DINO-rPPG: Remote photoplethysmography Measurement using Facial Representation from DINO Guidance

Jiho Choi[1], Sang Jun Lee[1,*]

[1]Jeonbuk National University, Republic of Korea

Abstract

Remote photoplethysmography (rPPG) is a camera-based technique that enables non-invasive monitoring of physiological signals such as heart rate (HR) and respiration rate (RR). In light of this advantage, many researchers have suggested deep learning-based methods to measure physiological signals from video data. The 3D convolutional neural network (3D CNN) has been widely applied to capture spatio-temporal features of subtle rPPG changes from facial video. However, the limited receptive fields of the convolution operation leave room for improvement in obtaining global facial features that are crucial for accurate rPPG estimation. Recently, vision transformer trained with self-supervised learning have emerged as powerful tools for extracting high-level features than CNN and supervised ViT. In this study, we propose DINO-rPPG, a method that utilizes a pre-trained DINO to obtain features relevant to the face without additional training. The DINO representation is extracted by a DINO-based semantic extractor (DSE), which effectively captures the high-level semantic features of the face region. The spatio-temporal feature is important for estimating accurate rPPG, therefore, we enhance the spatial DINO representation by incorporating it with features from a spatio-temporal extractor (STE). We conducted experiments using the V4V dataset for estimating HR values, and the results demonstrated that DINO representation guidance is effective for rPPG estimation.

Keywords

Remote photoplethysmography, Self-supervised vision transformer, Heart rate estimation,

1. Introduction

Remote photoplethysmography (rPPG) measurement is a non-contact physiological signal monitoring method based on a camera sensor. It operates through subtle movements of blood flow and changes in pixel values resulting from blood supply in facial images. Consequently, various physiological indicators such as heart rate (HR), respiration rate, and heart rate variability (HRV) can be extracted from the video data without the need for additional contact sensors. The rPPG technique addresses the limitations of conventional contact sensors that require physical attachment to the body. In particular, advances in rPPG research are achieved by the application of deep learning models, expanding applicability to fields such as telemedicine [1, 2], affective computing [3, 4, 5], and deepfake detection [6, 7, 8].

The 3rd Vision-based Remote Physiological Signal Sensing (RePSS) Challenge & Workshop, August 3–9, 2024, Jeju, South Korea

*Corresponding author.
jihochoi@jbnu.ac.kr (J. Choi); sj.lee@jbnu.ac.kr (S. J. Lee)
0000-0003-0771-149X (J. Choi); 0000-0002-9312-6299 (S. J. Lee)

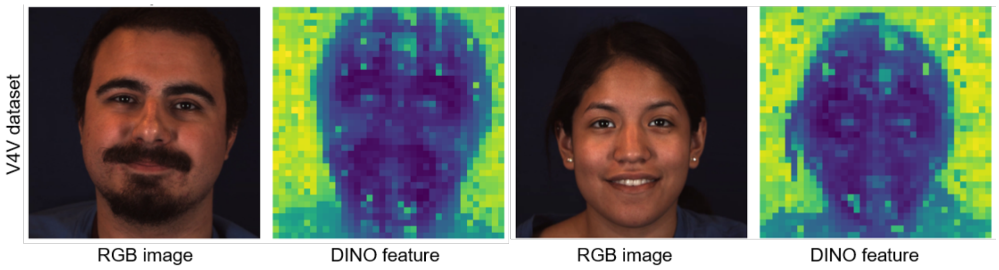

Figure 1: Visualization results of the features from pre-trained DINO. We extracted the features from randomly selected frames from videos in the V4V dataset and performed principal component analysis on these features.

The extraction of physiological signals from video data necessitates the capture of subtle changes in blood flow in facial areas. Early studies employed conventional signal processing methods [9, 10, 11] to calculate intensity changes in RGB images or applied color space transformations [12, 13]. Moreover, a comprehensive understanding of the spatial and temporal information and the corresponding signals is essential for a deep learning model to achieve accurate rPPG measurements. Recent studies preprocess spatio-temporal (ST) maps of the region of interest (ROI) to allow a model to estimate rPPG from ST map [14, 15]. Additionally, various deep learning architectures, such as 3D CNN and vision transformer (ViT), have been explored to effectively capture spatio-temporal features.

Challenges in measuring rPPG include subject movements and varying lighting conditions. The attention mechanisms was used to mitigate the impact of motion artifacts and improve ROI representation [16, 17]. This mechanisms enhance the robustness of deep learning model to noise and illumination changes by suppressing irrelevant information and focusing on important features. rPPGNet [16] introduced a skin-based attention module, and SAM-rPPGNet [17] proposed a spatial-temporal attention mechanism to reduce noise components and improve measurement accuracy. By incorporating attention mechanisms, the model can focus on face-relevant regions and spatio-temporal features that are crucial for estimating rPPG. This allows rPPG signals to be accurately estimated even in challenging environments.

Recently, a vision transformer trained in a self-supervised manner, called DINO [18], was proposed. DINO is trained on large datasets without ground-truth data and can extract meaningful representation from input image. By directly accessing the self-attention layer, it is possible to obtain features that capture high-level semantic information. Due to these properties, DINO has been utilized in downstream tasks and as a feature extractor. We propose a method to estimate rPPG by leveraging the representation of face region obtained from pre-trained DINO.

In this study, we propose a framework to guide rPPG measurement network using representations of ViT trained with self-supervised learning. We found that pre-trained DINO can successfully extract features for ROI regions without additional training, and the feature visualization results for the V4V dataset can be seen in Figure 1. Therefore, we construct a network that effectively combines visual and spatio-temporal features. Our contributions are as follows:

- We propose a framework for rPPG estimation, namely DINO-rPPG, leveraging the guidance of DINO representations.

- To the best of our knowledge, this is the first attempt to utilize DINO features in rPPG measurement.
- We demonstrate the effectiveness of the proposed method using the V4V database [19].

2. Related Works

2.1. Remote photoplethysmography measurement

Conventional rPPG methods using hand-crafted processes extract physiological signals using various facial regions [20] and specific color channels [21]. Additionally, signal decomposition methods such as independent component analysis [10, 11] and principal component analysis [22] have been employed to improve rPPG measurement. The CHROM algorithm [12] uses chrominance to suppress mirror distortion of image and is still widely used in recent research. However, these approaches tend to perform poorly when estimating in noisy or challenging environments, including those with motion artifacts.

Recently, deep learning methods such as convolutional neural networks (CNNs), have been applied to rPPG measurement. However, 2D CNN-based approaches [23, 24] have limitations in capturing temporal features, which are important for physiological signal analysis. To solve these shortcomings, 3D CNN models have been utilized to extract spatio-temporal features [25, 16, 26]. In particular, the 3D CNN has enabled promising rPPG measurement accuracy by leveraging spatio-temporal features and effectively dealing with motion artifacts in video data. Yu et al. proposed PhysNet [25], which is based on a 3D CNN with an encoder-decoder architecture, includes upsampling along the time axis to enhance temporal resolution. However, CNN-based model is limited to the local receptive fields, therefore, ViT architecture has been employed which aggregates global and local information. Yu et al. proposed PhysFormer [27], which aims to capture long-range spatio-temporal and local-global features using a video transformer architecture.

Additionally, ViT with trained by self-supervised manner can extract meaningful representations, and there have been attempts to utilize ViT without ground-truth data to obtain enhanced features. rPPG-MAE [28] was proposed as one of the rPPG method utilizing the representation of self-supervised learning. They utilized the masked autoencoder (MAE) to improve the network representation, robustness to noise, and demonstrated promising estimation accuracy on the VIPL-HR [29], PURE [30], and UBFC-rPPG [31] datasets. This suggests that deep learning models are capable of accurate signal estimation based on representations enriched with relevant feature information. Therefore, in this paper, we propose a method to extract and utilize meaningful information about facial regions from the ViT model DINO [18], trained by self-supervised learning.

2.2. Self-supervised vision transformer

The features of the vision translator provide powerful visual representations that can be used in downstream vision tasks. In particular, DINO is emerging due to its ability to capture high-level semantic information. DINO can obtain semantic information of images more effectively than CNN-based and ViT models trained with labels. This property is achieved by training

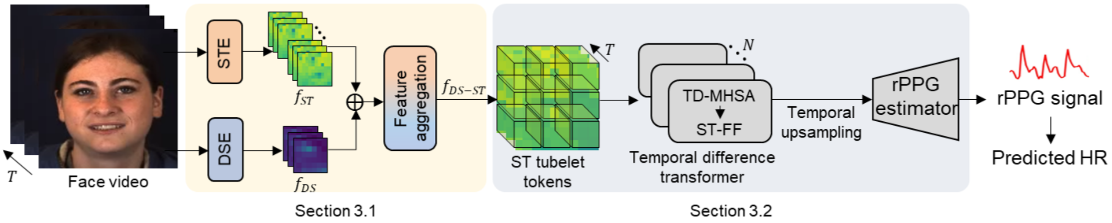

Figure 2: Overview of the proposed method. We input T frames of face video and extract features from DSE and STE. Each extractor outputs feature maps f_{DS} and f_{ST}, respectively. We then integrate them to obtain enhanced spatio-temporal representation, which is denoted as f_{DS-ST}. We use tokenized feature map f_{DS-ST} as input to the video transformer. Finally, the HR value can be estimated from the measured rPPG signal.

teacher and student networks with the same structure in a self-distilling manner. Specifically, augmentations are applied to the input images to generate different views, which are then passed through the networks. The student model is optimized using cross-entropy loss and the exponential moving average technique is applied to update the parameters. With self-supervised learning, ViT can serve as powerful feature extractors and has been applied in various vision tasks.

Another powerful ViT method, such as CLIP [32], provides meaningful features by estimating the similarity between images and text. However, DoesFS [33] demonstrated that DINO representation is superior to CLIP for facial regions using principal component analysis on tokens and keys in intermediate layers. Therefore, we leverage DINO as a feature extractor to improve the representation of the rPPG estimation model. The visualization of the features obtained from DINO is shown in Figure 1.

3. Method

In this section, we introduce DINO-rPPG, a DINO-guided physiological measurement method. The overall framework is shown in Figure 2. We describe the DINO-based semantic extractor (DSE), spatio-temporal extractor (STE), and feature aggregation process in Section 3.1. Then, the video transformer architecture in Section 3.2, and the loss functions for estimating rPPG and HR in Section 3.3.

3.1. DINO based spatio-temporal feature extraction

We leverage the pre-trained ViT-B model from DINOv2, which follows the transformer architecture and employs a self-supervised method trained on large scale datasets. During training, we first input video $v \in \mathbb{R}^{C\times T\times H\times W}$ into the feature extractors, DSE and STE. The DINO-based semantic extractor designed to capture high-level semantic information $f_{DS} = DSE(v)$ about the facial region in the input v that is important for rPPG signal estimation. The DSE features guide the model to learn enhanced representation of the facial regions, which is indicative of physiological signals and enables fast convergence.

However, the DINO feature $f_{DS} \in \mathbb{R}^{D \times T \times H/8 \times W/8}$ only provides the semantic information without considering the temporal aspects. Therefore, we leverage the spatio-temporal extractor to obtain spatial feature $f_{ST} = STE(v)$ over time in a local region, capturing spatio-temporal features $f_{ST} \in \mathbb{R}^{3 \times T \times H/8 \times W/8}$. Inspired by [27], we construct the STE with 3D convolution layers with kernel sizes of the 1×5×5, 1×3×3 and 1×3×3. The feature map f_{ST} is restricted in capturing global features due to its limited receptive field. Therefore, we aggregate low-dimensional local features from the STE and high-dimensional semantic features from the DSE to obtain an enhanced feature map with a DINO-based representation. The combined feature map f_{DS-ST} is obtained by concatenating the f_{ST} and f_{DS}, formulated as follows.

$$f_{DS-ST} = f_{ST} \oplus f_{DS}, \tag{1}$$

where $\oplus$ indicates concatenation process. To embed the combined feature map f_{DS-ST} from the input video, we generate spatio-temporal tubes by considering the temporal dimension. This process generates non-overlapping tubelet tokens from f_{DS-ST}.

3.2. Video transformer

We adopt the temporal difference transformer from PhysFormer [27], which is based on a video transformer architecture. We first embed the feature map f_{DS-ST} using the tubelet embedding method [34], which linearly embeds non-overlapping tubelets. The embedded spatio-temporal tubelet (ST tubelet) token sizes are defined as $T' = \left\lfloor \frac{T}{t} \right\rfloor, H' = \left\lfloor \frac{H/8}{h} \right\rfloor, W' = \left\lfloor \frac{W/8}{w} \right\rfloor$, with the tube size parameters t, h, w set to $(4, 4, 4)$ as in [27].

ST tubelet tokens are then input to the N transformer blocks, which extract local and global spatio-temporal rPPG features. The temporal difference multi-head self-attention block (TD-MHSA) captures local temporal difference features within the video frames, with its self-attention map representing the attention between key and query tube tokens. The self-attention map helps the network focus on the ROI in the spatial domain and locate the peak values of the estimated signal in the time domain. Specifically, the query and key are obtained by learning spatial differences between neighboring frames using temporal difference convolution [35], which captures the derivatives of the PPG signal crucial for rPPG estimation.

The spatio-temporal feed-forward (ST-FF) block refines local inconsistencies in the features extracted by the TD-MHSA block. The latent vector of the transformer blocks undergoes temporal upsampling to match the original video sequence length T, ensuring that the estimated rPPG signal has the same temporal resolution as the input frames. Finally, we extract the physiological signal using an rPPG estimator, which consists of a 1D convolution.

3.3. Loss functions

To optimize the proposed network, we define loss functions in the time domain, frequency domain, and by direct comparison of HR values. To extract accurate rPPG from facial video, we utilize a loss function in the time domain L_{time}. The accurate rPPG should exhibit a high trend similarity with the ground-truth PPG signal, and its peak values on the time axis should be close to the ground-truth peaks. Therefore, we define a loss function using the negative

Pearson correlation, formulated as follows.

$$L_{time} = 1 - \frac{T\sum_1^T xx' - \sum_1^T x \sum_1^T x'}{\sqrt{(T\sum_1^T x^2 - (\sum_1^T x)^2)(T\sum_1^T x'^2 - (\sum_1^T x')^2)}}, \quad (2)$$

where x and x' denote estimated rPPG and ground-truth PPG signals, respectively. The loss term L_{time} ensures that the predicted signal has a trend and peak values similar to the ground-truth signal on the time axis.

The PPG is recorded during a heartbeat and frequency analysis can reveal the periodicity of the signal. We obtain the power spectral density (PSD) by applying a fast Fourier transform to both the actual and predicted signals. Inspired by [26], we define L_{freq} as the root mean square error between the PSD of the signals.

$$L_{freq} = \|P(x) - P(x')\|_2 , \quad (3)$$

where $P(\cdot)$ represents PSD. Components such as motion noise have small amplitudes in the frequency domain and lack strong periodicity. Using L_{freq} enhances robustness to noise, aligns the rPPG signal more closely with the ground-truth, and minimizes errors in the frequency components.

From the rPPG signal, we can calculate the HR value and directly compare the estimated HR with the ground-truth HR. L_{hr} is defined in the time domain as follows.

$$L_{hr} = \|h - h'\|_2 , \quad (4)$$

where h and h' are estimated and actual heart rate value. The overall loss function is denoted as $L_{overall} = \lambda_1 L_{time} + \lambda_2 L_{freq} + \lambda_3 L_{hr}$, where hyper-parameters λ_1, λ_2, λ_3 are set to 1, 100, and 0.0001, respectively.

4. Experiments

4.1. Implementation details

We introduce preprocessing to prepare the input videos and signals for training our proposed model. Using the MTCNN algorithm, we find the bounding box of the face region and crop an area approximately 1.6 times the size of the box. The bounding box is determined in the first frame and applied to subsequent frames. The collected video is segmented into 160 frames and is used as input to the network. The frame rate of the video is maintained at 25 fps, and the signal is resampled to 25 Hz for synchronization.

Our proposed network is trained with an Adam optimizer and a learning rate of 1e-4. We set the batch size to 8 and train the model for 20 epochs on the NVIDIA RTX 4090. The parameters of transformer [27], N and D are set to 12 and 96, respectively. We adopted the mean absolute error (MAE), root mean squared error (RMSE), Pearson correlation coefficient (r) for evaluating the HR estimation task.

Table 1
Heart rate estimation results on V4V dataset. The best results are in bold.

Methods	MAE ↓	RMSE ↓	r ↑
DeepPhys [36]	10.2	13.25	0.45
PhysNet [16]	13.15	19.23	0.75
APNET [37]	4.89	7.83	0.74
DRPNET [26]	3.83	9.59	0.75
DINO-rPPG (Ours)	**3.09**	**7.05**	**0.83**

4.2. Datasets

Vision for Vitals (V4V) is a dataset released as part of the ICCV 2021 Vision for Vitals challenge. The dataset was collected from 179 subjects and consists of videos of them performing 10 tasks designed to induce different emotions. It contains a total of 1,358 video recordings along with accompanying physiological information such as heart rate, blood pressure, and PPG signals. The videos were collected at a resolution of 1280×720 and a frame rate of 25 fps, with varying lengths for each video.

4.3. Experimental results on V4V dataset

In this section, we evaluate the HR estimation performance of the proposed DINO-rPPG compared to previous methods on the V4V dataset. As shown in Table 1, our method consistently outperforms previous approaches across all metrics. DeepPhys [36] and PhysNet [16] perform poorly, with MAE and RMSE values exceeding 10. The recent state-of-the-art method, DRPNET [26], achieved the MAE of 3.83, RMSE of 9.59 and r of 0.75. However, our DINO-rPPG outperforms DRPNET with the MAE 3.09, demonstrating a significant reduction in error rate. We also achieved an r value of 0.83, which is higher than DRPNET, indicating a stronger linear relationship between the estimated HR and the ground-truth HR. Notably, the RMSE of the proposed method decreased from 7.83 in APNET [37] to 7.05. These results suggest that the proposed framework is effective in measuring rPPG and accurately estimating the HR values.

4.4. Ablation study

We also present ablation study results for the HR estimation task on the V4V dataset, as shown in Table 2. The representations of the self-supervised ViT model provides the high-level semantic information, allowing the model to be trained effectively based on abundant facial features and achieve fast convergence. However, the DINO representation provides spatial features without considering temporal information, which is important for sequence data. We report the MAE of 10.67 without using STE, indicating the need for spatio-temporal features. Moreover, when utilizing only STE, the network utilizes spatio-temporal features of the local region, which reduces HR estimation performance, with the MAE and RMSE values of 3.18 and 8.34, respectively. This highlights the limitations of relying solely on local features in 3D-CNN. Therefore, aggregating features from both the DSE and STE is crucial for estimating the rPPG signals, and we demonstrate it by achieving significant improvements in all metrics.

Table 2
Ablation study of DSE and STE on the V4V dataset. The best results are in bold.

Method	MAE ↓	RMSE ↓	r ↑
Ours	**3.09**	**7.05**	**0.83**
Ours w/o DSE	3.18	8.34	0.80
Ours w/o STE	10.67	13.97	0.13

Table 3
Ablation study of loss functions on the V4V dataset. The best results are in bold.

Method	MAE ↓	RMSE ↓	r ↑
Ours	**3.09**	**7.05**	**0.83**
Ours w/o L_{freq}	3.16	8.30	0.81
Ours w/o L_{hr}	3.30	8.74	0.79
Ours w/o L_{time}	4.63	10.71	0.69

Including pre-trained DINO features enhances the ability of network to capture global semantic information, resulting in more accurate prediction of rPPG signals and improved HR estimation.

We investigated the impact of the loss functions used in this study. The loss function L_{freq} is based on frequency domain analysis to measure the PSD difference of the signal. Without L_{freq}, the model report a slightly increased MAE of 3.16, indicating that the loss function effectively handles noise components of small amplitudes and poor periodicity in the frequency domain. Moreover, we conducted an ablation study on time domain loss functions L_{hr} and L_{time}. Comparing HR values directly enables accurate HR estimation, as evidenced by the observed increase in error without using L_{hr}. In addition, L_{time} applies penalty to the distance between the peak values of predicted and ground-truth rPPG to obtain a similar trend on the time axis. Omitting L_{time} degraded estimation performance, with the MAE of 4.63 and the r value of 0.69. We demonstrated the efficiency of the loss functions used in this study, highlighting its effectiveness in improving the accuracy and robustness of the model.

5. Conclusion

In this paper, we propose the deep learning model for rPPG estimation called DINO-rPPG, which utilizes the ViT model trained in a self-supervised manner. We found that DINO provides face-relevant representations without additional training, effectively capturing essential features for accurate rPPG estimation. The DSE uses pre-trained DINO to extract semantic representations of ROI regions, and is further enhanced by considering temporal information through spatio-temporal feature maps obtained from STE. Experimental results demonstrate that DINO-rPPG outperforms previous works in the HR estimation task, highlighting the important role of DINO guidance in improving model performance on the V4V dataset.

Acknowledgments

This research was supported by “Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE). (2023RIS-008)

References

[1] B. P. Yan, W. H. Lai, C. K. Chan, S. C.-H. Chan, L.-H. Chan, K.-M. Lam, H.-W. Lau, C.-M. Ng, L.-Y. Tai, K.-W. Yip, et al., Contact-free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals, Journal of the American Heart Association 7 (2018) e008585.

[2] Z. Sun, J. Junttila, M. Tulppo, T. Seppänen, X. Li, Non-contact atrial fibrillation detection from face videos by learning systolic peaks, IEEE Journal of Biomedical and Health Informatics 26 (2022) 4587–4598.

[3] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, F. Yang, Ubfc-phys: A multimodal database for psychophysiological studies of social stress, IEEE Transactions on Affective Computing 14 (2021) 622–636.

[4] W. Yu, S. Ding, Z. Yue, S. Yang, Emotion recognition from facial expressions and contactless heart rate using knowledge graph, in: 2020 IEEE International Conference on Knowledge Graph (ICKG), IEEE, 2020, pp. 64–69.

[5] Z. Yu, X. Li, G. Zhao, Facial-video-based physiological signal measurement: Recent advances and affective applications, IEEE Signal Processing Magazine 38 (2021) 50–58.

[6] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, A. Morales, Deepfakeson-phys: Deepfakes detection based on heart rate estimation. arxiv 2020, arXiv preprint arXiv:2010.00400 (2020).

[7] Y. Xu, R. Zhang, C. Yang, Y. Zhang, Z. Yang, J. Liu, New advances in remote heart rate estimation and its application to deepfake detection, in: 2021 International Conference on Culture-oriented Science & Technology (ICCST), IEEE, 2021, pp. 387–392.

[8] G. Boccignone, S. Bursic, V. Cuculo, A. D'Amelio, G. Grossi, R. Lanzarotti, S. Patania, Deepfakes have no heart: A simple rppg-based method to reveal fake videos, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 186–195.

[9] X. Li, J. Chen, G. Zhao, M. Pietikainen, Remote heart rate measurement from face videos under realistic situations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 4264–4271.

[10] M.-Z. Poh, D. J. McDuff, R. W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam, IEEE transactions on biomedical engineering 58 (2010) 7–11.

[11] M.-Z. Poh, D. J. McDuff, R. W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation., Optics express 18 (2010) 10762–10774.

[12] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE transactions on biomedical engineering 60 (2013) 2878–2886.

[13] W. Wang, A. C. Den Brinker, S. Stuijk, G. De Haan, Algorithmic principles of remote ppg, IEEE Transactions on Biomedical Engineering 64 (2016) 1479–1491.

[14] X. Niu, S. Shan, H. Han, X. Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, IEEE Transactions on Image Processing 29 (2019) 2409–2423.

[15] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, X. Chen, Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks, IEEE Transactions on Instrumentation and Measurement 69 (2020) 7411–7421.

[16] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, arXiv preprint arXiv:1905.02419 (2019).

[17] M. Hu, F. Qian, X. Wang, L. He, D. Guo, F. Ren, Robust heart rate estimation with spatial–temporal attention network from facial videos, IEEE Transactions on Cognitive and Developmental Systems 14 (2021) 639–647.

[18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).

[19] A. Revanur, Z. Li, U. A. Ciftci, L. Yin, L. A. Jeni, 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, 2021, pp. 2760–2767.

[20] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, N. Sebe, Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2396–2404.

[21] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light., Optics express 16 (2008) 21434–21445.

[22] G. Balakrishnan, F. Durand, J. Guttag, Detecting pulse from head motions in video, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3430–3437.

[23] X. Liu, J. Fromm, S. Patel, D. McDuff, Multi-task temporal shift attention networks for on-device contactless vitals measurement, Advances in Neural Information Processing Systems 33 (2020) 19400–19411.

[24] E. M. Nowara, D. McDuff, A. Veeraraghavan, The benefit of distraction: Denoising camera-based physiological measurements using inverse attention, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4955–4964.

[25] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. arxiv 2019, arXiv preprint arXiv:1905.02419 (????).

[26] G. Hwang, S. J. Lee, Phase-shifted remote photoplethysmography for estimating heart rate and blood pressure from facial video, arXiv preprint arXiv:2401.04560 (2024).

[27] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, G. Zhao, Physformer: Facial video-based physiological measurement with temporal difference transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4186–4196.

[28] X. Liu, Y. Zhang, Z. Yu, H. Lu, H. Yue, J. Yang, rppg-mae: Self-supervised pretraining with masked autoencoders for remote physiological measurements, IEEE Transactions on Multimedia (2024).

[29] X. Niu, H. Han, S. Shan, X. Chen, Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video, in: Computer Vision–ACCV 2018: 14th Asian Conference

on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14, Springer, 2019, pp. 562–576.

[30] R. Stricker, S. Müller, H.-M. Gross, Non-contact video-based pulse rate measurement on a mobile service robot, in: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, IEEE, 2014, pp. 1056–1062.

[31] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters 124 (2019) 82–90.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[33] Y. Zhou, Z. Chen, H. Huang, Deformable one-shot face stylization via dino semantic guidance, arXiv preprint arXiv:2403.00459 (2024).

[34] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6836–6846.

[35] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching, IEEE Signal Processing Letters 27 (2020) 1245–1249.

[36] W. Chen, D. McDuff, Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the european conference on computer vision (ECCV), 2018, pp. 349–365.

[37] D.-Y. Kim, S.-Y. Cho, K. Lee, C.-B. Sohn, A study of projection-based attentive spatial–temporal map for remote photoplethysmography measurement, Bioengineering 9 (2022) 638.