

Spatial Alignment and Temporal Matching Adapter for Video-Radar Remote Physiological Measurement

Qian Liang¹, Ruixu Geng¹, Jinbo Chen², Haoyu Wang¹, Yan Chen¹, Yang Hu^{1*}

¹University of Science and Technology of China, ²Nanyang Technological University

{qianliang, gengruixu, why1999}@mail.ustc.edu.cn, jinbo.chen@ntu.edu.sg

{eecyan, eeyhu}@ustc.edu.cn

Abstract

Remote physiological measurement (RPM) based on video and radar has made significant progress in recent years. However, unimodal methods based solely on video or radar sensor have notable limitations due to their measurement principles, and multimodal RPM that combines these modalities has emerged as a promising direction. Despite its potential, the lack of large-scale multimodal data and the significant modality gap between video and radar pose substantial challenges in building robust video-radar RPM models. To handle these problems, we suggest leveraging unimodal pre-training and present the Spatial alignment and Temporal Matching (SATM) Adapter to effectively fine-tune pre-trained unimodal backbones into a multimodal RPM model. Given the distinct measurement principles of video- and radar-based methods, we propose Spatial Alignment to align the spatial distribution of their features. Furthermore, Temporal Matching is applied to mitigate waveform discrepancies between video and radar signals. By integrating these two modules into adapters, the unimodal backbones could retain their modality-specific knowledge while effectively extracting complementary features from each other. Extensive experiments across various challenging scenarios, including low light conditions and head motions, demonstrate that our approach significantly surpasses the state-of-the-art methods.

1. Introduction

Remote physiological measurement has attracted a lot of attention recently owing to its ability to detect physiological signals without contact. A prominent line of research focuses on video-based remote photoplethysmography (rPPG), which uses ordinary cameras to extract vital signals by detecting periodic changes in skin color induced by heartbeats [4, 10, 24, 29, 49, 51]. Beyond video-based

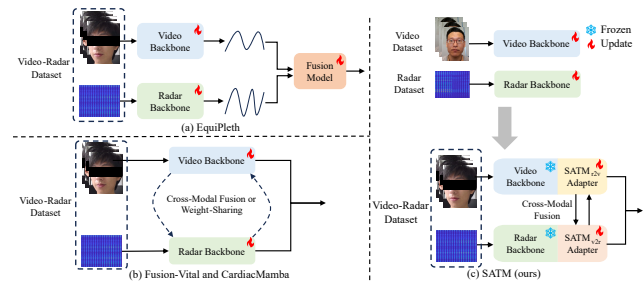


Figure 1. **Frameworks of video-radar RPM methods:** (a) EquiPleth [41] trains unimodal backbones on the multimodal dataset and leverage a late fusion model to ensemble their predictions. (b) Fusion-Vital [6] and CardiacMamba [46] train their end-to-end models from scratch on the multimodal dataset. (c) Our method makes full use of unimodal pre-training and propose the SATM adapter to enable effective multimodal adaptation.

methods, an alternative approach involves the use of radar sensors to capture cardiac activities through subtle mechanical motions of the human chest [3, 35, 42, 43].

However, both video- and radar-based methods have notable limitations. Video-based methods are highly sensitive to light conditions and may be influenced by subjects' skin tones [41]. On the other hand, radar-based methods are much more susceptible to subject movements, and most existing methods are designed for strict-constraint scenarios. Therefore, it is necessary and promising to combine video and radar for RPM measurement to overcome the limitations of unimodal approaches.

Nevertheless, constructing a robust video-radar RPM model remains a significant challenge. Due to the complexity of recording synchronized videos, radar data, and ground-truth PPG signals, video-radar RPM datasets are mostly limited in scale. Consequently, training a model from scratch on such datasets is prone to severe overfitting. Unfortunately, existing multimodal RPM works [6, 41, 46] overlook this problem, leading to limited robustness and generalization ability (See Sec. 4.3). In addition, video-

*Corresponding author

based methods focus primarily on **color changes of facial skin**, whereas radar-based methods capture **mechanical displacements of the chest**. This implies that these two modalities capture distinct signals from different regions of interest (ROIs), introducing a substantial modality gap and making it difficult to effectively fuse their features.

For the data scarcity challenge, we suggest making full use of unimodal pre-training, as unimodal data are easier to collect and large-scale unimodal datasets for remote physiological measurement are already available [31, 40, 53]. As illustrated in Fig. 1 (c), we first utilize large-scale unimodal datasets to pre-train our backbones. Subsequently, instead of fully fine-tuning the entire network on limited multimodal data, we propose leveraging cross-modal adapters to fine-tune unimodal backbones into a multimodal model while preserving the pre-trained knowledge. Several previous studies have introduced adapter techniques into multimodal learning [2, 28, 48]. For instance, UniAdapter [28] and MMA [48] design parameter-sharing adapters to facilitate modality alignment in vision-language models, while BAT [2] proposes a bi-directional adapter to directly extract cross-modal features for multimodal tracking. However, multimodal adapter for video-radar RPM has not been explored before. Furthermore, the aforementioned methods are suboptimal for multimodal RPM for two reasons: (1) crude modality alignment would hamper the learning of modality-specific features in this context, and (2) the significant modality gap between video and radar, as discussed earlier, impedes the direct fusion of their features.

To this end, we propose a novel SATM adapter to enable effective multimodal fine-tuning on the video-radar RPM dataset. Specifically, the Spatial Alignment (SA) and Temporal Matching (TM) modules are devised to manage the large modality gap between video and radar. To mitigate the gap introduced by their distinct ROIs, Spatial Alignment projects latent features from different modalities into a unified time series space and then utilizes cross-attention to align their spatial distributions. Additionally, to handle the gap caused by different waveforms of detected signals from video and radar, Temporal Matching further applies a matched filter to refine the spatially aligned features along the temporal dimension. These two modules narrow the modality gap collaboratively and are integrated into adapters to effectively extract cross-modal features. To validate the effectiveness of our approach, we collect a challenging dataset, MMRPM, which includes diverse light conditions and subject head motions. Comprehensive experiments conducted on the MMRPM and EquiPleth [41] datasets demonstrate the superiority of our method.

The key contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to highlight the dilemma of data hungry and data scarcity in

multimodal RPM. We propose leveraging unimodal pre-training and multimodal adapter to tackle this problem.

- To handle the significant modality gap between video and radar, we design the Spatial Alignment and Temporal Matching modules. These modules are further integrated into a novel SATM adapter, enabling effective multimodal fine-tuning while mitigating overfitting.
- Extensive experiments across various challenging scenarios show the superiority of the proposed method compared to existing approaches.

2. Related Works

2.1. Unimodal Remote Physiological Measurement

Most existing remote physiological measurement methods focus on video and radar sensors. Given that the optical absorption of hemoglobin molecules fluctuates along with the cardiac cycle, physiological signals can be remotely measured by detecting changes in skin color in RGB videos. Early works [7, 8, 36, 45] mainly utilize blind source separation and color space transformations to enhance the signal-to-noise ratio (SNR) of extracted rPPG signals. However, these methods are based on restrictive assumptions and struggle to handle complex scenarios. With the development of deep learning, researchers have employed 3D-CNN or modified vision transformer (ViT) to extract rPPG signals from raw videos [49, 51] or specifically designed STMaps [27, 32]. In addition to RGB videos, several studies have also explored rPPG measurement using near-infrared and thermal-infrared videos [5, 15, 34].

On the other hand, radar sensors can extract vital signals by detecting subtle movements of the human chest induced by physiological activities. [11] combines 4D beamforming and CNN to capture seismocardiography from radar sensors. [3] leverages the connection between cardiac mechanical activity and electrical activity to introduce the first radar-based electrocardiography (ECG) monitoring framework. [43] proposes a spatio-temporal network for radar-based Heart Rate Variability (HRV) detection. However, due to the minuscule scale of cardiac motions, radar-based physiological measurement often requires subjects to remain in a supine position and keep still, limiting its flexibility for real-world applications.

2.2. Video-Radar RPM Measurement

Given the complementary properties of vision and radar modalities, vision-radar fusion has been widely explored in tasks such as object detection [17, 23, 26, 30] and depth estimation [21, 25, 37]. In contrast, video-radar RPM measurement remains an understudied area, with only three related works reported recently [6, 41, 46]. EquiPleth [41] focuses on the skin tone bias problem of video-based methods. They design a late fusion model to ensemble predic-

tions from video and radar backbones and propose an adversarial training strategy to enhance the fairness of the final predictions. Fusion-Vital [6] first projects video and radar inputs into a shared time-difference domain and then employs attention mechanism for cross-modal fusion, showcasing comparable performance even in modality corruption scenarios. Similarly, CardiacMamba [46] adopts a weight-sharing VisionMamba [20] to develop an end-to-end video-radar RPM model. However, these works train their models from scratch on small-scale video-radar datasets, making them prone to overfitting the training sets. To overcome this limitation, we propose leveraging the relatively large-scale unimodal datasets for pre-training and design the SATM adapter to facilitate effective multimodal fine-tuning while preserving pre-trained knowledge.

2.3. Adapter Fine-Tuning

The increasing scale of foundation models has brought about the challenge of prohibitive cost and substantial data requirements to fully fine-tune these models. Therefore, parameter-efficient fine-tuning (PEFT) has attracted growing attention [12, 13, 16, 52]. Adapter [12], comprising a pair of projection matrices and a nonlinear activation function, is one of the most representative PEFT methods. Although originally developed for natural language processing (NLP), recent works have extended its application to multimodal settings [14, 28, 44, 48]. [28, 48] leverage parameter-sharing adapters to fine-tune vision-language models and promote modality alignment, which is particularly crucial in vision-text tasks. [2] proposes a bi-directional adapter to foster modality complementarity in multimodal tracking. Considering cross-attention is an effective, yet computationally inefficient mechanism for cross-modal fusion, [22] utilizes latent tokens to efficiently integrate cross-attention into adapters. However, these existing approaches are not well-suited for video-radar RPM. This is primarily because naive modality alignment would hinder the learning of modality-specific features, and the substantial modality gap between video and radar further impedes direct fusion of their features. To this end, we present the SATM adapter, specifically tailored for video-radar RPM fine-tuning.

3. Method

In this paper, we focus on constructing a robust multimodal RPM model on the video-radar RPM dataset. To this end, we first leverage existing large-scale unimodal datasets to pre-train our backbones. Afterwards, we freeze the backbones to preserve pre-trained knowledge and employ both vanilla and SATM adapters to fine-tune them into a video-radar RPM model. The proposed SATM adapter consists of two core components, Spatial Alignment and Temporal Matching, which can effectively narrow the modality gap and extract cross-modal features. The general framework

for video-radar fine-tuning is depicted in Fig. 2. We will present our method in more detail in the following parts of this section.

3.1. Backbone Design

3.1.1. Video Backbone

Following [19], we utilize the combination of DiffNorm and UniFormer [18] as our video backbone. DiffNorm is a lightweight module that utilizes difference operation and normalization to efficiently integrate the appearance and dynamic features of input videos. UniFormer is an enhanced version of the Vision Transformer (ViT) designed for video analysis. Each UniFormer block consists of Dynamic Position Embedding (DPE), Multi-Head Relation Aggregator (MHRA), and Feed Forward Network (FFN). MHRA is the core component that unifies convolution and attention in a concise transformer format. It employs convolution to tackle feature redundancy in shallow layers while utilizing self-attention for global modeling in deep layers.

Formally, given an input video $X_v \in \mathbb{R}^{3 \times T_1 \times H \times W}$, the DiffNorm module first transforms it into $F_v^{(0)} \in \mathbb{R}^{3 \times T_1 \times H \times W}$ as the input of UniFormer. Then the forward process of a UniFormer block can be expressed as:

$$\begin{aligned} Y_v^{(l)} &= \text{DPE}(F_v^{(l-1)}) + F_v^{(l-1)} \\ Z_v^{(l)} &= \text{MHRA}(\text{Norm}(Y_v^{(l)})) + Y_v^{(l)} \\ F_v^{(l)} &= \text{FFN}(\text{Norm}(Z_v^{(l)})) + Z_v^{(l)} \end{aligned} \quad (1)$$

where $F_v^{(l-1)}$ and $F_v^{(l)}$ are the output of the $l-1$ -th and the l -th UniFormer block, respectively. We will omit the normalization layers in the rest of the sections for brevity.

3.1.2. Radar Preprocessing and Backbone

The fundamental principle of radar-based physiological measurement involves capturing the subtle chest movement caused by cardiac activity. To extract displacement information from the radar intermediate frequency (IF) signal, we first perform range-FFT and coarse range gating to generate a range-time matrix and filter out irrelevant reflections. Afterwards, given that the 2D antenna array enables signal sampling in the spatial domain, we leverage beamforming to separate the reflections from different spatial positions:

$$\begin{aligned} B(\theta_A, \theta_E) &= \mathbf{W}^H(\theta_A, \theta_E) \mathbf{M} \\ \mathbf{W}(\theta_A, \theta_E) &= [\dots, \exp(-j\mathbf{k}(\theta_A, \theta_E)\mathbf{d}_r^{(i)}), \dots]^T \\ \mathbf{k}(\theta_A, \theta_E) &= \frac{2\pi}{\lambda} [\cos \theta_E \cos \theta_A, \cos \theta_E \sin \theta_A, \sin \theta_E] \end{aligned} \quad (2)$$

where $\mathbf{M} \in \mathbb{R}^{N_C \times R \times T_2}$ is the range-time matrix, with N_C, R, T_2 indicating the number of virtual channels, range bins and frames, respectively. $\mathbf{W}(\theta_A, \theta_E) \in \mathbb{R}^{N_C}$ is the steering vector in the direction of azimuth angle θ_A and elevation angle θ_E , and $\mathbf{d}_r^{(i)}$ is the position of the i -th virtual channel. The shape of the beamformed output B is

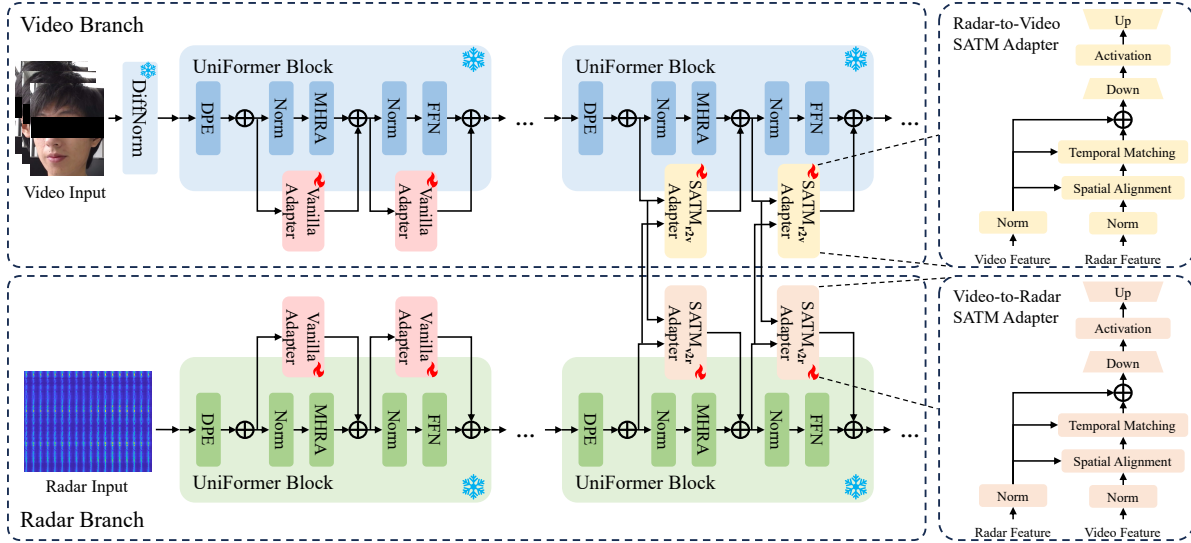


Figure 2. **The overall framework of our model:** Both backbones are based on UniFormer to facilitate cross-modal fusion. Vanilla adapters and SATM adapters are inserted into the shallow and deep layers of UniFormers, respectively, for multimodal fine-tuning. Each SATM adapter comprises two key components: Spatial Alignment and Temporal Matching. The Spatial Alignment module serves to align the spatial distributions of video and radar features, while the Temporal Matching module further refines the aligned features along the temporal dimension. Finally, the refined cross-modal features are passed through a vanilla adapter and added to the corresponding backbone as complementary features.

$(\Theta_A \times \Theta_E) \times R \times T_2$, where $(\Theta_A \times \Theta_E)$ is the number of angular bins. Following [3, 43], we further utilize a second-order differentiator to amplify cardiac movements while suppressing noise signals as follows:

$$f_{r,t} = (\psi_{t-3} + \psi_{t+3}) + 2(\psi_{t-2} + \psi_{t+2}) - (\psi_{t-1} + \psi_{t+1}) - 4\psi_t \quad (3)$$

where ψ_t is the phase of \mathbf{B} at the t -th frame. The output $F_r^{(0)} \in \mathbb{R}^{(\Theta_A \times \Theta_E) \times R \times T_2}$ is reshaped into $\mathbb{R}^{1 \times S \times T_2}$ to serve as the input of the radar backbone, where $S = \Theta_A \times \Theta_E \times R$ is the spatial dimension of the radar modality.

Considering the necessity to model the mechanical motion of the human chest across both spatial and temporal dimensions [43], we adopt 2D UniFormer as our radar backbone. Overall, our video and radar backbones are symmetrically structured, which greatly facilitates cross-modal fusion.

3.2. SATM Adapter

To enable effectively fine-tuning of the unimodal backbones and facilitate cross-modal modeling, we propose the SATM adapter. As illustrated in Fig. 2, Radar-to-Video SATM adapters and Video-to-Radar SATM adapters are inserted in parallel with the MHRA and FFN modules of the video and radar UniFormers respectively to extract complementary features from the other modality. As discussed in Sec. 1, the fundamental differences in measurement principles between video- and radar-based methods introduce

a significant modality gap between their features and implicate cross-modal fusion. Considering the different ROIs of video and radar (face vs. chest), we present the Spatial Alignment (SA) module to align the spatial distributions of their features. As for the differences in captured signals (color changes vs. mechanical displacements), we design the Temporal Matching (TM) module to reduce waveform discrepancies and refine the spatially aligned features. Next, we will provide a detailed explanation of these modules.

3.2.1. Spatial Alignment

The physiological information contained in the face and chest is inherently distributed in distinct forms, resulting in different spatial distributions of video and radar features. We would like to emphasize that this dissimilarity is not merely reflected in the different spatial dimensions of F_v and F_r (HW vs. S), and naively aligning their size through interpolation or linear projection is insufficient.

To address this challenge, we design the Spatial Alignment module. Given the video feature $F_v \in \mathbb{R}^{C \times t_1 \times h \times w}$ and radar feature $F_r \in \mathbb{R}^{C \times s \times t_2}$, the core idea involves utilizing cross-attention to align the spatial distributions of their spatial tokens, *i.e.* $f_v \in \mathbb{R}^{C \times t_1}$ and $f_r \in \mathbb{R}^{C \times t_2}$. Without loss of generality, we use the Radar-to-Video Spatial Alignment as an example for explanation, with its overall pipeline illustrated in Fig. 3 (a). First, bilinear interpolation is applied to align the temporal dimensions of F_v and F_r . The primary challenge then lies in effectively transforming the 2D spatial tokens for the calculation of the attention

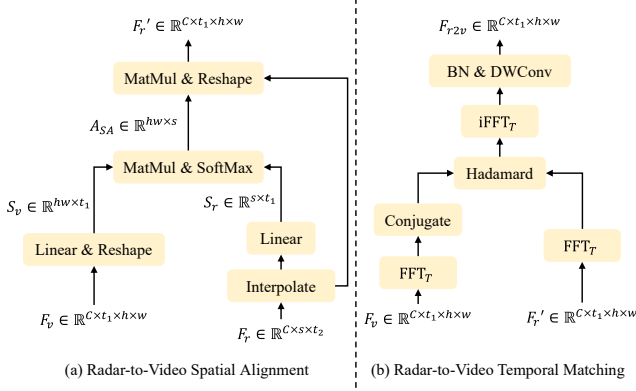


Figure 3. The pipelines of Radar-to-Video Spatial Alignment and Temporal Matching. “MatMul” and “Hadamard” denote matrix product and Hadamard product, respectively. FFT_T represents FFT along the temporal dimension.

scores. Considering that the final predicted 1D PPG signals are generated by the integration of the spatial and channel dimensions, we propose treating spatial tokens as distinct candidates for PPG signal prediction at different spatial locations and then projecting them from the 2D space into a unified 1D time series space by compressing their channel dimensions. Specifically, we leverage linear projections to transform F_v and F_r into the query matrix $S_v \in \mathbb{R}^{hw \times t_1}$ and the key matrix $S_r \in \mathbb{R}^{s \times t_1}$, respectively. Afterwards, the alignment between video and radar spatial tokens can be performed in the time series space and the resulting aligned radar features are computed as follows:

$$\text{SA}_{r2v}(F_v, F_r) = \text{Softmax}\left(\frac{S_v S_r^T}{\sqrt{t_1}}\right) F_r \quad (4)$$

Note that the matrix multiplication between the attention score and F_r is performed channel-wise.

Now let us revisit the way to transform the spatial tokens. Two alternative approaches involve compressing the temporal dimension or flattening the channel and temporal dimensions of spatial tokens. However, the former variant risks irreversible information loss due to excessive compression of temporal signals, which is paramount for the prediction of PPG signals. The second approach, while straightforward, fails to account for the inherent differences between the channel and temporal dimensions. In contrast, our spatial alignment, which operates in the time series space, adheres to the paradigm of PPG signal regression and has the advantage of preserving temporal information. A quantitative comparison is provided in Sec. 4.4.

3.2.2. Temporal Matching

Although video and radar features are temporally synchronized and ideally encompass consistent frequency information, *i.e.* heart rate, the inherent disparity between wave-

forms derived from skin color changes and chest mechanical movements introduces a significant modality gap in the temporal dimension. To this end, we further utilize Temporal Matching to refine the cross-modal features. According to the Matched Filter theory, given an input signal $r(t)$ composed of the target signal $s(t)$ and the noise $n(t)$, an output signal with the highest SNR at time step T could be obtained by applying the filter $h(t) = s(T - t)$. Inspired by this, we formulate the spatially aligned features as $r(t)$ and the waveform discrepancy as $n(t)$, and the refined feature could be obtained through matched filtering. Specifically, given the video feature $F_v \in \mathbb{R}^{C \times t_1 \times h \times w}$ and the spatially aligned radar feature $F'_r \in \mathbb{R}^{C \times t_1 \times h \times w}$, the refined radar-to-video feature can be computed by:

$$\text{TM}_{r2v}(F_v, F'_r) = \text{DWConv}(\text{BN}(\mathcal{F}_T^{-1}(\overline{\mathcal{F}_T(F_v)} \mathcal{F}_T(F'_r)))) \quad (5)$$

where $\mathcal{F}_T(\cdot)$ denotes Fast Fourier Transform (FFT) along the temporal dimension and $\overline{\mathcal{F}_T(\cdot)}$ represents FFT followed by the conjugate operation. DWConv and BN are depth-wise convolution and batch normalization, respectively.

3.2.3. Cross-Modal Fusion

Overall, the forward process of the Radar-to-Video SATM adapter can be represented as:

$$\begin{aligned} F_{r2v} &= \text{TM}_{r2v}(F_v, \text{SA}_{r2v}(F_v, F_r)) \\ \text{SATM}_{r2v}(F_v, F_r) &= W_{up}(\text{ReLU}(W_{down}(F_v + \alpha \cdot F_{r2v}))) \end{aligned} \quad (6)$$

where W_{down} , W_{up} are the down-projection and up-projection layers of the vanilla adapter [12] and α is a learnable scaling factor. Afterward, the forward process of a video UniFormer block can be reformulated as:

$$\begin{aligned} Z_v^{(l)} &= \text{MHRA}(Y_v^{(l)}) + Y_v^{(l)} + \text{SATM}_{r2v}(Y_v^{(l)}, Y_r^{(l)}) \\ F_v^{(l)} &= \text{FFN}(Z_v^{(l)}) + Z_v^{(l)} + \text{SATM}_{r2v}(Z_v^{(l)}, Z_r^{(l)}) \end{aligned} \quad (7)$$

The forward processes of the Video-to-Radar SATM adapter and the fine-tuned radar UniFormer blocks can also be computed similarly to Eq. (6) and Eq. (7).

In summary, Spatial Alignment aligns the spatial distributions of video and radar features in a unified time series space, while Temporal Matching utilizes a matched filter to handle the waveform difference between their signals. They complement each other by processing the cross-modal features along spatial and temporal dimensions separately, thereby effectively narrowing the substantial modality gap between video and radar features. According to [9], it is essential to balance the learning of unimodal and cross-modal features. Consequently, we only insert our SATM adapters into the last two stages of UniFormers while utilizing vanilla adapters in the first two stages. The output features of both backbones are integrated by a lightweight fusion stem and subsequently passed through a regression module to generate the predicted PPG signal.

4. Experiments

4.1. Dataset

We utilize the EquiPleth [41] dataset to evaluate our method, which contains 533 30-second samples collected from 28 light, 49 medium, and 14 dark skin tone volunteers. All samples are recorded under motionless conditions and with sufficient illumination.

In addition, considering that the EquiPleth dataset mainly focuses on constrained scenarios and shows limited practical diversity, we built a new dataset called MMRPM to promote a more comprehensive evaluation of our method. MMRPM comprises 180 60-second samples from 31 subjects, all of which are recorded under varying illumination conditions and subject motions.

Specifically, MMRPM encompasses three illumination conditions: dark, medium, and light. The “dark” and “medium” settings correspond to illumination levels of approximately 2.5 and 6.3 lux respectively, while “light” represents samples are collected under sufficient sunlight and indoor illumination. Furthermore, two types of subject activities are considered: sitting still and free head movement. Overall, MMRPM covers six scenarios, where low-light conditions are challenging to the video modality and head motions have a greater impact on the radar modality.¹

4.2. Implementation Details

Our proposed method is implemented using PyTorch. For video preprocessing, we first employ MTCNN to crop the face region in the initial frame and fix the region across subsequent frames. Then, we sample each video into clips with a time window of 160 frames and a step size of 80 frames, which are then resized to 96×96 pixels. In terms of radar preprocessing, the angular range selected for beamforming is $66^\circ \leq \Theta_A \leq 114^\circ$, $-24^\circ \leq \Theta_E \leq 24^\circ$.

We utilize UniFormer-S [18] and MAM in MAGIC [55] as our feature extractor and fusion stem respectively. As for the regression module, we follow the design in ADDP [19]. The video backbone is jointly pre-trained on VIPL [31, 32], MMPD [40], UBFC [1], PURE [39] and BUAA [47], and we use the dataset introduced in [53] for radar pre-training.

During multimodal fine-tuning, both backbones are frozen, and only the adapters, fusion stem and regression module are trained. We employ the loss functions introduced in [51] and utilize the Adam optimizer for training, with the learning rate and weight decay set to $5e-4$ and $5e-5$ respectively. The model is trained for 40 epochs with a batch size of 2. Following [38], we apply random horizontal flipping, spatially resized crop and random intensity noise to augment the input videos. For radar augmentation, we randomly shift the range bins of the range-time matrices.

¹For more details about our MMRPM dataset and pre-training, please refer to the supplemental materials.

Following [41], we perform our evaluation of the PPG signals over 300 samples windows with a stride of 128 samples. The heart rate corresponding to a specific signal clip is calculated using a combination of band-pass filtering, de-trending and peak detection in the frequency domain. Finally, the standard deviation of the error (Std), mean absolute error (MAE), root mean square error (RMSE), and Pearson’s correlation coefficient (R) are adopted to evaluate the predicted HR error. The unit of Std, MAE and RMSE is beats per minute (bpm). We split the MMRPM dataset into 5 subject-exclusive folds and use the official split for the EquiPleth dataset.

4.3. Main Results

We first compare our SATM with existing video-radar RPM methods [6, 41, 46] and several state-of-the-art multimodal adapter methods [2, 22, 28, 48] on the EquiPleth [41] dataset. To ensure a fair comparison, we provide extra results of improved versions of existing RPM methods by applying unimodal pre-training and replacing their original backbones with UniFormers. Since the radar used in the EquiPleth dataset activates only 1 transmitter and 1 receiver, we do not perform beamforming during radar pre-processing for this dataset. For existing video-radar RPM methods, we use the official implementation of EquiPleth, and re-implement Fusion-Vital and CardiacMamba, as the codes have not been released by the authors. As shown in Tab. 1, all of these methods exhibit RMSEs exceeding 2.9 bpm when trained from scratch. Their performance consistently improves when unimodal pre-training is applied, underscoring the significance of unimodal pre-training when video-radar data is limited. However, they still display MAEs greater than 0.64 bpm, regardless of whether using their original backbones or UniFormers, indicating that their full fine-tuning-based fusion strategies are suboptimal for the small-scale multimodal RPM dataset. In contrast, our method employs SATM adapters for multimodal fine-tuning, effectively leveraging pre-trained knowledge and enabling superior cross-modal fusion.

Furthermore, our SATM also demonstrates superior performance compared to existing multimodal adapter methods. The subpar performance of UniAdapter and MMA suggests that their parameter-sharing strategies are ill-suited for multimodal RPM, as they impede the learning of modality-specific features. To adapt BAT for video-radar RPM, we employ an additional linear projection layer to align the token numbers of the video and radar branches. However, it fails to handle the significant modality gap between video and radar, thereby resulting in an RMSE exceeding 1.65 bpm. While the best baseline, LAVISH, achieves an RMSE of 1.498 bpm, our SATM yields a considerably lower RMSE of 1.242 bpm.

We have also evaluated SATM on our self-collected

Methods	Original Modality	Unimodal Pre-trained	Std↓	MAE↓	RMSE↓	R↑	Trainable Params (M)↓
EquiPleth[41]	Video-Radar	×	3.274	1.063	3.283	0.957	3.95
Fusion-Vital[6]	Video-Radar	×	9.288	3.750	9.297	0.709	715.26
CardiacMamba[46]	Video-Radar	×	2.924	0.901	2.927	0.975	9.93
EquiPleth	Video-Radar	✓	2.979	0.968	2.990	0.965	3.95
Fusion-Vital	Video-Radar	✓	9.121	3.750	9.130	0.735	715.26
CardiacMamba	Video-Radar	✓	2.797	0.868	2.771	0.975	9.93
EquiPleth*	Video-Radar	✓	<u>1.435</u>	0.671	<u>1.436</u>	<u>0.992</u>	45.31
Fusion-Vital*	Video-Radar	✓	2.143	0.650	2.157	0.983	194.61
CardiacMamba*	Video-Radar	✓	2.313	0.744	2.324	0.980	33.62
UniAdapter[28]	Vision-Language	✓	1.653	0.590	1.656	0.990	10.60
MMA[48]	Vision-Language	✓	1.835	0.626	1.842	0.988	11.09
BAT[2]	RGB-TIR	✓	1.679	0.630	1.684	0.990	85.35
LAVISH[22]	Audio-Visual	✓	1.493	<u>0.557</u>	1.498	0.991	10.85
SATM (Ours)	Video-Radar	✓	1.217	0.540	1.242	0.994	10.78

Table 1. Overall performance on the EquiPleth dataset. The best and second best results are in bold and underlined respectively. * denotes that we replace the methods’ original backbones with UniFormers while preserving their fusion strategies. Blue and red indicate the existing video-radar RPM methods and existing adapter methods for other multimodal tasks, respectively.

Methods	Std↓	MAE↓	RMSE↓	R↑
EquiPleth	7.603	2.421	7.603	0.715
Fusion-Vital	9.139	5.165	9.315	0.501
CardiacMamba	6.099	2.160	6.107	0.790
EquiPleth*	5.964	1.739	5.974	0.806
Fusion-Vital*	5.550	1.662	5.589	0.827
CardiacMamba*	5.526	1.843	5.564	0.829
UniAdapter	<u>5.509</u>	1.588	5.510	0.831
MMA	5.694	1.722	5.694	0.819
BAT	6.432	2.151	6.433	0.760
LAVISH	5.660	<u>1.454</u>	<u>5.312</u>	<u>0.844</u>
SATM (Ours)	5.063	1.292	5.074	0.858

Table 2. Overall performance on the MMRPM dataset.

MMRPM dataset, which covers more challenging and diverse scenarios than the EquiPleth dataset. As illustrated in Tab. 2, existing video-radar RPM methods exhibit poor results on this dataset, highlighting their limited capability to handle complex scenarios. In comparison, our SATM consistently outperforms the baseline methods in all metrics.

We further conduct a cross-dataset evaluation by training the models on the EquiPleth dataset and evaluating them on the MMRPM dataset. As shown in Tab. 3, the improved versions of existing multimodal RPM methods exhibit subpar performance compared to adapter-based methods. This is primarily because they rely on full fine-tuning-based fusion strategies and are prone to overfitting small-scale training sets. In contrast, our method, built upon SATM adapters, achieves the best generalization performance and suffers

Methods	Std↓	MAE↓	RMSE↓	R↑
EquiPleth*	8.787	4.256	9.082	0.629
Fusion-Vital*	8.538	4.457	8.897	0.642
CardiacMamba*	9.048	5.038	9.308	0.603
LAVISH	<u>8.026</u>	<u>3.687</u>	<u>8.232</u>	<u>0.712</u>
SATM (Ours)	7.542	3.423	7.656	0.738

Table 3. Results of the cross-dataset evaluation from the EquiPleth dataset to the MMRPM dataset.

minimal performance degradation compared to other baselines. Although it still lags behind the performance in the intra-dataset evaluation, we argue that this is reasonable, given the significant differences between these two datasets.²

4.4. Ablation Study

Key Components. Our SATM adapter mainly consists of two key components: Spatial Alignment and Temporal Matching. In Tab. 4, we conduct an ablation study on the MMRPM dataset to assess the effectiveness of these modules. First, compared to the vanilla adapter (the first row), the adapter with Spatial Alignment reduces the MAE and RMSE by 0.087 bpm and 0.149 bpm respectively. This can be attributed to its capability to facilitate modality complementarity by extracting spatially aligned cross-modal features to the corresponding backbones. The MAE and RMSE can be further reduced to 1.292 bpm and 5.074 bpm when Temporal Matching is inserted, highlighting the necessity

²Details about cross-dataset evaluation are provided the appendix.

SA	TM	Std↓	MAE↓	RMSE↓	R↑
×	×	5.452	1.541	5.453	0.836
✓	×	<u>5.298</u>	<u>1.454</u>	<u>5.304</u>	<u>0.845</u>
✓	✓	5.063	1.292	5.074	0.858

Table 4. Ablation study on the key components. “SA” and “TM” denote Spatial Alignment and Temporal Matching respectively.

Variants	Std↓	MAE↓	RMSE↓	R↑
Interpolate	6.041	1.876	6.052	0.795
Linear	<u>5.579</u>	<u>1.628</u>	<u>5.582</u>	<u>0.828</u>
Channel	6.087	1.807	6.091	0.792
Flatten	5.588	1.693	5.595	0.826
Ours	5.298	1.454	5.304	0.845

Table 5. Ablation study on the design of Spatial Alignment.

to refine the aligned features along the temporal dimension.

Alternative Approaches for Spatial Alignment. We also investigate four alternative approaches for Spatial Alignment. Specifically, “Interpolate” and “Linear” refer to methods that directly align the sizes of video and radar features through interpolation and linear projection instead of cross-attention. The “Channel” and “Flatten” indicate that the alignment is performed in the channel space and on the flattened spatial tokens respectively. As shown in Tab. 5, both “Interpolate” and “Linear” exhibit suboptimal performance, achieving MAEs of 1.876 bpm and 1.628 bpm. These methods only align spatial sizes while neglecting inherent differences in spatial distributions between video and radar features, thereby failing to address the modality gap.

The “Channel” variant compresses the temporal dimensions of spatial tokens for the computation of the attention scores, focusing solely on the overall temporal period while neglecting fine-grained temporal variability. This results in degraded performance with an RMSE exceeding 6 bpm. The “Flatten” variant preserves temporal information, but conflates temporal and channel dimensions, overlooking their intrinsic differences. Therefore, it ultimately yields subpar performance with an MAE of 1.693 bpm and an RMSE of 5.595 bpm. In contrast, Spatial Alignment performed in the time series space effectively avoids temporal information loss and dimensional distortion, showing superior performance compared to these four variants.

Temporal Matching vs. Temporal Attention. Since attention is a straightforward method for cross-modal fusion, we compare our Temporal Matching (TM) with Temporal Attention (TA). Specifically, we replace the matched filtering operation with a cross-attention along the temporal dimension. As shown in Tab. 6, our TM consistently outperforms TA across all metrics. Moreover, since the time complexity of FFT is $\mathcal{O}(N \log N)$, TM is also more com-

Temporal	Std↓	MAE↓	RMSE↓	R↑
Attention	<u>5.155</u>	<u>1.367</u>	<u>5.130</u>	<u>0.853</u>
Matching	5.063	1.292	5.074	0.858

Table 6. Comparison between Temporal Matching and Temporal Attention.

Positions	Std↓	MAE↓	RMSE↓	R↑
SATM-4	5.168	1.386	5.179	0.852
SATM-3	<u>5.107</u>	<u>1.343</u>	<u>5.109</u>	<u>0.857</u>
SATM-2	5.063	1.292	5.074	0.858
SATM-1	5.526	1.614	5.526	0.830

Table 7. Ablation study on the position to add SATM adapters.

putationally efficient compared to the quadratic complexity of TA.

Position to Insert Our SATM Adapters. We also explore the impact of different positions to insert our SATM adapters. We present 4 positions for insertion, formulated as “SATM- n ”, where “ n ” denotes that we use SATM adapters in the last “ n ” stages of the UniFormers and employ vanilla adapters in earlier stages. As reported in Tab. 7, SATM-1 displays an MAE higher than the naive baseline (1.614 bpm vs. 1.541 bpm), indicating that inserting the cross-modal module too late would hinder performance due to the lack of cross-modal interaction. Meanwhile, the performance of SATM-3 and SATM-4 is also limited due to insufficient modality-specific modeling. In contrast, the results demonstrate that SATM-2 strikes a good trade-off between the learning of unimodal and cross-modal features.

5. Conclusion

In this paper, we focus on video-radar RPM measurement. To overcome the data scarcity challenge, we make full use of unimodal pre-training and present the SATM adapter to effectively fine-tune unimodal backbones into a multi-modal RPM model. The SATM adapter leverages Spatial Alignment and Temporal Matching to collaboratively narrow the modality gap between video and radar and effectively extract cross-modal features for both backbones. To thoroughly evaluate our method, we collect a challenging dataset that encompasses diverse illumination conditions and subject motions. Extensive experiments conducted on our self-collected and existing public datasets demonstrate that SATM outperforms existing methods.

Limitations. While SATM effectively constructs a robust video-radar RPM model, it primarily focuses on cross-modal interaction and still relies on vanilla adapters for unimodal modeling during fine-tuning. In the future, we hope to develop a unified framework that harmonizes cross-modal and unimodal modeling.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62172381.

References

- [1] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 6, 1
- [2] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bidirectional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 927–935, 2024. 2, 3, 6, 7
- [3] Jinbo Chen, Dongheng Zhang, Zhi Wu, Fang Zhou, Qibin Sun, and Yan Chen. Contactless electrocardiogram monitoring with millimeter wave radar. *IEEE Transactions on Mobile Computing*, 23(1):270–285, 2022. 1, 2, 4
- [4] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 1
- [5] Li-Wen Chiu, Yang-Ren Chou, Yi-Chiao Wu, and Bing-Fei Wu. Deep-learning based remote photoplethysmography measurement in driving scenarios with color and near-infrared images. *IEEE Transactions on Instrumentation and Measurement*, 2023. 2
- [6] Jae-Ho Choi, Ki-Bong Kang, and Kyung-Tae Kim. Fusion-vital: Video-rf fusion transformer for advanced remote physiological measurement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1344–1352, 2024. 1, 2, 3, 6, 7
- [7] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013. 2
- [8] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 2
- [9] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pages 8632–8656. PMLR, 2023. 5
- [10] Jingda Du, Si-Qi Liu, Bochao Zhang, and Pong C Yuen. Dual-bridging with adversarial noise generation for domain adaptive rppg estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10355–10364, 2023. 1
- [11] Unsoo Ha, Salah Assana, and Fadel Adib. Contactless seismocardiography via deep learning radars. In *Proceedings of the 26th annual international conference on mobile computing and networking*, pages 1–14, 2020. 2
- [12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 3, 5
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3
- [14] Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, XueQing Deng, Peng Wang, Zhao Zhang, Xiaohui Shen, and Jiashi Feng. Mv-adapter: Multimodal video transfer learning for video text retrieval. *arXiv preprint arXiv:2301.07868*, 2023. 3
- [15] Jitesh Joshi and Youngjun Cho. ibvp dataset: Rgb-thermal rppg dataset with high resolution signal quality labels. *Electronics*, 13(7):1334, 2024. 2
- [16] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 3
- [17] Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17615–17626, 2023. 2
- [18] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022. 3, 6, 2
- [19] Qian Liang, Yan Chen, and Yang Hu. Continual learning for remote physiological measurement: Minimize forgetting and simplify inference. In *European conference on computer vision*, pages 126–144. Springer, 2024. 3, 6
- [20] Zhu Lianghui, Liao Bencheng, Zhang Qian, Wang Xinlong, Liu Wenyu, and Wang Xinggong. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv. Org*, pages arXiv–org, 2024. 3
- [21] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10233–10240. IEEE, 2020. 2
- [22] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023. 3, 6, 7
- [23] Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. Rcbvdet: Radar-camera fusion in bird’s eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14928–14937, 2024. 2
- [24] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 1
- [25] Chen-Chou Lo and Patrick Vandewalle. Depth estimation from monocular images and sparse radar using deep ordinal regression network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3343–3347. IEEE, 2021. 2

- [26] Yunfei Long, Abhinav Kumar, Daniel Morris, Xiaoming Liu, Marcos Castro, and Punarjay Chakravarty. Radiant: Radar-image association network for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1808–1816, 2023. 2
- [27] Hao Lu, Hu Han, and S Kevin 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*, pages 12404–12413, 2021. 2
- [28] Haoyu Lu, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. *arXiv preprint arXiv:2302.06605*, 2023. 2, 3, 6, 7
- [29] Hao Lu, Zitong Yu, Xuesong Niu, and Ying-Cong Chen. Neuron structure modeling for generalizable remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18589–18599, 2023. 1
- [30] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2021. 2
- [31] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin 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*, pages 562–576. Springer, 2019. 2, 6, 1
- [32] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 2, 6, 1
- [33] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019. 1
- [34] Soyeon Park, Bo-Kyeong Kim, and Suh-Yeon Dong. Self-supervised rgb-nir fusion video vision transformer framework for rppg estimation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–10, 2022. 2
- [35] Vladimir L Petrović, Milica M Janković, Anita V Lupšić, Veljko R Mihajlović, and Jelena S Popović-Božović. High-accuracy real-time monitoring of heart rate variability using 24 ghz continuous-wave doppler radar. *Ieee Access*, 7: 74721–74733, 2019. 1
- [36] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 2
- [37] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9275–9285, 2023. 2
- [38] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14464–14474, 2023. 6
- [39] Ronny Stricker, Steffen Müller, and Horst-Michael 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*, pages 1056–1062. IEEE, 2014. 6, 1
- [40] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: multi-domain mobile video physiology dataset. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–5. IEEE, 2023. 2, 6, 1
- [41] Alexander Vilesov, Pradyumna Chari, Adnan Armouti, Anirudh Bindiganavale Harish, Kimaya Kulkarni, Ananya Deoghare, Laleh Jalilian, and Achuta Kadambi. Blending camera and 77 ghz radar sensing for equitable, robust plethysmography. *ACM Trans. Graph.*, 41(4):36–1, 2022. 1, 2, 6, 7
- [42] Fengyu Wang, Xiaolu Zeng, Chenshu Wu, Beibei Wang, and KJ Ray Liu. mmhrv: Contactless heart rate variability monitoring using millimeter-wave radio. *IEEE Internet of Things Journal*, 8(22):16623–16636, 2021. 1
- [43] Haoyu Wang, Jinbo Chen, Dongheng Zhang, Zhi Lu, Changwei Wu, Yang Hu, Qibin Sun, and Yan Chen. Contactless radar heart rate variability monitoring via deep spatio-temporal modeling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 111–115. IEEE, 2024. 1, 2, 4
- [44] Kai Wang, Yapeng Tian, and Dimitrios Hatzinakos. Towards efficient audio-visual learners via empowering pre-trained vision transformers with cross-modal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1837–1846, 2024. 3
- [45] Wenjin Wang, Sander Stuijk, and Gerard De Haan. Exploiting spatial redundancy of image sensor for motion robust rppg. *IEEE transactions on Biomedical Engineering*, 62(2): 415–425, 2014. 2
- [46] Zheng Wu, Yiping Xie, Bo Zhao, Jiguang He, Fei Luo, Ning Deng, and Zitong Yu. Cardiacmamba: A multimodal rgb-rf fusion framework with state space models for remote physiological measurement. *arXiv preprint arXiv:2502.13624*, 2025. 1, 2, 3, 6, 7
- [47] Lin Xi, Weihai Chen, Changchen Zhao, Xingming Wu, and Jianhua Wang. Image enhancement for remote photoplethysmography in a low-light environment. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 1–7. IEEE, 2020. 6, 1
- [48] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024. 2, 3, 6, 7
- [49] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from fa-

- cial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 1, 2
- [50] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, 27:1245–1249, 2020. 1
- [51] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196, 2022. 1, 2, 6
- [52] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 3
- [53] Bin-Bin Zhang, Dongheng Zhang, Yadong Li, Zhi Lu, Jinbo Chen, Haoyu Wang, Fang Zhou, Yu Pu, Yang Hu, Li-Kun Ma, et al. Monitoring long-term cardiac activity with contactless radio frequency signals. *Nature Communications*, 15(1):1–11, 2024. 2, 6, 1
- [54] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018. 1
- [55] Xu Zheng, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Centering the value of every modality: Towards efficient and resilient modality-agnostic semantic segmentation. In *European Conference on Computer Vision*, pages 192–212. Springer, 2024. 6