International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 3
ISSN: 2088-5334

Remote Heart Rate Estimation Using Attention-targeted Self-Supervised Learning Methods

Jaechoon Jo [a], Yeo-chan Yoon [b,*]

[a] *Department of Computer Education, Jeju National University, 63243 Jeju, Republic of Korea*
[b] *Department of Artificial Intelligence, Jeju National University, 63243 Jeju, Republic of Korea*
Corresponding author: [*]ycyoon@jejunu.ac.kr

***Abstract*—Heart rate measurement is a crucial factor for assessing the overall health status of an individual. Abnormal heart rates, whether lower or higher than baseline, can indicate potential pathological or physiological abnormalities. As a result, it is necessary to have reliable technology for monitoring heart rates in various fields, including medicine, biotechnology, and healthcare. With recent advancements in deep learning research, it is now possible to monitor heart rate conveniently and hygienically without specialized equipment, using facial video photo volume measurement. This new technology employs a deep learning-based video analysis method that requires a large data set to achieve high performance. However, collecting and labeling a vast amount of data is often impractical and costly. Therefore, researchers have been searching for alternative ways to achieve high performance with smaller datasets. This paper proposes a novel self-supervised learning approach suitable to the face video process. Our proposed method can effectively acquire a deep latent expression from a face image sequence and apply it to a target task through transfer learning. Using this method, we aim to improve the remote heart rate estimation performance in a limited-size dataset. Our proposed method is specialized for facial image sequences and focuses on the color change of the face to achieve high performance in existing attention-based deep learning models. The proposed self-supervised learning method has several advantages. First, it can learn useful features from unlabeled data, reducing the reliance on annotated datasets. Second, it can help overcome the problem of insufficient labeled data in specific domains, such as medical image analysis. Third, the proposed method can improve the performance of the target task using pre-trained models on different datasets. Finally, our approach improves the remote heart rate estimation performance by extracting useful features from facial images.**

***Keywords*— rPPG; self-supervised learning; facial video analysis.**

Manuscript received 24 Feb. 2022; revised 28 Mar. 2023; accepted 15 May 2023. Date of publication 30 Jun. 2023.

I. INTRODUCTION

The recent advancement of deep learning technology has led to the development of various healthcare applications, making it more convenient and effective. Among these applications, heart rate measurement is one of the most essential means for measuring an individual's health status. Tachycardia, with a heart rate greater than 100 bpm, and bradycardia, less than 60 bpm, are considered signs of health problems [1]. Photoplethysmography is a method of measuring heart rate by measuring blood flow in the face, and sensor-based measurement methods have been mainly used. However, it has become possible to measure the heart rate hygienically and conveniently using only a digital camera in a non-contact method using a deep learning-based facial image analysis method. This has significant implications for the medical field, as it provides an efficient way to monitor patients' heart rates without invasive techniques.

Deep learning models are powerful tools, but they require many data to train, and this data must be labeled, which can be time-consuming and expensive. To tackle this challenge, unsupervised learning methods, such as transfer and self-supervised learning, have become popular machine learning approaches. Unsupervised learning methods refer to machine learning methods using only readily available raw data without directly tagged data. Unsupervised learning can be used to learn in advance the latent representation of an image and use it to improve performance. Self-supervised learning is one of the unsupervised learning methods, and it is a method to learn latent representation by automatically and accurately generating labels without human intervention to build learning data.

Recently, many studies have been introduced to improve the performance of image analysis through self-supervised learning. Xiao et al. [2] and Dong et al. [3] introduced a self-supervised learning method that automatically builds training data by converting color images into gray images and then trains a CNN to recolor. Carl et al. [4] and Huang et al. [5] studied a self-supervised learning method that predicts the location of a patch after dividing one picture into multiple patches. Chun-Liang et al. [6] introduced a self-supervised learning approach that replaces small patches of images with other patches for image-based anomaly detection. These studies have demonstrated the potential of self-supervised learning in improving image processing performance.

Meanwhile, attention-based sequence processing methods and transformer-based algorithms [7], mainly used in natural language processing, have been applied to the visual field and show high performance [8]. The transformer-based learning algorithm splits an image into specific pieces and organizes them into sequences to apply attention-based methods and is applied to object detection tasks. Extending this to video [9], the transformer-based algorithm is also used for tasks such as motion detection. This has opened up new possibilities for using transformer-based algorithms in video processing applications.

We have proposed a new self-supervised learning method for face video processing. Our method is based on a video transformer architecture. This architecture involves dividing the face video into individual frames and then organizing these frames into sequences. The video transformer then processes these sequences using an attention-based mechanism, like the transformer algorithms used in natural language processing. The video transformer identifies important features in the sequence, allowing it to learn a robust latent representation of the video data. In order to assess the efficacy of our proposed method, we used it to build a high-performance rPPG (remote photo-plethysmography) model. rPPG is a non-invasive method of measuring heart rate using a digital camera and has numerous applications in the healthcare industry. Our rPPG model was trained using a dataset consisting of face videos and could accurately estimate individuals' heart rate in real-time. Our results show that our proposed self-supervised learning method outperformed other state-of-the-art methods, including traditional supervised learning methods, regarding accuracy and robustness.

The main advantage of our proposed approach is that it does not require huge amounts of labeled data. Labeled data is typically obtained through manual tagging by skilled personnel, which is time-consuming and expensive. Instead, our method uses unsupervised learning techniques to learn a robust latent representation of the video data, which can then be used to build accurate models for various tasks. Our method's ability to scale effectively and be cost-efficient has the potential to bring about a transformative shift in the development and deployment of healthcare applications.

In summary, our proposed self-supervised learning method for face video processing, which is based on a video transformer architecture, represents a significant breakthrough in remote photoplethysmography. Our method enables accurate and efficient heart rate measurement using only a digital camera, without invasive or cumbersome monitoring devices. We believe that our method has the potential to significantly impact the healthcare industry and improve the quality of care for patients worldwide.

II. Materials and Method

A. Related Works

This section presents recent research on remote photo-plethysmography(rPPG) and self-supervised learning approaches, which are relevant to the topic of our scientific paper. The following literature review summarizes the related works and identifies their contributions to the field.

1) Self-supervised learning

Self-supervised learning [10] is a form of unsupervised learning that utilizes abundant unlabeled data to learn valuable representations without manual annotation, which can be expensive, time-consuming, and impractical. The objective of self-supervised learning is to utilize the intrinsic structure and patterns of the input data to automatically generate supervisory signals, which are then used to train a model to learn valuable features. This technique has been extensively studied in various fields, including computer vision, natural language processing, and speech recognition, and has demonstrated encouraging results in improving the performance of various downstream tasks.

One of the critical advantages of self-supervised learning is its ability to learn transferable representations across different tasks or domains. By training a model on a significant amount of unlabeled data, the model can learn to extract high-level latent representation. This is known as transfer learning [11] and has been shown to be effective in numerous fields. For instance, transfer learning has been utilized in computer vision to improve the performance of various tasks, such as object detection, image classification, and segmentation. In natural language processing, transfer learning has been utilized to improve the performance of various language understanding tasks, including sentiment analysis, named entity recognition, and text classification.

Recently, various self-supervised learning techniques have been studied to enhance the performance of machine learning models. This paper reviews some of the recent studies in this area and highlights their contributions to the field. For example, Khare [12] utilized data from various sensors for self-supervised learning to predict human emotions. The researchers achieved this by training two networks simultaneously: a depth estimation network and a camera pose estimation network. The two networks were trained using the view synthesis task as the guiding signal. This means that the two networks were forced to learn to work together to generate realistic views of a scene from different perspectives. Once the two networks were trained, they could be used independently to estimate depth and camera pose from a single image.

Carl [4] used a convolutional neural network (CNN) to classify the relative positions of two image patches. This study investigates using spatial context as a free and abundant source of supervisory signals to train image embedding. They extracted random pairs of patches from each image in a large, unlabeled image collection and trained a convolutional neural network (CNN) to predict the position of the second patch

relative to the first. They claimed that the model must learn to identify objects and their components to succeed. They showed that the learned feature embedding, based on contextual information of an image, effectively detects visual similarity between various images. As an illustration, this representation allowed them to detect objects such as dogs and birds in the Pascal VOC 2011 object detection dataset without manual intervention. Moreover, they showcased that the learned ConvNet could be integrated into the R-CNN framework, surpassing the performance of a randomly initialized ConvNet and achieving state-of-the-art results compared to algorithms solely relying on annotations from the Pascal-provided training set.

Other studies have attempted to self-learn images by erasing, copying, or pasting patches of an image. For instance, some studies used image cropping [13] or randomly cutting out certain parts of the image [14] to increase the stability of CNNs, while others used long, thin squares cut and pasted in random colors to improve performance [15]. These studies demonstrate the effectiveness of self-supervised learning in improving the stability and performance of CNNs. Gidaris [16] proposed self-supervised learning to predict the rotation angle after rotating an image, improving object recognition technology's performance. The study proposed to learn image features by training ConvNets to recognize the 2D rotation applied to the input image. The researchers demonstrated, qualitatively and quantitatively, that this seemingly straightforward task provided a robust supervisory signal for learning semantic features.

This paper proposes a new method for analyzing and utilizing facial changes through self-supervised learning for facial expression recognition. Our method aims to learn representations that capture the intrinsic structure of facial changes, which can then be used to improve facial expression recognition performance. Our method involves training a deep learning model to predict the future frames of a facial video given a partial input sequence. By doing so, the model learns to capture the spatiotemporal dynamics of facial changes, which is essential for recognizing different facial expressions. We then transfer the learned latent representation to a heart rate recognition task, using facial regions such as the cheek and jaw as input data for robust recognition. In order to assess the efficacy of our approach, we performed experiments on several heart rate recognition datasets, including the COHFACE, UBFC1, UBFC2 datasets. Our results show that our method outperforms several state-of-the-art approaches, indicating the effectiveness and robustness of our approach. To gain a deeper understanding of the contributions of different components of our method, we perform ablation studies and study the consequences of different design options and learning strategies. Our experiments show that each component of our method contributes to the overall performance, including using facial video prediction as a pre-training task, choosing facial regions used as input to the heart rate recognition model, and using a multi-task learning approach.

2) rPPG estimation

In recent years, using digital cameras, remote photoplethysmography (rPPG) has emerged as a promising non-invasive technique for measuring physiological signals such as heart rate, respiratory rate, and blood pressure. Various machine learning-based approaches have been proposed to extract accurate and robust physiological signals from rPPG videos.

Liu [18] introduced the rPPG toolbox, which includes a multi-task temporal shift convolutional attention network (MTTS-CAN). This tool can measure heart rate and breathing rate on smartphones and other mobile devices without requiring specialized equipment. MTTS-CAN leverages temporal shift modules and an attention module to jointly remove noise to estimate pulse and respiration.

While supervised methods like MTTS-CAN require large amounts of labeled data, self-supervised methods like contrastive learning have shown promise in learning representative data features without labels. Nevertheless, current data augmentation methods used in contrastive learning are inadequate for capturing physiological signals from videos and struggle to capture delicate and periodic color/shape changes between frames. In response to this challenge, Wang [19] introduced a new self-supervised spatiotemporal learning framework specifically designed for learning rPPG representations, which leverages spatial and temporal augmentation to accurately detect subtle skin color fluctuations and periodic temporal changes in physiological signal features, respectively.

It is worth noting that transformer-based sequence analyzing approaches have achieved state-of-the-art performance in many vision-related tasks [20]-[22] and natural language processing tasks [23]-[25]. Yu [25] proposed a new framework called TransRPPG, a transformer-based algorithm that efficiently learns a representation of inherent vitality. This framework constructs multi-scale spatial-temporal maps (MSTmap) using facial skin and background regions. Subsequently, the transformer model extensively explores the global relationship within the MSTmaps to represent inherent vitality, which can be used to predict whether a 3D mask is being worn. In another work, Yu [26] proposed PhysFormer. PhysFormer is a transformer-based architecture for physiological measurement from facial videos. The framework incorporates temporal difference transformers, which improve quasi-periodic rPPG features through temporal difference-based global attention. Elaborate supervisions for PhysFormer are provided through label distribution learning and dynamic constraints in the frequency domain, effectively addressing the issue of overfitting. The important aspect of PhysFormer is that it can be trained from scratch on rPPG datasets, making it a promising transformer baseline for the rPPG community.

Kang [27] proposed a new video embedding method that creates a feature map called Multi-scale Adaptive Spatial and Temporal Map with Overlap (MAST_Mop) for each facial video sequence. This feature map contains both vital information and surrounding information as a reference, which helps to identify subtle changes occurring in an image sequence, such as variations in illumination stability. To extract heart rate (HR) information from MAST_Mop, the researchers propose a transformer-based model consisting of two distinct streams, where one stream primarily analyzes the pulse signal within the facial region, while the other stream identifies the subtle movements from the surrounding area. The model can perform adaptive noise cancellation by subtracting the two streams to enhance the HR signal.

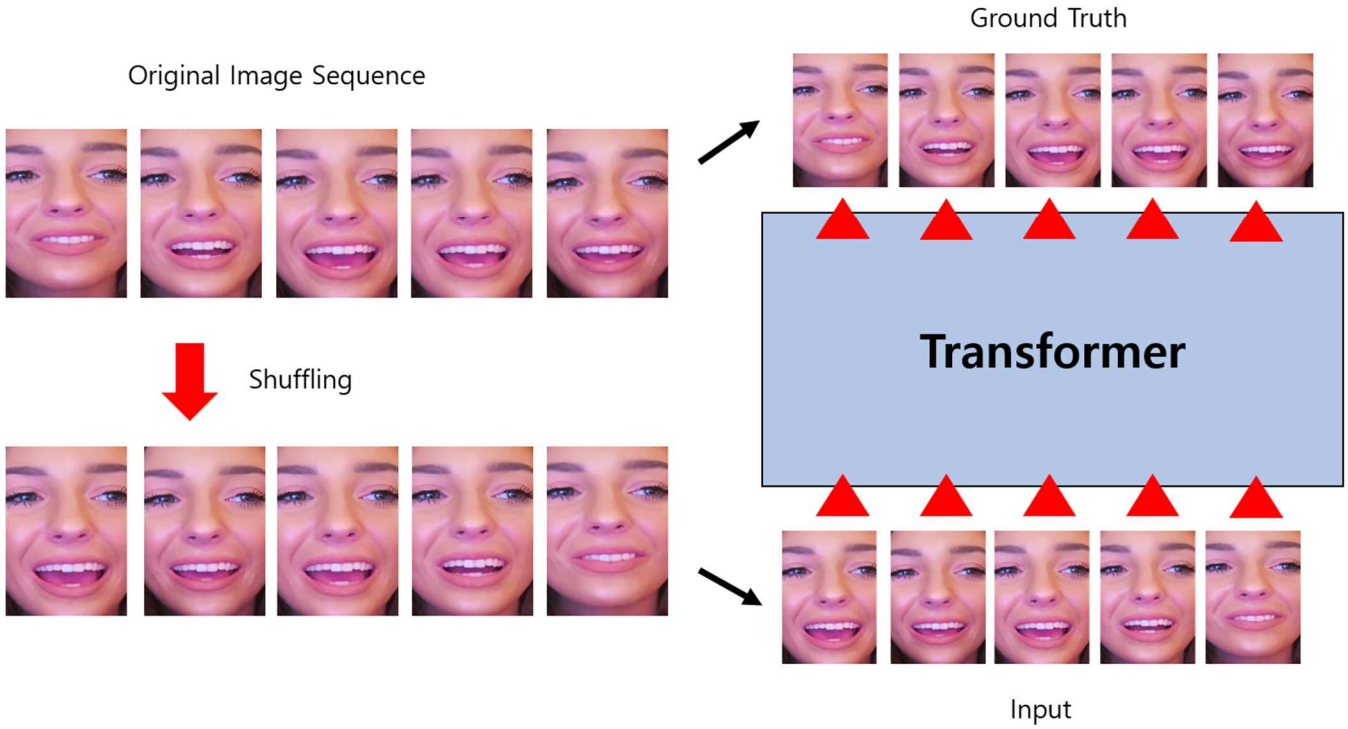

Fig. 1 Self-supervised learning for predicting the original sequence

In this paper, we introduce a region-based transformer model that builds upon the success of PhysFormer. Specifically, we fuse the latent vectors of three transformers, each trained on a different facial region (cheek, forehead, and entire face), to capture region-specific physiological signals and improve overall performance. Our experiments show that this approach yields better accuracy and robustness compared to using a single transformer on the entire face. Overall, our proposed approach contributes to the development of machine learning-based methods for extracting physiological signals from rPPG videos, which can have important applications in clinical settings and beyond.

B. Proposed Method

1) Self-supervised learning for predicting the original sequence.

The proposed method uses a self-supervised learning approach to learn deep representations of faces without requiring manual labeling. Specifically, we use a transformer-based deep learning model called Physformer as the backbone model. PhysFormer is a video architecture based on transformers that aggregate local and global spatiotemporal features adaptively to enhance the representation of rPPG[26]. To apply self-supervised learning, we first extract a sequence of face images from a video and randomly shuffle them. The shuffled sequence is then used as input to the transformer, which is trained to predict the original image sequence. The idea behind this method is that the transformer model would train to extract and encode useful representations from the face images, which can be useful for the rPPG prediction. We used publicly available video datasets [28] of subjects with various skin tones, captured under different lighting conditions and camera angles. Each video was manually inspected to remove any irrelevant frames, and then face detection and tracking algorithms were applied to extract a sequence of face images for each subject. These face images were then normalized to a consistent size and resolution to ensure consistency across the dataset.

Once the transformer is trained on the shuffled face image sequence, we use its learned parameters as the initial parameters for the transformer model used for rPPG prediction. This approach is known as transfer learning, which leverages the knowledge learned from one task to improve performance on another related task. By using the learned parameters as the initial parameters for rPPG prediction, our goal is improving the performance of the rPPG technology.

One advantage of using self-supervised learning for deep representation learning is that it can help detect subtle movements of the face as well as light reflected from the skin, which are important for accurate rPPG prediction. This is achieved through heart rate elasticity, which measures the variations in heart rate caused by movements of the face and changes in blood volume. By incorporating heart rate elasticity into the self-supervised learning approach, we aim to improve the accuracy of rPPG prediction further. Overall, the proposed method combines the strengths of self-supervised learning transfer learning, and heart rate elasticity to learn deep representations of faces without requiring any manual labeling and use these representations to improve the performance of rPPG technology. We believe that this approach has the potential to significantly advance the field of non-contact heart rate monitoring and contribute to the development of new applications in healthcare and beyond.

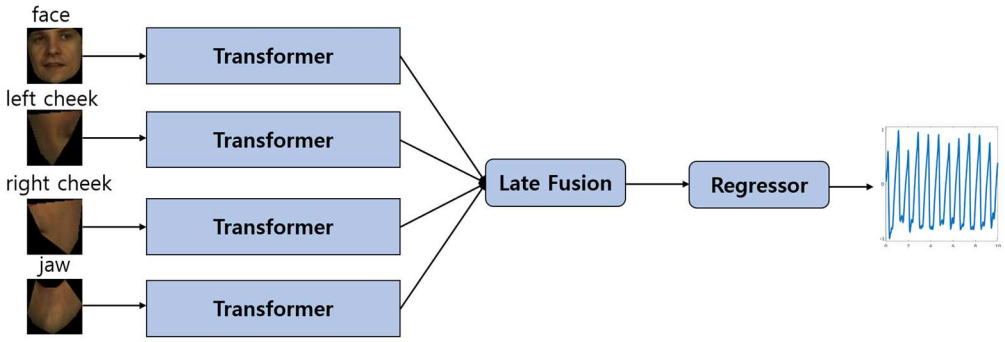

Fig. 2 Prediction of PPG with facial regions (face, left cheek, right cheek, jaw)

2) rPPG prediction with face regions

This study proposes a transformer-based architecture for predicting rPPG signals using four input channels: the entire face and three facial regions: the left cheek, right cheek, and jaw. The motivation behind considering multiple face regions is to enhance the robustness of the model against variations in illumination and facial expressions. Fig 2 shows the proposed architecture. Facial regions are chosen because they contain a dense network of blood vessels, and the pulsations in these vessels can be captured by analyzing the color changes in the skin. Multiple facial regions can provide more comprehensive information about the cardiovascular system, leading to better prediction performance.

The proposed architecture consists of a sequence of transformer decoder layers, followed by a late fusion layer and a regression layer. The transformer layers are responsible for extracting high-level features from the input signals and reconstructing the rPPG signals' temporal dynamics. The late fusion layer combines the encoded features of each channel, and the regression layer predicts the rPPG signal.

To train and evaluate the proposed model, we use a publicly available dataset containing videos of 30 subjects performing different activities, such as talking, reading, and exercising. The videos are captured using a high-resolution camera with a frame rate of 30 fps. The ground truth rPPG signals are extracted using a contactless imaging photoplethysmography technique.

III. Result and Discussion

We conducted experiments of rPPG-based physiological measurement for heart rate (HR) on three benchmark datasets, UBFC1 and 2[29], COHFACE [30]

A. Datasets and Performance Metrics

The datasets UBFC1 and UBFC2 both involve using a webcam to capture video at a resolution of 640x480 and 30 frames per second. The PPG signal is obtained through transmissive pulse oximetry at a frequency of 62 Hz. UBFC1 involves recording stationary participants, while UBFC2 involves participants playing a math game before a green screen. COHFACE is a dataset that includes facial video clips of 40 individuals captured at a resolution of 640x480 and a frame rate of 20 frames per second. PPG signals were also simultaneously acquired at a frequency of 256 Hz. Participants were instructed to remain seated and still in front of the camera during the recording, and the video clips were compressed heavily.

To train our model, we utilized the PURE dataset [31], which consists of 10 participants (8 males and 2 females), with each subject having 6 different setups of 1-minute videos totaling 60 video sequences. Each subject was recorded for a total of 6 minutes. The dataset presents a more challenging task as the videos were captured in six different settings: steady, talking, slow and fast translations, and small and medium rotations. We present the commonly employed evaluation metrics to assess the performance of our model, which includes Pearson's Correlation Coefficient (r), the Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), and

B. Experimental Results

Table 1 shows the performance of our systems and the base system Physformer[27] on three datasets – UBFC1, UBFC2, and COHFACE. The proposed systems are evaluated against the base system, which uses a single-channel face region as input. The proposed systems use face regions as multichannel inputs and incorporate self-supervised learning (SSL) for predicting the original sequence. Our results indicate that using face regions as multichannel inputs has improved the base model's performance on all three datasets. Specifically, the MAP performance on the COHFACE dataset has improved from 7.31 to 7.01, which is a statistically significant improvement. This suggests that multichannel inputs are beneficial for the rPPG prediction task.

TABLE I
EXPERIMENTAL RESULTS

Method	Dataset	MAE	RMSE	MAPE	r
Physformer	UBFC1	1.83	3.16	2.2	**0.97**
	UBFC2	1.7	3.6	1.94	0.98
	COHFACE	5.49	5.83	7.31	**1**
+Multi-channels	UBFC1	1.81	3.14	**2.08**	0.98
	UBFC2	1.63	3.62	1.91	**1**
	COHFACE	5.32	5.88	7.01	**1**
+SSL	UBFC1	**1.71**	**3.01**	2.13	0.99
	UBFC2	**1.62**	**3.58**	**1.77**	0.97
	COHFACE	**4.97**	**5.71**	**6.23**	**1**

Furthermore, incorporating self-supervised learning (SSL) has significantly improved the heart rate recognition performance of the proposed systems on all datasets except for Pearson's Correlation Coefficient. This finding indicates that SSL is an effective method for enhancing rPPG prediction performance, and it supports the notion that

unsupervised pre-training can improve supervised learning tasks. Overall, our experimental results demonstrate the efficacy of the proposed systems for rPPG prediction. These findings highlight the potential of incorporating self-supervised learning into other computer vision applications to improve their performance.

IV. Conclusion

In this paper, we introduce a new self-supervised learning method for face video processing that improves the performance of remote heart rate estimation using a limited dataset. Our proposed method specializes in facial image sequences and uses transfer learning to effectively acquire a deep latent expression from a face image sequence and apply it to a target task. We demonstrated that our proposed method outperformed existing attention-based deep learning models by focusing on the color change of the face. The experimental results showed that incorporating self-supervised learning into remote photoplethysmography can substantially enhance heart rate estimation performance, even with a limited dataset. Our proposed method achieved better performance on three publicly available datasets than the base method, confirming our approach's effectiveness.

In the future, we plan to explore various transfer learning techniques to enhance the performance of our proposed method further. We also plan to investigate the effectiveness of our method in real-world scenarios, where various factors such as lighting conditions, camera position, and facial expressions may affect heart rate estimation accuracy. Additionally, we will explore the possibility of applying our proposed method to other vital sign estimation tasks, such as blood pressure and respiratory rate. Finally, we will continue to improve our method's computational efficiency to enable real-time monitoring in practical applications.

Acknowledgment

This work was supported by the research grant of Jeju National University in 2023

References

[1] Kebe, Mamady, et al. "Human vital signs detection methods and potential using radars: A review." Sensors 20.5 (2020): 1454.

[2] Xiao, Yuxuan, et al. "Semantic-aware automatic image colorization via unpaired cycle-consistent self-supervised network." International Journal of Intelligent Systems 37.2 (2022): 1222-1238.

[3] Dong, Xuan, et al. "Self-supervised colorization towards monochrome-color camera systems using cycle CNN." IEEE Transactions on Image Processing 30 (2021): 6609-6622.

[4] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1422–1430, 2015.

[5] Huang, Ziwang, et al. "Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images." Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. Springer International Publishing, 2021.

[6] Li, Chun-Liang, et al. "Cutpaste: Self-supervised learning for anomaly detection and localization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[7] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[8] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[9] Arnab, Anurag, et al. "Vivit: A video vision transformer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[10] Schiappa, Madeline C., Yogesh S. Rawat, and Mubarak Shah. "Self-supervised learning for videos: A survey." ACM Computing Surveys (2022).

[11] Yoon, Yeo Chan. "Can We Exploit All Datasets? Multimodal Emotion Recognition Using Cross-Modal Translation." IEEE Access 10 (2022): 64516-64524.

[12] Khare, Aparna, Srinivas Parthasarathy, and Shiva Sundaram. "Self-supervised learning with cross-modal transformers for emotion recognition." 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021.

[13] Ren, Sucheng, et al. "A simple data mixing prior for improving self-supervised learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[14] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In AAAI, 2020.

[15] Bergmann, Paul, et al. "The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection." International Journal of Computer Vision 129.4 (2021): 1038-1059.

[16] Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).

[17] Ni, Aoxin, Arian Azarang, and Nasser Kehtarnavaz. "A review of deep learning-based contactless heart rate measurement methods." Sensors 21.11 (2021): 3719.

[18] Liu, Xin, et al. "Multi-task temporal shift attention networks for on-device contactless vitals measurement." Advances in Neural Information Processing Systems 33 (2020): 19400-19411.

[19] Wang, Hao, Euijoon Ahn, and Jinman Kim. "Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 2. 2022.

[20] Khan, Salman, et al. "Transformers in vision: A survey." ACM computing surveys (CSUR) 54.10s (2022): 1-41.

[21] Liu, Ze, et al. "Video swin transformer." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[22] Khan, Salman, et al. "Transformers in vision: A survey." ACM computing surveys (CSUR) 54.10s (2022): 1-41.

[23] Lund, Brady D., and Ting Wang. "Chatting about ChatGPT: how may AI and GPT impact academia and libraries?." Library Hi Tech News (2023).

[24] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." arXiv preprint arXiv:2203.02155 (2022).

[25] Trummer, Immanuel. "CodexDB: Synthesizing code for query processing from natural language instructions using GPT-3 Codex." Proceedings of the VLDB Endowment 15.11 (2022): 2921-2928.

[26] Yu, Zitong, et al. "Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection." IEEE Signal Processing Letters 28 (2021): 1290-1294.

[27] Yu, Zitong, et al. "PhysFormer: facial video-based physiological measurement with temporal difference transformer." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[28] Kollias, Dimitrios, et al. "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond." International Journal of Computer Vision 127.6-7 (2019): 907-929.

[29] Sabour, Rita Meziati, et al. "Ubfc-phys: A multimodal database for psychophysiological studies of social stress." IEEE Transactions on Affective Computing (2021).

[30] Tsou, Yun-Yun, et al. "Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos." Proceedings of the 35th annual ACM symposium on applied computing. 2020.

[31] Stricker, Ronny, Steffen Müller, and Horst-Michael Gross. "Non-contact video-based pulse rate measurement on a mobile service robot." The 23rd IEEE International Symposium on Robot and Human Interactive Communication. IEEE, 2014.