An Algorithmic Benchmark for Contactless Blood Oxygen Saturation Measurement from Facial Videos

Chun Hong Cheng[1], Zhikun Yuen[1], Wong Kwan Long[1,2], Jing Wei Chin[2]
Tsz Tai Chan[2], Richard So[1,2]
[1]*HKUST*
[2]*PanopticAI*

ABSTRACT

Blood oxygen saturation (SpO_2) is an important physiological sign for evaluating a person's health, where low levels of SpO_2 can indicate early signs of diseases such as COVID-19. While conventional SpO_2 measurement devices, such as pulse oximeters, require skin-contact, advanced computer vision approaches can enable remote SpO_2 monitoring through a regular camera. In this paper, we propose the first set of deep learning baselines for remote SpO_2 measurement from facial videos and evaluate them on a public benchmark database. We utilize a spatial-temporal representation to encode SpO_2 information recorded by conventional RGB cameras and directly pass them into various convolutional neural networks to predict SpO_2. The proposed deep learning-based approaches significantly outperform the existing statistical model for contactless SpO_2 measurement. We further analyze the impact of varying the spatial-temporal representation color space, subject scenarios, acquisition devices, and SpO_2 ranges to set the first benchmarks for the emerging research field.

CCS CONCEPTS

• **Applied computing → Health care information systems**.

KEYWORDS

non-contact monitoring, blood oxygen saturation measurement, deep learning, benchmark

ACM Reference Format:
Chun Hong Cheng[1], Zhikun Yuen[1], Wong Kwan Long[1,2], Jing Wei Chin[2], Tsz Tai Chan[2], Richard So[1,2]. 2022. An Algorithmic Benchmark for Contactless Blood Oxygen Saturation Measurement from Facial Videos. In *Proceedings of IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE'22).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Human vital signs, such as blood oxygen saturation (SpO_2), heart rate (HR), respiration rate, blood pressure, and body temperature, are standard parameters to illustrate a person's health status [7, 19]. Specifically, SpO_2 readings indicate whether a person has enough oxygen supply to operate efficiently and is a common metric for trauma management and early detection of diseases like hypoxemia [1].

The COVID-19 pandemic has critically affected many across the globe. According to [24, 46], 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 [32, 39]. Additionally, some patients experienced silent hypoxemia, in which they exhibit dangerously low SpO_2 readings without signs of respiratory distress [22]. Wide deployment of an accurate tool that can conveniently, quickly monitor SpO_2 in the general public would greatly enhance our ability to control inflammatory infectious diseases such as COVID-19.

Nowadays, SpO_2 is generally measured non-invasively through the use of pulse oximeters and other wearable devices [37, 10, 11]. However, contact-based devices have usability limitations and are impractical for long-term monitoring. Usage for extended periods can be uncomfortable and unsuitable for people who have sensitive skin [34]. Therefore, contactless approaches for SpO_2 measurement have emerged as an attractive alternative.

Over the last decade, several contactless SpO_2 measurement approaches have been proposed. Researchers have used a variety of cameras, from high-quality monochrome cameras equipped with special filters [43, 45, 16, 38, 44] to off-the-shelf webcams [3, 6], to estimate SpO_2 by capturing the subtle light intensity changes on the face. While pulse oximeters utilize red and infrared wavelengths for SpO_2 estimation, these methods replaced the infrared wavelength with the blue one since conventional cameras cannot capture it. Deep learning techniques have achieved state-of-the-art for remote measurement of physiological signs such as HR [9] and RR [5, 33]. However, remote SpO_2 measurement is still at its infancy, with only one deep learning-based paper using a 2D convolutional neural network (CNN) to predict SpO_2 from hand videos [23]. Additionally, existing methods are all evaluated on private self-collected datasets, preventing fair comparison of algorithmic performance.

In this paper, we utilize a spatial-temporal representation—that is, a spatial-temporal map (STMap) as proposed in [28]—to encode SpO_2 information from RGB 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. Moreover, we make use of a public benchmark dataset, VIPL-HR [28, 27], to conduct our experiments and analysis. The main contributions of our work are listed as follows:

CHASE'22, November 2022, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

- It is the first set of deep learning-based remote SpO_2 measurement methods that are 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 acts as a strong baseline for contactless SpO_2 measurement and allows future works to be benchmarked fairly, facilitating the research process of this emerging field.

2 RELATED WORKS

2.1 Contact-based SpO_2 Measurement

Today, pulse oximeters are one of the most commonly used devices for non-invasive monitoring of SpO_2. The principle underlying SpO_2 measurement through pulse oximetry is known as the Ratio of Ratios method. Pulse oximeters contain Light Emitter Diodes (LEDs) that generate two different light wavelengths, 660nm (red) and 940nm (infrared), to measure the different absorption coefficients of oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb) [20]. 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 %, can be determined from a table [2]. Healthy SpO_2 values generally range from 95% to 100% [25]. Equation 1 illustrates how pulse oximeters measure SpO_2.

$$SpO_2 = \frac{C_{HbO_2}}{C_{Hb} + C_{HbO_2}} \times 100\% \tag{1}$$

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

2.2 SpO_2 Measurement with RGB Camera

Since smartphones have become ubiquitous in our daily lives, researchers have explored the possibility of SpO_2 measurement through a smartphone camera [37, 10]. In 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. [37] proposed to replace the infrared component of the Ratio of Ratios method with the blue wavelength, since the difference between the absorption coefficient of HbO_2 and Hb are very similar at the two wavelengths [23, 10, 36, 41]. Equation 2 illustratres the Ratio of Ratios method for SpO_2 with an RGB camera.

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

where AC_{BLUE} and AC_{RED} represent the standard deviation of the blue and red color channels while DC_{BLUE} and DC_{RED} represent the mean of the blue and red color channels. 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.

2.3 Deep Learning-Based Remote Vital Signs Monitoring

During the last decade, many deep learning-based approaches have been developed for remote vital signs monitoring, with a majority of works focusing on HR [9, 8, 18, 49, 31, 13], followed by RR [5, 33]. 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 [40]. After extracting the rPPG signal, subsequent vital signs such as HR or RR can be obtained by further signal processing.

At the time of writing this paper, there is only one deep learning-based method for remote SpO_2 measurement [23]. It utilizes a 2D CNN to predict SpO_2 from a private dataset of hand videos. Novel approaches for remote SpO_2 measurement evaluated on a public benchmark dataset would be highly beneficial for the research community.

2.4 Spatial-temporal Representation for Vital Signs 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 [9], a spatial-averaging operation is generally performed on the region-of-interest (face) to improve the quality of the extracted signal. Niu et al. [28] proposed a spatial-temporal representation, spatial-temporal map (STMap), that is widely used for HR estimation as well as face anti-spoofing [28, 29, 48, 26, 30]. 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 are no existing works that have applied 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 METHODS

3.1 Spatial-temporal Maps Generation

As shown in Figure 1, we followed an approach similar to that proposed in [28] to generate spatial-temporal maps (STMaps). For each video, we randomly sampled 225 consecutive frames and used a face detector (OpenFace [4]) to obtain the subject's face location. The facial frames were downsampled to 128 x 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 x 8), and the average value of the color channels within each patch was extracted into a temporal sequence.

Other than the traditional RGB color space, an STMap can also be generated from different or a combination of multiple color spaces [29]. In this paper, we transformed the RGB color space to YUV and

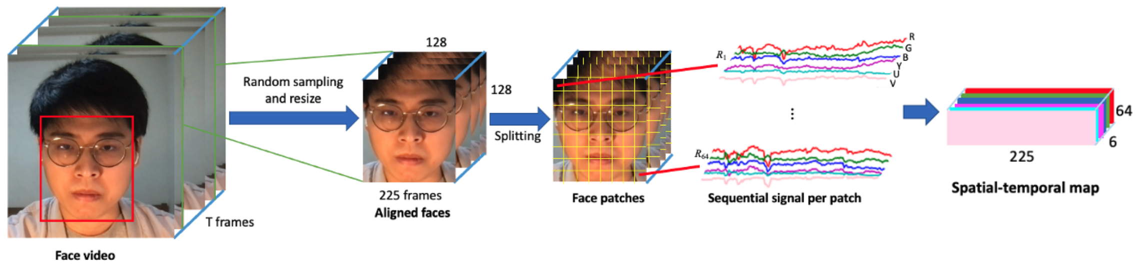


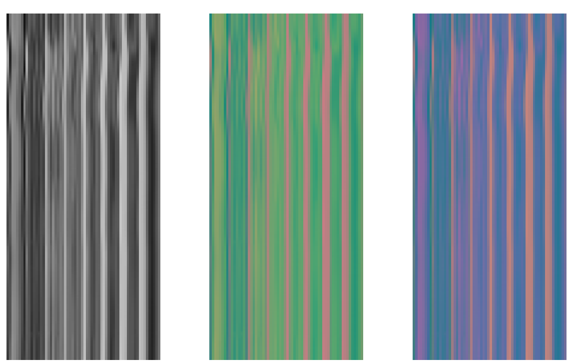
Figure 1: Process of generating a spatial-temporal map in RGB + YUV color spaces.

YCrCb through Equations 3 and 4 respectively:

$$\begin{aligned} Y &= 0.299 \times R + 0.587 \times G + 0.114 \times B \\ U &= -0.169 \times R - 0.331 \times G + 0.5 \times B + 128 \\ V &= 0.5 \times R - 0.149 \times G - 0.081 \times B + 128 \end{aligned} \tag{3}$$

$$\begin{aligned} Y &= 0.299 \times R + 0.587 \times G + 0.114 \times B \\ Cr &= (R - Y) \times 0.713 + 128 \\ Cb &= (B - Y) \times 0.564 + 128 \end{aligned} \tag{4}$$

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



(a) RGB map (b) YUV map (c) YCrCb map

Figure 2: Example of the spatial-temporal maps (STMaps) in RGB, YUV and YCrCb color spaces generated from the VIPL-HR dataset.

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 x 225 to match the input size of the CNNs. We selected and compared three state-of-the-art CNN architectures, including ResNet-50 [12], DenseNet-121 [14] and EfficientNet-B3 [42], that were pretrained with the ImageNet [35] dataset. Table 1 shows the model complexity of the selected models.

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

Model	Params	FLOPs
EfficientNet-B3 [42]	9.2M	1.0B
ResNet-50 [12]	26M	4.1B
DenseNet-121 [14]	8M	5.7B

3.3 Dataset

We trained and tested our methods on STMaps generated from the VIPL-HR dataset [28, 27], a public-domain dataset originally proposed for remote HR estimation. Since SpO_2 readings were also recorded during the data collection, VIPL-HR can be used for benchmarking contactless SpO_2 measurement methods as well. The dataset contains 2378 RGB and 752 near-infrared (NIR) facial videos of 107 subjects (79 males and 28 females) recorded by four acquisition devices (web camera, smartphone frontal camera, RGB-D camera, and NIR camera). The length of each video is around 30 seconds, with a frame rate of around 30 frames per second.

For our experiments, we utilized RGB videos of subjects in nine scenarios, including subjects sitting naturally: (1) at a distance of 1 meter, (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 meters instead of 1 meter), (7) after doing exercise for 2 minutes, (8) while holding the smartphone, and (9) while holding the smartphone and performing large head movements. Specific details of the data collection process is listed in [27]. The large variety of scenarios will contribute to the generalizability of the proposed methods 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 units of percent (%).

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 [42]	1.037	1.487	1.051	1.488	**1.012**	**1.473**	1.066	1.525
ResNet-50 [12]	1.109	1.568	1.098	1.532	**1.089**	**1.499**	1.099	1.525
DenseNet-121 [14]	1.118	1.579	1.110	1.579	**1.087**	**1.538**	1.104	1.589

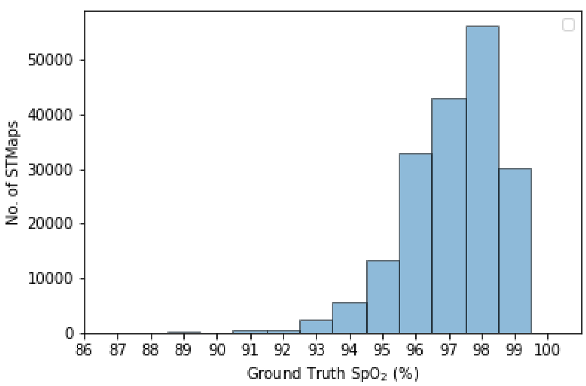


Figure 3: Ground truth SpO_2 (%) distribution of STMaps generated from the VIPL-HR dataset.

3.5 Training Settings

To ensure a fair evaluation process, we performed a 70:30 train-test split based on subjects. We randomly sampled 225 consecutive frames 70 times for each video in the train and test sets to generate STMaps. Figure 4 depicts the distribution of SpO_2 values of STMaps in the train and test sets.

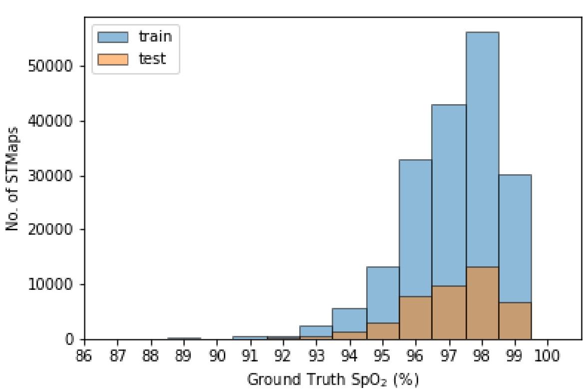


Figure 4: SpO_2 (%) distribution of STMaps in the train and test sets.

Table 3: Performance of deep learning (EfficientNet-B3 + RGB & YUV) and Ratio of Ratios methods for SpO_2 estimation.

Method	MAE (%)	RMSE (%)
EfficientNet-B3 + RGB & YUV	**1.012**	**1.473**
Ratio of Ratios (A = 125, B = 26) [16, 6]	3.334	5.137
Ratio of Ratios (A = 101.6, B = 5.834) [3]	1.838	2.489

For model training, we used the AdamW optimizer [21] 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.

4 RESULTS AND DISCUSSION

4.1 Performance on Different Color Spaces

As mentioned in [28, 47], selecting an appropriate color space of the spatial-temporal representation can reduce head motion artifacts and improve the overall signal quality. 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 proposed methods, EfficientNet-B3 trained on concatenated RGB and YUV STMaps (EfficientNet-B3 + RGB & YUV) achieved the lowest MAE and RMSE (Table 2). Although all models displayed the lowest MAE and RMSE when trained on concatenated RGB and YUV STMaps, the performance across different color spaces is very similar. Further investigation is required to evaluate whether there is a significant difference between a model's performance of SpO_2 estimation and the color space of the spatial-temporal representation.

4.2 Performance on Different Subject Scenarios and Acquisition Devices

As all models achieved a similar performance in the previous experiment, we used EfficientNet-B3 + RGB & YUV as a deep learning benchmark for subsequent analysis. We evaluated the performance of the 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 [3, 16, 6]. We further investigated the performance of the methods in different subject scenarios and acquisition devices in the VIPL-HR dataset.

Table 3 shows that the deep learning method significantly outperforms the conventional Ratio of Ratios method on the VIPL-HR dataset. 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 [15], indicating the potential of deep learning-based methods for real-world applications.

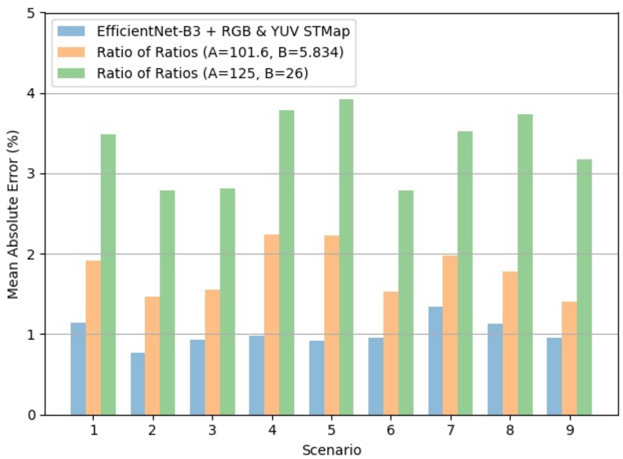


Figure 5: Mean Absolute Error (MAE) in percent (%) of remote SpO_2 estimation methods for different subject scenarios of the VIPL-HR dataset.

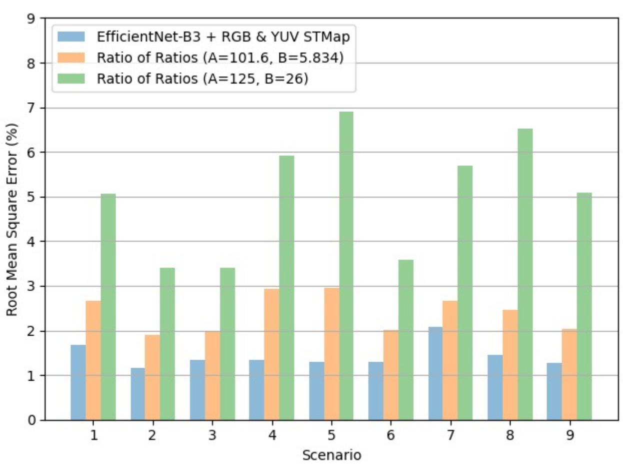


Figure 6: Root Mean Square Error (RMSE) in percent (%) of remote SpO_2 estimation methods for different scenarios of the VIPL-HR dataset.

Figure 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

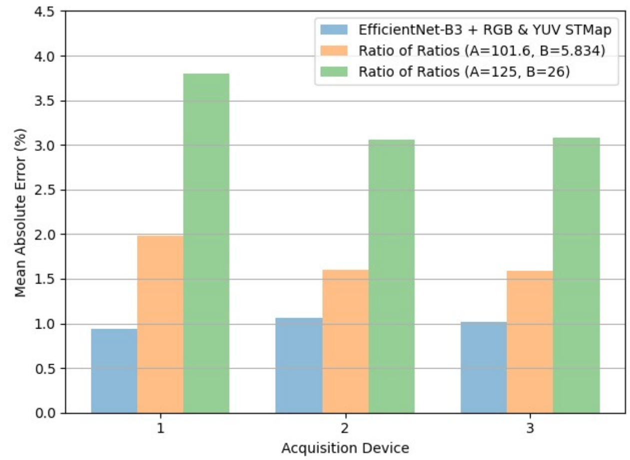


Figure 7: Mean Absolute Error (MAE) in percent (%) of remote SpO_2 estimation methods of different acquisition devices (1 = Web Camera, 2 = Smartphone Frontal Camera, 3 = RGB-D Camera) from the VIPL-HR dataset.

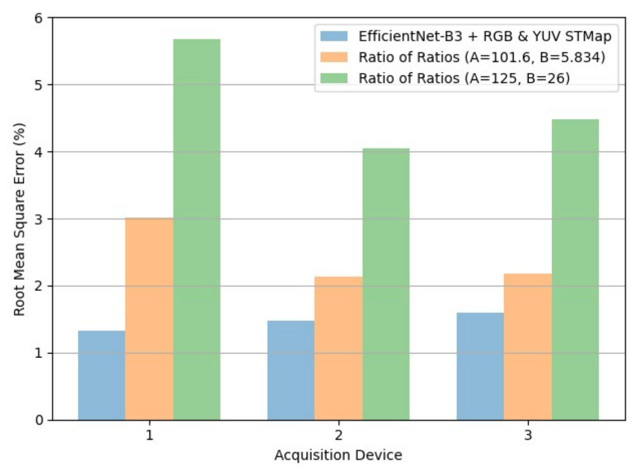


Figure 8: Root Mean Square Error (RMSE) in percent (%) of remote SpO_2 estimation methods of different acquisition devices (1 = Web Camera, 2 = Smartphone Frontal Camera, 3 = RGB-D Camera) from the VIPL-HR dataset.

Scenarios 4 and 5, indicating the deep learning method's potential to address illumination variations.

Figure 7 and 8 illustrate the performance of the tested methods on different acquisition devices in the VIPL-HR dataset, including: (1) Logitech C310 web camera (960 x 720, 25fps), (2) HUAWEI P9 frontal camera (1920 x 1080, 30fps), and (3) RealSense F200 RGB-D camera (1920 x 1080, 30fps). 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. Meanwhile, it can be seen that the conventional Ratio of Ratios

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

Method	Normal MAE RMSE (%) (%)	Abnormal MAE RMSE (%) (%)
EfficientNet-B3 + RGB & YUV	**0.978 1.288**	**3.077 3.563**
Ratio of Ratios (A = 125, B = 26) [16, 6]	3.140 4.972	6.798 7.496
Ratio of Ratios (A = 101.6, B = 5.834) [3]	1.690 2.264	4.482 5.034

method is likely affected by the resolution of the acquisition device, as shown in its mediocre performance when tested on videos captured by the web camera (lowest resolution).

4.3 Performance over Different SpO_2 Ranges

Inspired by Li et al. [17], 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% to 100%. Based on this classification, we separated the data into two groups: normal ($SpO_2 \geq 95\%$) and abnormal ($SpO_2 < 95\%$).

From Table 4, we observe that the deep learning method outperforms the Ratio of Ratios method 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 due to the distribution of the training dataset containing a smaller amount of low SpO_2 values (Figure 4). Similar to the conclusion drawn in [17] 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.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed the first deep learning benchmarks 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 the model inputs for training and testing. We then investigated the model performances using different STMap color spaces, on different subject scenarios, acquisition devices, and over different SpO_2 ranges. The proposed deep learning methods outperform the conventional Ratio of Ratios technique in all cases, setting a solid baseline for upcoming research.

For future work, we believe 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 representation. Furthermore, as demonstrated by Niu et al. [29], region-of-interest selection can be incorporated to capture areas that may contain a stronger physiological signal. Additionally, we would like to investigate the impact of resizing the STMaps to match the CNN's input dimensions, as this procedure may have introduced additional noise to the model. Last but not least, we would like to collect more data of subjects with abnormal SpO_2 readings or simulate low SpO_2 values through a similar approach in [23]. Additional data coverage of subjects with abnormal SpO_2 values can contribute to the development of more robust and accurate models for contactless SpO_2 measurement.

REFERENCES

[1] Felix Adochiei, Cristian Rotariu, Razvan Ciobotariu, and Hariton Costin. 2011. A wireless low-power pulse oximetry system for patient telemonitoring. In *2011 7th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*. IEEE, 1–4.

[2] Faisal Azhar, Ijlal Shahrukh, M Zeeshan-ul-Haque, Sarmad Shams, and Ahsan Azhar. 2009. An hybrid approach for motion artifact elimination in pulse oximeter using matlab. In *4th European Conference of the International Federation for Medical and Biological Engineering, Antwerp: Springer Berlin Heidelberg*. Vol. 22, 1100–1103.

[3] Ufuk Bal. 2015. Non-contact estimation of heart rate and oxygen saturation using ambient light. *Biomedical optics express*, 6, 1, 86–97.

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 59–66. DOI: 10.1109/FG.2018.00019.

[5] Dayi Bian, Pooja Mehta, and Nandakumar Selvaraj. 2020. Respiratory rate estimation using ppg: a deep learning approach. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 5948–5952.

[6] Gabriella Casalino, Giovanna Castellano, and Gianluca Zaza. 2020. A mhealth solution for contact-less self-monitoring of blood oxygen saturation. In *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–7.

[7] George Castledine. 2006. The importance of measuring and recording vital signs correctly. *British Journal of Nursing*, 15, 5, 285–285.

[8] Weixuan Chen and Daniel McDuff. 2018. Deepphys: video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 349–365.

[9] Chun-Hong Cheng, Kwan-Long Wong, Jing-Wei Chin, Tsz-Tai Chan, and Richard HY So. 2021. Deep learning methods for remote heart rate measurement: a review and future research agenda. *Sensors*, 21, 18, 6296.

[10] Xinyi Ding, Damoun Nassehi, and Eric C Larson. 2018. Measuring oxygen saturation with smartphone cameras using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 23, 6, 2603–2610.

[11] Illia Fedorin, Kostyantyn Slyusarenko, and Margaryta Nastenko. 2020. Respiratory events screening using consumer smartwatches. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, 25–28.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[13] Min Hu, Fei Qian, Dong Guo, Xiaohua Wang, Lei He, and Fuji Ren. 2021. Eta-rppgnet: effective time-domain attention network for remote heart rate measurement. *IEEE Transactions on Instrumentation and Measurement*, 70, 1–12.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

[15] 2011. International Organization for Standardization. *Particular requirements for basic safety and essential performance of pulse oximeter equipment.*

[16] Lingqin Kong et al. 2013. Non-contact detection of oxygen saturation based on visible light imaging device using ambient light. *Optics express*, 21, 15, 17464–17471.

[17] Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. 2020. The 1st challenge on remote physiological signal sensing (repss). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 314–315.

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

[19] Craig Lockwood, Tiffany Conroy-Hiller, and Tamara Page. 2004. Vital signs. *JBI reports*, 2, 6, 207–230.

[20] Santiago Lopez and RTAC Americas. 2012. Pulse oximeter fundamentals and design. *Free scale semiconductor*, 23.

[21] Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7.

[22] Christine Manta, Sneha S Jain, Andrea Coravos, Dena Mendelsohn, and Elena S Izmailova. 2020. An evaluation of biometric monitoring technologies for vital signs in the era of covid-19. *Clinical and Translational Science*, 13, 6, 1034–1044.

[23] Joshua Mathew, Xin Tian, Min Wu, and Chau-Wai Wong. 2021. Remote blood oxygen estimation from videos using neural networks. *arXiv preprint arXiv:2107.05087*.

[24] Biswadev Mitra, Carl Luckhoff, Rob D Mitchell, Gerard M O'Reilly, De Villiers Smit, and Peter A Cameron. 2020. Temperature screening has negligible value for control of covid-19. *Emergency Medicine Australasia*, 32, 5, 867–869.

[25] Meir Nitzan, Ayal Romem, and Robert Koppel. 2014. Pulse oximetry: fundamentals and technology update. *Medical Devices (Auckland, NZ)*, 7, 231.

[26] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. 2018. Synrhythm: learning a deep heart rate estimator from general to specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 3580–3585.

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

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

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

[30] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. 2019. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.

[31] Olga Perepelkina, Mikhail Artemyev, Marina Churikova, and Mikhail Grinenko. 2020. 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*, 288–289.

[32] Marco AF Pimentel, Oliver C Redfern, Robert Hatch, J Duncan Young, Lionel Tarassenko, and Peter J Watkinson. 2020. Trajectories of vital signs in patients with covid-19. *Resuscitation*, 156, 99–106.

[33] Vignesh Ravichandran, Balamurali Murugesan, Vaishali Balakarthikeyan, Keerthi Ram, SP Preejith, Jayaraj Joseph, and Mohanasankar Sivaprakasam. 2019. Respnet: a deep learning model for extraction of respiration from photoplethysmogram. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 5556–5559.

[34] Philipp V Rouast, Marc TP Adam, Raymond Chiong, David Cornforth, and Ewa Lux. 2018. Remote heart rate measurement using low-cost rgb face video: a technical literature review. *Frontiers of Computer Science*, 12, 5, 858–872.

[35] Olga Russakovsky et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 3, 211–252.

[36] JM Schmitt. 1986. Optical measurement of blood oxygenation by implantable telemetry. *Technical Report G558–15, Stanford*.

[37] Christopher G Scully, Jinseok Lee, Joseph Meyer, Alexander M Gorbach, Domhnull Granquist-Fraser, Yitzhak Mendelson, and Ki H Chon. 2011. Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Transactions on Biomedical Engineering*, 59, 2, 303–306.

[38] Dangdang Shao, Chenbin Liu, Francis Tsow, Yuting Yang, Zijian Du, Rafael Iriya, Hui Yu, and Nongjian Tao. 2015. Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system. *IEEE Transactions on Biomedical Engineering*, 63, 6, 1091–1098.

[39] Nichole Starr et al. 2020. Pulse oximetry in low-resource settings during the covid-19 pandemic. *The Lancet Global Health*, 8, 9, e1121–e1122.

[40] Yu Sun and Nitish Thakor. 2015. Photoplethysmography revisited: from contact to noncontact, from point to imaging. *IEEE transactions on biomedical engineering*, 63, 3, 463–477.

[41] Setsuo Takatani and Marshall D Graham. 1979. Theoretical analysis of diffuse reflectance from a two-layer tissue model. *IEEE Transactions on Biomedical Engineering*, 12, 656–664.

[42] Mingxing Tan and Quoc Le. 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[43] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, João Jorge, DA Clifton, and Chris Pugh. 2014. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35, 5, 807.

[44] Mark Van Gastel, Sander Stuijk, and Gerard De Haan. 2016. New principle for measuring arterial blood oxygenation, enabling motion-robust remote monitoring. *Scientific reports*, 6, 1, 1–16.

[45] Mark van Gastel, Wim Verkruysse, and Gerard de Haan. 2019. Data-driven calibration estimation for robust remote pulse-oximetry. *Applied Sciences*, 9, 18, 3857.

[46] Gary M Vilke, Jesse J Brennan, Alexandrea O Cronin, and Edward M Castillo. 2020. Clinical features of patients with covid-19: is temperature screening useful? *The Journal of Emergency Medicine*, 59, 6, 952–956.

[47] Yuting Yang, Chenbin Liu, Hui Yu, Dangdang Shao, Francis Tsow, and Nongjian Tao. 2016. Motion robust remote photoplethysmography in cielab color space. *Journal of biomedical optics*, 21, 11, 117001.

[48] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. 2021. Transrppg: remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE Signal Processing Letters*, 28, 1290–1294.

[49] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. 2019. 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*, 151–160.