UP-Phys: Exploring the Effect of Prior Knowledge in Unsupervised Remote Photoplethysmography

Yan Jiang[†], Mingyue Cao[†], Hao Yu, Xingyu Liu and Xu Cheng[*]

Nanjing University of Information Science and Technology, 219 Ningliu Road, Nanjing, Jiangsu, 210044, China

Abstract

Remote photoplethysmography (rPPG) is a non-contact method that estimates multiple physiological parameters according to facial videos. Although existing supervised rPPG methods have achieved remarkable performance, the success mainly benefits from massive and expensive annotated data. Fortunately, many unsupervised rPPG methods have emerged recently to solve this issue. However, we find that existing unsupervised rPPG methods are learn-from-scratch. Many downstream tasks in deep learning have achieved great success using fine-tuning strategies in the past decade. Inspired by this, we explore the effect of prior knowledge in unsupervised rPPG and proposed UP-Phys. Moreover, to regulate the backbone to prioritize regions rich in rPPG information, we propose a plug-and-play representation augmentation module (RAM). RAM dynamically enhances salient temporal-spatial information derived from extracted features, effectively reducing the effect of noise brought by lighting, motion, *etc.* Experiments on two widely used rPPG datasets UBFC-rPPG and PURE demonstrate the superiority of our proposed method. In addition, our method achieves 15.79 RMSE accuracy in the 3rd RePSS.

Keywords

Remote Photoplethysmography, Unsupervised Learning, Prior Knowledge

1. Introduction

Remote photoplethysmography (rPPG) estimates multiple physiological parameters that are important for healthcare including heart rate (HR), respiration frequency (RF), and heart rate variability (HRV) through videos captured by cameras [1]. Compared with traditional HR estimation approaches like electrocardiogram (ECG) [2] and photoplethysmography (PPG) [3] that require skin contact with subjects, rPPG is non-contact, thus avoiding discomfort and skin irritation caused by skin-contact sensors. To this end, rPPG technology has become intensively researched in recent years and plays an increasingly pivotal role in remote healthcare [1], affective computing [4, 5], spoof detection [6, 7], *etc.*

Existing rPPG methods [8, 9, 10, 11] have achieved remarkable performance with deep learning methods. However, the success mainly profits from supervised learning over massive human-labeled data. In fact, the process of collecting and annotating such data is prohibitively

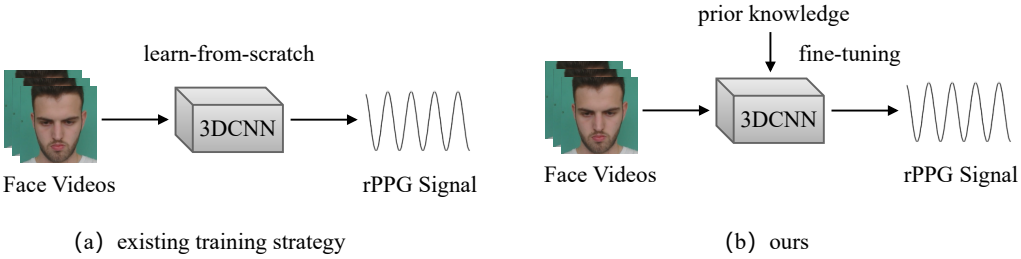
The 3rd Vision-based Remote Physiological Signal Sensing (RePSS) Challenge & Workshop

[*]Corresponding author.

[†]These authors contributed equally.

jiangyan@nuist.edu.cn (Y. Jiang); cmy@nuist.edu.cn (M. Cao); yuhao@nuist.edu.cn (H. Yu); xingyu@nuist.edu.cn (X. Liu); xcheng@nuist.edu.cn (X. Cheng)

0009-0002-2031-5627 (Y. Jiang); 0009-0005-7796-7484 (M. Cao); 0000-0002-8298-7181 (H. Yu); 0009-0009-6064-9104 (X. Liu); 0000-0003-2355-9010 (X. Cheng)



(a) existing training strategy (b) ours

Figure 1: Motivation of our proposed method. Existing methods adopt the learn-from-scratch strategy, which may introduce potential issues such as limited generalization, overfitting, reliance on the scale of data, *etc.* To solve this issue, our method adopts a fine-tuning strategy that introduces prior knowledge in rPPG, enhancing the robustness and efficacy of the learning process.

expensive, requiring not only the deployment of subjects equipped with contact PPG or ECG sensors but also careful consideration of various potential environmental factors such as lighting changes, motion, gestures, and so on while capturing data. In addition, existing supervised rPPG methods struggle to break through the bottleneck posed by unlabeled data due to their performance being positively corresponded to the scale of annotated data available, resulting in less applicability in real scenarios. Fortunately, some unsupervised rPPG methods have been proposed recently to solve this issue of expensive rPPG data annotations.

Existing unsupervised rPPG methods [12, 13, 14, 15] can be roughly divided into two categories: contrastive and non-contrastive. In the former category, Sun *et al.* [13] pioneered the introduction of contrastive learning into unsupervised rPPG methods with their proposal of Contrast-Phys. This method was developed based on four key observations: spatial similarity in rPPG signals, temporal similarity in rPPG signals, dissimilarity in rPPG signals across different videos, and HR range constraint. Crucially, Contrast-Phys eliminates the reliance on annotated data and achieves state-of-the-art in publicly available academic datasets. For the latter category, Speth *et al.* [14] extended unsupervised methods based on contrastive learning research lines into non-contrastive and proposed SiNC by discovering periodic signals in video data. SiNC considers that periodicity suffices for learning minuscule visual features corresponding to the blood volume pulse from unlabeled face videos, which brings novel inspirations into the rPPG community.

Despite achieving encouraging progress, the aforementioned unsupervised rPPG methods are learn-from-scratch, as shown in Fig.1 (a). This training strategy may introduce potential issues such as limited generalization, overfitting, and reliance on the scale of data. Moreover, the quality of predicted rPPG signals by the deep neural network has emerged as a pivotal challenge in elevating the performance ceiling of unsupervised rPPG, as it lacks effective supervision by label information. During the past decade, many downstream tasks in computer vision adopted the fine-tuning strategy [16, 17, 18] and achieved significant success. This success is attributed to the prior knowledge acquired through pretraining, which enables the network to adapt to various datasets more efficiently and attain superior performance. Inspired by this, in this paper, we explore the effect of prior knowledge in unsupervised rPPG and propose UP-Phys, as shown in Fig.1 (b). Specifically, we utilize the Contrast-Phys pre-trained on the MMSE-HR [19] dataset

and fine-tune other datasets. Compared with the official training protocol of 30 epochs, our UP-Phys undergoes only 1 epoch of fine-tuning, resulting in significant time savings during training. Furthermore, we design a plug-and-play representation augmentation module (RAM) that dynamically enhances salient temporal-spatial information derived from extracted features. This augmentation empowers the network to prioritize regions abundant in rPPG information, consequently reducing the effect of noise brought by lighting, motion, *etc.* Generally, the main contributions of this paper can be summarized as follows:

- We introduce a novel solution for unsupervised rPPG, termed UP-Phys, which leverages prior knowledge to reduce training time notably.
- We design a plug-and-play representation augmentation module (RAM) that dynamically enhances salient temporal-spatial information derived from extracted features for unsupervised rPPG.
- Experiments on PURE and UBFC-rPPG datasets demonstrate that our UP-Phys significantly outperforms existing unsupervised rPPG methods, and even surpasses some supervised counterparts. In addition, UP-Phys achieves 15.79 RMSE accuracy in 3rd RePSS.

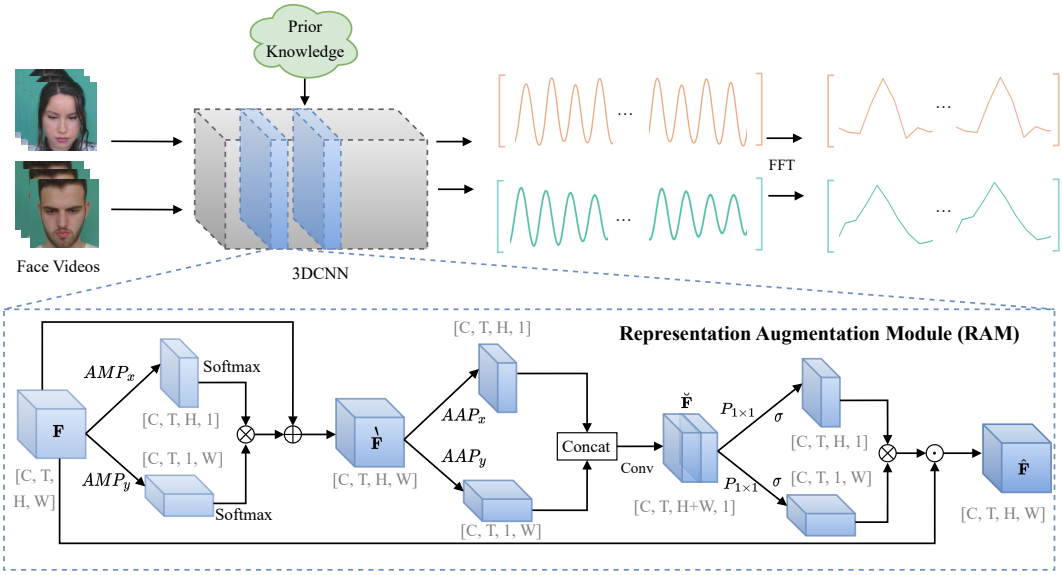


Figure 2: The pipeline of the proposed UP-Phys. RAM is our proposed Representation Augmentation Module.

2. Methodology

The overview of our proposed UP-Phys is shown in Fig.2.

Table 1
Experiments on different prior knowledge. V denotes the number of pre-training videos.

Index	V	Pretrain	UBFC-rPPG		
			MAE ↓	RMSE ↓	R ↑
1	0	✘	0.64	1.00	0.99
2	25	✓	0.50	1.47	0.99
3	50	✓	0.42	0.96	0.99
4	100	✓	0.33	0.65	0.99

2.1. Preprocessing

To reduce background noise and interference from irrelevant areas, we adopt the OpenFace toolkit to preprocess the video. Specifically, we begin by determining the minimum and maximum horizontal and vertical coordinates of generated landmarks to pinpoint the central facial point for each frame. The size of the bounding box is set to 1.2 times the range of the vertical coordinates of landmarks from the first frame, and this size remains constant for subsequent frames. Then, we crop the face from each frame and resize it to 128×128 according to the central facial point of each frame and bounding box. To minimize I/O overhead during training, we convert video files into Hierarchical Data Format (HDF5) format.

2.2. Prior Knowledge

Over the past decade, deep learning has achieved significant success, with many downstream tasks showing impressive results through fine-tuning pre-trained weights. Inspired by this, we introduce the fine-tuning strategy into unsupervised rPPG as existing methods are learn-from-scratch. Specifically, we utilize Contrast-Phys, pre-trained on the MMSE-HR dataset, and fine-tune it for just one epoch on the UBFC-rPPG dataset to investigate the impact of prior knowledge, as shown in Tab. 1.

With 25 pre-training videos, the MAE accuracy improves by 0.14 but RMSE accuracy increases by 0.47. This indicates that prior knowledge can help reduce the average error. The underlying reason for bad RMSE stems from less prior knowledge. When we increase the pre-training videos to 50, as shown in index 3, we can observe that both MAE and RMSE achieve significant improvement. Moreover, the pre-training with 100 videos shows the best performance with the lowest MAE of 0.33 and RMSE of 0.65. This indicates that larger prior knowledge significantly enhances the model's prediction accuracy and consistency. In summary, these results demonstrate a clear trend that as the number of pre-training videos increases, the accuracy of the model improves. This emphasizes the benefits of leveraging prior knowledge through pre-training in enhancing the performance of rPPG models.

2.3. Representation Augmentation Module

Existing unsupervised rPPG methods mainly design refreshing strategies to achieve robust training without annotated data. The quality of rPPG signal prediction by these methods heavily relies on the features extracted by the backbone. These unsupervised methods rely solely on

3DCNN and cannot accurately focus on regions with rich rPPG signals in complex environments such as head movement and lighting, resulting in difficulty in improving performance. Therefore, we propose a plug-and-play representation augmentation module (RAM) that dynamically enhances salient temporal-spatial information, helping the backbone focus on regions rich in rPPG information.

Specifically, given the input features $\mathbf{F} \in \mathbb{R}^{C\times T\times H\times W}$, we first apply 3D AdaptiveMaxPool to extract the most salient rPPG knowledge in both horizontal and vertical directions. Subsequently, we utilize a softmax function to transform this rPPG knowledge into a distribution ranging from 0 to 1. This distribution is then used to create the augmentation mask through multiplication. Finally, this augmented mask is added to the input features to enhance the rPPG information. It is written as follows:

$$\grave{\mathbf{F}} = \mathbf{F} + \text{Softmax}(\text{AMP}_x(\mathbf{F})) \otimes \text{Softmax}(\text{AMP}_y(\mathbf{F})), \tag{1}$$

where AMP_x and AMP_y denote the 3D AdaptiveMaxPool with pooling kernels $(T, H, 1)$ and $(T, 1, W)$, respectively. $\otimes$ is the multiplication operation.

After that, the augmented features $\grave{\mathbf{F}}$ are processed by 3D AdaptiveAvgPool to attain the directional rPPG knowledge. Then, we concatenate the two directional features along the spatial dimension to investigate the spatial rPPG information. In addition, A basic 3D convolutional block is employed to discover shared rPPG information and reduce channel dimension, which can be expressed as:

$$\breve{\mathbf{F}} = \text{Conv}(\text{Cat}(\text{AAP}_x(\grave{\mathbf{F}}), \text{AAP}_y(\grave{\mathbf{F}}))), \tag{2}$$

where AAP_x and AAP_y denote the 3D AdaptiveAvgPool with pooling kernels $(T, H, 1)$ and $(T, 1, W)$, respectively. $\text{Cat}(\cdot, \cdot)$ denotes the concatenation on the height dimension. Conv denotes the basic 3D convolutional block consisting of a pointwise convolution, batch normalization, and ELU activation.

Further, we split the $\breve{\mathbf{F}}$ along spatial dimension and get $\breve{\mathbf{F}}_h$ and $\breve{\mathbf{F}}_w$. Based on $\breve{\mathbf{F}}_h$ and $\breve{\mathbf{F}}_w$, a pointwise convolution is utilized to restore the channel dimension. Then, sigmoid normalization and multiplication are employed to generate a mask that discriminates against rPPG information. Finally, the mask is element-wise multiplicated with the input features to augment the features, thereby regulating the backbone sensitively concentrating on the regions rich in rPPG information.

$$\hat{\mathbf{F}} = [\sigma(\text{P}_{1\times1}(\breve{\mathbf{F}}_h)) \otimes \sigma(\text{P}_{1\times1}(\breve{\mathbf{F}}_w))] \odot \mathbf{F}. \tag{3}$$

where $\hat{\mathbf{F}} \in \mathbb{R}^{C\times T\times H\times W}$ is the augmented features; σ denotes the sigmoid function; $\text{P}_{1\times1}$ denotes the pointwise convolution; $\odot$ is the element-wise multiplication.

3. Experiments

3.1. Experimental Setup and Evaluation Protocol

Datasets. We evaluate the proposed method on the two widely used rPPG datasets UBFC-rPPG [20] and PURE [21]. In addition, we pretrain our method on the MMSE-HR [19] dataset. **UBFC-rPPG** contains 42 videos where subjects manipulate their heart rates by engaging in

mathematical games. Each video is recorded at 30 frames per second (fps), has a resolution of 640×480, and runs for approximately one minute. Ground truth data is collected synchronously using a CMS50E pulse oximeter at a sampling rate of 30 Hz. **PURE** records videos of 10 subjects across 6 different scenarios, including those with head movements. Each video maintains a one-minute duration, is captured at 30 fps, and boasts a resolution of 640×480. The ground truth is accurately recorded using a fingertip pulse oximeter at 60 Hz, specifically to capture the blood volume pulse (BVP) signal. **MMSE-HR** contains 102 videos from 40 subjects. Each video is 25fps, and the subject's emotional guidance ensures the heart rate changes. Physiological data were collected by the Boipac Mp150 data acquisition system at 1khz.

Evaluation Protocol. Following previous works [13, 14], we adopt mean absolute error (MAE), root mean squared error (RMSE), and person correlation coefficient (R) as the evaluation metrics.

Experimental Setup. We implement our UP-Phys on the PyTorch framework with two RTX 2080Ti GPUs. The Contrast-Phys [13] is utilized as our baseline. The proposed RAM is added after encoder 1 and encoder 2 of the backbone. We initially pre-train our UP-Phys model on the MMSE-HR dataset, utilizing the AdamW optimizer with a learning rate of 10^{-5} for 30 epochs. Subsequently, we only fine-tune the UP-Phys 1 epoch on the dataset to be evaluated. All other settings are maintained consistently with those of Contrast-Phys.

RePSS Setup. We first pre-train our UP-Phys on 209 videos collected by MMSE-HR and VIPL-HR [22] datasets. Subsequently, we fine-tune our method on the UBFC-rPPG and PURE datasets for 1 epoch. *We finally achieve 15.79 RMSE accuracy on the 3rd RePSS.*

3.2. Intra-Dataset Testing

We report 3 representative supervised and unsupervised methods for comparison.

Comparison with Unsupervised Methods. As reported in Tab. 2, the performance of our method surpasses current leading unsupervised methods. More precisely, our UP-Phys achieves 0.18 and 0.48 MAE accuracy on UBFC-rPPG and PURE datasets, respectively. It significantly outperforms SiNC [14] by 0.41 and 0.13 on these two datasets. Note that while our UP-Phys is based on Contrast-Phys [13], it significantly outperforms Contrast-Phys. This success is attributed to the pivotal role of prior knowledge and UP-Phys's keen ability to focus on regions abundant in rPPG information, simultaneously demonstrating the effectiveness of our proposed method.

Comparison with Supervised Methods. Supervised methods such as Dual-GAN [11] perform well on both datasets, particularly achieving excellent results with an MAE of 0.44 and an RMSE of 0.67 on UBFC-rPPG. This can be attributed to the ability of supervised methods to utilize labeled information in the dataset for training, facilitating the model to learn accurate heart rate estimation patterns. However, without the label information, our proposed UP-Phys significantly surpasses Dual-GAN. The excellent performance of our method benefits from the insightful design of the prior knowledge. Interestingly, our method shows the potential of unsupervised rPPG methods, and we believe this design can bring new insights to the rPPG community.

Table 2
Intra-dataset HR results. The best results are in bold. The Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson Correlation Coefficient (R) are reported.

Method Types	Methods	UBFC-rPPG			PURE		
		MAE ↓	RMSE ↓	R ↑	MAE ↓	RMSE ↓	R ↑
Supervised	PhysNet [9]	-	-	-	2.10	2.60	0.99
	PulseGAN [10]	1.19	2.10	0.98	-	-	-
	Dual-GAN [11]	0.44	0.67	0.99	0.82	1.31	0.99
Unsupervised	Gideon2021 [12]	1.85	4.28	0.93	2.30	2.90	0.99
	Contrast-Phys [13]	0.64	1.00	0.99	1.00	1.40	0.99
	SiNC [14]	0.59	1.83	0.99	0.61	1.84	1.00
	UP-Phys (Ours)	**0.18**	**0.45**	**0.99**	**0.48**	**0.69**	**1.00**

Table 3
Ablation studies for different components of the proposed UP-Phys on UBFC-rPPG. RAM denotes the proposed representation augmentation module. MAE, RMSE, and R are reported.

Index	Pretrain	RAM	MAE ↓	RMSE ↓	R ↑
1	✘	✘	0.64	1.00	0.99
2	✘	✓	0.58	1.50	0.99
3	✓	✘	0.33	0.65	0.99
4	✓	✓	**0.18**	**0.45**	**0.99**

3.3. Ablation Study

To evaluate the contribution of the designed component, we conduct an ablation experiment on the UBFC-rPPG dataset, as shown in Tab. 3.

Baseline in index 1 denotes that we directly train the Contrast-Phys [13]. It is observed that the baseline only achieves 0.64 MAE accuracy and 1.00 RMSE accuracy, showing the limited capability of the baseline to predict accurate HR.

Effectiveness of RAM. As shown in index 2, by only adding the RAM, the MAE slightly decreases to 0.58, but the RMSE increases to 1.50, indicating that the RAM module improves the prediction accuracy of the model on some samples but introduces large errors on other samples. With the help of knowledge, as shown in index 4, the MAE further decreases to 0.18 and the RMSE to 0.45, achieving a superior performance. This indicates that prior knowledge can help RAM significantly reduce prediction errors.

Effectiveness of Prior Knowledge. As shown in index 3, only directly adopting the pre-train can bring significant improvement. Specifically, the MAE drops from 0.64 to 0.33 and RMSE drops from 1.00 to 0.65. Meanwhile, this accuracy even surpasses existing unsupervised rPPG methods, showing the effectiveness of prior knowledge.

Generally, the above observation and analysis demonstrate the effectiveness of our proposed components.

4. Conclusion

This paper introduces a novel unsupervised method termed UP-Phys that leverages prior knowledge to reduce training time and improve HR estimate accuracy notably. Furthermore, we design a plug-and-play representation augmentation module (RAM) that dynamically enhances salient temporal-spatial information derived from extracted features. This augmentation empowers the network to prioritize regions abundant in rPPG information, consequently reducing the effect of noise brought by lighting, motion, *etc.* Experiments on PURE and UBFC-rPPG datasets demonstrate the effectiveness of our method. In addition, our method achieves 15.79 RMSE accuracy in the 3rd RePSS.

5. Acknowledgements

This research is funded in part by the National Natural Science Foundation of China (Grant No. 61802058, 61911530397), in part by the open Project Program of the State Key Laboratory of CAD&CG, Zhejiang University (under Grant A2318), and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX24_1514).

References

[1] J. Shi, I. Alikhani, X. Li, Z. Yu, T. Seppänen, G. Zhao, Atrial fibrillation detection from face videos by fusing subtle variations, IEEE Transactions on Circuits and Systems for Video Technology 30 (2019) 2781–2795.

[2] D. McDuff, E. Blackford, iphys: An open non-contact imaging-based physiological measurement toolbox, in: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2019, pp. 6521–6524.

[3] J. Allen, Photoplethysmography and its application in clinical physiological measurement, Physiological measurement 28 (2007) R1.

[4] 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.

[5] 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.

[6] L. Birla, P. Gupta, Patron: Exploring respiratory signal derived from non-contact face videos for face anti-spoofing, Expert Systems with Applications 187 (2022) 115883.

[7] L. Birla, P. Gupta, S. Kumar, Sunrise: Improving 3d mask face anti-spoofing for short videos using pre-emptive split and merge, IEEE Transactions on Dependable and Secure Computing (2022).

[8] 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.

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

[10] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, X. Chen, Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography, IEEE Journal of Biomedical and Health Informatics 25 (2021) 1373–1384.
[11] H. Lu, H. Han, S. K. Zhou, Dual-gan: Joint bvp and noise modeling for remote physiological measurement, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12404–12413.
[12] J. Gideon, S. Stent, The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3995–4004.
[13] Z. Sun, X. Li, Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast, in: European Conference on Computer Vision, Springer, 2022, pp. 492–510.
[14] J. Speth, N. Vance, P. Flynn, A. Czajka, Non-contrastive unsupervised learning of physiological signals from video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14464–14474.
[15] M. Cao, X. Cheng, X. Liu, Y. Jiang, H. Yu, J. Shi, St-phys: Unsupervised spatio-temporal contrastive remote physiological measurement, IEEE Journal of Biomedical and Health Informatics (2024).
[16] H. Yu, X. Cheng, W. Peng, Toplight: Lightweight neural networks with task-oriented pretraining for visible-infrared recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3541–3550.
[17] H. Yu, X. Cheng, W. Peng, W. Liu, G. Zhao, Modality unifying network for visible-infrared person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11185–11195.
[18] X. Liu, X. Cheng, H. Chen, H. Yu, G. Zhao, Differentiable auxiliary learning for sketch re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 3747–3755.
[19] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al., Multimodal spontaneous emotion corpus for human behavior analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3438–3446.
[20] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters 124 (2019) 82–90.
[21] 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.
[22] 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.