DiffRePSS: A Diffusion Model for Remote Physiological Signal Sensing

Wei Qian[1,†], Gaoji Su[1,†], Kun Li[4,*], Yuchen Ding[1], Xiangyuan Jia[1] and Dan Guo[1,2,3,*]

[1]School of Computer Science and Information Engineering, School of Artificial Intelligence, Hefei University of Technology (HFUT)
[2]Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China
[4]ReLER, CCAI, Zhejiang University, China

Abstract

This paper briefly introduces the solutions developed by our team, HFUT-VUT, for Remote Physiological Signal Sensing in the 4th Vision-based Remote Physiological Signal Sensing (RePSS) Challenge hosted at IJCAI 2025. Specifically, we present a diffusion-based model for remote physiological signal sensing, named DiffRePSS, inspired by diffusion models' noise distillation abilities. The proposed model takes the multi-scale temporal map (MSTmap) and its temporal difference representation as input of an alternated spatial-temporal Transformer backbone model to extract the rPPG signal from Gaussian noise using a diffusion process during training. The model then refines it using a reverse diffusion process. To overcome the HR distribution imbalance, we propose an effective data augmentation strategy that synthetically expands the HR distribution. As a result, our solutions achieved a remarkable RMSE score of 11.89505 on the test set, securing the Champion of this challenge.

Keywords

rPPG, Diffusion, Transformer, Data augmentation

1. Introduction

Remote physiological signal sensing (RePSS) [1, 2, 3, 4, 5, 6], particularly remote photoplethysmography (rPPG), has attracted increasing attention in recent years due to its contactless and low-cost nature, with broad applications in human-centric video understanding tasks [7, 8, 9, 10, 11, 12, 13, 14]. It enables heart rate (HR) and other vital signs to be estimated from facial videos by capturing subtle changes in skin color caused by blood volume variations. However, despite rapid advancements, robust and accurate rPPG estimation under real-world conditions remains challenging due to various factors such as head movement, illumination changes, and sensor noise.

Traditional hand-crafted algorithms [15, 16, 17, 18, 19] have attempted to suppress noise and isolate pulse signals using signal decomposition, color space transformation, or skin subspace modeling. More recently, deep learning-based approaches [20, 21, 22, 23, 5, 24, 25, 3, 26] have shown superior performance by learning end-to-end mappings from raw facial videos to physiological signals. Nevertheless, many of these models struggle to generalize in unconstrained environments, partly due to limited data diversity and a lack of robustness to noise and distribution shifts. In this work, we propose **DiffRePSS**, a diffusion-based model for remote physiological signal sensing, motivated by the generative capability and noise modeling strength of diffusion models. Specifically, we design a conditional diffusion framework to progressively recover clean rPPG signals from noise, guided by physiological priors extracted from facial videos. We introduce a multi-scale spatial-temporal map (MSTmap) and its temporal difference representation as inputs to our alternated spatial-temporal Transformer-based denoiser, which jointly models spatial and temporal dependencies in physiological features. Furthermore, to address the issue

The 4th RePSS - Multimodal Fusion Learning for Remote Physiological Signal Sensing, August 28th, 2025.
*Corresponding authors.
†These authors contributed equally.
qianwei.hfut@gmail.com (W. Qian); sugaojix@gmail.com (G. Su); kunli.hfut@gmail.com (K. Li); dingyuchen.hfut@gmail.com (Y. Ding); jiaxiangyuan9@gmail.com (X. Jia); guodan@hfut.edu.cn (D. Guo)
0009-0007-9467-6296 (W. Qian); 0009-0007-4876-6100 (G. Su); 0000-0001-5083-2145 (K. Li); 0009-0006-8058-0181 (Y. Ding); 0009-0005-1024-6412 (X. Jia); 0000-0003-2594-254X (D. Guo)

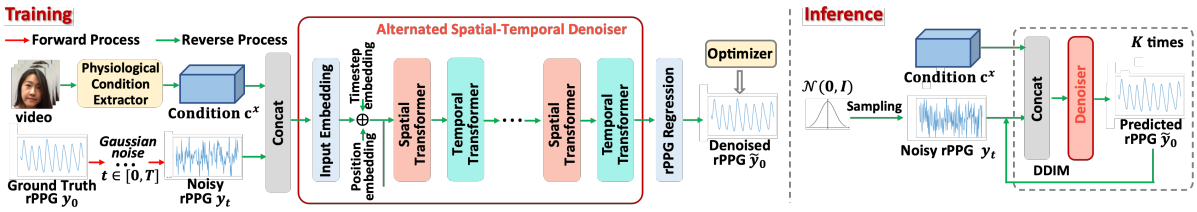

Figure 1: The overview of the proposed DiffRePSS.

of imbalanced HR distribution in existing datasets, we propose a novel HR distribution augmentation strategy by temporally resampling MSTmaps to simulate diverse HR values. Our proposed approach demonstrates strong performance on the challenging VIPL-HR and OBF datasets, achieving state-of-the-art results and winning the 4th RePSS Challenge.

The key contributions of this work are summarized as follows:

- We propose a diffusion-based framework for rPPG signal estimation, leveraging the DDIM mechanism for robust signal reconstruction.
- We introduce a temporal difference representation to capture fine-grained physiological dynamics.
- We introduce an alternated spatial-temporal Transformer denoiser to effectively model dependencies across space and time.
- We propose a simple yet effective HR distribution augmentation strategy to improve model generalization across HR ranges. The proposed method achieves state-of-the-art performance and wins 1st place in the 4th RePSS Challenge.

2. Methodology

We consider the problem of estimating accurate rPPG signals from facial videos. Formally, given an input facial video $\boldsymbol{v} \in \mathbb{R}^{T\times H\times W\times 3}$, our goal is to predict the one-dimensional quasi-periodic rPPG signal $\tilde{\boldsymbol{y}} \in \mathbb{R}^{T}$, where T, H, and W denote the frame number, height, and width of the video, respectively. As shown in Figure 1, we first send the facial video into *Physiological Condition Extractor* to extract the multi-scale spatial-temporal map (MSTmap) $\boldsymbol{x} \in \mathbb{R}^{T\times N\times C}$ and its temporal difference representation $\Delta\boldsymbol{x} \in \mathbb{R}^{T\times N\times C}$, where N and C correspond to the number of facial ROIs and channel dimensions, respectively. In the forward process, the ground truth rPPG signal $\boldsymbol{y} \in \mathbb{R}^{T}$ is corrupted by adding Gaussian noise, yielding the noisy rPPG signal $\boldsymbol{y}_t$ at the t-th timestep. During the reverse process, the Gaussian noises are fed to *Alternated Spatial-Temporal Denoiser*, generating a denoised rPPG signal.

2.1. Physiological Condition Extractor

Since periodic pulse signals arise from subtle light reflections caused by blood volume changes in skin regions, non-skin pixels and facial geometric structures are typically regarded as noise relative to skin chrominance features. To suppress such noise and emphasize physiological signals, we transform raw facial videos into Multi-scale Spatial-Temporal Maps (MSTmaps), which have been widely adopted in prior works [23, 26, 3, 27]. MSTmaps effectively encode physiological spatial-temporal information from videos. To further extract dynamic physiological variations, we compute frame-wise differences along the temporal dimension of the MSTmap, thereby capturing temporal changes related to the pulsatile signal. Given an input MSTmap $\boldsymbol{x} \in \mathbb{R}^{T\times N\times C}$, we compute the temporal difference as:

$$\Delta\boldsymbol{x}_{t,n,c} = \boldsymbol{x}_{t+1,n,c} - \boldsymbol{x}_{t,n,c}, \quad \text{for } t = 0, \ldots, T-2. \tag{1}$$

To preserve the original dimensionality, we apply zero-padding or temporal interpolation to $\Delta\boldsymbol{x}$, resulting in a temporal difference representation $\boldsymbol{x}^{\mathbf{D}} \in \mathbb{R}^{T\times N\times C}$. This representation accentuates the color fluctuations induced by cardiovascular activity, providing a clearer physiological cue.

Finally, we concatenate the original MSTmap $\boldsymbol{x}$ and the temporal difference representation $\Delta\boldsymbol{x}$ along the channel dimension, forming the complete physiological condition representation:

$$\mathbf{c}^{\boldsymbol{x}} = \text{Concat}(\boldsymbol{x}, \Delta\boldsymbol{x}) \in \mathbb{R}^{T\times N\times 2C}. \tag{2}$$

This fused physiological representation $\mathbf{c}^{\boldsymbol{x}}$ captures both the chrominance patterns and their dynamic variations, serving as a robust input for subsequent physiological signal estimation.

2.2. RePSS via Diffusion Model

Forward Process. The Forward Process is an approximate posterior that follows the Markov chain that gradually adds Gaussian noise $\mathcal{N}(0, I)$ to the original data $\boldsymbol{y}_0$. Followed by DDPM [28], the forward process q can be defined as:

$$q(\boldsymbol{y}_{1:K}|\boldsymbol{y}_0) = \prod_{k=1}^{K} q(\boldsymbol{y}_k|\boldsymbol{y}_{k-1}), \tag{3}$$

$$q(\boldsymbol{y}_k|\boldsymbol{y}_{k-1}) = \mathcal{N}\big(\boldsymbol{y}_k|\sqrt{\alpha_k}\boldsymbol{y}_{k-1}, (1-\alpha_k)\boldsymbol{I}\big), \tag{4}$$

where the scalars $\alpha_{1:K}$ are either predefined or learned variances, s.t. $1 > \alpha_1 > \alpha_2 > ... > \alpha_K > 0$. To simplify the training process, we sample $\boldsymbol{y}_k$ arbitrarily:

$$\boldsymbol{y}_k = \sqrt{\bar{\alpha}_k}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_k}\epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$, $\bar{\alpha}_k = \prod_{i=1}^{t} \alpha_i$, $\alpha_i = 1 - \beta_i$, β_i is a noise schedule. We adopt the cosine-schedule, which always increases as the sampling step k increases.

Reverse Process. In the training stage, $\boldsymbol{y}_k$ is sent to a denoiser $\mathscr{D}$ conditioned on physiological representation $\mathbf{c}^{\boldsymbol{x}}$ and timestep k to refine the rPPG signal $\tilde{\boldsymbol{y}}_0$ without noise:

$$\tilde{\boldsymbol{y}}_0 = \mathscr{D}(\boldsymbol{y}_k, \mathbf{c}^{\boldsymbol{x}}, k), \tag{6}$$

The entire framework is supervised by a standard Negative Pearson Correlation loss $\mathscr{L}_{rPPG}$ [26, 27]. At the inference stage, we first obtain an initial rPPG signal $\boldsymbol{y}_K$ by sampling noise from a unit Gaussian. Then $\tilde{\boldsymbol{y}}_0$ is predicted by passing $\boldsymbol{y}_K$ to the trained denoiser $\mathscr{D}$. Thereafter, $\tilde{\boldsymbol{y}}_0$ is used to generate the noisy rPPG signal $\tilde{\boldsymbol{y}}_{k'}$ as inputs to the denoiser for the next timestep via DDIM, which can be formulated as:

$$\boldsymbol{y}_{k'} = \sqrt{\alpha_{k'}} \cdot \tilde{\boldsymbol{y}}_0 + \sqrt{1-\alpha_{k'}-\sigma_k^2} \cdot \epsilon_k + \sigma_k\epsilon, \tag{7}$$

where k, k' are the current and next timesteps, respectively. The initial $k = K$. $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise independent of $\boldsymbol{y}_0$ and

$$\epsilon_k = (\boldsymbol{y}_k - \sqrt{\alpha_k} \cdot \tilde{\boldsymbol{y}}_0)/\sqrt{1-\alpha_k}, \tag{8}$$

$$\sigma_k = \sqrt{(1-\alpha_{k'})/(1-\alpha_k)} \cdot \sqrt{1-\alpha_k/\alpha_{k'}}, \tag{9}$$

where ϵ_k is the noise at timestep k. σ_k controls how stochastic the diffusion process is. Then, we can regenerate $\tilde{\boldsymbol{y}}_0$ using $\boldsymbol{y}_{k'}$ as inputs to the denoiser. This procedure will be iterated S times. Since we start from L at the beginning, the timestep of each iteration can be written as $k = K \cdot (1 - s/S)$, $s \in [0, S)$. This process gradually adds noise to the data with high probability.

2.3. Alternated Spatial-Temporal Denoiser

Due to the promising information interaction and global aggregation capabilities of Transformers [29, 30, 31, 32, 33, 34, 35, 36], we implement the Alternated Spatial-Temporal Denoiser (ASTDenoiser) $\mathscr{D}$ using a Transformer-like architecture. Specifically, assuming the input of l^{th} layer is $z^l \in \mathbb{R}^{T\times N\times D}$. In

the spatial transformer, we obtain spatial relationship $z^l_{spatial} \in \mathbb{R}^{T\times N\times}$ by directly applying multi-head self-attention (MSA) to D vectors of size N:

$$\hat{z}^l_{spatial} = \text{LayerNorm}(z^l + \text{MSA}(z^l, z^l, z^l)), \tag{10}$$

$$z^l_{spatial} = \text{LayerNorm}(\hat{z}^l_{spatial} + \text{FeedForward}(\hat{z}^l_{spatial})), \tag{11}$$

where $\hat{z}^l_{spatial}$ is intermediate variable. LayerNorm$(\cdot)$ denotes layer normalization, FeedForward denotes a multi-layer feedforward network, MSA$(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes the multi-head self-attention layer where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ serve as queries, keys and values. In the temporal transformer, we obtain temporal relationship $z^l_{temporal} \in \mathbb{R}^{N\times T\times D}$ by applying MSA to D vectors of size T:

$$\hat{z}^l_{temporal} = \text{LayerNorm}((z^l_{spatial})^{\text{T}} + \text{MSA}((z^l_{spatial})^{\text{T}}, (z^l_{spatial})^{\text{T}}, (z^l_{spatial})^{\text{T}})), \tag{12}$$

$$z^l_{temporal} = \text{LayerNorm}(\hat{z}^l_{temporal} + \text{FeedForward}(\hat{z}^l_{temporal})), \tag{13}$$

where $\hat{z}^l_{temporal}$ is intermediate variable. $(\cdot)^{\text{T}}$ denotes the transposition of matrices. Finally, we obtain the input of $(l+1)^{th}$ layer $z^{l+1} = z^l_{temporal}$. After L layer of the denoiser loop, a linear regression head is built to estimate the rPPG signal $\boldsymbol{y} \in \mathbb{R}^T$.

2.4. HR distribution Augmentation

Existing physiological measurement datasets are typically limited in scale and suffer from imbalanced HR distributions, with most samples concentrated in $[60, 90]$ beats per minute (bpm). This results in models overfitting to this dominant HR range while neglecting the tail ends of the HR range, particularly low and high HR values. However, normal human HR varies from approximately 40 to 180 bpm. To address this issue, we propose an effective data augmentation strategy that synthetically expands the HR distribution by directly applying upsampling and downsampling operations on MSTmaps.

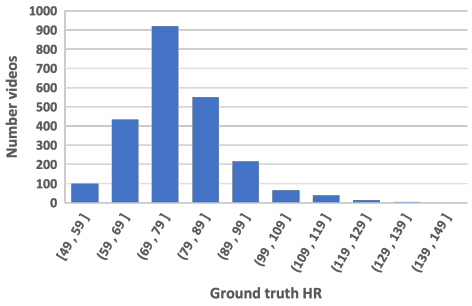

Figure 2: The ground truth HR distribution of the VIPL-HR dataset.

Specifically, unlike traditional image-based data augmentation conducted in the spatial domain, our method leverages the fact that HR estimation depends on the frequency of subtle color changes in facial videos. Therefore, we manipulate the temporal dimension of the MSTmap by applying downsampling or upsampling to adjust the frequency of these color variations, effectively simulating higher or lower HRs compared to the original video, and generating new MSTmaps for model training. As shown in Figure 2, we analyze the HR distribution in the large-scale VIPL-HR dataset and observe that most ground-truth HRs fall between 60 and 90 bpm, with only a small fraction of samples having HRs below 60 bpm or above 90 bpm.

Based on this observation, we perform the following operations to augment the original data. For MSTmaps with ground-truth HR below 80 bpm, we create a copy and concatenate it with the original along the temporal dimension, resulting in a 600-frame MSTmap. We then apply a downsampling operation with a sampling rate of 0.5 along the temporal axis to generate a new 300-frame MSTmap. This effectively doubles the HR, as the frequency of temporal variation is doubled $\text{HR}_{\text{aug}} = 2 \times \text{HR}_{\text{orig}}$. Following the same principle, for MSTmaps with ground-truth HR above 80 bpm, we first extract the initial 150 frames, then apply an upsampling operation with a sampling rate of 2 along the temporal axis to obtain a new 300-frame MSTmap. Consequently, the HR is halved $\text{HR}_{\text{aug}} = 0.5 \times \text{HR}_{\text{orig}}$. This data augmentation method enables us to construct a more diverse dataset with a broader HR distribution, covering both low and high HR ranges, thereby balancing the overall HR distribution.

3. Experiments

3.1. Experimental Setup

Datasets. Following the challenge requirements, we adopt the VIPL-HR dataset [37] to train and evaluate our model. VIPL-HR is a large-scale, multimodal, and challenging dataset for rPPG estimation. It contains recordings of 107 subjects under nine different scenarios, incorporating diverse head movements (e.g., stable, talking, and motion) and lighting conditions (e.g., lab, dark, bright), captured using three types of devices: Logitech C310 camera, RealSense F200, and smartphones. The dataset includes 3,130 visible-light facial video clips. The test set comprises the VIPL-HR [37] and OBF [2] datasets, each offering paired RGB-NIR videos for 100 subjects not included in the training set. For evaluation, the videos are segmented into 10-second clips. VIPL-HR captures a range of lighting conditions, while OBF features subjects with diverse skin tones.

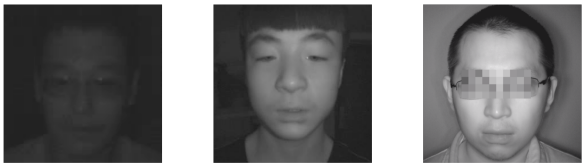

Figure 3: Compared with the NIR video in the test set, the video in the training set is much darker, and even the facial contours are not clear, making it very difficult to capture physiological signals.

Evaluation Metric. The root mean squared error (RMSE) is employed as the evaluation metric to measure the discrepancy between the predicted heart rate y_{pred} and the ground-truth heart rate y_{gt}.

3.2. Implementation Details

During preprocessing, facial region-of-interest (ROI) areas are extracted using the landmark detection module from OpenFace. Following the setup in [23], we generate MSTmaps using a sliding window of 300 frames (equivalent to 10 seconds) with a stride of 15 frames (0.5 seconds). For the alternated spatial-temporal denoiser, we set the feature dimension D to 128 and the number of layers L to 6. The model is trained using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 16. The maximum diffusion step K is set to 1000. The probability of applying heart rate distribution augmentation is set to 0.5. During inference, the number of reverse diffusion steps S is set to 5. To estimate heart rate from the predicted rPPG signal, we follow previous works [3, 4, 26] and apply a first-order Butterworth bandpass filter with a cutoff frequency range of [0.66 Hz, 3.0 Hz], corresponding to a heart rate range of [40, 180] beats per minute.

Table 1: The impact of different training/validation splits and video modalities on VIPL-HR.

Training	Validation	Modal	RMSE↓ (bpm)	Modal	RMSE↓ (bpm)
Folds 2,3,4,5	Fold 1	RGB	12.21363	NIR	17.46341
Folds 1,3,4,5	Fold 2	RGB	12.44439	NIR	18.44521
Folds 1,2,4,5	Fold 3	RGB	12.40643	NIR	15.14771
Folds 1,2,3,5	Fold 4	RGB	12.17062	NIR	16.94501
Folds 1,2,3,4	Fold 5	RGB	**11.89505**	NIR	**14.59949**

3.3. Experimental Results

We conducted experiments on the VIPL-HR dataset to evaluate different training/validation splits and video modalities (RGB and NIR). As shown in Table 1, the best performance was achieved using the training/validation split of Folds 1, 2, 3, 4 for training and Fold 5 for validation. We observed that models trained on RGB videos consistently outperformed those trained on NIR videos. To further investigate this performance gap, we visualized representative NIR videos from both the training and testing sets, as shown in Figure 3. It is evident that the training NIR videos are considerably darker than those in the test set, which severely compromises the already weak rPPG signals and

Table 2: The HR estimation results of the top-3 leaderboards on the test set in the 4th RePSS.

Team Name	Rank	RMSE↓ (bpm)
HFUT-VUT (Ours)	1	**11.89505**
IST	2	12.31846
xjgroupscu	3	12.70790

ultimately degrades model performance. Due to the significant distributional discrepancy between the training and testing NIR videos, we chose to utilize only the RGB modality for all experiments. As shown in Table 2, we report the top-3 results on the test dataset in the 4th RePSS challenge. Compared to other teams, we can see that our team achieves 1st place (RMSE of 11.89505 bpm), which is higher than the 2nd place (RMSE of 12.31846 bpm) by 3.4%.

4. Conclusion

We presented **DiffRePSS**, a diffusion-based method for remote physiological signal sensing. By combining a conditional diffusion framework with spatial-temporal representations and HR-aware augmentation, our model achieves robust rPPG estimation under real-world conditions. The proposed approach demonstrates state-of-the-art performance and ranked first in the 4th RePSS Challenge, highlighting the effectiveness of diffusion models in this field.

Acknowledgments

This work is supported by National Key R&D Program of China (NO.2024YFB3311602), Natural Science Foundation of China (62272144), the Anhui Provincial Natural Science Foundation (2408085J040), and the Major Project of Anhui Provincial Science and Technology Breakthrough Program (202423k09020001), and the Fundamental Research Funds for the Central Universities (JZ2024HGTG0309, JZ2024AHST0337).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 in order to: Grammar correction and language refinement. The authors independently reviewed, verified, and revised all generated content, and accept complete responsibility for the scientific validity and intellectual integrity of the publication.

References

[1] 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.

[2] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, G. Zhao, The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 242–249.

[3] W. Qian, D. Guo, K. Li, X. Zhang, X. Tian, X. Yang, M. Wang, Dual-path tokenlearner for remote photoplethysmography-based physiological measurement with facial videos, IEEE Transactions on Computational Social Systems (2024).

[4] Q. Li, D. Guo, W. Qian, X. Tian, X. Sun, H. Zhao, M. Wang, Channel-wise interactive learning for remote heart rate estimation from facial video, IEEE Transactions on Circuits and Systems for Video Technology (2023).

[5] X. Liu, B. Hill, Z. Jiang, S. Patel, D. McDuff, Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5008–5017.

[6] W. Qian, Q. Li, K. Li, X. Wang, X. Sun, M. Wang, D. Guo, Joint spatial-temporal modeling and contrastive learning for self-supervised heart rate measurement, arXiv preprint arXiv:2406.04942 (2024).

[7] D. Guo, K. Li, B. Hu, Y. Zhang, M. Wang, Benchmarking micro-action recognition: Dataset, methods, and applications, IEEE Transactions on Circuits and Systems for Video Technology 34 (2024) 6238–6252.
[8] K. Li, D. Guo, G. Chen, C. Fan, J. Xu, Z. Wu, H. Fan, M. Wang, Prototypical calibrating ambiguous samples for micro-action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 4815–4823.
[9] J. Zhou, X. Shen, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, Y. Zhong, Audio-visual segmentation with semantics, International Journal of Computer Vision (2024) 1–21.
[10] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, Y. Zhong, Audio–visual segmentation, in: European Conference on Computer Vision (ECCV), 2022, pp. 386–403.
[11] J. Zhou, L. Zheng, Y. Zhong, S. Hao, M. Wang, Positive sample propagation along the audio-visual event line, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8436–8444.
[12] J. Zhou, D. Guo, Y. Mao, Y. Zhong, X. Chang, M. Wang, Label-anticipated event disentanglement for audio-visual video parsing, in: European Conference on Computer Vision (ECCV), 2024, pp. 1–22.
[13] Z. Li, D. Guo, J. Zhou, J. Zhang, M. Wang, Object-aware adaptive-positivity learning for audio-visual question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024, pp. 3306–3314.
[14] Z. Li, J. Zhou, J. Zhang, S. Tang, K. Li, D. Guo, Patch-level sounding object tracking for audio-visual question answering, arXiv preprint arXiv:2412.10749 (2024).
[15] R. Špetlík, V. Franc, J. Matas, Visual heart rate estimation with convolutional neural network, in: Proceedings of the British Machine Vision Conference, 2018, pp. 3–6.
[16] 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.
[17] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE Transactions on Biomedical Engineering 60 (2013) 2878–2886.
[18] G. De Haan, A. Van Leest, Improved motion robustness of remote-ppg by using the blood volume pulse signature, Physiological Measurement 35 (2014) 1913–1913.
[19] 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.
[20] W. Chen, D. McDuff, Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 349–365.
[21] Z. Yu, W. Peng, X. Li, X. Hong, G. Zhao, 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, 2019, pp. 151–160.
[22] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. 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.
[23] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, G. Zhao, Video-based remote physiological measurement via cross-verified feature disentangling, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 295–310.
[24] Z. Sun, X. Li, Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast, in: Proceedings of the European Conference on Computer Vision, 2022, pp. 492–510.
[25] Z. Sun, X. Li, Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 1–18.
[26] W. Qian, K. Li, D. Guo, B. Hu, M. Wang, Cluster-phys: Facial clues clustering towards efficient

remote physiological measurement, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, p. 330–339.
[27] W. Qian, G. Su, D. Guo, J. Zhou, X. Li, B. Hu, S. Tang, M. Wang, Physdiff: Physiology-based dynamicity disentangled diffusion model for remote physiological measurement, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 6568–6576.
[28] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in neural information processing systems 33 (2020) 6840–6851.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).
[30] K. Li, J. Li, D. Guo, X. Yang, M. Wang, Transformer-based visual grounding with cross-modality interaction, ACM Transactions on Multimedia Computing, Communications and Applications 19 (2023) 1–19.
[31] K. Li, D. Guo, M. Wang, Vigt: proposal-free video grounding with a learnable token in the transformer, Science China Information Sciences 66 (2023) 202102.
[32] K. Li, P. Liu, D. Guo, F. Wang, Z. Wu, H. Fan, M. Wang, Mmad: Multi-label micro-action detection in videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 13225–13236.
[33] F. Wang, D. Guo, K. Li, M. Wang, Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 5345–5353.
[34] J. Gu, K. Li, F. Wang, Y. Wei, Z. Wu, h. Fan, M. Wang, Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition, in: Proceedings of the 33rd ACM International Conference on Multimedia, 2025, p. 5461–5470.
[35] K. Li, X. Peng, D. Guo, X. Yang, M. Wang, Repetitive action counting with hybrid temporal relation modeling, IEEE Transactions on Multimedia (2025).
[36] Y. Zhu, K. Li, Z. Yang, Exploiting efficientsam and temporal coherence for audio-visual segmentation, IEEE Transactions on Multimedia 27 (2025) 2999–3008.
[37] 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.