raw input frame and its down-sampled image are shown on the top two rows and the responses of different hidden layers in the Efficientnet-b3 model are shown sequentially. Here, only the first five convolutional layers are displayed as the feature maps of higher-level convolutional layers are hard to recognize. The untrained model is shown on the left as a sub-figure for comparison while the results for the trained model are shown on the right side. All the values were normalized between 0 and 1.

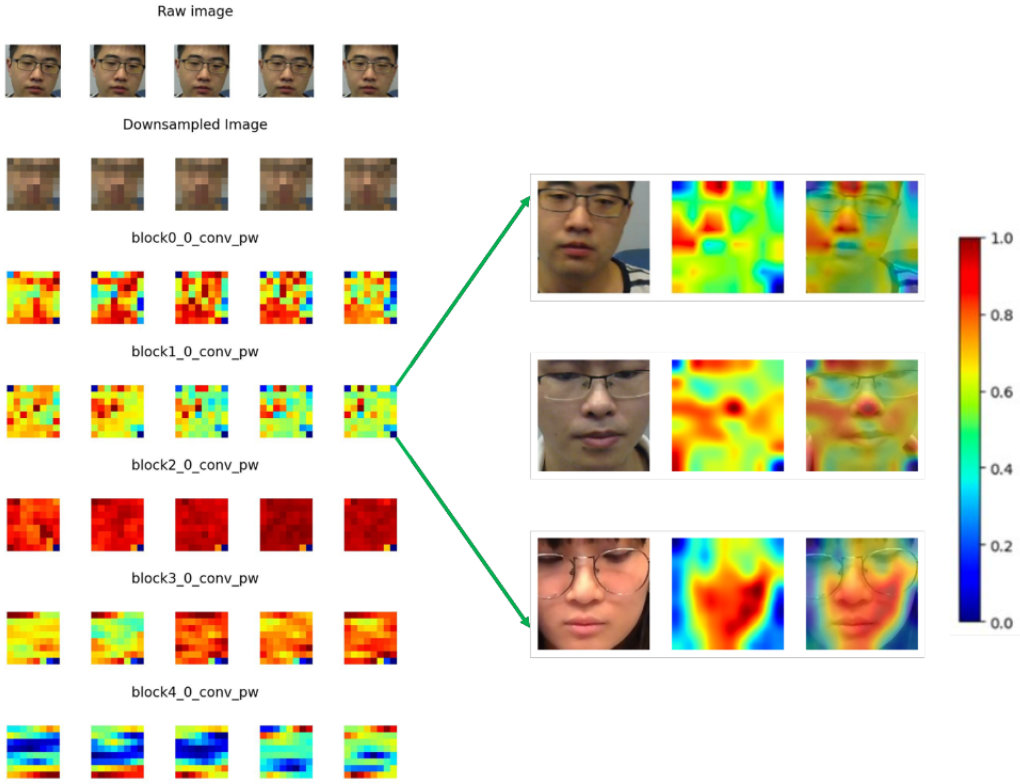

Figure 9. Visualization of feature maps. The left column illustrates the feature maps of hidden convolutional layers for the given input video stream after training for SpO_2 prediction. The first 5 convolutional layers were selected from the sequential blocks of Efficientnet-b3 model. The right column illustrates the raw image, overlayed with the interpolated feature maps extracted from hidden layer *block*1_0_*conv_pw* for 3 subjects in the VIPL-HR dataset.

It can be seen from the feature maps on the left side of Figure 9 that, in the initial *block*0_0_*conv_pw* layer, the outline of the subject is still recognizable by the human eye. For the *block*1_0_*conv_pw* layer, some regions are emphasized with larger weights and others are less stressed. To find the physical meaning of these regions, we aligned the feature map with the raw input frame by applying bicubic interpolation to retain the same

resolution as input raw images and overlayed them; the results of this process are displayed on the right side of Figure 9.

After interpolation, it can be seen clearly from the right side of Figure 9 that, in the *block*1_0_*conv_pw* layer, different face parts were assigned different weights. More specifically, the forehead, nose, and cheeks were assigned a larger weighting while other regions such as the torso or spaces without the human face carried less weight. This result is consistent with findings from many rPPG-related studies, where the forehead, left and right cheeks are often selected as the regions of interest (ROIs) as they carry more physiological information [21,22].

For the hidden convolutional layers in higher levels of the model, the patterns are illegible and therefore not discussed in our study.

5. Conclusions and Future Research Direction

In this paper, we proposed and evaluated a new deep learning method for remote SpO_2 measurement from facial videos in the VIPL-HR public database. We encoded the facial videos into STMaps, low-dimensional spatial–temporal representations containing physiological information of the subject, and directly used them as model inputs for training and testing. Our results indicate that the proposed deep learning method outperforms the conventional Ratio of Ratios technique by reducing the RMSE up to 66.7% when compared across different subject scenarios, acquisition devices, and SpO_2 ranges. This sets a new bench-marking baseline for upcoming research. The visualization of feature maps demonstrated that ROIs around the forehead, nose, and cheeks carry more weight for SpO_2 estimation. These findings increase the explainability of the models.

Regarding the direction of future research, we posit that improving the face detection process can generate more representative STMaps and enhance the model's robustness, especially for videos of subjects with large head movements. We expect that a face detector that operates on a per-frame basis, while taking into consideration the dimensional requirements to generate the STMap, can optimize the signal-to-noise ratio of the spatial–temporal representations. Furthermore, as demonstrated by Niu et al. [65], region-of-interest selection can be incorporated to capture areas that may contain stronger physiological signals. Additionally, further investigation could be directed toward assessing the impact of resizing the STMaps to match the CNN's input dimensions, as this procedure may introduce additional noise to the model. Other feature maps with hidden layers could be investigated to elucidate the mechanism of SpO_2 prediction. Moreover, most of the subjects that participated in the VIPL-HR dataset are Asians with Fitzpatrick Scale skin type III and IV [79]. Therefore, the proposed method may be biased to people with these skin types and may not perform considerably on darker skin tones (type VI), which is a common concern in remote vital sign monitoring [80–82]. Finally, we would like to collect more data of subjects with different skin tones and abnormal SpO_2 readings or to simulate low SpO_2 values through an approach similar to the one used in [59]. Additional data coverage of subjects with diverse skin tones and abnormal SpO_2 values can contribute to the development of more robust and accurate models for contactless SpO_2 measurement.

Author Contributions: C.-H.C., Z.Y., S.C. and K.-L.W.; methodology, C.-H.C., Z.Y. and S.C.; software, C.-H.C., Z.Y., S.C. and K.-L.W.; validation, C.-H.C., Z.Y., S.C. and K.-L.W.; formal analysis, C.-H.C., Z.Y., S.C. and K.-L.W.; investigation, C.-H.C. and Z.Y.; resources, C.-H.C. and Z.Y.; data curation, C.-H.C., Z.Y., S.C. and K.-L.W.; writing—original draft preparation, J.-W.C., T.-T.C. and R.H.Y.S.; writing—review and editing, C.-H.C., Z.Y. and S.C.; visualization, R.H.Y.S.; supervision, K.-L.W.; project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the Innovation Technology Commission of Hong Kong (project number TSSSU/HKUST/21/10 and PsH/053/22). Authors are also grateful for the support of HKSTP incubation program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available dataset was analyzed in this study. This dataset can be found here: https://vipl.ict.ac.cn/resources/databases/201811/t20181129_32716.html, accessed on 26 February 2024.

Conflicts of Interest: Authors Shutao Chen, Kwan-Long Wong, Jing-Wei Chin, Tsz-Tai Chan and Richard H. Y. So were employed by the company PanopticAI. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Castledine, G. The importance of measuring and recording vital signs correctly. *Br. J. Nurs.* **2006**, *15*, 285. [CrossRef]
2. Molinaro, N.; Schena, E.; Silvestri, S.; Bonotti, F.; Aguzzi, D.; Viola, E.; Buccolini, F.; Massaroni, C. Contactless Vital Signs Monitoring from Videos Recorded with Digital Cameras: An Overview. *Front. Physiol.* **2022**, *13* , 160. [CrossRef]
3. Adochiei, F.; Rotariu, C.; Ciobotariu, R.; Costin, H. A wireless low-power pulse oximetry system for patient telemonitoring. In Proceedings of the 2011 7th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, 12 May 2011; pp. 1–4.
4. Dumitrache-Rujinski, S.; Calcaianu, G.; Zaharia, D.; Toma, C.L.; Bogdan, M. The role of overnight pulse-oximetry in recognition of obstructive sleep apnea syndrome in morbidly obese and non obese patients. *Maedica* **2013**, *8*, 237.
5. Ruangritnamchai, C.; Bunjapamai, W.; Pongpanich, B. Pulse oximetry screening for clinically unrecognized critical congenital heart disease in the newborns. *Images Paediatr. Cardiol.* **2007**, *9*, 10.
6. Mitra, B.; Luckhoff, C.; Mitchell, R.D.; O'Reilly, G.M.; Smit, D.V.; Cameron, P.A. Temperature screening has negligible value for control of COVID-19. *Emerg. Med. Australas.* **2020**, *32*, 867–869. [CrossRef]
7. Vilke, G.M.; Brennan, J.J.; Cronin, A.O.; Castillo, E.M. Clinical features of patients with COVID-19: Is temperature screening useful? *J. Emerg. Med.* **2020**, *59*, 952–956. [CrossRef]
8. Pimentel, M.A.; Redfern, O.C.; Hatch, R.; Young, J.D.; Tarassenko, L.; Watkinson, P.J. Trajectories of vital signs in patients with COVID-19. *Resuscitation* **2020**, *156*, 99–106. [CrossRef]
9. Starr, N.; Rebollo, D.; Asemu, Y.M.; Akalu, L.; Mohammed, H.A.; Menchamo, M.W.; Melese, E.; Bitew, S.; Wilson, I.; Tadesse, M.; et al. Pulse oximetry in low-resource settings during the COVID-19 pandemic. *Lancet Glob. Health* **2020**, *8*, e1121–e1122. [CrossRef]
10. Manta, C.; Jain, S.S.; Coravos, A.; Mendelsohn, D.; Izmailova, E.S. An Evaluation of Biometric Monitoring Technologies for Vital Signs in the Era of COVID-19. *Clin. Transl. Sci.* **2020**, *13*, 1034–1044. [CrossRef]
11. Scully, C.G.; Lee, J.; Meyer, J.; Gorbach, A.M.; Granquist-Fraser, D.; Mendelson, Y.; Chon, K.H. Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Trans. Biomed. Eng.* **2011**, *59*, 303–306. [CrossRef]
12. Ding, X.; Nassehi, D.; Larson, E.C. Measuring oxygen saturation with smartphone cameras using convolutional neural networks. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 2603–2610. [CrossRef] [PubMed]
13. Rouast, P.V.; Adam, M.T.; Chiong, R.; Cornforth, D.; Lux, E. Remote heart rate measurement using low-cost RGB face video: A technical literature review. *Front. Comput. Sci.* **2018**, *12*, 858–872. [CrossRef]
14. Stogiannopoulos, T.; Cheimariotis, G.A.; Mitianoudis, N. A Study of Machine Learning Regression Techniques for Non-Contact SpO_2 Estimation from Infrared Motion-Magnified Facial Video. *Information* **2023**, *14*, 301. [CrossRef]
15. Tarassenko, L.; Villarroel, M.; Guazzi, A.; Jorge, J.; Clifton, D.; Pugh, C. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiol. Meas.* **2014**, *35*, 807. [CrossRef]
16. Kong, L.; Zhao, Y.; Dong, L.; Jian, Y.; Jin, X.; Li, B.; Feng, Y.; Liu, M.; Liu, X.; Wu, H. Non-contact detection of oxygen saturation based on visible light imaging device using ambient light. *Opt. Express* **2013**, *21*, 17464–17471. [CrossRef]
17. Shao, D.; Liu, C.; Tsow, F.; Yang, Y.; Du, Z.; Iriya, R.; Yu, H.; Tao, N. Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 1091–1098. [CrossRef]
18. Liao, W.; Zhang, C.; Sun, X.; Notni, G. Oxygen saturation estimation from near-infrared multispectral video data using 3D convolutional residual networks. In Proceedings of the Multimodal Sensing and Artificial Intelligence: Technologies and Applications III. SPIE, Munich, Germany, 9 August 2023; Volume 12621, pp. 177–191.
19. Freitas, U.S. Remote camera-based pulse oximetry. In Proceedings of the 6th International Conference on eHealth, Telemedicine, and Social Medicine, Barcelona, Spain, 23–27 March 2014; pp. 59–63.
20. Guazzi, A.R.; Villarroel, M.; Jorge, J.; Daly, J.; Frise, M.C.; Robbins, P.A.; Tarassenko, L. Non-contact measurement of oxygen saturation with an RGB camera. *Biomed. Opt. Express* **2015**, *6*, 3320–3338. [CrossRef]
21. Bal, U. Non-contact estimation of heart rate and oxygen saturation using ambient light. *Biomed. Opt. Express* **2015**, *6*, 86–97. [CrossRef]
22. Casalino, G.; Castellano, G.; Zaza, G. A mHealth solution for contact-less self-monitoring of blood oxygen saturation. In Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 7 July 2020; pp. 1–7.
23. Cheng, J.C.; Pan, T.S.; Hsiao, W.C.; Lin, W.H.; Liu, Y.L.; Su, T.J.; Wang, S.M. Using Contactless Facial Image Recognition Technology to Detect Blood Oxygen Saturation. *Bioengineering* **2023**, *10*, 524. [CrossRef]

24. Cheng, C.H.; Wong, K.L.; Chin, J.W.; Chan, T.T.; So, R.H. Deep Learning Methods for Remote Heart Rate Measurement: A Review and Future Research Agenda. *Sensors* **2021**, *21*, 6296. [CrossRef]
25. Chen, W.; McDuff, D. Deepphys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8 September 2018; pp. 349–365.
26. Liu, X.; Fromm, J.; Patel, S.; McDuff, D. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19400–19411.
27. Yu, Z.; Peng, W.; Li, X.; Hong, X.; Zhao, G. Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October 2019; pp. 151–160.
28. Perepelkina, O.; Artemyev, M.; Churikova, M.; Grinenko, M. HeartTrack: Convolutional neural network for remote video-based heart rate monitoring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14 June 2020; pp. 288–289.
29. Hu, M.; Qian, F.; Guo, D.; Wang, X.; He, L.; Ren, F. ETA-rPPGNet: Effective time-domain attention network for remote heart rate measurement. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12. [CrossRef]
30. Birla, L.; Shukla, S.; Gupta, A.K.; Gupta, P. ALPINE: Improving remote heart rate estimation using contrastive learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2 January 2023; pp. 5029–5038.
31. Li, B.; Zhang, P.; Peng, J.; Fu, H. Non-contact PPG signal and heart rate estimation with multi-hierarchical convolutional network. *Pattern Recognit.* **2023**, *139*, 109421. [CrossRef]
32. Sun, W.; Sun, Q.; Sun, H.M.; Sun, Q.; Jia, R.S. ViT-rPPG: A vision transformer-based network for remote heart rate estimation. *J. Electron. Imaging* **2023**, *32*, 023024. [CrossRef]
33. Speth, J.; Vance, N.; Flynn, P.; Czajka, A. Non-contrastive unsupervised learning of physiological signals from video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17 June 2023; pp. 14464–14474.
34. Ouzar, Y.; Djeldjli, D.; Bousefsaf, F.; Maaoui, C. X-iPPGNet: A novel one stage deep learning architecture based on depthwise separable convolutions for video-based pulse rate estimation. *Comput. Biol. Med.* **2023**, *154*, 106592. [CrossRef]
35. Wang, R.X.; Sun, H.M.; Hao, R.R.; Pan, A.; Jia, R.S. TransPhys: Transformer-based unsupervised contrastive learning for remote heart rate measurement. *Biomed. Signal Process. Control* **2023**, *86*, 105058. [CrossRef]
36. Gupta, K.; Sinhal, R.; Badhiye, S.S. Remote photoplethysmography-based human vital sign prediction using cyclical algorithm. *J. Biophotonics* **2024**, *17*, e202300286. [CrossRef]
37. Othman, W.; Kashevnik, A.; Ali, A.; Shilov, N.; Ryumin, D. Remote Heart Rate Estimation Based on Transformer with Multi-Skip Connection Decoder: Method and Evaluation in the Wild. *Sensors* **2024**, *24*, 775. [CrossRef] [PubMed]
38. Wu, C.; Chen, J.; Chen, Y.; Chen, A.; Zhou, L.; Wang, X. Pulse rate estimation based on facial videos: An evaluation and optimization of the classical methods using both self-constructed and public datasets. *Tradit. Med. Res.* **2024**, *9*, 2. [CrossRef]
39. Liu, X.; Zhang, Y.; Yu, Z.; Lu, H.; Yue, H.; Yang, J. rPPG-MAE: Self-supervised Pretraining with Masked Autoencoders for Remote Physiological Measurements. *IEEE Trans. Multimed.* **2024**. [CrossRef]
40. Bian, D.; Mehta, P.; Selvaraj, N. Respiratory rate estimation using PPG: A deep learning approach. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20 July 2000; pp. 5948–5952.
41. Ravichandran, V.; Murugesan, B.; Balakarthikeyan, V.; Ram, K.; Preejith, S.; Joseph, J.; Sivaprakasam, M. RespNet: A deep learning model for extraction of respiration from photoplethysmogram. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23 July 2019; pp. 5556–5559.
42. Liu, Z.; Huang, B.; Lin, C.L.; Wu, C.L.; Zhao, C.; Chao, W.C.; Wu, Y.C.; Zheng, Y.; Wang, Z. Contactless Respiratory Rate Monitoring for ICU Patients Based on Unsupervised Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17 June 2023; pp. 6004–6013.
43. Yue, Z.; Shi, M.; Ding, S. Facial Video-based Remote Physiological Measurement via Self-supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13844–13859. [CrossRef]
44. Brieva, J.; Ponce, H.; Moya-Albor, E. Non-Contact Breathing Rate Estimation Using Machine Learning with an Optimized Architecture. *Mathematics* **2023**, *11*, 645. [CrossRef]
45. Lee, H.; Lee, J.; Kwon, Y.; Kwon, J.; Park, S.; Sohn, R.; Park, C. Multitask Siamese Network for Remote Photoplethysmography and Respiration Estimation. *Sensors* **2022**, *22*, 5101. [CrossRef] [PubMed]
46. Vatanparvar, K.; Gwak, M.; Zhu, L.; Kuang, J.; Gao, A. Respiration Rate Estimation from Remote PPG via Camera in Presence of Non-Voluntary Artifacts. In Proceedings of the 2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN), Ioannina, Greece, 27 September 2022; pp. 1–4.
47. Ren, Y.; Syrnyk, B.; Avadhanam, N. Improving video-based heart rate and respiratory rate estimation via pulse-respiration quotient. In Proceedings of the Workshop on Healthcare AI and COVID-19, Baltimore, MD, USA, 22 July 2022; pp. 136–145.
48. Hu, M.; Wu, X.; Wang, X.; Xing, Y.; An, N.; Shi, P. Contactless blood oxygen estimation from face videos: A multi-model fusion method based on deep learning. *Biomed. Signal Process. Control* **2023**, *81*, 104487. [CrossRef]

49. Hamoud, B.; Othman, W.; Shilov, N.; Kashevnik, A. Contactless Oxygen Saturation Detection Based on Face Analysis: An Approach and Case Study. In Proceedings of the 2023 33rd Conference of Open Innovations Association (FRUCT), Zilina, Slovakia, 24 May 2023; pp. 54–62.
50. Akamatsu, Y.; Onishi, Y.; Imaoka, H. Blood Oxygen Saturation Estimation from Facial Video Via DC and AC Components of Spatio-Temporal Map. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4 June 2023; pp. 1–5.
51. Gupta, A.; Ravelo-Garcia, A.G.; Dias, F.M. Availability and performance of face based non-contact methods for heart rate and oxygen saturation estimations: A systematic review. *Comput. Methods Programs Biomed.* **2022**, *219*, 106771. [CrossRef]
52. Niu, X.; Shan, S.; Han, H.; Chen, X. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Trans. Image Process.* **2019**, *29*, 2409–2423. [CrossRef]
53. Niu, X.; Han, H.; Shan, S.; Chen, X. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 562–576.
54. Severinghaus, J.W. Takuo Aoyagi: Discovery of pulse oximetry. *Anesth. Analg.* **2007**, *105*, S1–S4. [CrossRef]
55. Tian, X.; Wong, C.W.; Ranadive, S.M.; Wu, M. A Multi-Channel Ratio-of-Ratios Method for Noncontact Hand Video Based SpO_2 Monitoring Using Smartphone Cameras. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 197–207. [CrossRef]
56. Lopez, S.; Americas, R. Pulse oximeter fundamentals and design. *Free. Scale Semicond.* **2012**, 23.
57. Azhar, F.; Shahrukh, I.; Zeeshan-ul Haque, M.; Shams, S.; Azhar, A. An Hybrid Approach for Motion Artifact Elimination in Pulse Oximeter using MatLab. In Proceedings of the 4th European Conference of the International Federation for Medical and Biological Engineering, Antwerp, Belgium, 23 November 2008; Springer: Berlin/Heidelberg, Germnay, 2009; Volume 22, pp. 1100–1103.
58. Nitzan, M.; Romem, A.; Koppel, R. Pulse oximetry: Fundamentals and technology update. *Med. Devices* **2014**, *7*, 231. [CrossRef]
59. Mathew, J.; Tian, X.; Wu, M.; Wong, C.W. Remote Blood Oxygen Estimation From Videos Using Neural Networks. *arXiv* **2021**, arXiv:2107.05087.
60. Schmitt, J. *Optical Measurement of Blood Oxygenation by Implantable Telemetry*; Technical Report G558–15; Stanford University: Stanford, CA, USA, 1986.
61. Takatani, S.; Graham, M.D. Theoretical analysis of diffuse reflectance from a two-layer tissue model. *IEEE Trans. Biomed. Eng.* **1979**, *26*, 656–664. [CrossRef]
62. Sun, Y.; Thakor, N. Photoplethysmography revisited: From contact to noncontact, from point to imaging. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 463–477. [CrossRef]
63. Xiao, H.; Liu, T.; Sun, Y.; Li, Y.; Zhao, S.; Avolio, A. Remote photoplethysmography for heart rate measurement: A review. *Biomed. Signal Process. Control* **2024**, *88*, 105608. [CrossRef]
64. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13 August 2016; pp. 785–794.
65. Niu, X.; Yu, Z.; Han, H.; Li, X.; Shan, S.; Zhao, G. Video-based remote physiological measurement via cross-verified feature disentangling. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 295–310.
66. Yu, Z.; Li, X.; Wang, P.; Zhao, G. Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE Signal Process. Lett.* **2021**, *28*, 1290–1294. [CrossRef]
67. Niu, X.; Han, H.; Shan, S.; Chen, X. Synrhythm: Learning a deep heart rate estimator from general to specific. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20 August 2018; pp. 3580–3585.
68. Niu, X.; Zhao, X.; Han, H.; Das, A.; Dantcheva, A.; Shan, S.; Chen, X. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14 May 2019; pp. 1–8.
69. Baltrusaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L.P. OpenFace 2.0: Facial Behavior Analysis Toolkit. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), Xi'an, China, 15 May 2018; pp. 59–66. [CrossRef]
70. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27 June 2016; pp. 770–778.
71. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21 July 2017; pp. 4700–4708.
72. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9 June 2019; pp. 6105–6114.
73. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
74. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April 2018.
75. Yang, Y.; Liu, C.; Yu, H.; Shao, D.; Tsow, F.; Tao, N. Motion robust remote photoplethysmography in CIELab color space. *J. Biomed. Opt.* **2016**, *21*, 117001. [CrossRef]

76. Stricker, R.; Müller, S.; Gross, H.M. Non-contact video-based pulse rate measurement on a mobile service robot. In Proceedings of the The 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25 August 2014; pp. 1056–1062.
77. International Organization for Standardization. *Particular Requirements for Basic Safety and Essential Performance of Pulse Oximeter Equipment*; International Organization for Standardization: Geneva, Switzerland, 2011.
78. Li, X.; Han, H.; Lu, H.; Niu, X.; Yu, Z.; Dantcheva, A.; Zhao, G.; Shan, S. The 1st challenge on remote physiological signal sensing (repss). In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13 June 2020; pp. 314–315.
79. Fitzpatrick, T.B. The validity and practicality of sun-reactive skin types I through VI. *Arch. Dermatol.* **1988**, *124*, 869–871. [CrossRef]
80. Nowara, E.M.; McDuff, D.; Veeraraghavan, A. A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13 June 2020; pp. 284–285.
81. Shirbani, F.; Hui, N.; Tan, I.; Butlin, M.; Avolio, A.P. Effect of ambient lighting and skin tone on estimation of heart rate and pulse transit time from video plethysmography. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20 July 2020; pp. 2642–2645.
82. Dasari, A.; Prakash, S.K.A.; Jeni, L.A.; Tucker, C.S. Evaluation of biases in remote photoplethysmography methods. *NPJ Digit. Med.* **2021**, *4*, 91. [CrossRef]