Article

A Study of Projection-based Attentive Spatial-Temporal Map for Remote Photoplethysmography Measurement

Dae-Yeol Kim 1,3, Soo-Young Cho2, Chae-Bong Sohn 3,\* and Kwangkee Lee 1,\*

|  |
| --- |
| **Citation:** Lastname, F.; Lastname, F.; Lastname, F. Title. *Biosensors* **2022**, *12*, x. https://doi.org/10.3390/xxxxx  Received: date  Accepted: date  Published: date  **Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.    **Copyright:** © 2022 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |

1 Tvstorm, Sunghyun Building, 255 Hyorung-ro, Secho-gu, Seoul 13875, Korea; wagon0004@tvstorm.com

2 Department of Information Contents, Kwangwoon University, Seoul 01897, Korea; mba4711@kw.ac.kr

3 Department of Electronics and Communications Engineering, Kwangwoon University, Seoul 01897, Korea;

**\*** Correspondence: [cbsohn@kw.ac.kr](mailto:cbsohn@kw.ac.kr), kklee@tvstorm.com

**Abstract:** The Photoplethysmography (PPG) signal contains various information related to CVD (cardiovascular disease), such as information on acute heart attack and arrhythmia, and renin-angiotensin system Variables. The remote PPG (rPPG) method is a method that can measure PPG using a face image taken with a camera without a PPG device. The rPPG method is being studied using deep learning to achieve performance close to PPG. Deep Learning based rPPG methods can be classified into three main categories. First, there is a 3D CNN approach with facial image video as input, which focuses on the spatio-temporal changes of the facial video. The second method is a method using a spatio-temporal map (STMap), and the video image was pre-processed using the point that it is easier to analyze changes in blood flow in time order. The last method used a preprocessing model using a dichromatic reflection model. This study proposes the concept of the Axis Project Network (APNet) that complements the point that the 3D CNN method requires a large memory, the STMap method requires a preprocessing method, and the dyschromatic reflection model (DRM) method does not learn long-term temporal characteristics. Finally, the proposed APNET reduced the network memory size by more than 100% and confirmed that the notch was observed in the inferred PPG, suggesting that it can give meaningful results to the study when developing the rPPG algorithm.

**Keywords:** remote Photoplethysmography, CVD, Axis Projection

1. Introduction

PPG is a method of measuring photoplethysmogram (PPG) by analyzing the physical characteristics of hemoglobin in blood generated by light transmission and reflection. By using PPG information, various vital signals such as pulse, oxygen saturation, blood pressure, and respiration rate can be acquired. Furthermore, stroke symptoms are detected through frequency analysis, renin-angiotensin-based variables, sympathetic and parasympathetic nervous system components, and respiratory arrhythmias can be detected. The disease information included according to the frequency band of the PPG is shown in Table 1 [1]. PPG is divided into an oximeter, a contact-type device that uses the tip of a finger as the measurement area, and the rPPG method, which observes changes in blood flow in the face with video data. Among the two methods, rPPG is emerging as a new method of measuring PPG because it can be easily measured with a smartphone camera and has the strength of a non-contact measurement method [2].

**Table 1.** Disease information according to frequency band.

|  |  |  |
| --- | --- | --- |
| Parameter | Frequency Band | Description |
| ULF |  | Associated with acute heart attack and arrhythmias |
| VLF |  | Variables dependent on the renin–angiotensin system |
| LF |  | Controlled by the sympathetic and parasympathetic nervous systems |
| HF |  | There is a heart rate variability related to the respiratory system, called respiratory arrhythmias |

Heart diseases such as myocardial infarction and heart failure can be detected by continuous monitoring of vital signs. In addition, according to the Early Warning Score (EWS) in Table 2, which is used in medical services to determine the disease level of a patient at an early stage, deterioration of health and cardiac arrest can be detected using vital signs [3]. rPPG is a video-based measurement method that measures changes in blood flow by analyzing the amount of light diffused reflection information generated by changes in blood flow in the face. This approach was approached by benchmarking the principle of the PPG sensor, but PPG may not be suitable for continuous monitoring in that a measuring device must be carried with you at all times. Since the rPPG method can be measured using a camera, it can be effectively measured in any space according to the spread of mobile devices and IOT devices with cameras.

**Table 2.** Early Warning Score [3].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Score | 0 | 1 | 2 | 3 |
| Heart Rate (beats per minute) | 51 – 100 | 40-50 101-110 | <40 111-130 | >130 |
| Systolic blood pressure (mmHg) | 101 – 160 | 81-100 161-200 | 70-80 >200 | <70 |
| Respiratory rate (per minute) | 9 – 14 | 15-20 | <9 21-30 | >30 |
| Temperature (degrees Celsius) | 36.1 – 37.5 | 35.0-36.0 >37.5 | <35 |  |
| Consciousness level | Alert | Responds to voice | Responds to pain | Unresponsive |

With the development of Deep Learning, rPPG models using various methods have been researched. Among the rPPG models, representative methods used as baselines include “PhysNet”, “RhytmNet”, and “Deephys” methods, and each has a different approach [4-6]. “PhysNet” proposed a method to analyze the spatio-temporal characteristics of video pixel changes with a 3D convolutional network, and “RhythmNet” proposed a method to learn the time series characteristics of PPG by introducing a preprocessing process that converts video into STMap. Finally, “DeepPhys” proposed a differential image learning method that removes specular reflection information and learns the amount of change in the PPG waveform based on the DRM model. “PhysNet” has a problem with using a lot of memory because it uses 3D convolution, and “RhythmNet” has a problem with a large proportion of preprocessing that occurs because it uses STMap. Finally, since “DeepPhys” infers two image groups, it cannot learn long-term time-series features, resulting in a problem that requires post-processing.

In this study, we propose the concept of Axis Projection Network that complements the various problems of representative models. Suggestions are as follows:

* Axial Projection:  
  A proposal for a new time series feature analysis method using axial projection. By projecting the video to various axes, the information obtained from the data in each direction is acquired.
* End-to-End rPPG inference model:  
  Analyze facial features and perform spatio-temporal feature analysis with performance similar to STMap at the network level.
* Memory saving model:  
  Minimize the parameters of the deep learning model by using only the minimum number of 3D convolutions.

**2. Related Works**

Various rPPG methods are being studied with the goal of having similar performance to contact PPG (cPPG), the result of an Oximeter. Although mathematical models such as CHROM using the existing color difference information and POS using light reflection models showed excellent performance, they showed poor performance in various bad environments depending on motion artifact by human movement and light distortion. to overcome these problems, research is underway to create an rPPG model using Deep Learning [7-8].

2.1. Deep Learing-based rPPG methods

Various Deep Learning methods are being studied for rPPG task, which can be divided into three main methods according to the data input type.

The first is a method of having differential information of an image or video as a model’s input. “HR Track” analyzes the difference image spatially using 3D CNN as video information, and “DeepPhys”, a 2D CNN model that uses a single difference image as input. In particular, the “DeepPhys” is modeled based on the DRM, and it consists of a motion model that extracts features from the degree of movement between two frames from a face video, and an appearance model that extracts facial features through normalization values between two frames [6,9]. Derived from this method, a “High-order” study was conducted [10]. This study suggested obtaining first derivative PPG (VPG) and second derivative PPG(APG), which can provide additional information such as blood pressure estimation and heart rate fibrillation from facial images [11].

The second method is to have the video as an input without preprocessing. “PhysNet” devised a 3D CNN-based network to learn the spatio-temporal features of the face and measure the PPG signal. It was proposed to use the Neg-Pearson loss, and through this, improved experimental results were derived. Afterwards, the performance was improved through heart rate inference using the Spatial Temporal Efficient Network (STVEN) using the modified 3D CNN [4,12].

The last method is to use the STMap format, which preprocess the video in 2D image, as an input. “RhythmNet” proposed a preprocessing method using STMap. PPG is measured through 2D CNN by converting video into STMap showing spatio-temporal features [5].

2.2. Transformer network in rPPG tasks

The transformer module was first proposed in the field of NLP and showed good results. Recently, as a proposal of Vision Transformer (ViT), it is widely used as a method for image learning in vision tasks. An image patch-based inference method was proposed for image classification, and various variants of ViT were proposed to solve the characteristic that ViT lacks inductive bias [13-16]. Various ViTs showed better performance than existing CNNs in image analysis and recently introduced Vits for motion recognition, detection, and super resolution. In the rPPG task, there are related studies using ViT for rPPG inference, such as “Transrppg”, “EfficentPhys”, and “PhysFormer” [17-19]. “Transrppg” proposed a method of extracting rPPG from the space preprocessed through Vit for face attack detection in rPPG. “EfficientPhys” extracted spatial-temporal features using the Swin layer, and “Physformer” proposed a method for measuring tube Tokenizer-based long-distance spatial-temporal rPPG in face video and showed good results.

**3. Proposed Method**

In this study, we propose the concept of APNET that does not require any other preprocessing and analyzes video features from various directions using axial projection.

Graphical user interface

Description automatically generated

**Figure 1.** APNet Architecture

Figure 1 is a APNET architecture. APNET is composed of Axis Feature Extractor, Feature mixer, and PPG Block, which are feature extractors for each axis that have the same shape. The transformerlearns global features but lacks inductive bias and tends to overfit the train set. CNNs learn inductive bias but do not learn global features. We used Maxvit because it has the characteristic of learning local features at the same time as learning global features [20]. Specifically, Axis Feature Extractor is designed to extract features that are specialized in one axis. Given an RGB face video input, each feature is extracted through . The output of the Feature Extractor is combined through the Feature Mixer, and at this time, it is designed to create the optimal feature through various calculations. The optimal feature is transferred to a one-dimensional PPG block for generating PPG. PPG Block functions to decode the processed feature into PPG of T length.

3.1 Axis Feature Extractor: axis projection-based time-series feature extractor module

Axis Feature Extractor is designed to have the same model structure with only different input types for each axis. Stream is composed of 2D feature extraction block, ST (Spatial Temporal) translator, and ST feature extraction block. The input of the model proceeds in the form of a three-dimensional video arrangement, but each input is reconstructed in a two-dimensional form and used for the input of the 2D feature extraction block.

|  |  |
| --- | --- |
|  | (1) |

2D Feature Block consists of two Conv2d operations using kernel size (3x3) and stride size 2.

|  |  |
| --- | --- |
|  | (2) |

ST translator functions to change 2D image into STMap format.

|  |  |
| --- | --- |
|  | (3) |

ST Feature extraction block is composed of kernel (1, T), Maxvit layer with dilation (1, T), Spatial Attention module, and 3D translator Cascade. ST Feature extraction block gives the effect of learning spatio-temporal features using STMap form data.

|  |  |
| --- | --- |
|  | (4) |

3.2 Feature Mixer: spatio-temporal feature combination module

Graphical user interface, diagram

Description automatically generatedEach of the three features extracted from Axis Feature Extractor has specific features for the time axis, W axis, and H axis. It is difficult to expect high accuracy when learning PPG using only each feature, but high accuracy can be expected by combining the facial characteristics extracted from the time axis, the PPG characteristics of the W axis, and the PPG characteristics of the H axis. Interpolation is required because the results extracted from the stream are processed according to the input size of each axis. After interpolation, the processed feature map was processed, which is an Attentive 3D feature including PPG information.

**Figure 2.** Type of Feature mixer

Figure 2 is Type of Feature Mixer. Attentive feature is divided into 10 cases according to the processing method.

3.3 PPG Decoder: Generate PPG Signal from attentive feature

The attentivefeature requires a 3Dconvolution operation to extract time series information in a three-dimensional form. However, it was changed to STMap and then Vit was applied to achieve a similarperformance to 3D convolution. To learn the features of all areas, Max Vit was selected and used, and the last feature to extract the PPG was generated by spatially averaging, and the result was obtained through the Residual operation of Conv1d.

3.4 Loss function

The PPG signal contains various physical information. In order to simply infer the heart rate, information extraction is possible only with neg-Pearson.

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

Where is the ground truth PPG signals, and is the inference value of APNet. A and P refer to the amplitude and phase of the PPG signal. FFT denotes a fast Fourier transform. The FFT loss has the effect of learning a fine waveform by comparing the size of each frequency band [21].

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

where is the length of the signal. The neg-Pearson loss has the effect of learning the trend similarity and peak position. In previous studies, neg-Pearson was used to target the heart rate measurement using the trend and peak of PPG. In this study, neg-Pearson x FFT Loss was used as the loss, paying attention to the fact that the peak and notch of PPG are needed to extract various information.

Chart, line chart

Description automatically generated

**Figure 3.** waveform of PPG

Figure 3 shows the waveform of PPG. There are three important points in PPG: Systolic Peak, Diastolic Peak, and Notch, and the notch is an important point for inferring blood pressure.

**4. Experiment**

4.1 Dataset

The V4V dataset provides training data consisting of 82 female participants and 58 male participants [22]. Each participant performs 10 tasks, including smiling and squinting. The ground truth is provided with PPG information, heart rate, and respiration rate sampled at 1000 kHz. In the actual challenge, the model trained with the training data was evaluated as a test set, but the data set was composed by dividing the train set in a ratio of 80:20 because the data set did not contain label information for the test set.

4.2 Assessment Metric of Proposed Method

4.2.1 Quantitative Assessment Methods

* Pearson-Correlation (PCC; R) : PCC is a method of interpreting the linear relationship between two given signals. The closer the PCC result is to 1, the more positive the linear relationship.
* HR-Mean Average Error (MAE): HR-MAE is used to verify the accuracy of HR reprocessed with rPPG results.

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

* HR-Root Mean Square Error: Used to see the standard mean error of HR.

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

* Network Memory: The larger the network memory, the more unusable in a low-spec environment. The smaller the memory, the better the performance of other evaluation indicators.

4.2.2 Subjective Assessment Methods

* Roi Assessment: The importance of each pixel of the input image is determined using a Backpropagation based Method (BBM) [23].

A picture containing chart

Description automatically generated

**Figure 4.** PPG extraction accuracy evaluation according to face area. (Top-5: Yellow, Bot-5: Blue)

Figure 4 is an image of PPG extraction accuracy evaluation according to face area. according to [24], Upper Medial Forehead, Lower Medial Forehead, Glabella, Right Malar, Left Malar pixel information is important.

4.3 Experiment result

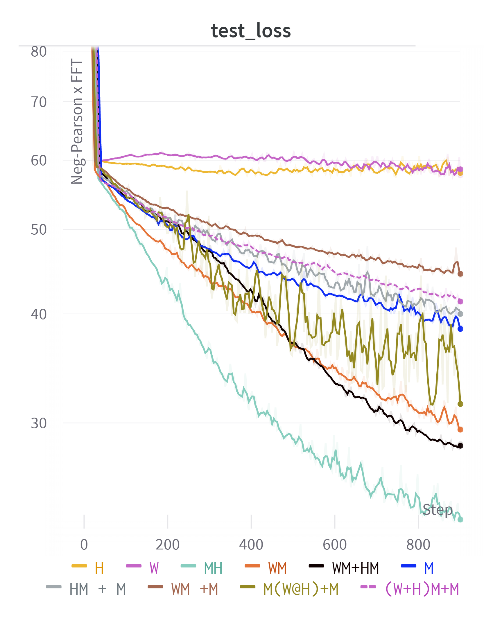
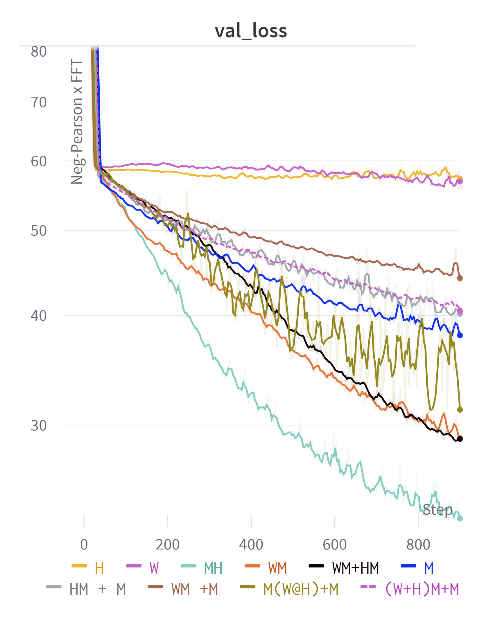
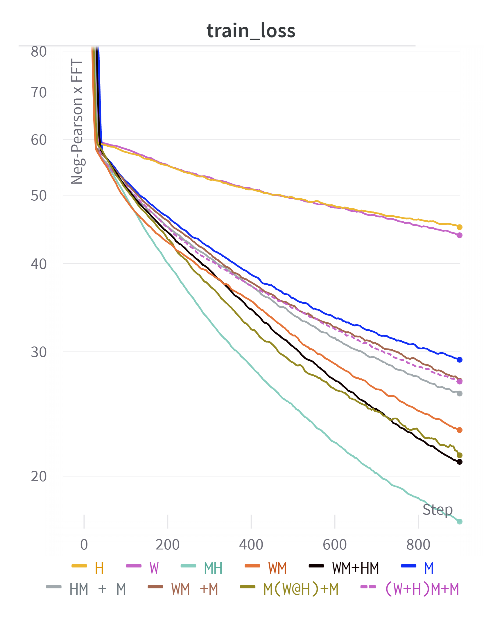
4.3.1 Quantitative assessment results

Using the three-evaluation metrics mentioned in Section 4.2, we evaluated 10 feature mixer types and two loss functions.

**Table 3.** This is a table. Tables should be placed in the main text near to the first time they are cited.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature Mixer** | **HR-MAE** | **HR-RMSE** | **R** | **HR-MAE** | **HR-RMSE** | **R** |
| **Neg-Pearson Loss** | | | **Neg-Pearson Loss x FFT Loss** | | |  | entry 1  data  data |
| M(W+H)+M | 80.64 | 82.65 | 0.12 | 81.04 | 83.05 | 0.00 |
| M(WxH)+M | 80.87 | 82.83 | 0.37 | 10.63 | 14.01 | 0.24 |
| WM+M | 80.16 | 82.10 | 0.48 | 8.49 | 11.74 | 0.53 |
| HM+M | 80.20 | 82.18 | 0.46 | 8.27 | 11.35 | 0.57 |
| M | 80.53 | 82.50 | 0.41 | 8.72 | 11.67 | 0.53 |
| WM+HM | 80.27 | 82.30 | 0.1 | **5.37** | 8.48 | **0.67** |
| WM | 80.17 | 82.17 | 0.13 | 6.96 | 9.88 | 0.65 |
| HM | 80.74 | 82.70 | 0.33 | 5.79 | **8.46** | 0.64 |
| H | 80.46 | 82.48 | 0.05 | 11.13 | 14.55 | 0.33 |
| W | 79.11 | 81.23 | 0.01 | 11.81 | 14.81 | 0.38 |

Table 3 shows the evaluation results for 10 feature mixers. HR-MAE and HR-RMSE were evaluated for heart rate per second. In the case of using the Neg-Pearson loss function, overall learning was not performed. On the other hand, when Neg-Pearson and FFT loss were used together, excellent convergence performance was shown. In particular, the type of feature mixer of WM+HM showed the best performance for HR-MAE and R among the three items, and HR-RMSE also showed the second-best performance.



**Figure 4.** This is a figure. Schemes follow the same formatting.

Figure 4 is a graph of neg-Pearson loss. In the early stage of learning, learning is conducted focusing on the trend and peak of the target under the influence of neg-Pearson. After that, using the FFT loss, learn the fine waveforms other than the peak. When only the final loss values were compared, the HM feature mixer showed the best performance, followed by WM+HM.

**Table 4.** Comparison of Memory Size with Representative Algorithm.

|  |  |
| --- | --- |
| Model | Memory Size |
| Deepphys | 5.6MB |
| PhysNet | 3.0MB |
| APNet(Proposed) | 1.1MB |

Table 4 shows the model size of a representative rPPG algorithm. Compared to Deepphys, memory decreased by about 4MB and compared to PhysNet by 1.9MB.

4.3.2 Subjective assessment results

LRP (Layer-wise Relevance Propagation), one of the representative BBMs, decomposes the output of DNN to obtain relevance scores for each feature [25]. In CNN, validity is propagated between layers, and it is a method that can indicate the importance of each pixel by calculating the degree of error through backpropagation.

Chart

Description automatically generated with medium confidence

**Figure 5.** APNET LRP Results.

Figure 5 is the result of LRP. The closer the pixel value is to red, the higher the importance value of the deep learning network is evaluated. As a result of LRP, it can be seen that the forehead, Glabella and Malar region was selected as the region of interest among the regions showing good performance in extracting the 5 types of PPG mentioned in 4.2.2, and it can be seen that the nose, which has a negative effect on learning, was evaluated as the region of interest.

A picture containing line chart

Description automatically generated

**Figure 6.** APNet's inference results.

Figure 6 is a graph of the inference result. As a result of the corresponding graph, it can be seen that the notch of the PPG is well inferred.

5. Conclusions

In summary, in this paper we have proposed:

* An AxisProjection method that can replace the existing spatio-temporal analysis method.
* APNet using AxisProjection method and evaluation of 10 feature mixers.
* A loss function so that the notch of PPG can be learned and analyze the importance of the input pixel using the BBM method.

In this study, the concept of AxisProjection method, a new spatiotemporal analysis method, was proposed. The AxisProjection method has the advantage of using less memory by converting a 3D shape into a 2D shape. In addition, APNet using the AxisProjection method was proposed, and a comparative evaluation was performed through 10 feature mixers to obtain good performance in APNet. Blood flow in the face of a person has a high correlation with respect to the nose and tends to flow from arteries to capillaries. Therefore, it is possible to extract physiological characteristic information from each block divided into H, W, and T, and excellent results can be obtained if each physiological characteristic information is well extracted.

In addition, we proposed a new loss function that PPG's notch can learn. Notch is an important biomarker that can detect LVET (Left Ventricular Ejection Time) and blood pressure. Existing PPG research tends to focus on heartbeat only, and notch, an important biomarker of PPG, is considered. As there is a tendency to not be able to do this, we tried to solve this problem, and after learning, it was confirmed that malar and forehead were recognized as important parts of learning through BBM analysis.

The rPPG task requires various physiological knowledge, and it is not easy to apply it. From the point of view of comparative evaluation of the rPPG algorithm, a standardized evaluation method is not yet prepared, and the dataset used in some prominent papers is inaccessible. In this study, the goal was to infer accurate PPG in a short time using the AxisProjection method, and the V4V dataset was selected due to the problem of not being able to access multiple datasets. In the case of V4V data, only the train data was disclosed, and validation data could not be obtained, so the training data was divided and used for learning and evaluation. In future research, we will improve the performance of the proposed AxisProjection method and create a general-purpose model using various data. In addition, we intend to propose a baseline for the rPPG method by comparing and evaluating several representative deep learning-based models with the same dataset and the same evaluation metric.

**Supplementary Materials:** The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

**Author Contributions:** Conceptualization, D.-Y.K. and K.L.; methodology, D.-Y.K.; software, D.-Y.K.; validation, S.-Y.C.; formal analysis, K.L. and C.-B.S.; investigation, K.L. and C.-B.S.; resources, D.-Y.K., S.-Y.C; data curation, D.-Y.K.; writing—original draft preparation, D.-Y.K.; writing—review and editing, D.-Y.K., K.L. and C.-B.S.; visualization, D.-Y.K.; supervision, K.L. and C.-B.S.; project administration, K.L.; funding acquisition, K.L,S.-Y.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT). (No. 2021-0-00900, Adaptive Federated Learning in Dynamic Heterogeneous Environment).

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Acknowledgments:** This work was supported by the Research Grant of Kwangwoon University in 2022.

**Data Availability Statement:** The V4V dataset is available at https://vision4vitals.github.io (accessed date 19 September 2022).

**Conflicts of Interest:** The authors declare no conflict of interest

References

1. Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. Frontiers in public health, 258.
2. Kim, D., Kim, J., & Lee, K. (2021). Real-time Vital Signs Measurement System using Facial Image Data. Journal of Broadcast Engineering, 26(2), 132-142.
3. Mydin, H. H., Murphy, S., Clague, H., Sridharan, K., & Taylor, I. K. (2013). Anemia and performance status as prognostic markers in acute hypercapnic respiratory failure due to chronic obstructive pulmonary disease. International journal of chronic obstructive pulmonary disease, 8, 151.
4. Yu, Z., Li, X., & Zhao, G. (2019). Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. arXiv preprint arXiv:1905.02419.
5. Niu, X., Shan, S., Han, H., & Chen, X. (2019). Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing, 29, 2409-2423.
6. Chen, W., & McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the european conference on computer vision (ECCV) (pp. 349-365).
7. De Haan, G., & Jeanne, V. (2013). Robust pulse rate from chrominance-based rPPG. IEEE Transactions on Biomedical Engineering, 60(10), 2878-2886.
8. Wang, W., Den Brinker, A. C., Stuijk, S., & De Haan, G. (2016). Algorithmic principles of remote PPG. IEEE Transactions on Biomedical Engineering, 64(7), 1479-1491.
9. Perepelkina, O., Artemyev, M., Churikova, M., & Grinenko, M. (2020). HeartTrack: Convolutional neural network for remote video-based heart rate monitoring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 288-289).
10. Hill, B. L., Liu, X., & McDuff, D. (2021). Learning Higher-Order Dynamics in Video-Based Cardiac Measurement. arXiv preprint arXiv:2110.03690.
11. Mohebbian, M. R., Dinh, A., Wahid, K., & Alam, M. S. (2020). Blind, cuff-less, calibration-free and continuous blood pressure estimation using optimized inductive group method of data handling. Biomedical Signal Processing and Control, 57, 101682.
12. Yu, Z., Peng, W., Li, X., Hong, X., & Zhao, G. (2019). Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 151-160).
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
14. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6836-6846).
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).
16. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2022). A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence.
17. Yu, Z., Li, X., Wang, P., & Zhao, G. (2021). Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. IEEE Signal Processing Letters, 28, 1290-1294.
18. Liu, X., Hill, B. L., Jiang, Z., Patel, S., & McDuff, D. (2021). Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement. arXiv preprint arXiv:2110.04447.
19. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P. H., & Zhao, G. (2022). PhysFormer: facial video-based physiological measurement with temporal difference transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4186-4196).
20. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). Maxvit: Multi-axis vision transformer. arXiv preprint arXiv:2204.01697.
21. Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., & Chen, X. (2021). PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. IEEE Journal of Biomedical and Health Informatics, 25(5), 1373-1384.
22. Revanur, A., Li, Z., Ciftci, U. A., Yin, L., & Jeni, L. A. (2021). The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2760-2767).
23. Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.
24. Kim, D. Y., Lee, K., & Sohn, C. B. (2021). Assessment of ROI Selection for Facial Video-Based rPPG. Sensors, 21(23), 7923.
25. Binder, A., Montavon, G., Lapuschkin, S., Müller, K. R., & Samek, W. (2016, September). Layer-wise relevance propagation for neural networks with local renormalization layers. In International Conference on Artificial Neural Networks (pp. 63-71). Springer, Cham.