JBO Journal of Biomedical Optics

RESEARCH PAPER

Leveraging 3D convolutional neural network and 3D visible-near-infrared multimodal imaging for enhanced contactless oximetry

Wang Liao,[a,*] Chen Zhang,[a] Belmin Alić,[b] Alina Wildenauer,[c] Sarah Dietz-Terjung,[c] Jose Guillermo Ortiz Sucre,[d] Sivagurunathan Sutharsan,[d] Christoph Schöbel,[c] Karsten Seidl,[b] and Gunther Notni[a,e]

[a]Ilmenau University of Technology, Department of Mechanical Engineering, Ilmenau, Germany
[b]University of Duisburg-Essen, Chair of Electronic Components and Circuits, Duisburg, Germany
[c]University Medicine Essen, Ruhrlandklinik, Chair of Sleep and Telemedicine, Essen, Germany
[d]University Medicine Essen, Ruhrlandklinik, Department of Pneumology, Essen, Germany
[e]Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany

ABSTRACT. **Significance:** Monitoring oxygen saturation (SpO_2) is important in healthcare, especially for diagnosing and managing pulmonary diseases. Non-contact approaches broaden the potential applications of SpO_2 measurement by better hygiene, comfort, and capability for long-term monitoring. However, existing studies often encounter challenges such as lower signal-to-noise ratios and stringent environmental conditions.

Aim: We aim to develop and validate a contactless SpO_2 measurement approach using 3D convolutional neural networks (3D CNN) and 3D visible-near-infrared (VIS-NIR) multimodal imaging, to offer a convenient, accurate, and robust alternative for SpO_2 monitoring.

Approach: We propose an approach that utilizes a 3D VIS-NIR multimodal camera system to capture facial videos, in which SpO_2 is estimated through 3D CNN by simultaneously extracting spatial and temporal features. Our approach includes registration of multimodal images, tracking of the 3D region of interest, spatial and temporal preprocessing, and 3D CNN-based feature extraction and SpO_2 regression.

Results: In a breath-holding experiment involving 23 healthy participants, we obtained multimodal video data with reference SpO_2 values ranging from 80% to 99% measured by pulse oximeter on the fingertip. The approach achieved a mean absolute error (MAE) of 2.31% and a Pearson correlation coefficient of 0.64 in the experiment, demonstrating good agreement with traditional pulse oximetry. The discrepancy of estimated SpO_2 values was within 3% of the reference SpO_2 for ~80% of all 1-s time points. Besides, in clinical trials involving patients with sleep apnea syndrome, our approach demonstrated robust performance, with an MAE of less than 2% in SpO_2 estimations compared to gold-standard polysomnography.

Conclusions: The proposed approach offers a promising alternative for non-contact oxygen saturation measurement with good sensitivity to desaturation, showing potential for applications in clinical settings.

Keywords: oxygen saturation; contactless oximetry; multimodal imaging; deep learning

Paper 240087SSR received Mar. 29, 2024; revised Jul. 30, 2024; accepted Aug. 1, 2024; published Aug. 21, 2024.

*Address all correspondence to Wang Liao, wang.liao@tu-ilmenau.de

1 Introduction

Vital signs, such as body temperature, heart rate, respiratory rate, and blood pressure are standard indicators of an individual's physiological functions in most medical settings.[1] Monitoring these vital parameters is crucial for early diagnosis, medical treatment, risk assessment, and patient recovery monitoring.[2,3] With the advancement of medical measurement technology, oxygen saturation has increasingly become recognized as an indispensable fifth vital sign.[4] Oxygen saturation indicates the percentage of oxygenated hemoglobin (HbO_2) and hemoglobin (Hb) in the blood, in which the artery should be in the range of 95% to 100% in healthy individuals.[5] Many pulmonary diseases cause abnormalities in oxygen saturation values, such as acute pneumonia, chronic obstructive pulmonary disease (COPD), and sleep apnea syndrome (SAS). Furthermore, the outbreak of coronavirus (COVID-19) has further underscored the critical importance of oxygen saturation measurement.

The gold standard for measuring arterial oxygen saturation (SaO_2) is the invasive arterial blood gas (ABG) test,[6] which is performed by medical professionals. Mixed venous oxygen saturation (SvO_2) is normally measured via a pulmonary artery catheter. Non-invasive methods based on near-infrared spectroscopy are developed to measure tissue oxygen saturation (StO_2), which directly provides an assessment of the oxygenation status of tissues. Time-domain near-infrared spectroscopy (TD-NIRS) is an established technique, which allows the estimation of StO_2 at multiple depths, including beyond 2 cm deep.[7,8] This capability opens a range of applications, such as determining StO_2 in the brain.[9] The estimation of SaO_2 at peripheral capillary is called SpO_2. A non-invasive pulse oximeter is known for its convenience for real-time SpO_2 estimation. Polysomnography (PSG) systems [10] used in sleep monitoring also incorporate pulse oximeter to record SpO_2 overnight. A typical pulse oximeter employs a light source that projects red and infrared light onto fingertips or earlobes. Oxygenated hemoglobin and deoxygenated hemoglobin exhibit distinct characteristics of absorption spectra. By contrasting the transmitted light intensities at 660 and 940 nm wavelengths captured by the photoelectric sensor, the pulse oximeter determines the SpO_2 by utilizing the ratio-of-ratios (RR) method.[11] However, contact-based methods face challenges for patients with infectious diseases or allergies,[12] especially during long-term measurements such as sleep monitoring. To overcome these limitations of contact-based methods, there is an increasing focus on camera-based SpO_2 measurement. Bui et al.[13] and Ding et al.[14] utilized a camera-based approach, where participants placed a finger over the smartphone's camera and flash, diverging from true contactless methods. Many studies on contactless SpO_2 measurements usually use red, green, and blue (RGB) cameras to capture hands[15,16] or faces[17–19] with ambient light and extract weak pulsatile temporal features from remote photoplethysmogram (rPPG) signals through different analytical filtering techniques[20–22] or neural networks[23–25] to calculate SpO_2. Acquiring high-quality rPPG signals is a challenging task, which can be affected by factors like illumination conditions, sampling rate, and sensor noise, along with disruptions from facial movements such as smiles or blinks, which compromise SpO_2-related information. The spatial encoded patterns of the captured skin regions have been proven by Wieringa et al.[26] and Rosa and Betini[27] to contain oxygen saturation information. Hu et al.[28] employed a 2D residual cascade and coordinate attention mechanism to analyze feature channel correlations of spatial data, using neural networks to extract and concatenate spatial features for estimation. Few studies simultaneously consider both spatial and temporal features. To fill in the gap, in our previous work,[29] 3D convolutional network (3D CNN) are used to extract spatial-temporal information from the near-infrared multispectral videos for SpO_2 estimation. Besides, in our literature review scope, we observed that current research gaps of camera-based contactless SpO_2 measurement include region of interest (ROI) tracking, acquiring datasets with significant SpO_2 fluctuations, and validation in clinical settings. We noted that most studies are based on datasets containing only a few instances of low SpO_2 levels and the overwhelming majority of SpO_2 ranges between 95% and 100%. To address these challenges, in this work, we propose a 3D convolutional neural networks-based approach to estimate SpO_2 from videos captured by our 3D visible-near-infrared (VIS-NIR) multimodal camera system. The performance is verified through both short-term daytime measurements on healthy participants and

continuous long-term nighttime monitoring of patients with sleep apnea. The contributions of this work include the following:

1. We utilized a 3D VIS-NIR multimodal camera system to capture multimodal facial videos and proposed steps including multimodal image registration, 3D ROI tracking, spatial and temporal preprocessing, and 3D CNN-based spatial-temporal features extraction to enable oxygen saturation estimation in both during day and night.
2. We conducted a breath-holding study on 23 healthy participants with different skin types, achieving an MAE of 2.31 and a Pearson correlation coefficient of 0.64 compared to the reference oxygen saturation ranging from 80% to 99% measured by pulsed oximeter on the fingertip. In addition, our approach was also validated by a trial study involving long-term overnight monitoring of four real sleep disorder patients, demonstrating good agreement with the gold standard PSG.
3. We discussed various feature extraction strategies, different image channel combinations, and diverse neural network architectures (including light-weight networks) for their capability and performance to estimate SpO_2 from 3D VIS-NIR multimodal videos.

2 Proposed Approach Based on Multimodal Imaging

Multimodal imaging refers to the integration of various imaging modalities such as 3D imaging, multispectral imaging, and thermal imaging. It allows for enhanced and more dependable analysis to realize intricate tasks[30–33] based on diverse feature combinations from different imaging modalities. In our work, we use four imaging modalities, which include images from color (RGB) cameras, NIR 780 and NIR 940 nm cameras, and disparity maps produced by active stereo matching based on two NIR 850 nm cameras and GOBO projector.[34] The details of our camera system setup will be introduced in Sec. 3. In this section, the proposed approach will be introduced, detailing how to regress SpO_2 using 3D CNN from multimodal video sequences after multimodal image registration, 3D ROI tracking, and spatial and temporal preprocessing.

2.1 Multimodal Image Registration

For the purpose of pixel-wise fusion of information from different 2D modalities, the 2D images were registered together using 3D information. Camera calibration is always the initial step. The intrinsic parameters of the two NIR cameras for stereo matching and also other 2D cameras are calibrated using Zhang's algorithm.[35] Simultaneously, the extrinsic camera parameters are calculated with respect to a reference 2D camera, for example, the RGB camera, using the method introduced in Ref. 36. Based on the NIR 850 nm camera parameters, a disparity map can be converted to a 3D point cloud. Assume (u_i, v_i) is the projection of one 3D point (x_i, y_i, z_i) of the point cloud on the image plane of one of the 2D cameras (RGB, NIR 780 nm or NIR 940 nm), the transformation can be calculated as follows:

$$s \cdot \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = K_c \cdot \left(R_c \cdot R_{\text{rect}}^{-1} \cdot \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} + T_c \right), \tag{1}$$

where s is a factor for the normalization of homogeneous 2D points, K_c is the intrinsic parameters matrix of this camera, R_c and T_c are the rotation matrix and translation vector of this camera, and R_{rect} is the rotation matrix of the reference camera for stereo rectification. When the projected image point does not align precisely with a pixel, bilinear interpolation among adjacent pixels is performed. Through this method, each 2D image captured by the cameras can be accurately mapped to the corresponding 3D point cloud. In this way, once an ROI is selected on a 2D image modality, it can be converted to the corresponding 3D ROI. The 3D ROI can be projected onto the images from other 2D cameras to assign gray values to these 3D points. In our work, the forehead region was used as the ROI for SpO_2 estimation because of good blood flow, thin epidermis, and no hair.[37,38] As shown in Fig. 1, we select a forehead region with width h and height w as ROI $(h, w, 3)$ on the color face image, and it can be converted to a 3D ROI. This 3D ROI is then projected to NIR 780 and NIR 940 nm images to obtain registered NIR 780 nm ROI $(h, w, 1)$ and NIR 940 nm ROI $(h, w, 1)$, from which corresponding gray values can be obtained.

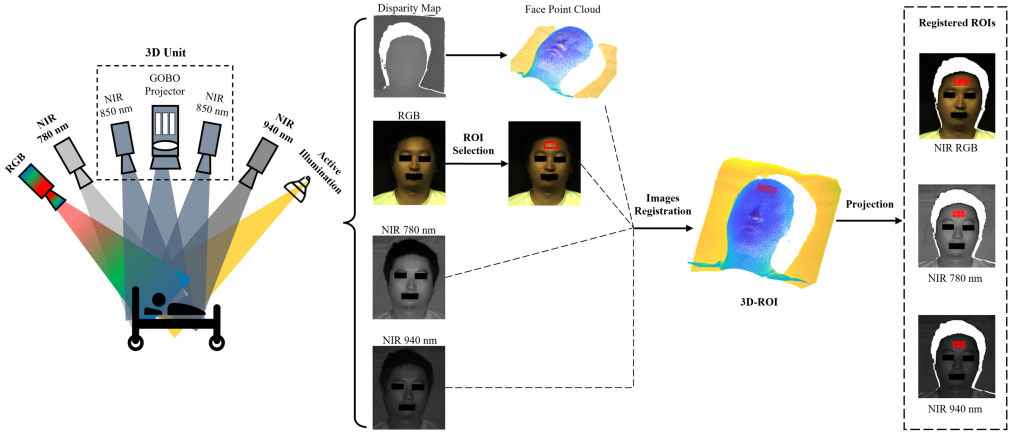

Fig. 1 Schematic of 3D information-based multimodal facial image registration.

2.2 Face Analysis and 3D ROI Tracking

For a continuous, registered multimodal facial video, we firstly utilized the MediaPipe Face Mesh framework,[39] a pretrained, light-weight deep learning model, for high-precision facial feature extraction, leveraging its capability to identify and track 468 distinct landmarks across various facial regions on RGB video. Each landmark, along with its image coordinates, is uniquely indexed, enabling us to perform automatic video anonymization. This is achieved by pinpointing the landmarks of the eyes and mouth regions in each frame and overlaying black rectangles over these areas across all registered imaging modalities. Subsequently, the image coordinates of landmarks on the forehead region in the first frame of the RGB video are used to define a forehead 2D ROI, which is then converted to 3D ROI. From the second frame onwards, the 3D ROI was tracked based on the 2D coordinates of the facial landmarks and the 3D point cloud as shown in Fig. 2. At each frame, the facial landmarks will be converted to the registered point cloud as 3D facial landmarks. Let the set of 3D facial landmarks on the first video frame be denoted as $P_1 = \{p_{1i} \in \mathbb{R}^3 | i = 1,2,\ldots,n\}$, where each p_{1i} is a 3D point represented as a column vector in homogeneous coordinates $p_{1i} = [x_{1i}, y_{1i}, z_{1i}, 1]^T$. Similarly, for the kth frame, the set of corresponding 3D facial landmarks is $P_k = \{p_{ki} \in \mathbb{R}^3 | i = 1,2,\ldots,n\}$, with each landmark p_{ki} also represented in homogeneous coordinates $p_{ki} = [x_{ki}, y_{ki}, z_{ki}, 1]^T$. Assume that the head is a rigid body, which means that the participant's facial expression was unchanged over the video period. To model the current 3D head pose relating to the 3D face pose on the first frame, the rigid body transformation with six degrees of freedom (DoF) from P_1 to P_k described by a rotation R_k and a translation t_k can be estimated as follows:

$$(R_k, t_k) = \arg\min_{R,t} \sum_{i=1}^{n} \|R p_{1i} + t - p_{ki}\|^2. \tag{2}$$

Thus, by employing the rotation R_k and the translation t_k, all points within the 3D ROI defined on the first frame can be transformed to their corresponding positions on the kth frame. Head movements typically occur in three dimensions, not confined to a single plane. Tracking a fixed skin area is evidently more suitable using 3D information, whether there is significant movement or subtle involuntary motion. As shown in Fig. 3, we demonstrate the tracking effectiveness when projecting the tracked 3D ROI back into an RGB 2D ROI. One of the participants is instructed to remain as still as possible for 4 min. However, slight involuntary head movements are inevitable. Whether assessing reference regions visually or evaluating by structural similarity (SSIM), the proposed 3D-based tracking method can more exactly track the ROI throughout the video.

2.3 Spatial and Temporal Preprocessing

As shown in Fig. 4, the tracked 3D ROI of the head in a video can be projected onto each modality to obtain 2D ROI videos. When these modalities are concatenated, a registered multimodal forehead ROI video is formed, encompassing five channels including R, G, B, 780, and 980 nm.

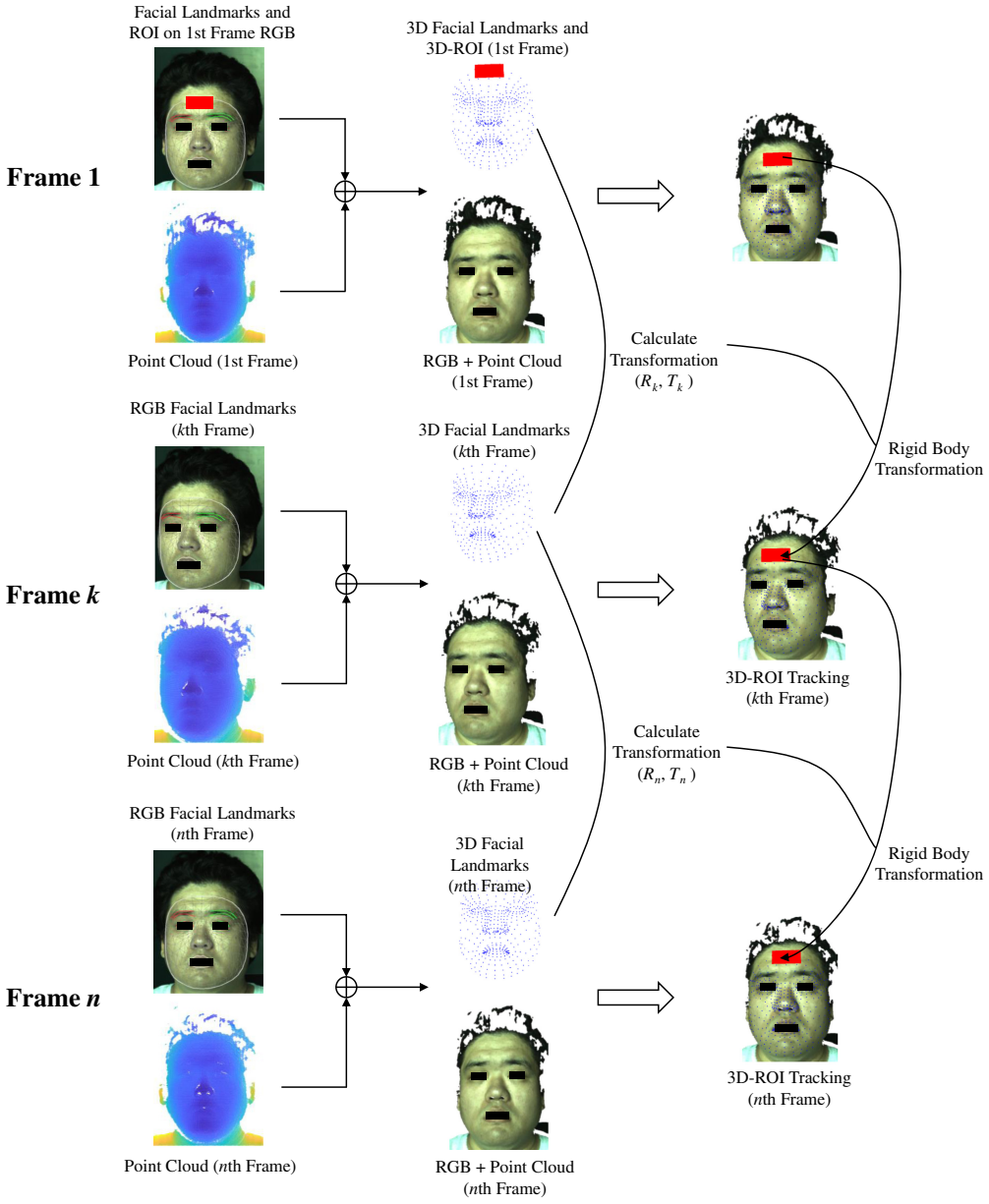

Fig. 2 Illustrative example of the 3D ROI tracking across sequential frames.

Then, spatial and temporal preprocessing is applied. Assuming there is a multimodal forehead ROI video V, and for a given channel, each frame has a height h and width w. The videos are spatially partitioned into $m \times n$ block videos, with each block video spatially sized $\lfloor \frac{h}{m} \rfloor \times \lfloor \frac{w}{n} \rfloor$, discarding residual pixels at the edges. For the ith block video in a specific channel, each of its pixel values can be represented as $B_i(x, y, t)$, where x and y denote spatial coordinates and t denotes time. A cubic polynomial $P_i = a_i t^3 + b_i t^2 + c_i t + d_i$ can be fitted as the temporal trend of B_i:

$$(a_i, b_i, c_i, d_i) = \underset{a_i, b_i, c_i, d_i}{\arg\min} \sum_t \left(\frac{1}{|B_i|} \sum_{x,y \in B_i} B_i(x, y, t) - P_j \right)^2. \tag{3}$$

Thus, for a certain pixel value of this B_i, it can be decomposed into the trend part $T_i(x, y, t) = P_i(t)$ and the detrended part $B_i'(x, y, t) = B_i(x, y, t) - P_i(t)$. This blockwise temporal detrending is replicated across all blocks and five channels, decomposing the multimodal forehead ROI video V into two components: one devoid of temporal trend, presumably carrying

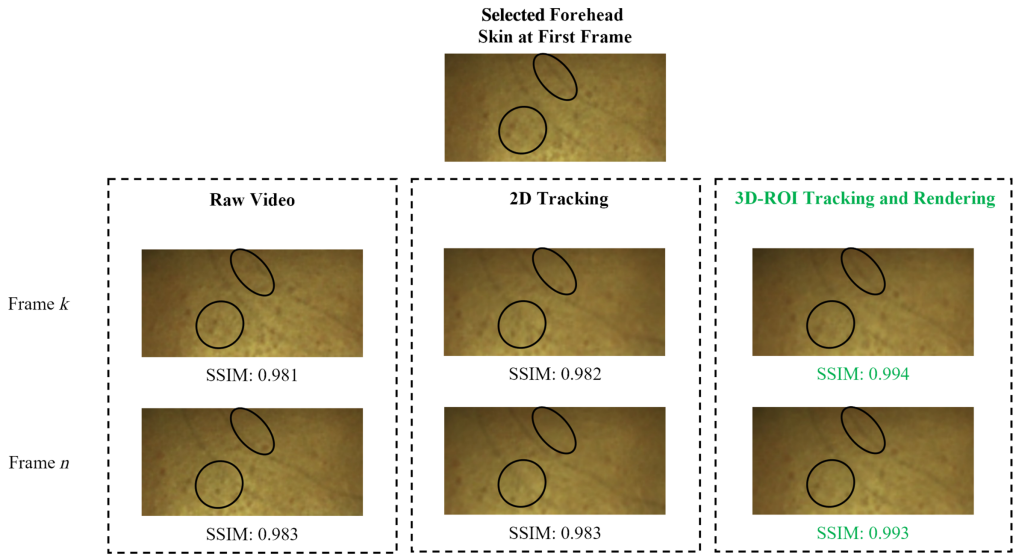

Fig. 3 Comparative analysis of ROI tracking for a forehead region initialized in the first frame of a video sequence. The black elliptical outlines in the ROI highlight reference features such as hair and skin hyperpigmentation, serving as markers to intuitively observe the tracking performance. Structural similarity (SSIM) is calculated to quantitatively assess the tracking performance.

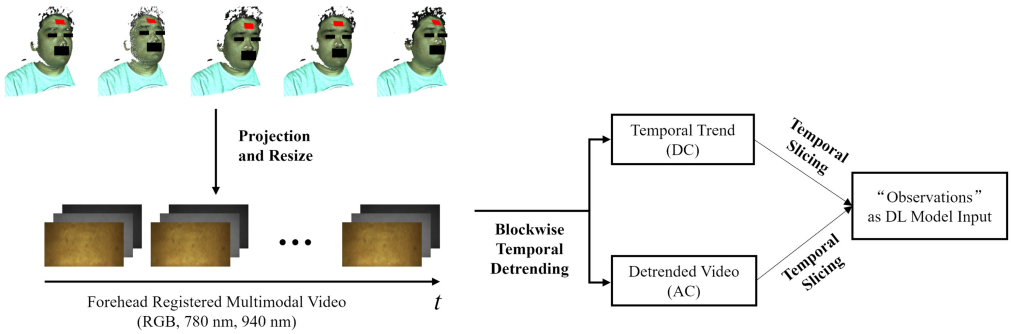

Fig. 4 Process flow from 3D ROIs of a video sequence to input of the deep learning model.

more information similar to the AC component in traditional methods, and the trend component, encapsulating more DC component information. Then, these two parts of the video are temporally sliced into 15 frames detrended video sequences and trend video sequences of 1-s time length, respectively. Concatenating a detrended video sequence and a trend video sequence forms an "observation," which serves as the input to the deep learning model.

2.4 Oxygen Saturation Regression with 3D CNN

These observations serve as input of spatial-temporal convolutional layers for feature extraction. Spatial-temporal convolution, also known as 3D convolution, enhances the feature extraction ability on volumetric data, thereby integrating information across various spatial dimensions and the temporal axis.[40] The 3D convolutional kernel slides across the input "observation," computing a dot product between its learnable weights and the corresponding local regions of the input at each position.

As shown in Fig. 5, we use a ResNet 18[41]-like structure with 3D convolution as a feature extractor. The input "observation" is firstly fed into a 3D convolutional layer with a kernel size of [7,7,7] and then forwarded to four residual blocks with a convolutional kernel size of [3,3,3]. To accentuate global feature representation while diminishing the focus on local textural details, a 3D global average pooling layer is situated before the residual blocks. The extracted features are flattened to the feature vector as the input of the regressor, which is composed of two fully connected layers (FC). The output of the regressor is normalized to be between 0 and 1, which is

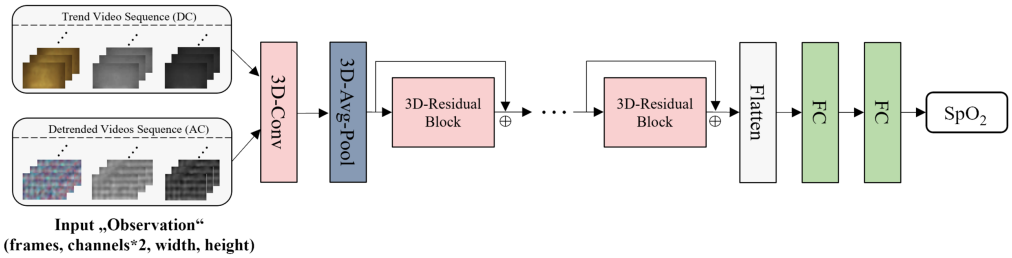

Fig. 5 Neural network structure for oxygen saturation estimation.

estimated SpO_2 after scaling. Every "observation" is associated with one SpO_2 output from the regressor and a reference value. For training the neural network, mean square error (MSE) is used as the loss function and Adam[42] is chosen as the optimizer. We use both dropout and early stopping to prevent overfitting. Hyperparameters are set empirically. Neither commercial pulse oximeters nor clinical devices used for oximetry analysis provide decimal values, so we obtained oxygen saturation reference values as integers. Although neural networks are capable of producing outputs with decimals, we have rounded the outputs as the only post-processing step.

3 Experiment Setting and Data Acquisition

3.1 Multimodal Imaging Camera System

We utilized a multimodal imaging system manufactured by the Fraunhofer Institute for Applied Optics and Precision Engineering in our previous work[43] and established an experimental setup at University Medicine Essen as shown in Fig. 6.

The sensor head of this camera system contains a real-time 3D sensor unit composed of two NIR 850 nm high-speed cameras with a full width of half maximum (FWHM) of 50 nm and a high-speed GOBO projector[34] at the same light wavelength. Besides the 3D sensor, a color camera, two NIR cameras at 780 and 940 nm, and a thermal camera are integrated into the housing. In this study, the thermal camera is inactive, which is integrated for the estimation of other vital signs. The frame rates of these 2D cameras are 15 Hz, and they are hardware-triggered and synchronized with the 3D video stream. The spatial resolution of these active 2D cameras is 896×704. The system utilizes a light-emitting diode (LED) array for homogeneous illumination, comprising one LED operating at 780 nm and three LEDs at 940 nm. Each LED in the array has a beam angle within half-maximum intensity ranging from 90 deg to 120 deg, with an output power of 1 W. The camera system encompasses a lateral measurement field of ~500 mm by 400 mm when positioned at an intermediate distance of 1.5 m, and the cumulative irradiation from this LED array configuration is $\sim 1.255\ \mu W/mm^2$, thereby adhering to the safety standards for ocular exposure.[44]

Fig. 6 Multimodal camera system with sensor head composed of a GOBO projector (1), two NIR cameras at 850 nm (2, 3), NIR camera at 780 nm (4), NIR camera at 940 nm (5), thermal camera (6), LED array with LEDs at 780 and 940 nm (7), and color camera (8).

3.2 Video Data Acquisition and Reference Value Recording

To validate our approach, a total of 23 cardiopulmonary healthy participants (numbered Par#1 to Par#23) were recruited for a breath-holding study. The study is approved by the Ethics Committee of the Faculty of Medicine, University of Duisburg-Essen (approval no. 21-10312-BO). Informed consent was obtained from all individual participants included in this experiment. Their Fitzpatrick skin types[45] range from type II to type V. To obtain video data with low SpO_2 values, participants were expected to exhale as much as possible and then hold their breath for a while during a video shoot. For comfort and health reasons, the duration of breath-holding was determined by the participants themselves. When they felt they could not tolerate breath-holding anymore, they would breathe normally for a period of time. Participants repeated the cycle of exhalation, breath-holding, inhalation, and normal breathing three times in ~4 min. While we advised participants to face the camera system with the front view, we could not constrain their head movements. Especially, the breath-holding can lead to momentary discomfort, resulting in some unavoidable involuntary movements. Participants were engaged in two separate measurements, interspersed with a 5-min break to regulate their breathing. During the video capture, a Pulox PO-200 pulse oximeter was clipped to the fingertip to measure the participant's reference SpO_2 values. A webcam was used to capture the pulse oximeter display with a frame rate of 1 Hz. Pre-trained optical character recognition model by EasyOCR was used to read the SpO_2 reference from captured displays. By holding breath, the SpO_2 value can drop below 95%, which is considered the lower limit of the normal range,[46] and some participants can even drop to 80%. After the participant resumes normal breathing, the SpO_2 will quickly return to the healthy range. Captured video and reference recording were first synchronized by time-stamps. It is worth mentioning that since the face and fingertips are different parts of the body, the synchronization in recording time does not mean that the SpO_2 obtained from facial videos and those obtained from fingertip pulse oximeters are synchronized. According to Refs. 47 and 48, the SpO_2 obtained from facial videos are ~20 s. faster than those obtained from fingertip pulse oximeters. Therefore, in our subsequent experiments, for training, we applied a constant 20 s time advance to the reference values from the pulse oximeter. For evaluating the inference, we shifted the reference time trace within a range of 20 s $\pm$ 5 s to maximize its correlation with the estimated time trace.

Through our experiment, a total of 168.5 min of multimodal videos were captured, equating to 10,112 1-s "observations" after preprocessing. Each 1-s "observation" corresponds to a specific SpO_2 value. The reference SpO_2 values ranged from 80% to 99%. In our literature review, we observed that open-source datasets for camera-based SpO_2 estimation are scarce, and no studies utilizing 3D VIS-NIR multimodal imaging have been found. Most of the datasets do not focus on SpO_2 but rather on heart rate and respiration. Within our research scope, we found the PURE,[49] VIPL-HR,[50] and UBFC-rPPG[51] datasets. In the PURE dataset, the researchers used an RGB camera to record 10 healthy subjects. The VIPL-HR dataset includes 107 healthy participants, mostly recorded with RGB modality, and a few with both RGB and NIR multispectral videos. The UBFC-rPPG dataset has only a few participants with SpO_2 reference values and includes only RGB videos. It can be seen from Table 1, that within the limited camera-based benchmark datasets available, reference SpO_2 values rarely drop below the healthy range, with almost no instances falling below 90%. Despite the challenges associated with obtaining data on

Table 1 Comparison of reference SpO_2 coverage and distribution of benchmark datasets with ours.

	SpO_2 (%) Coverage				Centile of SpO_2 (%)		
Dataset	[80, 85]	[86, 90]	[91, 95]	[96, 100]	5th percentile	Q1	Median
PURE[49]	-	0.55	14.19	85.26	95	96	98
VIPL-HR[50]	-	0.20	13.02	86.78	94	96	97
UBFC-rPPG[51]	-	-	2.27	97.73	96	96	97
Ours	2.82	11.66	31.78	53.74	87	92	96

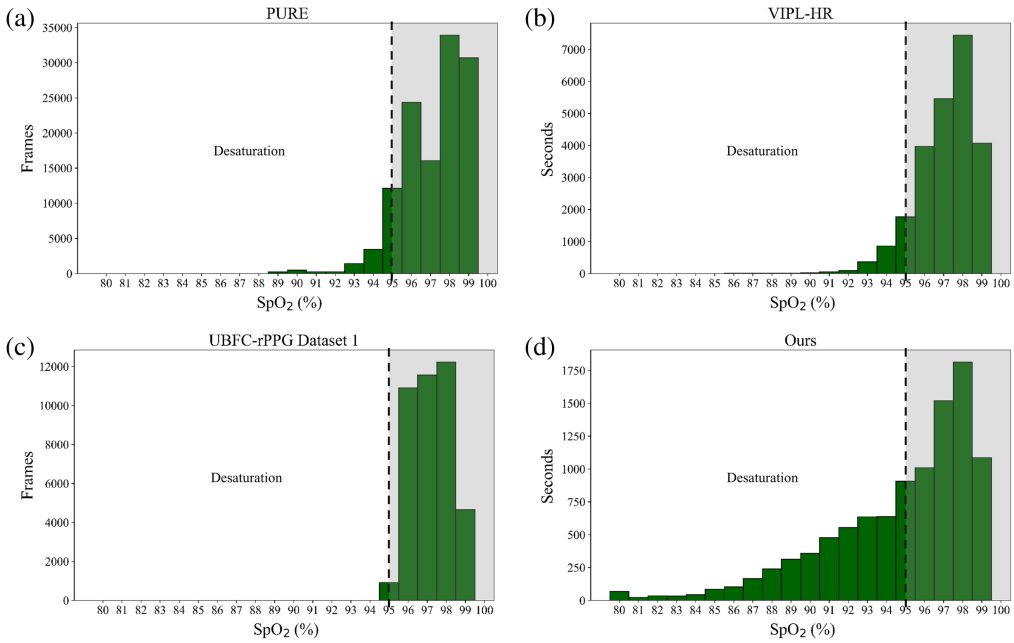

Fig. 7 Histograms of reference SpO_2 distributions in benchmark and our datasets. In PURE and UBFC-rPPG dataset 1, each sample corresponds to one frame, while in VIPL-HR and our dataset, each sample represents 1 s. The dashed line to the left indicates desaturation, that is, SpO_2 values below 95%.

low SpO_2 levels, our dataset successfully includes 40% of "observations" with desaturation. Specifically, it encompasses 2.82% of "observations" with SpO_2 from 80 to 85 and 11.66% within the SpO_2 range 86% to 90%. In addition, the 25th percentile (Q1) of SpO_2 reference in our dataset is located at 92%. A comparison of four histograms has been shown in Fig. 7, representing the distribution of SpO_2 values in different datasets. Unlike the other datasets, which show a steep decline in instances frequency below 95%, ours maintains a more gradual decrease, including many lower values. This suggests that our dataset captures a broader instances spectrum of SpO_2 values, potentially offering richer insights for desaturation scenarios.

4 Results and Discussion

As introduced in the previous section, this work involved 23 participants, each of whom was recorded in two separate around 4-min measurement sessions. We embarked on our validation by addressing a "participant-dependent" scenario, also referred to as "precision healthcare" validation, which means using one measurement from each participant as training data, with the subsequent measurement serving as the test data. This scenario emphasizes personalized analysis. Our focal point of result discussion shifts towards a more practical and generalizable scenario known as the "participant-independent" scenario or "leave-one-participant-out" validation. We systematically designate the two measurements of each participant as the test data while utilizing all available measurements from the remaining 22 participants as the training dataset. This strategy is aimed at validating the robustness and generalizability of our approach across different subjects. We will also explore the performance of various feature extraction strategies and the corresponding network architectures. In addition, we will also present test results and application scenarios with different input modalities. Finally, a clinical trial involving sleep apnea patients will be introduced to demonstrate the transferability and potential applications of our approach.

4.1 Performance Metrics

To evaluate the performance of the proposed approach, we employed two standard metrics commonly utilized in regression analyses: mean absolute error (MAE) and Pearson's correlation coefficient (ρ). If $y_i \in Y$ denotes the from proposed approach estimated SpO_2 and $\hat{y}_i \in \hat{Y}$ denotes their corresponding reference values, the MAE and ρ can be defined as

$$\mathrm{MAE}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{4}$$

$$\rho(Y,\hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \overline{Y})(\hat{y}i - \overline{\hat{Y}})}{\sqrt{\sum i = 1^{n}(y_i - \overline{Y})^2 \sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{Y}})^2}}. \tag{5}$$

In addition, we introduced bias (B), also known as "mean," to represent the average discrepancy between all estimated SpO_2 values and their corresponding reference values. Meanwhile, the 95% limits of agreement (95% LoA) are defined as the range covering 1.96 times the standard deviation of these discrepancies, offering an insight into the consistency of our estimations.

4.2 Results with Proposed Approach

Table 2 summarizes the performance of SpO_2 estimation with the proposed approach in the aforementioned two validation scenarios. Across both scenarios, the average correlation coefficients of all test measurements (Avg. ρ) remain robust, suggesting a strong correlation between estimated and actual SpO_2 values. The bias (B) is generally low, indicating minimal systematic underestimation or overestimation. In the "Precision Healthcare" scenario, the overall MAE stands at 2.12%, with a slightly higher MAE of 2.41% observed during desaturation events. The "leave-one-participant-out" scenario exhibits an overall MAE of 2.31%, with desaturation events resulting in a higher MAE of 3.26%. No significant deterioration in results is noted across any specific skin type. However, skin type V displayed a notably better MAE compared to others, potentially due to the data obtained with narrow reference SpO_2 values distribution from participants with this skin type.

Considering the generalizability of the proposed approach and the prospect of practical applications, all the results presented and discussed next will be based on the more complex "leave-one-participant-out" scenario.

As shown in Fig. 8, we have introduced the percentage of time the discrepancy between the estimation and reference values falls within a certain range (PERC) and the Bland–Altman plot[52] to analyze the agreement between our proposed approach and pulse oximeter recordings in the

Table 2 Results summary of the performance of the proposed approach.

	MAE (%)					
Skin type	All	Normal	Desaturation	Avg. ρ	B (%)	95% LoA (%)
Precision healthcare scenario						
II	2.06	1.63	2.45	0.72	−0.63	[−5.44, 4.17]
III	2.04	1.81	2.20	0.70	−0.13	[−5.59, 5.33]
IV	2.48	1.67	3.02	0.80	−1.16	[−7.77, 5.44]
V	1.38	0.84	1.83	0.55	−0.50	[−4.24, 3.23]
All	**2.12**	**1.71**	**2.41**	**0.72**	**−0.46**	**[−6.19, 5.27]**
Leave-one-participant-out scenario						
II	2.04	2.02	2.08	0.54	0.71	[−4.14, 5.55]
III	2.42	1.76	3.40	0.62	−0.51	[−6.76, 5.75]
IV	2.23	1.56	3.46	0.72	−0.05	[−6.15, 6.05]
V	1.80	1.98	1.28	0.68	1.31	[−2.33, 4.95]
All	**2.31**	**1.74**	**3.26**	**0.64**	**−0.20**	**[−6.29, 5.88]**

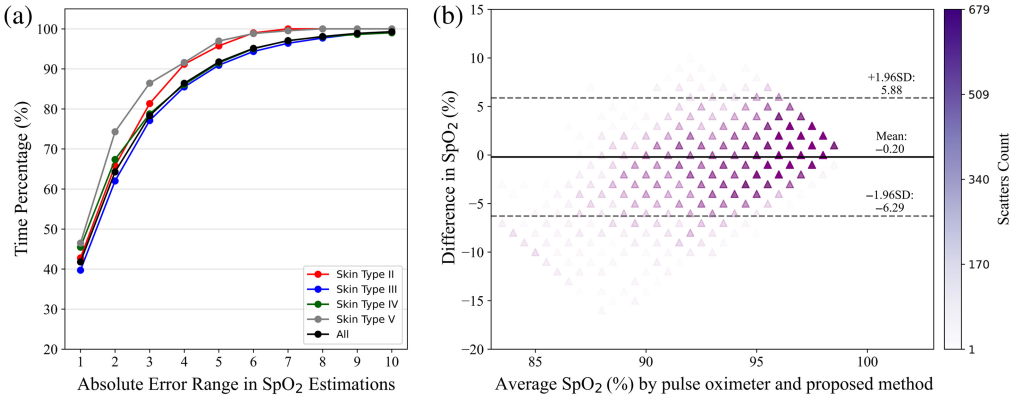

Fig. 8 Performance visualization of the proposed approach in the "leave-one-participant-out" scenario. (a) The percentage of time (PERC) within the range of absolute error 1% to 10% between reference SpO_2 and estimated SpO_2. (b) The Bland–Altman plot shows the agreement between the proposed approach and the commercial pulse oximeter. The *y*-axis represents the differences between the estimated and reference SpO_2, while the *x*-axis represents the average of the two values. Three lines represent respectively the mean difference (bias) and upper and lower 95% limit of agreement. The transparency of the triangle markers reflects the number of overlapping scatters.

"leave-one-participant-out" scenario. It is observed in Fig. 8(a) that the discrepancy of estimated SpO_2 values is within 3% of the reference values for ~80% of all time points. Both Table 2 and Fig. 8(a) demonstrate that our approach does not perform significantly worse in estimating SpO_2 for any specific skin type. Furthermore, as shown in Fig. 8(b), the vast majority of the data points in the Bland–Altman plot lie within the 95% LoA, suggesting a strong agreement between the two SpO_2 measurement approaches. But the 95% LoA range from −6.29 to 5.88, which is higher compared to those reported in some classic works.[27,53] This can be due to the wide distribution of SpO_2 values in our dataset, which ranges from 80% to 99%, and includes a significant number of low-oxygen saturation values, with nearly 3% of values falling below 85%. Besides, the higher 95% LoA reflects the suboptimal performance of our approach in extreme cases, which may be due to the imbalance in the training data. Capturing more data with a low SpO_2 level for supervised learning could be expected to improve this situation. The estimated SpO_2 signals and reference signals in the "leave-one-participant-out" scenario for all the participants are presented in Fig. 9. We spliced two videos for one participant so that the SpO_2 curves of each participant should contain several dips that result from breath-holding. The MAE of estimated and reference signals across the participants ranges from 1.57% to 3.53%, and the Pearson correlation coefficient varies from 0.49 to 0.73. For the majority of the time, even during the desaturation events, the estimated SpO_2 values track closely with the reference SpO_2 values, although some variations exist. As shown for Par#14, Par#19, and Par#22, when the reference SpO_2 values are exceptionally low, typically below 85%, the estimated values indicate a downward trend but do not reach those low levels. This could be attributed to the scarcity of extremely low data points involved during the training.

4.3 Discussion of Feature Extraction Strategies and Network Structures

In our proposed approach, we treat a 1-s "observation," namely a 15-frame video sequence of preprocessed multimodal videos with both DC and AC components, as the input to a 3D CNN feature extractor for simultaneous extraction of temporal and spatial features. In this subsection, we compare the performance of various feature extraction strategies and corresponding network architectures, which are schematically depicted in Fig. 10, with the proposed approach.

1. Strategy A: we process the DC and AC components of multimodal videos by averaging them spatially to obtain multimodal DC and AC signals. These signals are then sliced into 1-s sequences, each containing 15 time points, to serve as inputs for shallow 1D-CNN feature extractors. This strategy only focuses on temporal feature extraction but not spatial features.

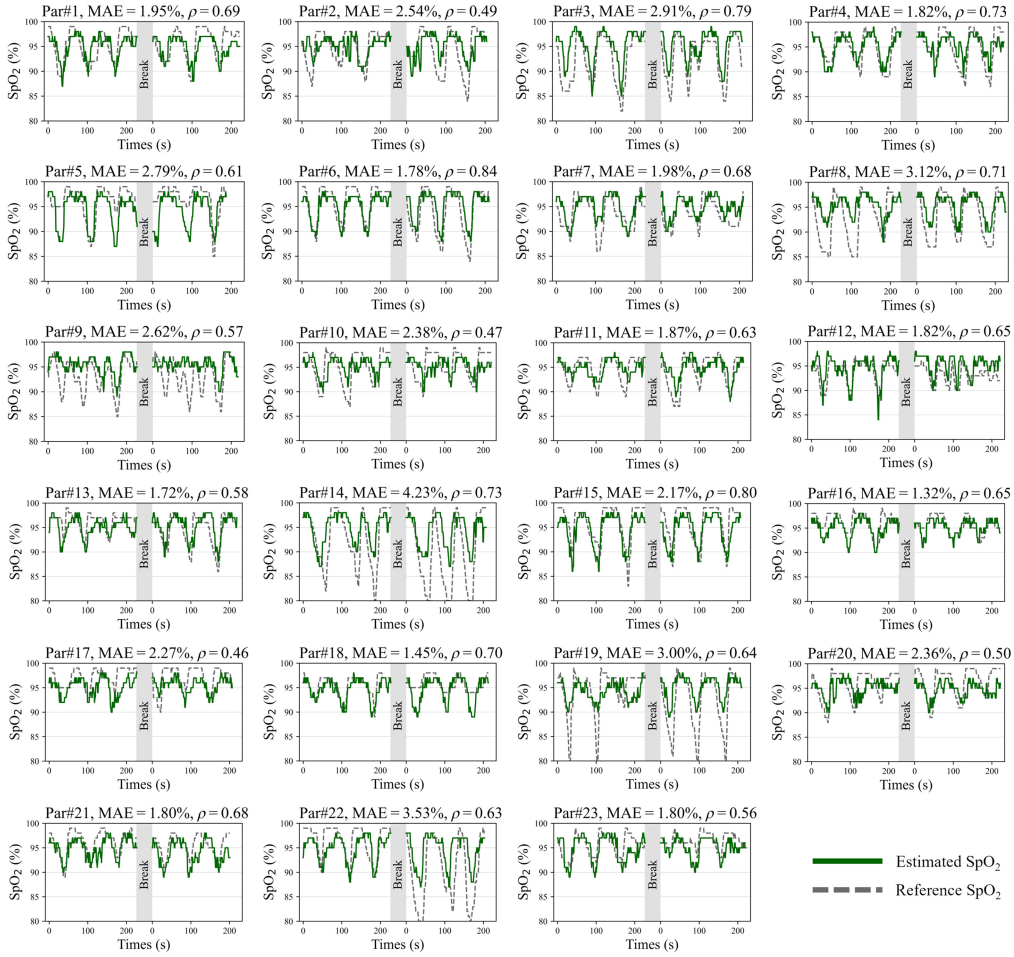

Fig. 9 Estimated SpO_2 values and pulse oximeter measured reference SpO_2 values of 23 participants in the "leave-one-participant-out" scenario. For each participant, the model is trained by all measurements from the other 22 participants and tested on the left two measurements of this participant. Between two test measurements, there is a break. The green lines represent the estimated values, while the reference signals are dashed gray lines.

2. Strategy B: similar to strategy A, this strategy begins by spatially averaging the DC and AC components of multimodal videos to obtain multimodal spatially averaged signal sequences. However, unlike strategy A, each time point within a sequence is flattened and fed into a long short-term memory (LSTM) network[54] as a one-time step. The LSTM model outputs one SpO_2 estimation value after processing all time points within the sequence. This strategy aims to capture the dependencies between different time points within the signal sequence. But there is also spatial feature neglect.
3. Strategy C: we blockwise spatial average the DC and AC components of multimodal videos and concatenate them to obtain multiple multimodal signals corresponding to the number of blocks. These multiple signals are then sliced into sequences as inputs for shallow 2D CNN feature extractor. In this way, both temporal features and some spatial features between the blocks are concurrently considered.
4. Strategy D: similar to strategy C, we can obtain multiple multimodal signals. Then, at each time point, multiple signal values from different channels are first processed through a 1D CNN to extract spatial features, which are then flattened and fed into an LSTM as a one-time step. The LSTM extract then temporal dependencies related features within the sequence.
5. Strategy E: each frame of an "observation" is processed through a 2D CNN feature extractor to get spatial features. Subsequently, the features of each frame are flattened and serve as a one-time step for the LSTM. This approach initially extracts features in the spatial domain and then analyzes temporal dependencies within these spatial features.

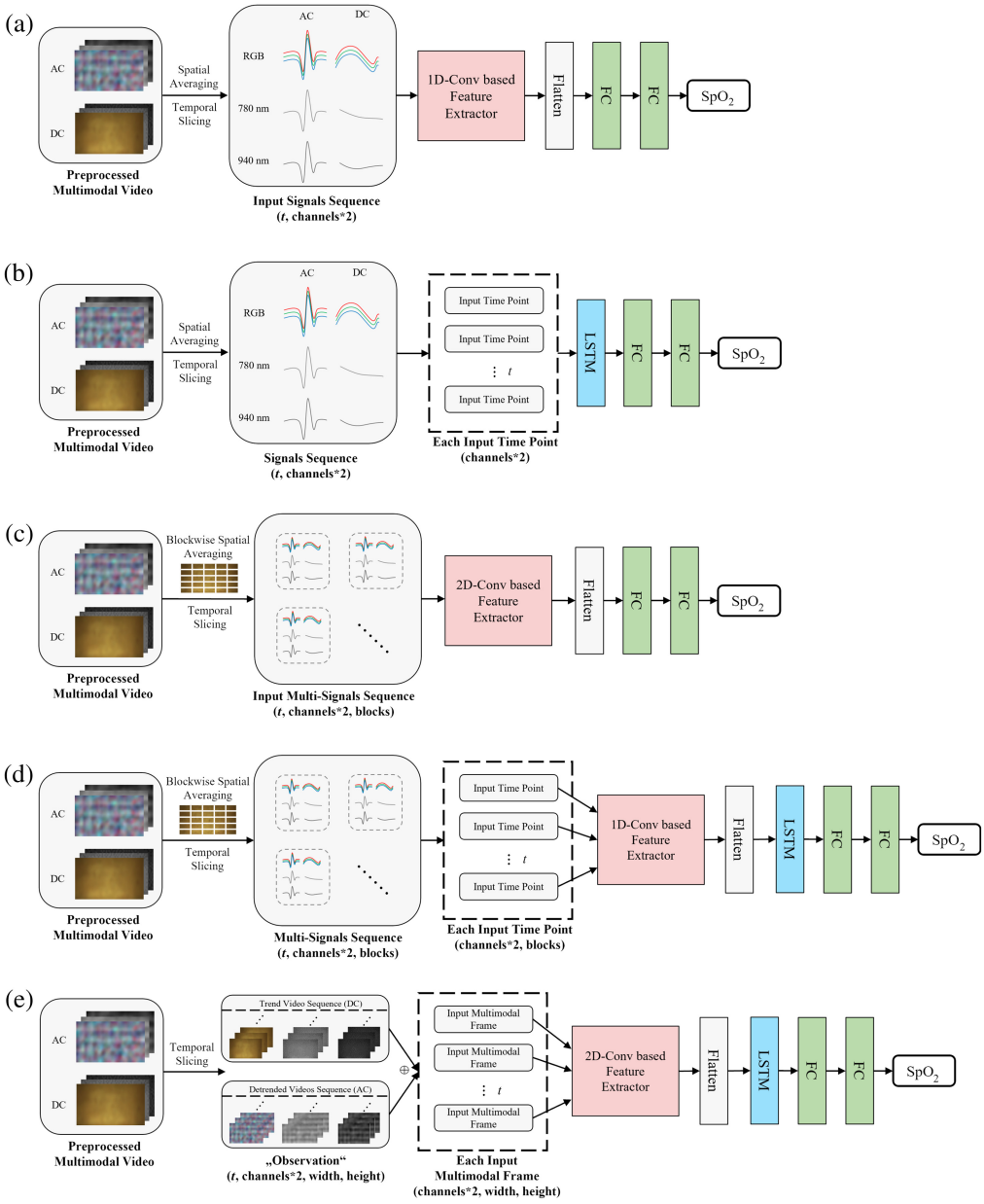

Fig. 10 Schematic of different feature extraction strategies and corresponding network architectures. (a) Strategy A. (b) Strategy B. (c) Strategy C. (d) Strategy D. (e) Strategy E.

As shown in Table 3, the proposed strategy (strategy F), in which temporal and spatial features are simultaneously extracted by 3D CNN, yields the best regression outputs. The distributions of MAE and Pearson correlation coefficients for the estimation SpO_2 using different strategies compared to reference values across 23 participants are presented in Fig. 11. It is noteworthy that in terms of both MAE and Pearson correlation coefficients, the proposed strategy demonstrates better results statistics (median, Q1, Q3) and distribution. Besides, strategies C and E exhibited similar performances and both achieved an MAE below 2.5 and an average Pearson correlation coefficient above 0.6.

We also compared different 3D CNN structures for feature extraction, considering both regression performance and the complexity of the models (size and computational load). Therefore, MACs, inference time, and the number of learnable parameters are introduced for a comprehensive evaluation of the network structure's performance. As shown in Table 4, 3D ResNet 10, 3D ResNet 18, and 3D ResNet 34 have no significant difference in regression performance for our task, while 3D AlexNet performs comparatively worse. Our proposed

Table 3 Result comparison between different feature extraction strategies.

Index	Feature extraction strategy	Network as feature extractor	MAE (%)	Avg. ρ
A	Only temporal features	1D CNN	3.14	0.45
B	Only temporal features	LSTM	3.24	0.46
C	Simultaneously extract temporal features and spatial features between blocks	2D CNN	2.49	0.61
D	Extract firstly spatial features between blocks and then temporal features	1D CNN + LSTM	2.63	0.59
E	Extract firstly spatial features on each frame and then temporal features	2D CNN + LSTM	2.43	0.63
F (proposed)	**Simultaneously extract spatial and temporal features**	**3D CNN (ResNet 18)**	**2.31**	**0.64**

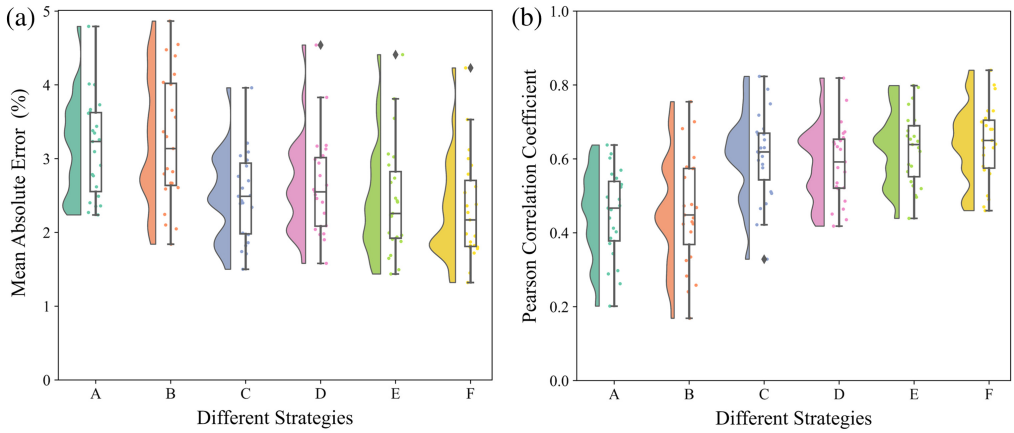

Fig. 11 Raincloud plots, which combine elements of box plots, violin plots ("cloud" part), and scatter plots ("rain" part), for performance metrics across different feature extraction strategies. The "cloud" part represents result distribution, while the "rain" indicates individual results of 23 participants. For each metric, boxes are used to describe the interquartile range (IQR) of the "leave-one-participant-out" test results on 23 participants with different strategies, which spans from the 25th percentile (Q1) to the 75th percentile (Q3). The whiskers extending from the boxes represent non-outlier results within 1.5 times IQR. The lines inside the boxes are medians. (a) Mean absolute error (MAE). (b) Pearson correlation coefficient (ρ).

Table 4 Performance comparison between different 3D CNN-based networks as feature extractor.

Model	MACs (G)	Inference Time (ms)	Learnable Parameters (M)	MAE (%)	Avg. ρ
3D AlexNet	0.73	28.10	2.07	2.81	0.52
3D ResNet 10	3.49	34.90	14.56	2.33	0.64
3D ResNet 18	4.27	45.80	33.36	2.31	0.64
3D ResNet 34	5.55	69.10	63.67	2.36	0.63
3D MobileNet V1	0.14	29.90	3.31	2.61	0.59
3D MobileNet V2	0.22	44.40	2.36	2.73	0.57
3D ShuffleNet V1	0.11	52.30	0.95	2.87	0.52
3D ShuffleNet V2	0.11	30.90	1.30	2.57	0.59

approach is just to choose the best network structure based on MAE and Avg. ρ, which is 3D ResNet 18. Lightweight networks such as 3D MobileNet V1, 3D MobileNet V2, 3D ShuffleNet V1, and 3D ShuffleNet V2 significantly reduce model complexity without a noticeable loss in regression performance.[55] Among them, 3D ShuffleNet V2 performs the best in these lightweight networks, achieving 2.57% MAE and 0.59 average Pearson correlation coefficient, which provides valuable reference for potential applications on mobile and embedded platforms.

4.4 Discussion of Image Modalities

After image registration, our method permits the combination of different imaging modalities for SpO_2 regression. Utilizing only NIR 780 and NIR 940 nm allows for overnight measurement, thereby broadening the applicability of this approach, such as in sleep monitoring scenarios. Table 5 illustrates the overall results when employing different modalities. Although the concurrent use of both RGB and NIR modalities yields the best estimation performance, relying solely on RGB or NIR does not lead to a collapse but only a slight MAE increase and an acceptable decrease of the Pearson correlation coefficient. From Fig. 12, it can be seen that estimations using only NIR resulted in a slightly higher MAE distribution for several participants and presented completely outlying Pearson correlation coefficients for two participants. However, in general, the distribution of the estimation results is similar to that when only RGB is used.

4.5 Clinical Validation on Sleep Apnea Patients

To clinically validate our method, we conducted a patient study in cooperation with the Center for Sleep and Telemedicine, University Medicine Essen, and recruited four patients with suspected SAS. SAS is a sleep-related breathing disorder characterized by repetitive breathing interruptions during sleep, resulting in daytime drowsiness, concentration difficulties, and increased

Table 5 Comparison of different input 2D imaging modalities.

Input modalities	Overnight measurement	MAE (%)	Avg. ρ
RGB	-	2.58	0.55
NIR	✓	2.68	0.51
RGB + NIR	-	2.31	0.64

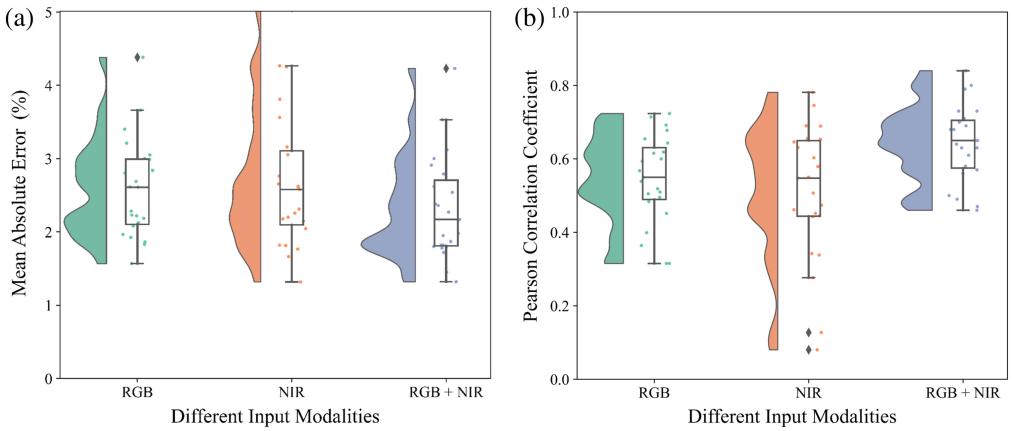

Fig. 12 Raincloud plots, which combine elements of box plots, violin plots ("cloud" part), and scatter plots ("rain" part), for performance metrics across different input modalities. The "cloud" part represents result distribution, while the "rain" indicates individual results of 23 participants. For each metric, boxes are used to describe the interquartile range (IQR) of the "leave-one-participant-out" test results on 23 participants with different strategies, which spans from the 25th percentile (Q1) to the 75th percentile (Q3). The whiskers extending from the boxes represent non-outlier results within 1.5 times IQR. The lines inside the boxes are medians. (a) Mean absolute error (MAE). (b) Pearson correlation coefficient (ρ).

Table 6 Results of SpO_2 estimation in trial clinical validation on SAS patients.

	Age (years)	AHI	ODI	Sleep time (h)	Available data (h)	MAE (%)
Patient #1	51	29.0	34.6	6.23	3.61	2.17
Patient #2	56	69.9	78.0	5.62	4.76	1.97
Patient #3	80	20.9	10.7	5.40	3.69	1.34
Patient #4	58	59.7	62.8	6.18	1.16	1.19

risk of cardiovascular diseases. Furthermore, recurrent breathing interruptions lead to a decrease in blood oxygen levels and eventually hypoxemia. The age of the included patients ranged from 51 to 58, while their apnea-hypopnea index (AHI) ranged from 29 to 69.9 and the oxygen desaturation index (ODI) from 10.7 to 62.8. AHI measures the severity of sleep apnea by calculating the number of apnea and hypopnea events per hour of sleep, while ODI quantifies the frequency of oxygen desaturation events, specifically drops of 3% or more, per hour of sleep.[56,57] Each patient is assigned a unique identifier, ranging from patients #1 to #4. The study is approved by the Faculty of Medicine, University of Duisburg-Essen (approval no. 21-10312-BO). Informed consent is obtained from all individual patients. These patients spent one night in the sleep laboratory, being simultaneously monitored by our camera system and the PSG system for reference. The color camera of our system is inactive during the measurement. Thus, the previously RGB-based facial landmark extraction and forehead ROI definition have been shifted to operate on NIR 780 images. In this experimental phase, our camera system's sensor head cannot move or rotate, resulting in a fixed field of view. We can check that the patient's face is within the camera's view at the start of recording, but patients might turn or move their heads after falling asleep. Therefore, in Table 6, we list some information about these four patients with their total sleep hours, the corresponding duration of available data, and MAE between estimated SpO_2 and reference in this duration.

To provide a more intuitive demonstration of the clinical results, we demonstrate in Fig. 13 the dynamic response of the estimated and reference SpO_2 signals during periods with

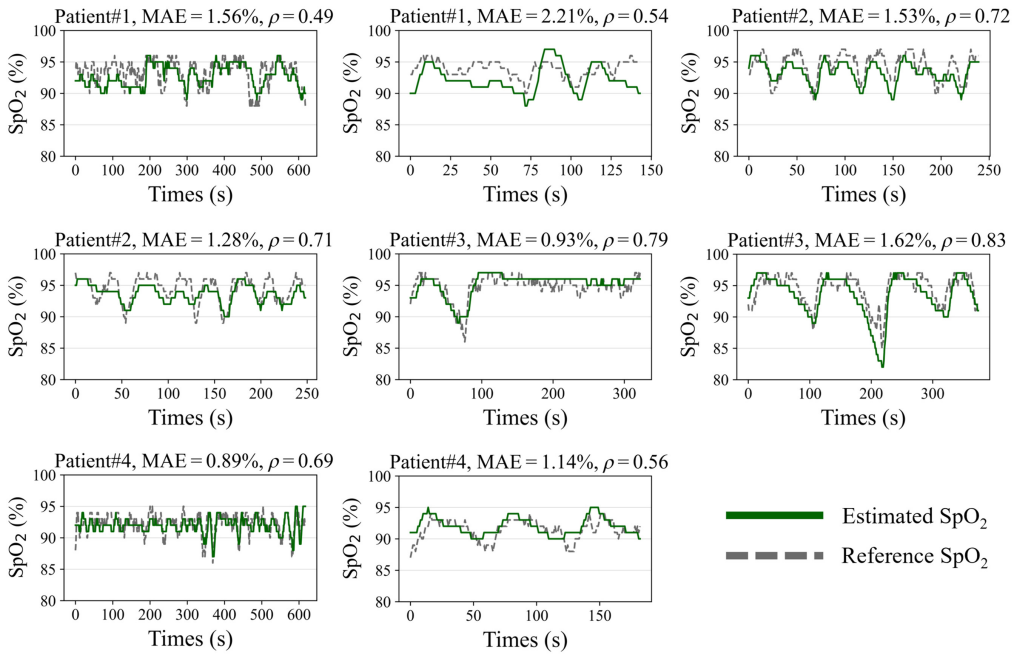

Fig. 13 Estimated SpO_2 values and PSG measured reference SpO_2 values of 4 SAS patients (two periods with desaturation events for each patient). The model is trained by all measurements from 23 healthy participants. The test was conducted at night and used only the infrared channels. The green lines represent the estimated values, while the reference signals are dashed gray lines.

desaturation and resaturation events. For each patient, two separate time periods with desaturation events are presented in two consecutive subplots. In a previous article of our research group,[58] we showed that we can distinguish periods with and without desaturation events in SAS patients, however without estimating the SpO_2 value. In this study, we show that we are able to accurately estimate the SpO_2 value in patients with a highly dynamic SpO_2 behavior with low MAE and high Pearson correlation coefficient. Furthermore, we have shown that the approach developed on healthy awake subjects can be applied to symptomatic SAS patients during sleep.

5 Conclusion and Future Work

This study introduced a contactless approach for SpO_2 estimation using 3D CNN and 3D VIS-NIR multimodal imaging. Through multimodal image registration, accurate 3D ROI tracking, multimodal video preprocessing, and spatial-temporal feature extraction, oxygen saturation can be accurately estimated from facial videos. The approach exhibited promising results, achieving an MAE of 2.31% and a Pearson correlation coefficient of 0.64 in a breath-holding study on healthy participants during short-term daytime measurements, showing a strong response to desaturation events and good agreement with recordings from contact-based commercial pulse oximeters. In clinical trials involving patients with sleep apnea syndrome, our approach demonstrated robust performance, with an MAE of less than 2% in SpO_2 estimations compared to gold-standard polysomnography (PSG). For the further improvement of SpO_2 estimation, we plan to utilize 3D information to incorporate illumination correction, aiming to further reduce distortions that are unrelated to oxygen saturation. Besides, future studies will focus on expanding the dataset to include a broader range of real patients, including varied skin types and more extensive pathological conditions (both stationary and ambulatory settings), to further validate the approach's effectiveness and generalizability. Furthermore, we aim to combine other non-contact measured vital signs, such as heart rate, respiration, and oxygen saturation, for correlation analysis to enhance disease diagnosis and patient recovery process monitoring.

Disclosures

The authors have no relevant financial interests in this article and no potential conflicts of interest to disclose.

Code and Data Availability

Due to privacy or ethical restrictions, the raw data used in this article is not publicly available but can be made available by the authors upon reasonable request.

Acknowledgments

This work is funded through a research grant (NO 416/4-1 675252) from the German Research Foundation. The experimental and clinical study is approved by the Ethics Committee of the Faculty of Medicine, University of Duisburg-Essen (approval no. 21-10312-BO). Specially, we express our great gratitude to all the participants and patients for their invaluable contribution to this work.

References

1. D. Evans, B. Hodgkinson, and J. Berry, "Vital signs in hospital patients: a systematic review," *Int. J. Nurs. Stud.* **38**(6), 643–650 (2001).
2. I. Brekke et al., "The value of vital sign trends in predicting and monitoring clinical deterioration: a systematic review," *Plos One* **14**(1) (2019).
3. C. Downey et al., "The impact of continuous versus intermittent vital signs monitoring in hospitals: a systematic review and narrative synthesis," *Int. J. Nurs. Stud.* **84**, 19–27 (2018).
4. W. R. Mower et al., "Pulse oximetry as a fifth pediatric vital sign," *Pediatrics* **99**, 681–686 (1997).
5. A. Mohyeldin, T. Garzón-Muvdi, and A. Quiñones-Hinojosa, "Oxygen in stem cell biology: a critical component of the stem cell niche," *Cell Stem Cell* **7**(2), 150–161 (2010).
6. B. Brown and B. Eilerman, "Understanding blood gas interpretation," *Newborn Infant Nurs. Rev.* **6**(2), 57–62 (2006).

7. A. Sudakou et al., "Time-domain NIRS system based on supercontinuum light source and multi-wavelength detection: validation for tissue oxygenation studies," *Biomed. Opt. Express* **12**(10), 6629–6650 (2021).
8. A. Sudakou et al., "Two-layered blood-lipid phantom and method to determine absorption and oxygenation employing changes in moments of DTOFs," *Biomed. Opt. Express* **14**(7), 3506–3531 (2023).
9. F. Lange and I. Tachtsidis, "Clinical brain monitoring with time domain NIRS: a review and future perspectives," *Appl. Sci.* **9**(8), 1612 (2019).
10. V. Ibáñez, J. Silva, and O. Cauli, "A survey on sleep assessment methods," *PeerJ* **6**, e4849 (2018).
11. J. E. Sinex, "Pulse oximetry: principles and limitations," *Am. J. Emerg. Med.* **17**(1), 59–66 (1999).
12. T. Tamura et al., "Wearable photoplethysmographic sensors—past and present," *Electronics* **3**(2), 282–302 (2014).
13. N. Bui et al., "Smartphone-based spO_2 measurement by exploiting wavelengths separation and chromophore compensation," *ACM Trans. Sens. Networks* **16**(1), 1–30 (2020).
14. X. Ding, D. Nassehi, and E. C. Larson, "Measuring oxygen saturation with smartphone cameras using convolutional neural networks," *IEEE J. Biomed. Health Inf.* **23**(6), 2603–2610 (2019).
15. X. Tian et al., "A multi-channel ratio-of-ratios method for noncontact hand video based SpO_2 monitoring using smartphone cameras," *IEEE J. Sel. Top. Signal Process.* **16**(2), 197–207 (2022).
16. T. Pursche et al., "Video-based oxygen saturation measurement," in *IEEE Int. Conf. Consum. Electron. (ICCE)*, IEEE, pp. 1–4 (2022).
17. M. van Gastel and W. Verkruysse, "Contactless SpO_2 with an RGB camera: experimental proof of calibrated SpO_2," *Biomed. Opt. Express* **13**(12), 6791–6802 (2022).
18. L. Kong et al., "Non-contact detection of oxygen saturation based on visible light imaging device using ambient light," *Opt. Express* **21**(15), 17464–17471 (2013).
19. N. H. Kim et al., "Non-contact oxygen saturation measurement using YCGCR color space with an RGB camera," *Sensors* **21**(18), 6120 (2021).
20. B. Wei et al., "Analysis and improvement of non-contact SpO_2 extraction using an RGB webcam," *Biomed. Opt. Express* **12**(8), 5227–5245 (2021).
21. Q. He et al., "Spatiotemporal monitoring of changes in oxy/deoxy-hemoglobin concentration and blood pulsation on human skin using smartphone-enabled remote multispectral photoplethysmography," *Biomed. Opt. Express* **12**(5), 2919–2937 (2021).
22. A. R. Guazzi et al., "Non-contact measurement of oxygen saturation with an RGB camera," *Biomed. Opt. Express* **6**(9), 3320–3338 (2015).
23. Y. Akamatsu, Y. Onishi, and H. Imaoka, "Blood oxygen saturation estimation from facial video via DC and AC components of spatio-temporal map," in *ICASSP 2023-2023 IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP)*, IEEE, pp. 1–5 (2023).
24. T. Stogiannopoulos, G.-A. Cheimariotis, and N. Mitianoudis, "A study of machine learning regression techniques for non-contact SpO_2 estimation from infrared motion-magnified facial video," *Information* **14**(6), 301 (2023).
25. J. Mathew et al., "Remote blood oxygen estimation from videos using neural networks," *IEEE J. Biomed. Health Inf.* **27**, 3710–3720 (2023).
26. F. P. Wieringa, F. Mastik, and A. V. D. Steen, "Contactless multiple wavelength photoplethysmographic imaging: a first step toward "SpO_2 camera" technology," *Ann. Biomed. Eng.* **33**, 1034–1041 (2005).
27. A. D. F. G. Rosa and R. C. Betini, "Noncontact SpO_2 measurement using eulerian video magnification," *IEEE Trans. Instrum. Meas.* **69**(5), 2120–2130 (2019).
28. M. Hu et al., "Contactless blood oxygen estimation from face videos: a multi-model fusion method based on deep learning," *Biomed. Signal Process. Control* **81**, 104487 (2023).
29. W. Liao et al., "Oxygen saturation estimation from near-infrared multispectral video data using 3D convolutional residual networks," *Proc. SPIE* **12621**, 126210O (2023).
30. A. Chromy and O. Klima, "A 3D scan model and thermal image data fusion algorithms for 3D thermography in medicine," *J. Healthc. Eng.* **2017**, 5134021 (2017).
31. Y. Zhang et al., "Point cloud hand–object segmentation using multimodal imaging with thermal and color data for safe robotic object handover," *Sensors* **21**(16), 5676 (2021).
32. J. Kim et al., "3D multi-spectrum sensor system with face recognition," *Sensors* **13**(10), 12804–12829 (2013).
33. M. Landmann et al., "High-speed 3D thermography," *Opt. Lasers Eng.* **121**, 448–455 (2019).
34. S. Heist et al., "High-speed three-dimensional shape measurement using GOBO projection," *Opt. Lasers Eng.* **87**, 90–96 (2016).
35. Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(11), 1330–1334 (2000).
36. M. Rosenberger et al., "3D high-resolution multimodal imaging system for real-time applications," *Proc. SPIE* **11397**, 1139704 (2020).
37. H. J. Park et al., "Ultrasonography analysis of vessels around the forehead midline," *Aesthet. Sur. J.* **41**(10), 1189–1194 (2021).

38. K. M. Jeong et al., "Ultrasonographic analysis of facial skin thickness in relation to age, site, sex, and body mass index," *Skin Res. Technol.* **29**(8), e13426 (2023).
39. C. Lugaresi et al., "Mediapipe: a framework for building perception pipelines," arXiv:1906.08172 (2019).
40. D. Tran et al., "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 4489–4497 (2015).
41. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
42. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
43. C. Zhang et al., "Enhanced contactless vital sign estimation from real-time multimodal 3D image data," *J. Imaging* **6**(11), 123 (2020).
44. International Commission on Non-Ionizing Radiation Protection, "Guidelines of limits of exposure to broad-band incoherent optical radiation (0.38 to 3 μm)," *Health Phys.* **73**, 539–554 (1997).
45. T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Arch. Dermatol.* **124**(6), 869–871 (1988).
46. B. K. Peterson, "Chapter 22 - vital signs," in *Physical Rehabilitation*, M. H. Cameron and L. G. Monroe, Eds., pp. 598–624, W.B. Saunders, Saint Louis (2007).
47. W. Verkruysse et al., "Calibration of contactless pulse oximetry," *Anesth. Analg.* **124**(1), 136–145 (2017).
48. M. van Gastel, W. Wang, and W. Verkruysse, "Reducing the effects of parallax in camera-based pulse-oximetry," *Biomed. Opt. Express* **12**(5), 2813–2824 (2021).
49. R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *23rd IEEE Int. Symp. Rob. and Hum. Interact. Commun.*, IEEE, pp. 1056–1062 (2014).
50. X. Niu et al., "VIPL-HR: a multi-modal database for pulse estimation from less-constrained face video," *Lect. Notes Comput. Sci.* **11365**, 562–576 (2019).
51. S. Bobbia et al., "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognit. Lett.* **124**, 82–90 (2019).
52. J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet* **327**(8476), 307–310 (1986).
53. J.-C. Cobos-Torres and M. Abderrahim, "Simple measurement of pulse oximetry using a standard color camera," in *40th Int. Conf. Telecommun. and Signal Process. (TSP)*, IEEE, pp. 452–455 (2017).
54. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).
55. O. Köpüklü et al., "Resource efficient 3D convolutional neural networks," in *IEEE/CVF Int. Conf. Comput. Vision Workshop (ICCVW)*, IEEE, pp. 1910–1919 (2019).
56. D. A. Pevernagie et al., "On the rise and fall of the apnea-hypopnea index: a historical review and critical appraisal," *J. Sleep Res.* **29**(4), e13066 (2020).
57. G. Ernst et al., "Difference between apnea-hypopnea index (AHI) and oxygen desaturation index (ODI): proportional increase associated with degree of obesity," *Sleep Breath.* **20**, 1175–1183 (2016).
58. B. Alić et al., "Contactless camera-based detection of oxygen desaturation events and ODI estimation during sleep in SAS patients," in *Proc. 17th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, Vol. 1, pp. 599–610 (2024).

Wang Liao received the BEng degree in mechanical engineering from Beijing Jiaotong University in 2017 and the MSc degree in mechanical engineering from Leibniz University Hannover in 2021. He is working in the Group for Quality Assurance and Industrial Image Processing at the Ilmenau University of Technology. His current research focuses on multimodal and multispectral image processing and AI-based feature fusion for industrial and biomedical scenarios.

Chen Zhang received his BSc, MSc, and PhD degrees from the Ilmenau University of Technology in 2013, 2014, and 2024, respectively. Since 2015, he has been a member of the Group for Quality Assurance and Industrial Image Processing at the Ilmenau University of Technology. He is now working on the development and implementation of advanced 3D, multi-spectral, and multimodal imaging systems with applications in industrial and biomedical fields.

Belmin Alić received his BSc degree in electrical and electronics engineering from the International Burch University in Sarajevo, Bosnia and Herzegovina, in 2016, and his MSc degree in embedded systems engineering from the University of Duisburg-Essen in Duisburg, Germany, in 2018. He is currently pursuing his PhD at the University of Duisburg-Essen, Germany. His research focus includes biosignal analysis and feature engineering for camera-based vital sign monitoring and detection of sleep-related breathing disorders.

Alina Wildenauer is a researcher and doctoral candidate at the University Medicine Ruhrlandklinik in Essen, and received her BA and MA in sociology with specialization in

statistics and empirical research in 2016 and 2020, respectively. Her current research interests include sleep research and digitalization in health care, as well as patient-reported outcome measures and clinical trials.

Sarah Dietz-Terjung is a biotechnologist and medical physicist. Since 2015, she has been researching sensor development and the use of AI in pneumology and sleep medicine at the University Medical Center Essen, Ruhrlandklinik, and has also completed her doctorate in this field.

Jose Guillermo Ortiz Sucre graduated from the Universidad Central de Venezuela in 2010. He completed his specialization in radiology in 2015 in Caracas, Venezuela, and his specialization in pulmonology in June 2023 in Essen, Germany. He currently works as a research associate at the Ruhrlandklinik in Essen, focusing on the study of cystic fibrosis and pulmonary fibrosis patients, fulfilling the role of a study physician.

Sivagurunathan Sutharsan is a senior physician in the clinic for pneumology at the Ruhrlandklinik. His expertise lies in the areas of bronchiectasis, cystic fibrosis, respiratory physiology, interventional pulmonology, and pleural tuberculosis. In addition to his clinical work, he is involved in various projects not only in the field of basic research but also in the development of innovative sensor technology for the long-term monitoring of patients with chronic lung disease.

Christoph Schöbel holds Germany's first university professorship for sleep and telemedicine. In addition to scientific work on the cardiovascular effects of sleep disorders, Prof. Schöbel is involved in the further development of telemedical approaches in interdisciplinary collaborative projects and is also developing new care approaches in the field of sleep medicine in collaboration with funding bodies, incorporating smart sensor technology and new digital methods including self-tracking.

Karsten Seidl is a full professor of micro and nanosystems for medical technology at the University of Duisburg-Essen and head of Business Unit Health at the Fraunhofer Institute for Microelectronic Circuits and Systems, Duisburg (Germany). He studied Electrical Engineering and Information Technology at Ilmenau University of Technology (Germany) and did his PhD at the University of Freiburg/IMTEK (Germany). Before his current position, he has been working at the Robert Bosch and Bosch Healthcare Solutions (BHCS) GmbH.

Gunther Notni studied physics at the Friedrich Schiller University in Jena and works at the Fraunhofer Institute for Applied Optics and Precision Engineering IOF in Jena and Ilmenau University of Technology where he is appointed to the professorship of the "Quality Assurance and Industrial Image Processing" department. His work focuses on the development of optical 3D sensors and the principles of multimodal and multispectral image processing and their application in human-machine interaction, quality assurance, and medicine.