IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 January 2024, accepted 6 February 2024, date of publication 9 February 2024, date of current version 21 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365055

RESEARCH ARTICLE

A Study on the Implementation of Temporal Noise-Robust Methods for Acquiring Vital Signs

SEONGCHAN PARK [1], HEEJUN YOUN [1], SEUNGHYUN LEE [2], AND SOONCHUL KWON [3]

[1]Department of Plasma Bio Display, Kwangwoon University, Seoul 01897, South Korea
[2]Department of Ingenium College Liberal Arts, Kwangwoon University, Seoul 01897, South Korea
[3]Department of Interdisciplinary Information System, Graduate School of Smart Convergence, Kwangwoon University, Seoul 01897, South Korea

Corresponding author: Soonchul Kwon (ksc0226@kw.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information and Communication Technology (ICT) Challenge and Advanced Network of Human Resource Development (HRD) (ICAN) Program, Supervised by the Institute of Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-RS-2022-00156215; and in part by the Research Grant of Kwangwoon University, in 2024.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT There has been a surge in research focused on the analysis of vital signs using remote photoplethysmography (rPPG) sensors, as opposed to traditional photoplethysmography (PPG) methods. Unlike PPG, rPPG imposes no spatial constraints and employs a straightforward measurement technique, making it increasingly prevalent. Its integration into image processing, harnesses the remarkable advances in artificial-intelligence technology, achieving accuracy that is comparable to that of traditional PPG sensors. In prior studies, obtaining vital signs often necessitated an unnecessary and procedural fixation of facial positions within frames to enhance predictive accuracy. Despite such fixation, achieving notably high accuracy remained elusive. Here, we introduce a simple yet robust approach utilizing videos captured by an rPPG sensor, ensuring both high accuracy and resilience to noise. We propose a convolutional neural network model meticulously designed to resist interference from noise data that may arise in the initial stages, coupled with effective preprocessing techniques to attain superior predictive accuracy. Data extracted by a facial extractor undergoes preprocessing via normalization. Leveraging the Temporal Shift Module (TSM), this normalization efficiently captures temporal relationships without incurring additional computational overhead. Mitigating noise signal interference from non-facial data through the use of multiple attention masks and augmenting prediction accuracy via skip connections. Moreover, we compile a specially tailored dataset for pulse rate and breath rate data, catering specifically to the East Asian population. The proposed process demonstrates outstanding performance in predicting both pulse rate and breath rate.

INDEX TERMS Attention network, convolutional neural network (CNN), remote PPG (rPPG), skip connection, unconstrained sensor, vital signs.

I. INTRODUCTION

In recent decades, the acquisition of human vital signs has become a prominent area of study [1], [2], [3]. Monitoring vital signs plays a crucial role in patient health management and has proven valuable in aspects such as disease prevention, especially for senior populations [4], [5]. The primary vital signs of interest include pulse rate (PR), breath rate (BR), and stress index (SI). While previous research primarily focused on PR and BR, recent studies [6] have introduced methods to measure SI based on PR.

Methods for acquiring vital signs can be broadly categorized into those that employ constrained sensors and those that employ unconstrained sensors. Constrained sensor

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

methods, such as photoplethysmography (PPG) and electrocardiography (ECG), are well-established for their high accuracy [7]. ECG involves the attachment of electrodes to the skin to record the electrical activity of the heart, providing pulse rate (PR) information. PPG sensors, typically placed on the fingertip or other areas, emit light into the blood to capture PR information through its interaction with the blood. Despite their accuracy, constrained sensors require direct skin contact, which can be uncomfortable for users [8].

To overcome these limitations, a measurement method based on remote photoplethysmogram (rPPG) has been proposed [9]. It is the most researched computer vision-based approach for extracting vital signs from video sequences. This method capitalizes on the phenomenon of rPPG, which results from periodic changes in local tissue blood volume due to heartbeats. Importantly, it does not require any specialized equipment; rather, it uses video data from Red, Green, Blue (RGB) cameras readily available in daily life, like smartphone cameras or webcams. The post-COVID era has significantly heightened the interest in remote physiological measurement through rPPG [10], [11]. Nevertheless, existing methods for vital-sign analysis through traditional rPPG encounter limitations, particularly concerning noise such as variations in lighting conditions and facial movements [12].

The integration of deep learning (DL) into computer vision, owing to the remarkable advances in artificial intelligence, has yielded numerous achievements [13], [14], [15]. DL has been instrumental in improving conventional computer-vision technology and achieving high accuracy [16]. Various DL models have emerged, with some incorporating convolutional neural networks (CNNs) in rPPG analysis [17], [18], [19]. Physnet [17] used a 3D-CNN to harness spatiotemporal information, whereas rPPGNet [18] also achieved excellent performance using a multi-layered 3D-CNN architecture. Both models demonstrated outstanding results. Deepphys [19] used 2D-CNN and a skin reflection mechanism to extract information from various facial regions, outperforming existing methods. However, DL still poses a significant challenge in dealing with spatial noise in critical conditions, which is considered an ongoing obstacle [20], [21]. To mitigate the impact of such noise, approaches involving multi-task mechanisms [22] have been proposed.

Recent studies have proposed approaches to address noise caused by facial movements using state-of-the-art Face Detectors. By extracting only the region of interest (ROI) - the face - through facial detection, researchers aim to overcome the drawbacks induced by facial movements for rPPG analysis. However, since facial detectors operate based on deep learning CNN algorithms, they cannot guarantee 100% accuracy, occasionally detecting non-facial areas. Consequently, during the process of frame extraction, images from non-facial regions might enter the model as input. The inclusion of noisy data alongside regular data can cause confusion in the model's inference, leading to a decline in inference accuracy.

This study presents a straightforward yet highly accurate and robust approach using videos captured by an rPPG sensor. Our approach focuses on extracting facial data through facial detection from sensor-recorded videos without constraints, emphasizing on preprocessing and utilizing CNN algorithms for vital signs monitoring. To ensure robustness against non-facial noise data mixed with regular data via facial detection, we structured the process accordingly. The extracted facial data undergoes a series of calculations during preprocessing to identify relationships between frames. Subsequently, in the proposed CNN model, we employ the Temporal Shift Module (TSM) to transform the data into time-series data. Attention networks are utilized to emphasize facial features during intermediate training stages, incorporating skip connections to retain attention masks and features concerning temporal changes. Additionally, we constructed a dataset for vital signs monitoring targeting individuals of East Asian descent. Videos from 30 East Asian participants were recorded, and using a unconstrained sensor, we measured vital signs (PR, BR) to calculate rPPG.

This paper is structured as follows: Section II presents related works. Section III explains the proposed vital-sign acquisition process. Section IV provides details about the dataset, evaluation methods, and training. Section V conducts experiments covering the model's training results through convergence speed, Model's Ablation experiments proposed by us, comparative performance experiments with other models, and investigations into the relationship between PR and SI. Finally, Sections VI and VII discuss the findings and draw conclusions.

II. RELATED WORKS

A. TRADITIONAL METHODS OF RPPG

The rPPG study was first conducted by Verkruysee et al [24]. It was initially proposed as a contactless method for measuring PR by capturing variations in skin color using RGB cameras, based on the principles of the existing PPG method. Over time, many researchers made significant efforts to enhance the accuracy of PR analysis using rPPG [25], [26], [27], [28], [29], resulting in three distinct methods: methods based on blind source separation (BSS)techniques, optical model, and data.

BSS involves the separation of independent source signals from mixed input signals. Independent component analysis (ICA) and principal component analysis are the most commonly used techniques in BSS. With ICA, researchers can isolate the RGB signal most independent from the rest and identify the signal most closely related to PR. High accuracy in PR calculation using ICA has been achieved [30], [31].

The optical-model-based method estimates PR and blood flow information based on optical properties. This method calculates color-difference signal features by combining RGB signals between frames, and uses these features to estimate PR. Changes in the color-difference signal are attributed to variations in skin reflection and blood absorption spectra, enabling the measurement of PR. This approach is widely employed in non-invasive heart-rate monitoring, and it allows for accurate PR measurement through a combination of optical properties and data analysis [32], [33].

Data-based methods estimate PR using extensive training data. This approach leverages machine-learning and deep-learning techniques to train models that estimate PR and blood flow information from data. By using diverse and abundant datasets, these models recognize intricate patterns and features, providing accurate PR predictions. Data-based methods enable remote, non-invasive vital-sign monitoring and find applications in various fields, including medical, sports, and smart wearable devices [34], [35], [36].

B. USING ATTENTION NETWORK FOR DEEP-LEARNING TRAINING

In the realm of DL, CNNs have demonstrated remarkable achievements in image processing tasks [37], [38], [39]. However, with the surge in demand for more efficient techniques to highlight essential data and suppress irrelevant information, recent research [40], [41], [42] has witnessed increasing interest in the integration of attention networks into CNNs. These endeavors have led to noticeable performance enhancements by focusing on specific regions or objects within images. For instance, in tasks like image-caption generation, the introduction of visual attention, which assigns weight to crucial image regions, has significantly improved accuracy and semantic consistency. Furthermore, the fusion of CNNs with attention networks is gaining momentum in tasks such as object detection and image classification [43], [44], [45].

The recently proposed Multi-Task Attention Network (MTAN) [46] enables simultaneous multi-task training and the sharing of common features. Each task is equipped with a soft-attention module, which takes the shared features of the network as input and is trained to emphasize task-specific features. This emphasis on features allows the benefits of shared network features while also enabling individualized training for each task. MTAN has proven to be an effective approach, delivering significant results in multi-task training. Moreover, it is user friendly and amenable to end-to-end training.

C. EFFICIENT TEMPORAL INFORMATION ANALYSIS

In the realm of video-data processing, 3D-CNNs have proven to be efficient in incorporating temporal modeling, leading to substantial improvements in performance and accuracy [23], [47], [48], [49]. However, it is worth noting that a network that uses temporal information, like 3D-CNN, tends to come with increased computational complexity and more parameters compared to using a combination of 2D-CNN and recurrent neural networks (RNN) [49]. To match the performance of efficient temporal modeling achieved by 3D-CNN, the inclusion of temporal information within each frame becomes imperative. TSM [50] enables the seamless exchange of temporal information across multiple frames by shifting the tensor along the temporal dimension. This facilitates the exchange of information between frames, resulting in outstanding performance improvements, both in terms of reduced latency and increased accuracy. TSM has proven particularly effective when applied to real-time video object detection.

D. EXTRACTION OF REGION OF INTEREST (ROI) USING FACE DETECTOR

Most of the current rPPG research focuses on acquiring vital signs through subtle changes in micro-blood perfusion in the face [20], [21]. Studies in this domain can be categorized into those that extract facial landmarks to analyze specific Regions of Interest (ROIs) such as the cheeks or forehead's blood flow variations and those that analyze overall blood flow characteristics across the entire face. Therefore, in analyzing vital signs from images captured by RGB cameras, the Region of Interest (ROI) for vital sign analysis is predominantly defined as the face, with other areas outside the face considered irrelevant to the analysis. Hence, the use of a face detector to define the ROI is crucial. However, when extracting N frames using a face detector, there is no guarantee that facial data will be detected in all N frames. Even when employing high-performing CNN algorithms, achieving 100% accuracy is not guaranteed. If non-facial data infiltrates among regular data due to the CNN model's performance, it introduces noise, subsequently leading to a decline in prediction accuracy during signal analysis.

III. VITAL SIGNS ACQUISITION PROCESS

Figure 1 illustrates the proposed vital-sign acquisition process proposed in this study. It leverages an RGB camera to selectively crop the facial region from a recorded video, which serves as the input data. This focused cropping ensures that the location of the face within the video does not affect the results. Moreover, by isolating the face image from the surrounding background, noise is minimized, resulting in improved accuracy. The face image is preprocessed, normalized, and then fed into an object vital signs acquisition network (OVSA-Net)—the model introduced in this study. The integration of TSM [50] enhances the ability of the model to discern temporal differences within 2D data. Additionally, the multi-attention network further boosts the emphasis on the region of interest (ROI). The outcomes generated by the model are then processed to derive values for PR, BR, and SI through signal processing.

A. PREPROCESSING

$$Img_i = \frac{Img_i - Img_{i-1}}{Img_i + Img_{i-1}}, Img_i = \frac{Img_i}{std(Img)} \tag{1}$$

$$Img_i = Img_i - mean(Img), Img_i = \frac{Img_i}{std(Img)} \tag{2}$$

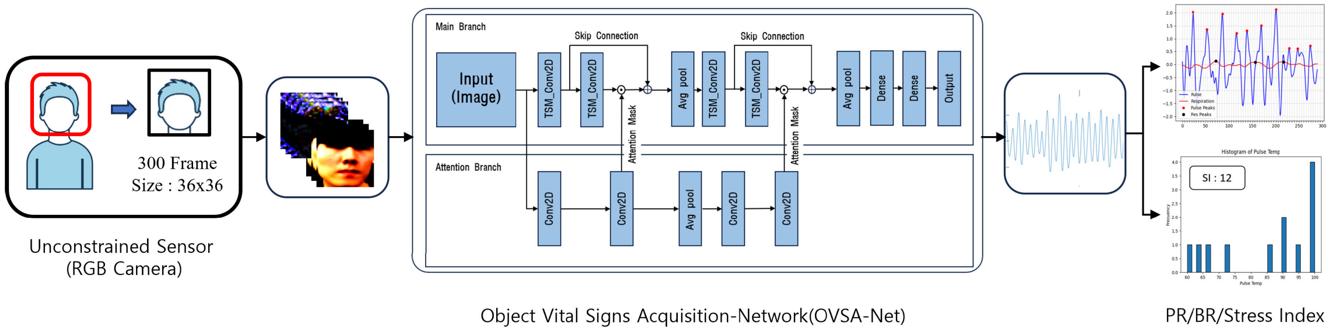


FIGURE 1. The process of acquiring vital signs (PR, BR, SI) from video measured by an unconstrained sensor.

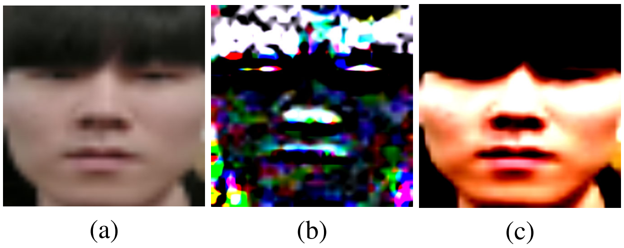


FIGURE 2. Comparison of the original image and the preprocessed image.

Figure 2 illustrates the raw, main branch, and attention branch inputs, respectively. Subfigures (b) and (c) depict the normalized images, which are crucial in enhancing training and inference accuracy. Formulas 1 and 2 represent the normalization equations for input to the main and attention branches, respectively. Img_i means frame i, and $Img_i - 1$ means frame i-1. Since the methods for measuring rPPG involve capturing the PPG signal from the face, it is essential to monitor the RGB values of the skin. The proposed model predicts vital signs by analyzing the variations in RGB values of the face corresponding to photoplethysmographic changes. In the main branch, a normalization process is applied to calculate the RGB differences between frames comprising time-series data. As shown in Formula 1, this involves computing the differences across all frames in the image, followed by normalization to the standard deviation. In the attention branch, the input image is used to emphasize the weight of ROI within the frame. To enhance the importance of ROI, the input image is normalized based on the mean value of the entire input frame. As demonstrated in Formula 2, this process entails calculating the difference from the average across all frames in each image, followed by normalization to the standard deviation.

B. PROPOSED CNN MODEL

Figure 3 illustrates the architecture of the proposed CNN-based OVSA-Net in this study. The model was designed based on the VGG16 Network [51]. VGG16 possesses a deep and concise structure composed of 3 × 3 convolution filters and pooling layers, making it relatively simple and easy to implement and understand. Despite its simplicity, VGG16 has demonstrated high performance on large-scale datasets like ImageNet and proved to be effective in various types of image classification tasks. Accordingly, in this paper, the model was constructed by mimicking the structure of VGG16. Efforts were made to reduce the Computation Cost by resizing the input images to 36 × 36. Consequently, the model was configured with only eight layers, including Fully Connected Layers, achieving excellent performance. The structure involved repeating a combination of two convolutional layers with 3 × 3 filters followed by an Average Pooling layer twice. Subsequently, it was constructed with two passes through Fully Connected Layers. Our focus in designing the model for acquiring biosignals was on accuracy, convergence speed, and reducing noise within the videos. We aimed to ensure real-time applicability for use in applications or web-based research by implementing a low-cost model, avoiding high expenses. To achieve this, our design was tailored to encompass specific functionalities: 1) The ability to effectively perceive RGB changes within the face over time, and to effectively learn by emphasizing the importance of the facial region, which is the source of vital signs acquisition. 2) The ability to quickly process 3D data captured by video data without much computational cost, while effectively perceiving the temporal relationship information between frames. 3) The ability to maintain the accuracy of the model even if noise signals are mixed in the input video data. 4) The ability to achieve high accuracy with a small amount of training, with a fast convergence speed. To achieve the four abilities, this architecture uses the following three elements: TSM 2D convolution, multi-attention network, and skip connection.

1) TSM 2D CONVOLUTION

When comparing the existing 2D-CNN and 3D-CNN models, 2D-CNN stands out for its relatively efficient computational cost and its ability to effectively leverage dimensional information [52]. However, it has a drawback: the inability to seamlessly incorporate temporal information. In contrast, 3D-CNN excels in exploiting temporal information, but it comes at the cost of significantly higher computational demands, roughly quadratic compared to 2D-CNN.

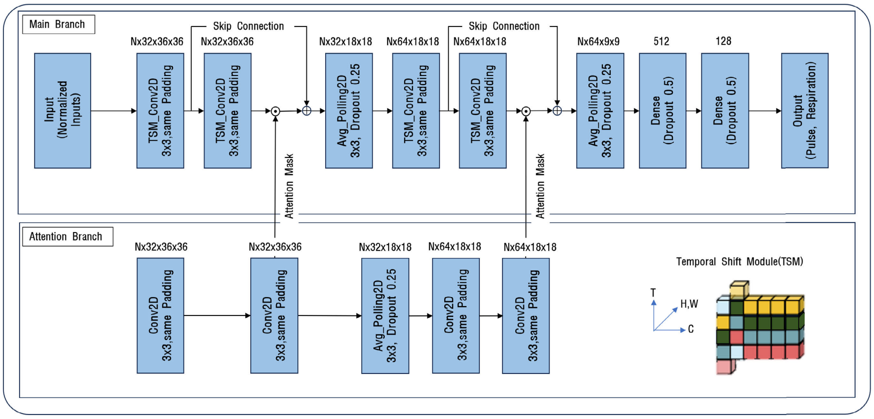


FIGURE 3. Object vital signs acquisition network(OVSA-Net) architecture.

We adopted a 2D-CNN approach using TSM proposed by Ji Lin [50] to strike a balance between fast computational speed and temporal information utilization. TSM offers the advantage of capturing efficient temporal information by exchanging information among channels of 2D images, thereby blending information along the time axis [50]. Leveraging this advantage, the integration of 2D-CNN with TSM enabled us to create a model with faster computation than 3D-CNN, while also achieving improved accuracy through the inclusion of temporal information.

$$X_i^{-1} = X_{i-1},\ X_i^0 = X_i,\ X_i^{+1} = X_{i+1} \tag{3}$$

$$Y_i = w_1 X_{i-1} + w_2 X_i + w_3 X_{i+1} \tag{4}$$

Formula 3 and 4 are the expressions used in the TSM_Conv2D operation. Where X_i^{-1}, X_i^0, and X_i^{+1} represent the -1, 0, and 1 channels of the current frame X_i, respectively. And w is the weight. Y_i denotes the convolution operator. The structure of the 2D-CNN incorporating TSM is as follows: Initially, a shift operation is performed for each channel without any additional computation, as described in Formula 3. Formula 4 demonstrates a process similar to multiply-accumulate, seamlessly integrating it with a 2D convolution, effectively eliminating the need for additional computations found in traditional 2D-CNN models.

2) MULTI-ATTENTION NETWORK

The main branch undergoes a preprocessing step that uses the difference in RGB values between frames and standard deviation within the input image, with the value depicted in Figure 2-(b) serving as input. Internally, various temporal information is integrated by shifting to the time axis of specific channels using TSM. However, this temporal shifting and integration process has its drawbacks - during the extraction of facial data through the face extraction model, the introduction of noise, which comprises data other than the face, can occur. This noise can pose difficulties in the connection of pixel information between frames, potentially leading to the loss of feature information. Consequently, this becomes a contributing factor to the decrease in prediction accuracy [55]. To address this, we incorporate a multi-attention network [23]. Within the multi-attention network, we use the value from Figure 2-(c) as input. This value is normalized using the standard deviation of the subtracted values as the mean value. The attention branch shares the same structure as the main branch, and the attention mask is transmitted to the main branch after every two convolution layers.

$$F_{mul}(A, B) = A \odot B \tag{5}$$

Formula 5 represents the connection formula between the Main branch and the attention mask. A denotes the output of TSM_Conv2D, while B represents the values of the attention mask. The operation involves an element-wise multiplication between A and B. By boosting the weight of the facial region within the main branch, this approach enables clear identification of differences in RGB values related to the PPG of the face.

3) SKIP CONNECTION

In the process of designing a deep-learning model, practitioners often grapple with issues related to accuracy, generalization, and the resolution of gradient vanishing problems [53]. In the OVSA-Net model, after the multiply operation within the attention mask component, we explored various patterns to connect spatial features to reduce the number of parameters of the model and the number of tasks the model is required to perform. Among these strategies, this model achieved improvements in training convergence rates while introducing spatial features and complexity through short skip connections [54].

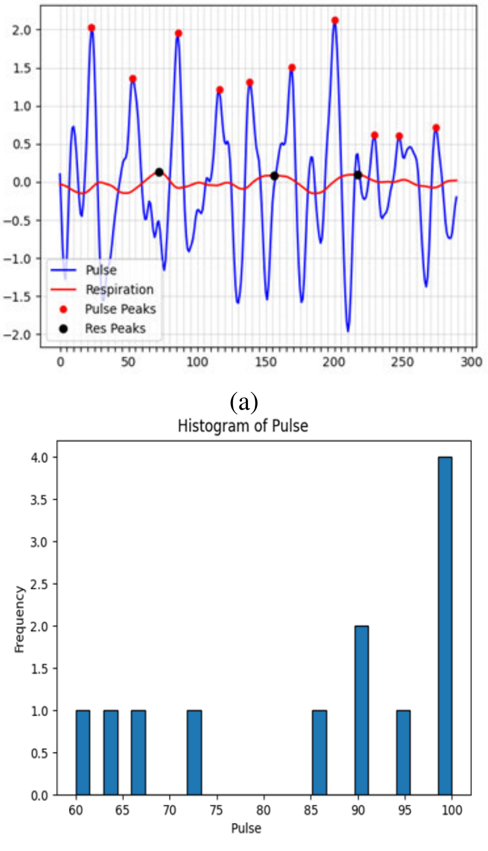
$$F_{sum}(X, M) = X + M \tag{6}$$

This represents the formula for the skip connection used in this paper. X signifies the value to be retrieved from the previous layer through the skip connection, while M represents the output value denoted as F_{mul} in Formula 5.

The operation involves an element-wise summation between X and M.

C. VITAL SIGNS ANALYSIS METHOD

Figure 4-(a) shows the PR and BR values, which are the outputs of the model obtained through signal processing, involving techniques such as Detrend and Butterworth filtering. In contrast, Figure 4-(b) presents a histogram used to calculate SI based on the PR values. Formulas 7 and 8 outline the calculations for the average PR and BR per minute, respectively. In a pulse graph, the interval between peaks is commonly referred to as peak to peak. In this context, we can compute PR_i, representing the predicted PR per minute, by dividing peak to peak by 60. If there are N peak-to-peak intervals, taking their sum and dividing by N yields PR, the average PR per minute. Similarly, in the context of respiration, the interval between breaths is known as the RR interval. For BR estimation, we calculate BR_i, the predicted BR per minute, by dividing the RR interval by 60. Summing N RR intervals and dividing by N provides BR, the average breath rate per minute. Figure 4-(b) involves converting the pulse values from Figure 4-(a) into a histogram to facilitate the calculation of SI. To compute SI, we referred to Baevsky's SI methodology, and the resulting value was obtained as the output.



(a)

(b)

FIGURE 4. Experimental results of vital signs(PR, BR, SI). (a) Experimental results of PR and BR, (b) Experimental results of SI.

$$PR_i = \frac{60}{Peak\ to\ Peak\ interval}, PR = \frac{1}{N}\sum_{i=1}^{N} PR_i \tag{7}$$

$$BR_i = \frac{60}{RR\ interval}, BR = \frac{1}{N}\sum_{i=1}^{N} BR_i \tag{8}$$

Figure 5 illustrates Amo, Mo, and MxDMn as depicted in the histogram used to compute SI. Formula 9 is employed for calculating SI. In this context, Amo represents the mode amplitude expressed as a percentage, Mo denotes the most frequent PR interval, and MxDMn signifies the difference between the longest and shortest PR intervals. Amo, the mode amplitude, refers to the normalized height of peaks in the pulse rate histogram, typically around 50 ms. MxDMn represents the disparity between the longest and shortest peak-to-peak pulse rate intervals.

$$Stress\ Index(SI) = \sqrt{\frac{(Amo)100}{2(Mo)(MxDMn)}} \tag{9}$$

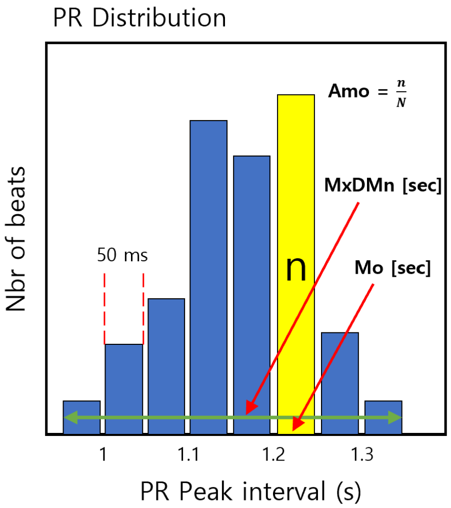


FIGURE 5. Computation of geometric measures of Baevsky's stress index.

TABLE 1. Stress states based on SI ranges.

$0 < \mathrm{SI} < 10$	Low Stress
$10 \leq \mathrm{SI} \leq 15$	Normal Stress
$15 < \mathrm{SI}$	High Stress

The graph for all generated PR from Figure 4-(a) is depicted in Figure 4-(b) and transformed into a histogram similar to Figure 5. From this histogram, Amo, Mo, and MxDMn are computed, followed by employing formula 9. This process leads to the derivation of the SI. Table 1 provides an overview of stress states classified based on SI. It's important to note that since SI utilizes the peak-to-peak pulse rate intervals, variations in PR values can lead to corresponding changes in SI values.

IV. EXPERIMENTAL ENVIRONMENT

A. DATASET

For our experiment, we developed a custom dataset consisting of East Asian participants. A total of 30 healthy adults, aged between 20 and 30, took part in the study. This dataset comprises video recordings of the subjects and concurrent PR and BR data collection. All participants involved in the development of the dataset were obtained with proper consent before proceeding.

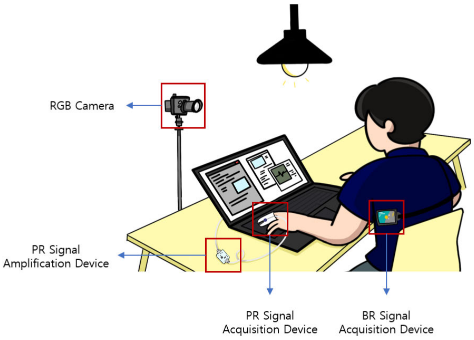


FIGURE 6. Dataset measurement method.

TABLE 2. Dataset measurement setup information.

Camera (Image sensor)	ELP-USB4K03-MFV(IMX415 image sensor, 3840 * 2160, 30fps, Variable focal length lens 5-50mm)
Contact PPG sensor	PSL-iPPG2C(Acquisition, Range 0-3.3V, reflection PPG) PSL-DAQ(Amplification, 60HZ Notch Filter, 7500V/V) (PhysioLab, Republic of Korea)
Contact Breathing sensor	Go Direct Respiration Belt(Range 0-50N, Resolution 0.01) (Vernier, USA)
Image resolution / fps	Original : 1920 x 1080, Face cropped : 200 x 300 / 30fps
Experiment setup	Camera fixed on upper body, allowing facial movement
Illumination setup	Indoor LED lighting set at 5700K, including settings for both illuminated and off conditions
Number of subjects	30 subjects (male : 25, female : 5)
Recording time	Total 120 min (60 samples, 2min per sample)
Recording program	Self-designed recording program based on Python

Figure 6 and Table 2 provide an overview of the measurement environment for each dataset. During the experiments, the participants were seated in chairs. The camera used was ELP-USB4K03-MFV, PR values were recorded using a PhysioLab's PPG sensor, PSL-IPPG2C Module, and PSL-DAQ Module. Simultaneously, BR was measured using Verniner's Go Direct Respiration Belt, a constrained sensor. We employed a self-designed Python-based program for data construction, enabling simultaneous recording of video and sensor data. The video data was recorded at 30 frames per second (fps), while the sensor data was recorded at a rate of 30Hz per second. The program was configured to record two sensor data inputs within a single frame, enhancing the precision of dataset construction. Following the recording, a separate process cropped the facial region from the captured video for input into the deep learning model. In our study, we watched visual materials of various genres to obtain a wide range of Pulse Rate (PR) and Breath Rate (BR) data. Additionally, illumination adjustments were made to create a dataset robust to changes in lighting conditions. Each session consisted of two filming periods, resulting in a total duration of 2 minute, from which we collected a total of 60 samples of video data.

B. TRAINING DETAIL

In this experiment, we applied a stochastic gradient descent optimizer with a maximum of 30 epochs and adaptive learning rate, initialized at 0.001. A mean squared error (MSELoss) function was employed as the loss function. The network components were implemented using the PyTorch library, and the Nvidia RTX 3090 graphics processing unit was used. Activation functions were based on the hyperbolic tangent (tanh). Dropout was applied at a rate of 0.5 for the fully connected layer and 0.25 after employing average pooling. Since the proposed process relies on cropped data of only the facial area as input, a separate detection model was used. Cropped facial data were resized to a size of 36 × 36x3 and used as input to the model. The batch size (N) was set to 16, as it provided the most stable results based on experimental findings. TSM was applied using an offline method. The set of 60 images from the dataset constructed in this study was divided into training, testing, and validation sets at an 80%, 10%, and 10% ratio, respectively, ensuring a comprehensive representation of PR and BR.

C. EVALUATION METHOD

$$MSE = \frac{\sum_{i=1}^{N}(V_{preducted_i} - V_{gt_i})^2}{N} \tag{10}$$

Formula 10 represents the Mean Squared Error (MSE), which computes the average of the squared differences between predicted values, denoted as $V_{Predicted}$, and actual values, denoted as V_{gt}. N signifies the number of samples. A smaller value indicates that the model's predictions are closer to the actual values, serving as a metric for evaluating the model's performance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(V_{preducted_i} - V_{gt_i})^2}{N}} \tag{11}$$

Formula 11 represents the Root Mean Squared Error (RMSE), which calculates the square root of the mean of the squared differences between predicted values and actual values. A smaller RMSE value indicates that the model's predictions are more accurate and is used as a metric for evaluating the model's performance.

$$\rho = \frac{\sum_{i=1}^{n}(V_{gt_i} - \overline{V_{gt}})(V_{predicted_i} - \overline{V_{predicted}})}{\sqrt{\sum_{i=1}^{n}(V_{gt_i} - \overline{V_{gt}})^2}\sqrt{\sum_{i=1}^{n}(V_{predicted_i} - \overline{V_{predicted}})^2}} \tag{12}$$

Formula 12 represents the Pearson Correlation Coefficient, which is an indicator of the linear relationship between two variables. $\overline{V_{predicted}}$ and $\overline{V_{gt}}$ represent the predicted and actual values, respectively. It yields values between -1 and 1. A value closer to 1 indicates a positive correlation, while a value closer to -1 indicates a negative correlation between the variables. This coefficient helps in assessing the association between two variables.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{V_{gt_i} - V_{predicted_i}}{V_{gt_i}}\right| \tag{13}$$

Formula 13 represents the Mean Absolute Percentage Error (MAPE), which indicates the average of the percentage errors between predicted values and actual values. It calculates the percentage error to intuitively assess the model's prediction accuracy, making it useful for relative performance evaluation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|V_{gt_i} - V_{predicted_i}| \tag{14}$$

Formula 14 represents the Mean Absolute Error (MAE), which signifies the average of the absolute value of errors between predicted values and actual values. Since it uses absolute values, it evaluates prediction accuracy based solely on the magnitude of errors, exhibiting a characteristic of being less sensitive to outliers.

$$Mean = \frac{1}{n}\sum_{i=1}^{n} Val_i \tag{15}$$

Formula 15 represents a metric used to calculate the mean of a dataset. The variable Val can interchangeably represent either $V_{Predicted}$ or V_{gt}. It is employed to understand the central tendency of the data and compute representative values, showcasing where the overall data is centered.

$$std = \sqrt{\frac{\sum_{i=1}^{N}(Val_i - \overline{Val})^2}{N}} \tag{16}$$

Formula 16 represents the standard deviation (std), which indicates how far the data is spread out from the mean. $\overline{Val}$ can interchangeably represent either $\overline{V_{predicted}}$ or $\overline{V_{gt}}$. It helps in understanding the dispersion of data, showing how consistent predictions are or how widely distributed the data is from the average.

$$Min = Min(Val) \tag{17}$$

Formula 17 represents a metric used to calculate the minimum value within a dataset. By determining the minimum value, it allows for an understanding of the overall range of the data. It is effective for identifying outliers and assessing whether data points fall outside the typical range.

$$Max = Max(Val) \tag{18}$$

Formula 18 represents a metric used to calculate the maximum value within a dataset. Determining the maximum value helps understand the overall range of the data. It is effective for identifying outliers and assessing whether data points fall outside the typical range.

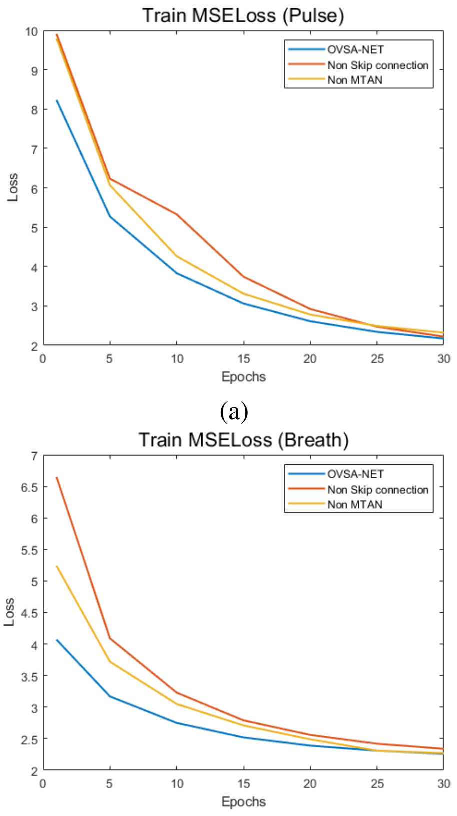
$$Error\ Rate = \frac{|V_{gt} - V_{predicted}|}{V_{gt}} \tag{19}$$

Formula 19 represents an evaluation metric indicating how much the model's predictions deviate from the actual results. A low error rate signifies the accuracy of the model, while a high error rate indicates its inaccuracy. This metric provides an intuitive understanding of the model's prediction accuracy and allows for easy comparison of model performance using relative values.

V. EXPERIMENTAL RESULTS

A. MODELS TRAINING RESULTS

Figure 7 compares the convergence speeds by evaluating the loss using Formula 10, which utilizes Mean Squared Error (MSE) as the objective function, across 30 epochs



(a)

(b)

FIGURE 7. Comparing train loss of pulse and breath for each of the our models.

TABLE 3. Comparison of PR Values among PPG sensor measured ground truth, proposed method (Ours), models without skip connection, and models without MTAN.

Video Num	Method	Mean PR (std)	Min PR	Max PR	Error Rate
1	PPG sensor (ground truth)	110.12 (7.71)	67	139	-
	Ours	106.3 (29.5)	49	180	3.4%
	Ours (Non skip connection)	96.7 (26.6)	53	180	12.1%
	Ours (Non MTAN)	99.6 (27.1)	41	180	9.4%
2	PPG sensor (ground truth)	110.5 (10.18)	41	150	-
	Ours	105.0 (30.9)	50	180	4.93%
	Ours (Non skip connection)	96.2 (29.0)	43	180	12.8%
	Ours (Non MTAN)	100.7 (31.1)	45	180	8.8%
3	PPG sensor (ground truth)	93.1 (10.2)	49	164	-
	Ours	96.9 (32.5)	40	180	4.0%
	Ours (Non skip connection)	101.2 (28.6)	47	180	8.7%
	Ours (Non MTAN)	98.4 (30.2)	45	180	5.6%
4	PPG sensor (ground truth)	96.9 (7.9)	53	129	-
	Ours	95.3 (31.7)	45	180	1.5%
	Ours (Non skip connection)	85.5 (29.1)	45	180	11.7%
	Ours (Non MTAN)	85.7 (31.6)	37	180	11.5%
5	PPG sensor (ground truth)	101.8 (11.6)	53	150	-
	Ours	97.3 (30.6)	41	180	4.4%
	Ours (Non skip connection)	93.2 (32.6)	48	180	8.4%
	Ours (Non MTAN)	92.9 (30.5)	44	180	8.7%
6	PPG sensor (ground truth)	97.9 (8.4)	67	129	-
	Ours	95.7 (30.5)	48	180	2.2%
	Ours (Non skip connection)	109.7 (30.3)	52	180	12.0%
	Ours (Non MTAN)	110.0 (31.6)	50	180	12.3%

for Our (OVSA-Net), Our (non-skip connection), and Our (non-MTAN) models. These comparisons were made using the dataset constructed within this paper for training. Since all three models output two metrics (pulse, breath), the loss values for each metric were compared. Upon observing both Figure 7-(a) and (b), models without skip connections exhibit slower convergence rates compared to those with skip connections. This observation confirms that skip connections enhance convergence speed. Additionally, Our OVSA-Net, which incorporates both skip connections and MTAN, starts with lower loss values and demonstrates rapid convergence. This indicates that using skip connections in conjunction with MTAN leads to higher convergence rates compared to using them individually.

B. RESULTS OF OUR MODEL'S ABLATION EXPERIMENT

Table 3 compares the ground truth values measured by PPG sensors with the predicted values of three deep learning models—our proposed model, the model without skip connections, and the model without MTAN—using a random selection of six videos from the validation data constructed within this paper. Mean PR represents the average of the ground truth and predicted PR values, serving as the final model output. The standard deviation (std) measures the consistency of predictions across the model's data and shows how widely distributed the data is. The min and max values demonstrate the potential range of predicted PR values, indicating the overall spectrum of the data. This is beneficial for identifying outliers and assessing whether values fall outside the typical range. Error Rate is a crucial evaluation metric used to gauge the model's performance. It calculates the percentage difference between the ground truth values measured by PPG sensors and the Mean PR values predicted by the deep learning models. Lower Error Rate indicates higher prediction accuracy of the deep learning model. While the models using only skip connections or only MTAN exhibit an average Error Rate of approximately 10%, the OVSA-Net employing both skip connections and MTAN demonstrates an average Error Rate of 4%. This observation confirms that OVSA-Net, employing all modules, outperforms individual module usages in terms of outstanding performance.

Table 4 compares the ground truth values measured by the Breathing sensor with the predicted values of three deep learning models—our proposed model, the model without skip connections, and the model without MTAN—using a random selection of six videos from the validation data constructed in this paper. Mean BR represents the average of the ground truth and predicted BR values, serving as the final model output. Standard deviation (std) measures the consistency of predictions across the model's data and shows how widely distributed the data is. The min and max values demonstrate the potential range of predicted BR values, indicating the overall spectrum of the data. This aids in identifying outliers and assessing whether values fall outside the typical range. Error Rate is a critical evaluation metric used to assess the model's performance. It calculates the percentage difference between the ground truth values measured by the Breathing sensor and the Mean BR values predicted by the deep learning models. Lower Error Rate indicates higher prediction accuracy of the deep learning model. While the models using only skip connections or only MTAN exhibit an average Error Rate of around 20-30%, the OVSA-Net employing both skip connections and MTAN

TABLE 4. Comparison of BR values among breathing sensor measured ground truth, proposed method (Ours), models without skip connection, and models without MTAN.

Video Num	Method	Mean BR (std)	Min BR	Max BR	Error Rate
1	Breathing sensor (ground truth)	30 (5.4)	17	47	-
	Ours	28.3 (8.0)	13	48	5.5%
	Ours (Non skip connection)	28.1 (8.8)	12	59	6.2%
	Ours (Non MTAN)	30.7 (10.3)	7	59	2.5%
2	Breathing sensor (ground truth)	23.3 (3.4)	14	30	-
	Ours	24.3 (8.7)	10	50	3.9%
	Ours (Non skip connection)	27 (9.5)	14	53	15.4%
	Ours (Non MTAN)	30.0 (9.5)	13	52	28.4%
3	Breathing sensor (ground truth)	23.7 (3.0)	13	28	-
	Ours	22.4 (9.7)	4	32	5.7%
	Ours (Non skip connection)	20.6 (9.8)	3	40	13.1%
	Ours (Non MTAN)	23.8 (9.9)	5	52	0.4%
4	Breathing sensor (ground truth)	22.4 (3.4)	9	28	-
	Ours	21.5 (8.4)	4	36	4.0%
	Ours (Non skip connection)	20.7 (8.5)	3	26	7.4%
	Ours (Non MTAN)	8.2 (8.0)	2	20	63.2%
5	Breathing sensor (ground truth)	31.4 (4.7)	17	36	-
	Ours	29.9 (9.7)	6	43	4.5%
	Ours (Non skip connection)	23.7 (9.7)	4	30	24.3%
	Ours (Non MTAN)	21.6 (9.9)	5	36	31.1%
6	Breathing sensor (ground truth)	31.1 (1.9)	27	36	-
	Ours	32.8 (14.5)	2	57	5.4%
	Ours (Non skip connection)	25.4 (8.1)	3	24	18.1%
	Ours (Non MTAN)	26.4 (8.2)	7	36	14.9%

demonstrates an average Error Rate of 4.5%. This observation confirms that OVSA-Net, employing all modules, exhibits outstanding performance compared to individual module usages.

TABLE 5. Experimental results for PR and BR on the Our dataset.

Method	Value	RMSE	MAE	ρ	MAPE
Ours	PR	3.77	3.53	0.92	3.43
	BR	1.66	1.56	0.92	5.67
Ours (Non Skip connection)	PR	11.46	11.24	-0.13	10.98
	BR	4.46	3.93	0.47	14.14
Ours (Non MTAN)	PR	9.85	9.61	0.16	9.43
	BR	7.78	6.02	0.40	23.45

Table 5 presents the comparison results for four parameters of our proposed model, OVSA-Net, the model without skip connections, and the model without MTAN. For comparison, we utilized a dataset consisting of video data from East Asian participants, which we created ourselves. The models were trained using a set of 50 videos, each containing diverse PR and BR values necessary for learning. Subsequently, the models were evaluated on 10 test data samples. Our proposed model exhibited outstanding performance in RMSE, MAE, Pearson's correlation, and MAPE compared to the other models considered in the comparison. These results indicate that combining skip connections and MTAN in our proposed model delivers better performance than using them separately.

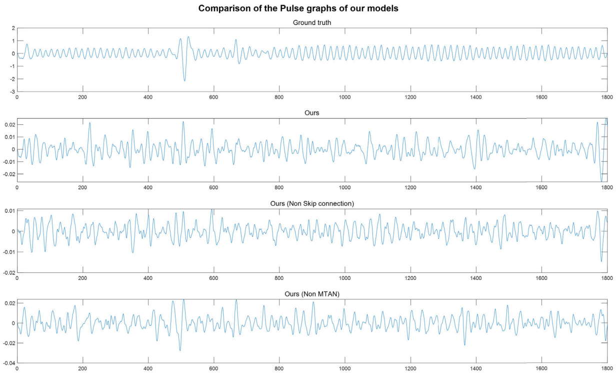


FIGURE 8. Comparison of PR graph among breathing sensor measured ground truth, proposed method (Ours), models without skip connection, and models without MTAN.

Figure 8 shows the comparison among the ground truth values measured by the PPG sensor and the predictions of three different deep learning models: our proposed model, the model without skip connections, and the model without MTAN, using a randomly selected video from the validation dataset constructed in this paper. Comparing the ground truth values with the predictions of the three models, it is noticeable that they all exhibit relatively similar graph shapes. However, when observing the x-axis peak positions, the proposed OVSA-Net utilizing both skip connection and MTAN demonstrates peak positions closer to the x-axis peaks of the ground truth compared to the other models. This indicates the higher predictive accuracy of our proposed model.

Figure 9 shows a graph comparing the ground truth values measured by the Breathing sensor with predictions made by three different deep learning models: our proposed model, the model without skip connections, and the model without MTAN, using a randomly selected video from the validation dataset constructed in this paper. When comparing the ground truth values with the three models, it is evident that the models

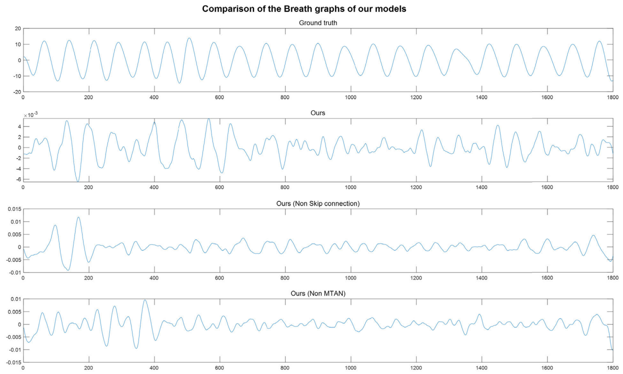


FIGURE 9. Comparison of BR values among breathing sensor measured ground truth, proposed method (Ours), models without skip connection, and models without MTAN.

solely utilizing skip connection or MTAN demonstrate unsatisfactory results. There is a noticeable discrepancy in the x-axis positions of the peaks when compared to the ground truth. Additionally, the gentle slopes of the peaks might hinder peak detection, potentially impacting accuracy. In contrast, the OVSA-Net, which employs both skip connection and MTAN, shows peak positions on the x-axis similar to the ground truth, and exhibits more pronounced peaks. This suggests that our proposed OVSA-Net model resembles the ground truth in graph shape and demonstrates higher predictive accuracy.

C. RESULTS OF COMPARATIVE EXPERIMENTS WITH OTHER MODELS

Table 6 shows a comparison of four parameters among five vital signs acquisition models, including our proposed model, OVSA-Net. The comparison was conducted using a dataset comprising videos of East Asian participants, specifically collected for this study. The models were trained using 50 videos containing various PR and BR values necessary for learning. Following the training phase, the models were evaluated using 10 test data samples. Our proposed model, OVSA-Net, exhibited outstanding performance in terms of RMSE, MAE, Pearson's correlation, and MAPE compared to the other models examined in this comparison.

TABLE 6. Experimental results for PR and BR on the Our dataset.

Method	Value	RMSE	MAE	ρ	MAPE
MTTS-CAN[23]	PR BR	11.20 4.85	9.16 3.22	-0.05 0.53	9.43 11.45
POS[9]	PR BR	31.90 -	29.69 -	-0.30 -	28.73 -
PBV[24]	PR BR	12.86 -	10.74 -	0.31 -	10.38 -
FlowNet2.0[61]	PR BR	- 9.44	- 9.13	- 0.82	- 33.39
Ours	PR BR	3.77 1.66	3.53 1.56	0.92 0.92	3.43 5.67

D. RELATIONSHIP BETWEEN PR AND SI

Table 7 represents the predicted PR and SI values for six randomly selected videos from the validation dataset. The predicted PR values for each video can be displayed as histograms similar to Figure 4-(b), enabling the calculation of Bavesky's Stress Index (SI) using formula 9. Through this table, it is observable that as PR increases, SI also increases, while a decrease in PR leads to a decrease in SI. This implies that SI is influenced by PR values.

TABLE 7. Stress index(SI) measured based on the PR value.

Video Num	PR	Stress Index(SI)
1	82	9
2	93	11
3	92	11
4	75	8
5	108	14
6	102	13

VI. DISCUSSION

A. THE PERFORMANCE OF THE PROPOSED PROCESS

In this paper, a performance comparison was conducted regarding the preprocessing and CNN model proposed in the process. Table 3 and 4 clearly shows the difference between employing skip connection and MTAN separately and using both through ablation experiments on our proposed model. The error rates for PR and BR range from 3.4% to 4.9% and 3.9% to 5.7%, respectively, when both skip connection and MTAN are applied in OVSA-Net. These error rates signify a significant enhancement in prediction accuracy. Figure 8 visually depicts the model's performance as part of the ablation experiments. Although they possess similar waveforms, OVSA-Net exhibits fewer errors in peak locations (x-axis) compared to other models concerning the ground truth. Figure 9 clearly shows the superiority of OVSA-Net's performance. The peak locations (x-axis) and the shape of the graph closely resemble the ground truth, showcasing the robustness and high accuracy achieved by using skip connection and MTAN together, particularly in mitigating noise introduced when using TSM. However, we acknowledge that while the peak locations (y-axis) of PR and BR resemble the ground truth, they do not achieve a notably high level of accuracy. Additionally, intermittent abnormal peaks are observed, emphasizing the need for a more sophisticated preprocessing stage and stronger weights in the training process. This indicates a necessity for further research in the future. Table 6 compares our proposed OVSA-Net with similar vital signs acquisition models using the dataset developed in this paper. Through this comparison, it becomes evident that our proposed OVSA-Net exhibits outstanding performance when compared with other models in terms of the four parameters (RMSE, MAE, Pearson's correlation, and MAPE).

B. COMPARISON WITH PRIOR STUDY

Previous studies have pursued vital-sign acquisition by either integrating spatiotemporal information or handling temporal and spatial information separately. In the context of spatial information, researchers have primarily focused on designating ROI on the cheek and forehead, areas where PPG signals can be effectively measured [60]. While spatial information can be beneficial, it often fails to fully exploit temporal information, resulting in relatively high error rates. Furthermore, addressing noise-related challenges when using temporal information has been an ongoing concern. When spatiotemporal information is used, it may enhance accuracy; however, this approach often comes at the cost of increased computational speed and computational load compared to separately handling time and space. The rPPG-based vital-sign acquisition process adopted in our study, which employs the CNN-based OVSA-Net, follows a distinct path. It crops only the face portion of the image and uses it as input, eliminating the need for separate ROI classification. Furthermore, it exclusively uses temporal information and achieves a low MAPE without introducing additional computational overhead typical in conventional 2D-CNN models. This underscores the ability to achieve high accuracy through the combined use of the attention mask and skip connection while focusing solely on temporal information.

C. CONSTRUCTING THE DATASET

Research findings have demonstrated that skin histograms exhibit variations among different racial groups [56]. This phenomenon implies that a model trained using data from a specific racial group may experience increased error rates when applied to a different racial group. Notably, many of the existing vital-sign datasets predominantly feature data from individuals of Western descent [57], [58], [59]. Consequently, training models on datasets specifically tailored to the East Asian population has proven to be a pivotal step in achieving high accuracy. OVSA-Net is structured to accommodate the training of raw data obtained from measurement equipment, simplifying the training process by eliminating the need for additional preprocessing steps.

VII. CONCLUSION

We present a novel process designed to estimate three vital signs: PR, BR, and SI using input video data. Our proposed process, which encompasses the preprocessing of the input video data and the subsequent generation of prediction values for the three vital signs, plays a crucial role in enhancing the convergence rate of model training and accuracy. A key component of our model is TSM, which efficiently incorporates temporal data into 2D-CNN without the need for additional computational resources. This feature paves the way for the development of programs and applications that can leverage real-time performance in the future. Moreover, the attention branch within the model enables the model to focus on the face region, preventing accuracy degradation caused by the spread of noise data from outside the face through TSM. The inclusion of a skip connection further improves the convergence rate by reintroducing spatial information after applying an attention mask. The combined use of the attention mask and skip connection enhances accuracy, even in the presence of noise. Additionally, the dataset we constructed in this study has the advantage of primarily targeting the East Asian population. The performance of the proposed model was significantly enhanced through training with this dataset, ultimately increasing accuracy for the East Asian demographic.

REFERENCES

[1] M. M. Rodgers, V. M. Pai, and R. S. Conroy, "Recent advances in wearable sensors for health monitoring," *IEEE Sensors J.*, vol. 15, no. 6, pp. 3119–3126, Jun. 2015.

[2] M. Salem, A. Elkaseer, I. A. M. El-Maddah, K. Y. Youssef, S. G. Scholz, and H. K. Mohamed, "Non-invasive data acquisition and IoT solution for human vital signs monitoring: Applications, limitations and future prospects," *Sensors*, vol. 22, no. 17, p. 6625, Sep. 2022.

[3] F. T. Z. Khanam, A. Al-Naji, and J. Chahl, "Remote monitoring of vital signs in diverse non-clinical and clinical scenarios using computer vision systems: A review," *Appl. Sci.*, vol. 9, no. 20, p. 4474, Oct. 2019.

[4] W. Q. Mok, W. Wang, and S. Y. Liaw, "Vital signs monitoring to detect patient deterioration: An integrative literature review," *Int. J. Nursing Pract.*, vol. 21, no. 2, pp. 91–98, May 2015.

[5] M. J. Rantz, M. Skubic, M. Popescu, C. Galambos, R. J. Koopman, G. L. Alexander, L. J. Phillips, K. Musterman, J. Back, and S. J. Miller, "A new paradigm of technology-enabled 'vital signs' for early detection of health change for older adults," *Gerontology*, vol. 61, no. 3, pp. 281–290, 2014.

[6] R. M. Baevsky and A. G. Chernikova, "Heart rate variability analysis: Physiological foundations and main methods," Cardiometry, Russian New Univ., Tech. Rep. 10, May 2017, pp. 66–76, doi: 10.12710/cardiometry.2017.6676.

[7] D. T. Weiler, S. O. Villajuan, L. Edkins, S. Cleary, and J. J. Saleem, "Wearable heart rate monitor technology accuracy in research: A comparative study between PPG and ECG technology," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, Sep. 2017, vol. 61, no. 1. Newbury Park, CA, USA: Sage, pp. 1292–1296.

[8] F. Yanowitz, P. Kinias, D. Rawling, and H. A. Fozzard, "Accuracy of a continuous real-time ECG dysrhythmia monitoring system," *Circulation*, vol. 50, no. 1, pp. 65–72, Jul. 1974.

[9] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017.

[10] D. Qiao, F. Zulkernine, R. Masroor, R. Rasool, and N. Jaffar, "Measuring heart rate and heart rate variability with smartphone camera," in *Proc. 22nd IEEE Int. Conf. Mobile Data Manag. (MDM)*, Jun. 2021, pp. 248–249.

[11] A. E. Müller, R. C. Berg, P. S. J. Jardim, T. B. Johansen, and S. S. Ormstad, "Can remote patient monitoring Be the new standard in primary care of chronic diseases, post-COVID-19?" *Telemedicine e-Health*, vol. 28, no. 7, pp. 942–969, Jul. 2022.

[12] K. Lee, S. Kim, B. An, H. Seo, S. Park, and E. C. Lee, "Noise-assessment-based screening method for remote photoplethysmography estimation," *Appl. Sci.*, vol. 13, no. 17, p. 9818, Aug. 2023.

[13] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, p. 113, Feb. 2018.

[14] A. Z. da Costa, H. E. H. Figueroa, and J. A. Fracarolli, "Computer vision based detection of external defects on tomatoes using deep learning," *Biosystems Eng.*, vol. 190, pp. 131–144, Feb. 2020.

[15] Y. Jiang, W. Wang, and C. Zhao, "A machine vision-based realtime anomaly detection method for industrial products using deep learning," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2019, pp. 4842–4847.

[16] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Advances in Computer Vision*. Berlin, Germany: Springer, 2019, pp. 128–144.

[17] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," 2019, *arXiv:1905.02419*.

[18] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 151–160.

[19] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 349–365.

[20] Z. Yang, H. Wang, and F. Lu, "Assessment of deep learning-based heart rate estimation using remote photoplethysmography under different illuminations," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 6, pp. 1236–1246, Dec. 2022.

[21] S. Chen, "Deep learning-based image enhancement for robust remote photoplethysmography in various illumination scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2023, pp. 6077–6085.

[22] Z.-K. Wang, Y. Kao, and C.-T. Hsu, "Vision-based heart rate estimation via a two-stream CNN," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3327–3331.

[23] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Proc. NIPS*, 2020, pp. 19400–19411.

[24] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Opt. Exp.*, vol. 16, no. 26, p. 21434, Dec. 2008.

[25] L. Xi, W. Chen, C. Zhao, X. Wu, and J. Wang, "Image enhancement for remote photoplethysmography in a low-light environment," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 1–7.

[26] Z. Yue, S. Ding, S. Yang, H. Yang, Z. Li, Y. Zhang, and Y. Li, "Deep super-resolution network for rPPG information recovery and noncontact heart rate estimation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[27] Y. Ba, Z. Wang, K. D. Karinca, O. D. Bozkurt, and A. Kadambi, "Style transfer with bio-realistic appearance manipulation for skin-tone inclusive rPPG," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, Aug. 2022, pp. 1–12.

[28] B.-F. Wu, P.-W. Huang, D.-H. He, C.-H. Lin, and K.-H. Chen, "Remote photoplethysmography enhancement with machine leaning methods," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 2466–2471.

[29] L. Q. Abdulrahaman, "Two-stage motion artifact reduction algorithm for rPPG signals obtained from facial video recordings," *Arabian J. Sci. Eng.*, Apr. 2023, doi: 10.1007/s13369-023-07845-2.

[30] R. Macwan, Y. Benezeth, and A. Mansouri, "Remote photoplethysmography with constrained ICA using periodicity and chrominance constraints," *Biomed. Eng. OnLine*, vol. 17, no. 1, pp. 1–22, Dec. 2018.

[31] H. Xiao, W. Zhang, K. Chen, X. Zhu, Y. Li, D. Zhang, and D. Liu, "Facial video heart rate detection based on fast-ICA and LA-Res2Net," *Proc. SPIE*, vol. 12707, pp. 460–467, Jun. 2023.

[32] A. H. M. Z. Karim, M. S. Miah, G. R. A. Jamal, R. A. Fahima, and M. T. Rahman, "Application of chrominance based rPPG in estimation of heart rate from video signal," in *Proc. 24th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2021, pp. 1–6.

[33] R.-Y. Huang and L.-R. Dung, "A motion-robust contactless photoplethysmography using chrominance and adaptive filtering," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2015, pp. 1–4.

[34] F. Schrumpf, P. Frenzel, C. Aust, G. Osterhoff, and M. Fuchs, "Assessment of non-invasive blood pressure prediction from PPG and rPPG signals using deep learning," *Sensors*, vol. 21, no. 18, p. 6022, Sep. 2021.

[35] G. B. Papini, P. Fonseca, M. M. van Gilst, J. W. M. Bergmans, R. Vullings, and S. Overeem, "Wearable monitoring of sleep-disordered breathing: Estimation of the apnea–hypopnea index using wrist-worn reflective photoplethysmography," *Sci. Rep.*, vol. 10, no. 1, p. 13512, Aug. 2020.

[36] N. N. Sahoo, B. Murugesan, A. Das, S. Karthik, K. Ram, S. Leonhardt, J. Joseph, and M. Sivaprakasam, "Deep learning based non-contact physiological monitoring in neonatal intensive care unit," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 1327–1330.

[37] P. Arena, A. Basile, M. Bucolo, and L. Fortuna, "Image processing for medical diagnosis using CNN," *Nucl. Instrum. Methods Phys. Res. Sect. A, Accel., Spectrometers, Detectors Associated Equip.*, vol. 497, no. 1, pp. 174–178, Jan-2003.

[38] M. A. Parab and N. D. Mehendale, "Red blood cell classification using image processing and CNN," *Social Netw. Comput. Sci.*, vol. 2, no. 2, p. 70, Apr. 2021.

[39] A. Ramdani, A. Virgono, and C. Setianingsih, "Food detection with image processing using convolutional neural network (CNN) method," in *Proc. IEEE Int. Conf. Ind. 4.0, Artif. Intell., Commun. Technol. (IAICT)*, Jul. 2020, pp. 91–96.

[40] N. Tong, Y. Tang, B. Chen, and L. Xiong, "Representation learning using attention network and CNN for heterogeneous networks," *Exp. Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115628.

[41] N. Wang, M. Chen, and K. P. Subbalakshmi, "Explainable CNN-attention networks (C-attention network) for automated detection of Alzheimer's disease," 2020, *arXiv:2006.14135*.

[42] Y. Geng, Z. Yu, Y. Long, L. Qin, Z. Chen, Y. Li, X. Guo, and G. Li, "A CNN-attention network for continuous estimation of finger kinematics from surface electromyography," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6297–6304, Jul. 2022.

[43] W. Li, K. Liu, L. Zhang, and F. Cheng, "Object detection based on an adaptive attention mechanism," *Sci. Rep.*, vol. 10, no. 1, p. 11307, Jul. 2020.

[44] K. Hara, M.-Y. Liu, O. Tuzel, and A.-M. Farahmand, "Attentional network for visual object detection," 2017, *arXiv:1702.01478*.

[45] Z. Wu, B. Hou, and L. Jiao, "Multiscale CNN with autoencoder regularization joint contextual attention network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1200–1213, Feb. 2021.

[46] S. Liu, E. Johns, and A. J. Davison, "End-To-End multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1871–1880.

[47] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3D convolutional neural network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3054–3062.

[48] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3D convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sens.*, vol. 10, no. 2, p. 75, Jan. 2018.

[49] J. Fu, Y. Yang, K. Singhrao, D. Ruan, F. Chu, D. A. Low, and J. H. Lewis, "Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging," *Med. Phys.*, vol. 46, no. 9, pp. 3788–3798, Sep. 2019.

[50] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7083–7093.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[52] J. Yu, B. Yang, J. Wang, J. Leader, D. Wilson, and J. Pu, "2D CNN versus 3D CNN for false-positive reduction in lung cancer screening," *J. Med. Imag.*, vol. 7, no. 5, Oct. 2020, Art. no. SP-051202.

[53] R. Sun, "Optimization for deep learning: Theory and algorithms," 2019, *arXiv:1912.08957*.

[54] G. Mazaheri, N. C. Mithun, J. H. Bappy, and A. K. Roy-Chowdhury, "A skip connection architecture for localization of image manipulations," in *Proc. CVPR Workshops*, 2019, pp. 119–129.

[55] X. Zou, L. Yang, D. Liu, and Y. Jae Lee, "Progressive temporal feature alignment network for video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16443–16452.

[56] D. McDuff, "Camera measurement of physiological vital signs," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–40, Sep. 2023.

[57] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognit. Lett.*, vol. 124, pp. 82–90, Jun. 2019.

[58] G. Biagetti, P. Crippa, L. Falaschetti, L. Saraceni, A. Tiranti, and C. Turchetti, "Dataset from PPG wireless sensor for activity monitoring," *Data Brief*, vol. 29, Apr. 2020, Art. no. 105044.

[59] G. Heusch, S. Marcel, and A. Anjos, "A reproducible study on remote heart rate measurement," Idiap Res. Inst., Tech. Rep., Dec. 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1709.00962

[60] S. Kwon, J. Kim, D. Lee, and K. Park, "ROI analysis for remote photoplethysmography on facial video," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 4938–4941.

[61] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

SEONGCHAN PARK received the B.S. degree in computer science from Kwangwoon University, Seoul, South Korea, where he is currently pursuing the M.S. degree. His research interests include object tracking and recognition, and processing systems using artificial intelligence based on medical images in the medical field.

HEEJUN YOUN received the B.S. degree in computer science from Kwangwoon University, where he is currently pursuing the M.S. degree. His research interests include artificial intelligence-based biosignal processing and medical image processing.

SEUNGHYUN LEE received the B.E. degree from the Department of Applied Electronic Engineering, Kwangwoon University, in 1984, and the M.E. and Ph.D. degrees from the Department of Electronic Engineering, Kwangwoon University, in 1986 and 1993, respectively. His research interests include holography and immersive content.

SOONCHUL KWON received the bachelor's degree in industrial engineering and the master's and Ph.D. degrees in computer vision. He was a Mobile Core Network Engineer with LG Telecom, South Korea, from 2005 to 2008, and a Network Optimization Researcher with IST International Inc., USA, from 2009 to 2010. He is currently an Associate Professor with the Department of Information Systems and an Advisor with the Spatial Computing Laboratory, Kwangwoon University. His research interests include intelligent image processing based on cloud edge computing. In 2017, he received a commendation from the President of Korea in recognition of his outstanding achievements in academic-industrial collaboration.