Fusion-Vital: Video-RF Fusion Transformer for Advanced Remote Physiological Measurement

Jae-Ho Choi[1], **Ki-Bong Kang**[2], **Kyung-Tae Kim**[3]
[1]Stanford University, CA, USA
[2]Samsung Electronics, South Korea
[3]Pohang University of Science and Technology, South Korea
jhochoi@stanford.edu, kb131.kang@samsung.com, kkt@postech.ac.kr

Abstract

Remote physiology, which involves monitoring vital signs without the need for physical contact, has great potential for various applications. Current remote physiology methods rely only on a single camera or radio frequency (RF) sensor to capture the microscopic signatures from vital movements. However, our study shows that fusing deep RGB and RF features from both sensor streams can further improve performance. Because these multimodal features are defined in distinct dimensions and have varying contextual importance, the main challenge in the fusion process lies in the effective alignment of them and adaptive integration of features under dynamic scenarios. To address this challenge, we propose a novel vital sensing model, named Fusion-Vital, that combines the RGB and RF modalities through the new introduction of pairwise input formats and transformer-based fusion strategies. We also perform comprehensive experiments based on a newly collected and released remote vital dataset comprising synchronized video-RF sensors, showing the superiority of the fusion approach over the previous single-sensor baselines in various aspects.

Introduction

Human physiological signs, such as respiration and cardiograms, are representative indicators that can directly reflect one's physical and mental conditions. For example, continuous monitoring of human physiology enables an overall diagnosis of general health (Revanur et al. 2021) as well as mental fatigue (e.g., sleep status (Zhao et al. 2017) or stress level (Zhang et al. 2012; McDuff et al. 2016)). Traditionally, cardiopulmonary measurements have relied on information from contact sensors; however, their constraints for direct interaction with the skin precipitate great inconvenience to users, disturbing continuous monitoring in everyday life. To alleviate this discomfort and achieve ubiquitous sensing, recent approaches have focused on remote solutions that can extract human vital signs without the need for physical contact.

General non-contact physiology systems typically exploit the remote photoplethysmography (rPPG) characteristics of a camera: the RGB spectrum of the skin vibrates along with the blood volume pulse (BVP), which directly involves the human vital signs (Wang, Kao, and Hsu 2019; Lu, Han, and Zhou 2021). Despite its potentiality for capturing vital motion, the fundamental weaknesses of RGB-reflected physiology, such as the great variability in surrounding illumination and vulnerability to dark settings, remain open challenges. Moreover, the algorithmic necessity for consistent face tracking (Estepp, Blackford, and Meier 2014; Bobbia et al. 2019; Revanur et al. 2021; Choi, Kang, and Kim 2022) incurs degradation in global motion settings.

An effective solution involves the utilization of alternative sensory systems that maintain functional robustness to the aforementioned issues. A representative alternative is a radio frequency (RF) sensor that measures the radial depth near the chest vibrating in response to the vital cycle of the individual. Unlike video-based physiology, which relies on the RGB pixel intensity as a core information source (Zheng et al. 2020; Park et al. 2019), RF sensors infer radial depth information through periodic transmission and reception of electromagnetic signals, inherently mitigating the influence of surrounding illumination. Nevertheless, RF sensors also suffer from their own disadvantages, such as poor angular resolution that makes them susceptible to lateral motions, or the greater difficulty they pose for data acquisition compared to cameras (Boyer 2011; Choi et al. 2020; Choi, Kim, and Kim 2022). Consequently, most previous RF physiological approaches have depended upon learning-free methods and have been limited to controlled setups (Li and Lin 2018; Mercuri et al. 2018, 2019; Obadi et al. 2022).

In this study, we aim to explore the potential for enhancing cardiopulmonary measurements through the multimodal fusion of video and RF reflections. Our fundamental premise is that the RGB and radio data can serve as complementary information, particularly in terms of physiological monitoring. Video and RF sensors capture human vital signs based on disparate physical signatures (RGB intensity and radial distance) that emanate from different body regions (facial skin and upper front of the body), respectively. More importantly, while cameras can maintain high resolution in the lateral direction but lack depth information (Long et al. 2021b), RF sensors have a fine depth resolution but suffer from poor lateral resolution (Fogle and Rigling 2012; Choi, Kim, and Kim 2021), meaning that the strengths of each sensor can compensate for the weaknesses of the other. To fully exploit such complementarity between RGB and RF

data, we present a novel model, called Fusion-Vital, which combines video and RF reflections for advanced physiological measurement. We first introduce new input modalities for RGB and RF to project them in a shared time-difference domain, wherein the minute physiological signature can effectively be captured while avoiding the interference from global motions. They are subsequently fed into our end-to-end network, comprising parallel encoding branches that leverage two different pipelines matching the specific domain knowledge of each sensor. In addition, we introduce a novel transformer-based multi-level fusion strategy that aligns domain discrepancies while guaranteeing a complementary/adaptive fusion of both sensory branches.

To the best of our knowledge, this is the first attempt for deep multimodal fusion of video and RF data to implement advanced remote physiology. Given the lack of a video-RF calibrated dataset for vital measurement, we created the first video-RF rPPG dataset, which will be publicly available to support future research. We validated the effectiveness of the Fusion-Vital model for respiration and BVP estimation tasks. The experimental results indicate that the proposed method can predict both respiration and BVP more accurately than current state-of-the-art approaches. Furthermore, we demonstrate that the combination of RGB and RF modalities brings great robustness in challenging scenarios, such as darkness or occlusions.

Related Work

Video-Based Remote Physiology. Given that the reflectance spectrum of the human skin vibrates along with vital movements (Verkruysse, Othar Svaasand, and Stuart Nelson 2008), human physiology can remotely be reconstructed using RGB sequences reflected from exposed skin, particularly those from the facial area. However, such vital motions are subtle and often contaminated by external factors, such as global movements and illumination changes (Xu, Sun, and Rohde 2014; Chen and McDuff 2018). Traditional approaches have relied upon signal decomposition methods, such as principal component analysis (PCA) (Balakrishnan, Durand, and Guttag 2013) and independent component analysis (ICA) (Poh, McDuff, and Picard 2010; Monkaresi, Calvo, and Yan 2014), to restore the desired vital signals under such a low signal-to-noise ratio (SNR) condition. With the advent of deep learning, some approaches have exploited its nonlinear modeling capability to train direct mappings from RGB sequences to gold-standard signals (Chen and McDuff 2018; Spetlik et al. 2018; Yu et al. 2019; Wang, Kao, and Hsu 2019; Lu, Han, and Zhou 2021). More recent techniques have introduced inverse attention (Nowara, McDuff, and Veeraraghavan 2021) or temporal shift modules (Liu et al. 2020) to effectively suppress the interference caused from head movements.

RF-Based Remote Physiology. RF sensor is characterized by offering superior depth resolvability as well as Doppler information (see supplementary materials for details), allowing it to capture even microscopic oscillations modulated by human physiology (Jiang et al. 2020; Zhang et al. 2022). Since such depth vibration is most prominent in the vicinity of the chest, RF-based approaches track vital movements based primarily on sequential radial ranges detected around the torso (Mercuri et al. 2019; Ha, Assana, and Adib 2020; Choi et al. 2021). Given the difficulty in interpreting radio signals and acquiring data, most RF physiology techniques rely on learning-free frameworks, using signal decomposition methods such as frequency analysis (Li and Lin 2008; Tu, Hwang, and Lin 2016), wavelet decomposition (Li and Lin 2018; Mercuri et al. 2018, 2019), and fuzzy logic (Choi et al. 2021). However, recent approaches have achieved improved performance by leveraging the capability of learning schemes, triggering the use of deep learning in the area of RF-based physiology. Ha *et. al.* (Ha, Assana, and Adib 2020) proposed an encoder-decoder architecture that reconstructs vital signatures from raw RF phase reflections, and Zheng *et. al.* (Zheng et al. 2021) adopted a variational inference approach. Furthermore, (Choi, Kang, and Kim 2022) succeeded in extracting respiration from a moving person by introducing a multi-task adversarial learning framework. Unlike conventional techniques relying solely on a single modality for vital estimation, this study explores advanced physiological measurements by fusing RGB and RF data in a complementary manner.

RGB-RF Fusion. To enhance robustness in dark or adverse weather conditions (Qian et al. 2021), several studies have investigated the fusion of RGB and RF modalities (Long et al. 2021b,a; Bijelic et al. 2020; Nabati and Qi 2021; Cheng, Xu, and Liu 2021; Hwang et al. 2022). Most of these studies have focused on outdoor sensing applications, such as autonomous driving (Nabati and Qi 2021; Cheng, Xu, and Liu 2021; Dong et al. 2021; Hwang et al. 2022), where RGB images and 2D RF bird-eye-view (BEV) images have been spatially fused. Early fusion models were based on object-level fusion, which coupled RGB and RF modalities through the fusion of independent object-level outputs detected from each modality, using statistical association algorithms (Ji and Prokhorov 2008; Janda et al. 2013; Wang et al. 2016). With the emergence of deep learning, there has been a recent surge in research on deep-level RGB-RF fusion. Some works (Nabati and Qi 2021; Dong et al. 2021) have proposed deep feature-level fusion of convolutional representations encoded from RGB and RF BEV images for advanced object detection in autonomous vehicles. Cheng *et. al.* (Cheng, Xu, and Liu 2021) developed an attention-based camera-radar fusion architecture for small object detection, whereas Long *et. al.* expanded the RGB-RF fusion scheme to pixel-wise depth (Long et al. 2021b) or velocity (Long et al. 2021a) completion tasks. However, previous methods have mainly focused on the spatial fusion of RGB and RF data. In contrast, our module is designed specifically towards the temporal fusion of RGB and RF, along with newly proposed temporal input formats.

Methodology

Preliminaries

Reflection Model in RGB Modality. In this section, we present a mathematical model for RGB spatiotemporal variations induced from human physiology (Wang et al. 2017; Chen and McDuff 2018). Let the RGB intensity of the m-th

image pixel at time t be defined as

$$\mathbf{C}_m^{(\mathrm{RGB})}(t) = I(t)\left(\boldsymbol{v}_d(t) + \boldsymbol{v}_s(t)\right) + \mathrm{n}^{(\mathrm{RGB})}(t), \quad (1)$$

where $I(t)$ denotes the luminance level, $\boldsymbol{v}_d(t)$ and $\boldsymbol{v}_s(t)$ the diffuse and specular reflections, respectively, and $\mathrm{n}^{(\mathrm{RGB})}(t)$ the quantization noise. $I(t)$, $\boldsymbol{v}_d(t)$, and $\boldsymbol{v}_s(t)$ can further be decomposed into static and time-varying components as

$$I(t) = I_0\left(1 + \Phi\left(g(t), p(t)\right)\right), \quad (2)$$

$$\boldsymbol{v}_d(t) = d_0\boldsymbol{u}_d + p(t)\boldsymbol{u}_p, \quad (3)$$

$$\boldsymbol{v}_s(t) = \left(s_0 + \Psi\left(g(t), p(t)\right)\right)\boldsymbol{u}_s, \quad (4)$$

where I_0, d_0, and s_0 are the stationary components of the luminance intensity, diffuse reflection, and specular reflection, respectively; $\Phi\left(g(t), p(t)\right)$ and $\Psi\left(g(t), p(t)\right)$ refer to the time-dependent components nonlinearly modulated both by non-physiological variations (e.g., facial expressions or global movements) $g(t)$ and desired vital source $p(t)$. $\boldsymbol{u}_d$ and $\boldsymbol{u}_s$ are the unit color vectors for the skin tissue and light source spectrum, respectively, and $\boldsymbol{u}_p$ is the relative pulsatile change induced by hemoglobin and melanin absorption. Since the products of time-varying terms are much smaller than the static components (Chen and McDuff 2018; Liu et al. 2020), Eq. (1) can be simplified by disregarding such terms:

$$\mathbf{C}_m^{(\mathrm{RGB})}(t) \approx c_0 I_0 \boldsymbol{u}_c + c_0 I_0 \Phi\left(g(t), p(t)\right)\boldsymbol{u}_c + I_0 \Psi\left(g(t), p(t)\right)\boldsymbol{u}_s + I_0 p(t)\boldsymbol{u}_p + \mathrm{n}^{(\mathrm{RGB})}(t). \quad (5)$$

where $c_0\boldsymbol{u}_c = s_0\boldsymbol{u}_s + d_0\boldsymbol{u}_d$ with $\boldsymbol{u}_c$ representing the unit color vector for skin reflection.

Reflection Model in RF Modality. Unlike cameras, which involve reflected RGB light, RF sensors emit a periodic electromagnetic signal that bounces off the human body and returns back to spatially-separated receiver channels. After pre-processing, the channel-wise RF reflection $\mathbf{C}^{(\mathrm{RF})}(t)$ is expressed as follows (see supplementary for details):

$$\mathbf{C}^{(\mathrm{RF})}(t) = \boldsymbol{\alpha}(t)\exp\left(j\boldsymbol{\theta}(t)\right) + \mathrm{n}^{(\mathrm{RF})}(t). \quad (6)$$

The amplitude $\alpha(t)$ and phase $\theta(t)$ components of the signal are further decomposed into

$$\boldsymbol{\alpha}(t) \approx \sqrt{\frac{P_{\mathrm{Tx}} G \sigma \lambda^2}{(4\pi)^3 \left(\boldsymbol{\Theta}\left(g(t), p(t)\right)\right)^4}}, \quad (7)$$

$$\boldsymbol{\theta}(t) = \frac{4\pi}{\lambda}\boldsymbol{\Theta}\left(g(t), p(t)\right), \quad (8)$$

where P_{Tx}, G, σ, and λ, all of which are approximately static over time, denote the radio transmission power, antenna gain, electromagnetic reflectivity of the human body, and signal wavelength, respectively. Note that the non-physiological fluctuations $g(t)$ and the desired vital motions $p(t)$ are also involved in the RF magnitude and phase, which are modulated by the radial projection function $\boldsymbol{\Theta}\{\cdot\}$ (Li and Stoica 2008).

Motivation

Our Fusion-Vital network is motivated by the distinctive domain properties of the RGB and RF modalities, where the physiological signal of interest, $p(t)$, is shared in the same time domain but is involved through different media (light and electromagnetic waves), with different information sources (RGB intensity and radial range). Moreover, since the camera projects the surrounding reflections onto a 2D plane that is perpendicular to the RF LoS dimension, whereas the RF sensor captures signals along the LoS dimension itself, their fusion enables the multidimensional analysis of human movements, leading to enhanced analysis of $p(t)$ in the presence of undesired $g(t)$.

Fusion-Vital Model

Overview. Fig. 1 illustrates the overall pipeline of the proposed Fusion-Vital model. The model firstly projects raw RGB and RF reflections into a shared time-difference domain. Subsequently, a two-branch parallel architecture coupled with multi-level feature fusion modules is adopted to effectively utilize the complementary signatures of the RGB and RF modalities. One branch encodes the RGB modality, which is further split into two sub-branches for facial and motional modeling, respectively. The other branch is responsible for electromagnetic extraction from the RF modality. The physiological sequences are finally reconstructed based on the adaptive temporal fusion of the multimodal representations.

Time-Difference Alignment of RGB-RF. For successful extraction of micro-scale $p(t)$ from the input in the presence of contamination from $g(t)$, an effective solution is to process the input to involve the time-difference domain of $g(t)$ and $p(t)$ (i.e., $g'(t)$ and $p'(t)$) instead of direct use of it (Chen and McDuff 2018). Namely, given a video clip $\left\{\mathbf{C}^{(\mathrm{RGB})}(t), \cdots, \mathbf{C}^{(\mathrm{RGB})}(t+T)\right\}$ in the case of RGB branch, we can generate the motional sequence $\{\mathbf{M}(t), \cdots, \mathbf{M}(t+T-1)\} \in \mathbb{R}^{T\times 3\times H^{in}\times W^{in}}$ in the time-difference domain, where $\mathbf{M}(t) = (\mathbf{C}^{(\mathrm{RGB})}(t+1) - \mathbf{C}^{(\mathrm{RGB})}(t))/(\mathbf{C}^{(\mathrm{RGB})}(t+1) + \mathbf{C}^{(\mathrm{RGB})}(t))$, and adopt it as a basical RGB input. However, the problem lies in RF modality, where $p(t)$ is intertwined in non-linear and ambiguous manners within the raw RF reflections (see supplementary material), rendering it challenging to directly project $\mathbf{C}^{(\mathrm{RF})}(t)$ to associate the time-difference domain of $g(t)$ and $p(t)$. This, in turn, yields temporal misalignment between RGB and RF, as well as unstable vital tracking in RF branch.

As an effective solution for this, we focus on the Doppler characteristics of RF signals (see supplementary for details): the level of Doppler frequency shift in RF reflection can serve as an alternative indicator linearly representing the time-difference domain of $g(t)$ and $p(t)$. Specifically, unlike the previous approaches, which tries to directly extract $p(t)$ from $\mathbf{C}^{(RF)}(t)$ using a series of heuristic and complicated processes (Mercuri et al. 2019), we perform a short-time Fourier transform (STFT) on $\boldsymbol{\alpha}(t)$ and $\exp(j\boldsymbol{\theta}(t))$ to form pairwise time-frequency images $\mathbf{F}_\alpha \in \mathbb{C}^{N_{\mathrm{Rx}}\times T\times F}$ and $\mathbf{F}_\theta \in \mathbb{C}^{N_{\mathrm{Rx}}\times T\times F}$, where N_{Rx} is the number of receiver channels. Note that this time-frequency modality follows time-difference trajectory of each body part as 2D format, maintaining superiority in the context of robustness from the burden of non-linear estimation as well as time-difference

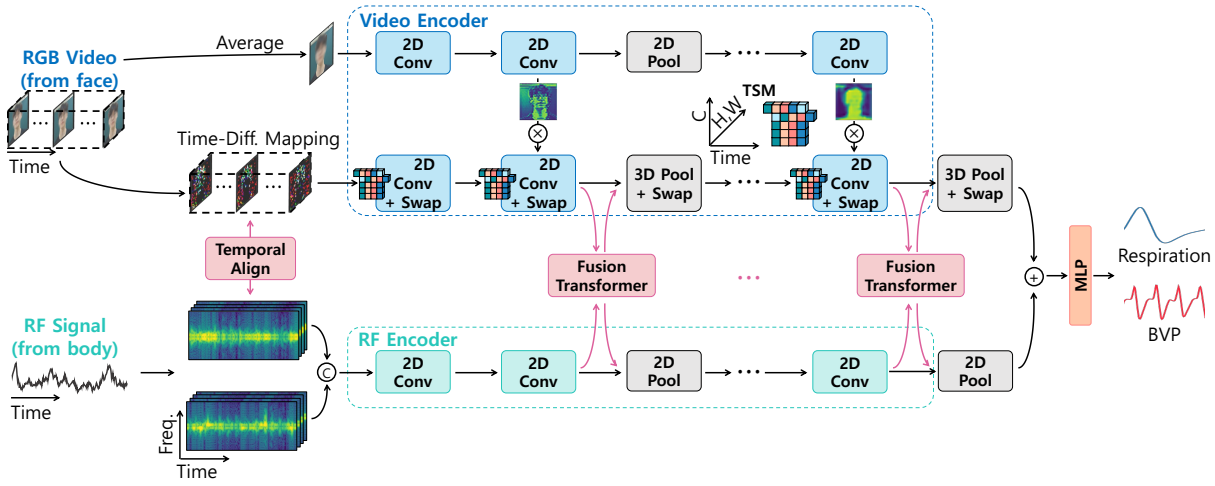

Figure 1: Overall Fusion-Vital architecture. The input RGB and RF streams are first transformed into motional clips and RJTF maps under the shared time-difference domain, which are then fed into parallelized video and RF encoders, respectively. During training, our novel fusion transformer modules adaptively integrate the RGB and RF modalities to generate the final respiratory or BVP signals.

domain alignment with RGB modality.

RGB Encoding. The holistic RGB branch of our Fusion-Vital network follows the concept proposed in (Chen and McDuff 2018), which involves parallel encoding of motional and spatial features, as well as the bridge network based on spatial soft attention. Given a time-difference video clip $\{\mathbf{M}(t), \cdots, \mathbf{M}(t+T-1)\}$ and the appearance image $\mathbf{A}$, where $\mathbf{A}$ is the average of the clip in the time domain, the two types of RGB inputs are embedded in parallel using quasi-symmetric convolutional pipelines.

For appearance modeling, our architecture employs general 2D convolutional embedding. Unlike the appearance input $\mathbf{A}$, which incorporates only the spatial features of the facial area, the motional time-difference sequence involves both facial and temporal signatures. This property, in turn, mandates that motional embedding considers spatiotemporal 3D convolutions, which can substantially inflate the computational overhead. Inspired by the encoding of motion representations based on the temporal shift module (TSM) (Liu et al. 2020; Lin et al. 2022), which allows spatiotemporal modeling even without leveraging 3D operations, our RGB motional branch leverages 2D convolutions combined with TSMs. As illustrated in Fig. 1, the TSM is plugged in every 2D convolution, shifting the channel of its input tensor in the time direction. Specifically, the tensors are subdivided into three portions across the channel, two of which are shifted by $+1$ or -1 frame along the temporal dimension, while the rest remain unshifted. Additionally, in the case of pooling in the motional branch, we adopt 3D pooling (in width, height, and time) such that multi-level temporal resolution can be contemplated in vital estimation.

A soft-attention module is adopted as a bridge between the appearance and time-difference branches. The soft-attention mask, formed from the appearance domain, attends to the physiology-related pixels (typically the facial area) within the intermediate motional representations, thereby focusing the network more on the desired signals while excluding spurious information induced from the temporal tensor shift (Liu et al. 2020). Attention masking is introduced right before each pooling layer, which can formally be achieved through the element-wise product between the time-difference feature and the corresponding mask as

$$\mathbf{M}_l \odot \frac{H_l W_l \cdot \gamma\left(\omega_l \mathbf{A}_l + b_l\right)}{2\left\|\gamma\left(\omega_l \mathbf{A}_l + b_l\right)\right\|_1}, \tag{9}$$

where $\mathbf{M}_l$ and $\mathbf{A}_l$ refer to the l-th layer representations from the time-difference and appearance branches, respectively; ω_l denotes the 1×1 convolution filter and $\gamma(\cdot)$ is the sigmoid activation function.

RF Encoding. We concatenate the log-magnitudes of $\mathbf{F}_\alpha$ and $\mathbf{F}_\theta$ in a channel-wise manner, generating the final RF input of $\mathbf{F} \in \mathbb{R}^{2N_{\mathrm{Rx}} \times T \times F}$. Considering the 2D format of the RJTF input and the parallelism with the RGB branch, the RF embedding branch follows a 2D convolutional architecture that is equivalent to that of the appearance embedding in the RGB branch, except for the first convolution module, which is modified to be compatible with the N_{Rx}-channel input.

Multimodal Fusion in Time-Difference Domain. To combine the 3D representations from the RGB branch with the 2D representations from the RF branch, the two modalities must be coordinated along the same dimension. Noting that our RGB-RF input modalities share the time-difference dimension, we define the temporal fusion between the two modalities.

For achieving adaptive fusion of the temporal context within each modality, we proposes to use the capability of cross-attention (CA) in the temporal fusion of RGB and RF, whose overall pipeline is illustrated in Fig. 2 Given the intermediate representations $\mathbf{R}_l^{(\mathrm{RGB})} \in \mathbb{R}^{C^l \times T^l \times H^l \times W^l}$ and $\mathbf{R}_l^{(\mathrm{RF})} \in \mathbb{R}^{C^l \times T^l \times F^l}$ from the parallelized RGB-RF encod-

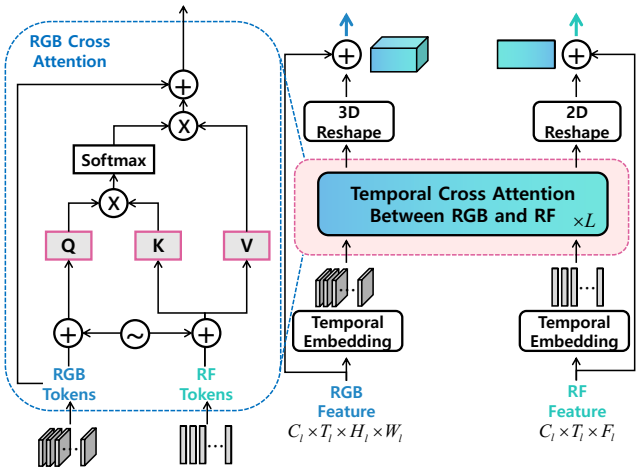

Figure 2: Transformer block for the temporal fusion of the RGB and RF signals.

ing branches, we first unfold them on a basis of modality-wise time-centric patches $\bar{\mathbf{R}}_l^{(\mathrm{RGB})} \in \mathbb{R}^{T_l \times (C_l H_l W_l)}$ and $\bar{\mathbf{R}}_l^{(\mathrm{RF})} \in \mathbb{R}^{T_l \times (C_l F_l)}$, generating flattened tokens in the shared time-difference domain. The tokens are then embedded with independent projection $f_l^{(\mathrm{RGB})}(\cdot)$ and $f_l^{(\mathrm{RF})}(\cdot)$ for sensor-wise alignment, followed by normalization and temporal embedding.

The module conducts CA between $\mathbf{x}_l^{(\mathrm{RGB})} = f_l^{(\mathrm{RGB})}\left(\bar{\mathbf{R}}_l^{(\mathrm{RGB})}\right)$ and $\mathbf{x}_l^{(\mathrm{RF})} = f_l^{(\mathrm{RF})}\left(\bar{\mathbf{R}}_l^{(\mathrm{RF})}\right)$, trying to capture the contextual dependency between them at each time instant. This can mathematically be expressed as

$$\begin{aligned} \mathbf{q} = \mathbf{x}_l^{(m_1)}\mathbf{W}_q, \quad \mathbf{k} = \mathbf{x}_l^{(m_2)}\mathbf{W}_k, \quad \mathbf{v} = \mathbf{x}_l^{(m_2)}\mathbf{W}_v, \\ \mathrm{CA}\left(\mathbf{x}_l^{(m_1)}, \mathbf{x}_l^{(m_2)}\right) = \mathrm{softmax}\left(\mathbf{q}\mathbf{k}^T/\sqrt{d/h}\right)\mathbf{v}, \end{aligned} \tag{10}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v \in \mathbb{R}^{d \times d/h}$ are trainable parameters; d and h are the hidden dimension and the number of heads, respectively. $m1$ and $m2$ represent the selection of sensor modality: for example, if $m1$ is selected as RGB, then $m2$ becomes RF, and vice versa. This CA mechanism is conducted in multiple heads, denoted as multi-head CA (MCA).

The aggregated information is added to the identity shortcut, and then projected again with sensor-wise normalization and feed-forward network (i.e., $e_l^{(m_1)}(\cdot)$ or $e_l^{(m_2)}(\cdot)$). This MCA cycle can be repeated multiple times throughout the fusion block. As illustrated in Fig. 2, the fused representation is upsampled back with respect to the dimensions of the original RGB/RF representations, which is reincorporated into the primary branch via element-wise summation. Note that the multimodal fusion of the RGB and RF branches is performed before every pooling layer, achieving temporal fusion with multiple resolutions.

Finally, the overall network is optimized based on the $\mathcal{L}_1$ distance between the estimated output $\hat{p}(t)$ and the gold-standard physiological signal $p(t)$.

Experimental Results

Experimental Setup

Datasets. To evaluate the effectiveness of the proposed model, we performed extensive experiments on two datasets: the publicly available RRM-static dataset (Choi, Kang, and Kim 2022) and our newly collected physiological dataset, named the Multimodal Database for rPPG (MMD-rPPG). The RRM-static dataset comprises approximately 2.4 h of synchronized video clips captured at a frame rate of 30 fps, RF reflected signals recorded at 1000 fps, and ground-truth respiration signals, which correspond to 13 stationary participants. The MMD-rPPG dataset, which is the first multimodal dataset in the objective of cardiac estimation, includes 3 h of synchronized video and RF (a fps of 30 for RGB and a fps of 1000 for RF) reflections from 15 participants, complemented with gold-standard BVP signals recorded using a Neulog BVP sensor (NeuLog 2017). The dataset also provides reflections registered under challenging scenarios, such as dark settings for RGB and occluded scenarios in which a person's body is blocked by an object for RF, allowing for more practical evaluations of remote physiology. The detailed configurations for each dataset can be found in the supplementary material.

Implementation Details The proposed Fusion-Vital model was trained using the ADAM optimizer with a batch size of 64 and a learning rate of 0.0001. The inputs for the RGB branch consisted of video clips that were center-cropped and resized to 36×36 pixels to facilitate constant facial tracking and reduce camera noise (Chen and McDuff 2018). As for the RF branch, the received complex radio signals were transformed to the RJTF format using a STFT with a Hann window 300 ms long, hop size of 60 ms, and 256-point FFT. The resulting images were then resized to fit the temporal dimension of the RGB inputs. To ensure a fair comparison, all temporal models were configured with a window size of 10 frames.

For the numerical evaluation of the estimated physiological signs, we followed the protocols described in (Liu et al. 2020; Choi, Kang, and Kim 2022). We post-processed non-overlapped 10-s windows from the outputs, followed by band pass filtering with a passband of [0.08 Hz, 0.6 Hz] for respiration and [0.75 Hz, 2.5 Hz] for heartbeat. The resulting vital rates estimated in beats per minute (BPM) were compared with the gold-standard using four standard metrics: mean absolute error (MAE), root mean squared error (RMSE), Pearson's correlation coefficient (ρ), and standard deviation (Std). To ensure subject-independent cross-validation, the datasets were divided into person-wise subfolds. More details regarding the evaluation protocols are available in the supplementary material.

Quantitative Results

Comparison with the State-of-the-Art. We compared the performance of the proposed Fusion-Vital model with that of several state-of-the-art remote physiology models (Chen and McDuff 2018; Nowara, McDuff, and Veeraraghavan 2021; Liu et al. 2020; Tu, Hwang, and Lin 2016; Mercuri et al. 2019; Zheng et al. 2021; Ha, Assana, and Adib 2020; Choi,

Method	Input	Respiration Rate (BPM)				Heart Rate (BPM)			
		MAE↓	RMSE↓	ρ↑	Std↓	MAE↓	RMSE↓	ρ↑	Std↓
CAN (Chen and McDuff 2018)	RGB	3.16	5.83	0.57	5.21	3.43	6.42	0.80	5.26
Nowara *et al.* (Nowara, McDuff, and Veeraraghavan 2021)	RGB	2.51	4.58	0.67	4.25	2.66	5.14	0.85	5.10
MTTS-CAN (Liu et al. 2020)	RGB	2.65	4.13	0.69	4.04	2.41	5.27	0.88	4.44
Tu *et al.* (Tu, Hwang, and Lin 2016)	RF	5.46	7.31	0.19	4.86	5.50	11.68	0.64	9.79
Mercuri *et al.* (Mercuri et al. 2019)	RF	2.52	5.64	0.54	5.47	4.73	9.60	0.70	8.55
Zheng *et al.* (Zheng et al. 2021)	RF	1.68	3.82	0.72	3.45	2.50	5.32	0.88	4.29
Ha *et al.* (Ha, Assana, and Adib 2020)	RF	1.37	3.36	0.75	3.21	2.83	5.48	0.86	4.80
RF-Vital (Choi, Kang, and Kim 2022)	RF	0.66	1.44	0.88	1.43	2.19	4.75	0.90	4.31
Fusion-Vital (Ours)	RGB+RF	**0.44**	**1.07**	**0.93**	**1.19**	**1.61**	**3.05**	**0.97**	**3.02**

Table 1: Quantitative comparison of the proposed Fusion-Vital model and eight baseline methods based on their performance on the RRM-static (for respiration) and MMD-rPPG (for heartbeat) datasets.

Method	Input	Heart Rate–Dark (BPM)				Heart Rate–Occluded (BPM)			
		MAE↓	RMSE↓	ρ↑	Std↓	MAE↓	RMSE↓	ρ↑	Std↓
MTTS-CAN (Liu et al. 2020)	RGB	Not Applicable				2.72	5.17	0.85	4.51
RF-vital (Choi, Kang, and Kim 2022)	RF	2.35	4.36	0.89	4.12	Not Applicable			
Fusion-Vital (Ours)	RGB+RF	2.39	4.40	0.89	4.05	2.68	5.11	0.86	4.32

Table 2: Quantitative comparison of the Fusion-Vital model and single-sensor methods based on their estimation of heart rate under challenging scenarios. Note that Fusion-Vital shows stable predictions under both scenarios even with corrupted inputs.

Input Modality		MAE↓	RMSE↓	ρ↑	Std↓
RGB	RF				
Time	Time	1.75	3.48	0.94	3.43
Time-Diff.	Time	2.12	4.42	0.91	4.04
Time	Time-Diff.	1.86	3.67	0.93	3.51
Time-Diff.	Time-Diff.	**1.61**	**3.05**	**0.97**	**3.02**

Table 3: Heart rate estimation performance using different combinations of RGB-RF input modalities.

Kang, and Kim 2022) which depend on a single modality (either video or RF data). Each model was trained on the RRM-static dataset for respiration and on the MMD-rPPG dataset for heartbeat, and tested only on the general samples (i.e., not including the challenging dark nor occluded scenarios).

Regarding the breathing rate estimation results (presented on the left side of Table 1), our method outperforms the previous baseline models by a large margin, achieving an 83.4% reduction in MAE and a 74.1% decrease in RMSE compared to the best output among the RGB-only models, as well as a reduction of 33.3% in MAE and 25.7% in RMSE compared to the best RF-only model. As for heart rate estimation (right side of Table 1), our fusion-based approach also exhibited superior performance, achieving a 33.2% reduction in MAE and a 42.1% reduction in RMSE compared with the best RGB-only model, and a 26.5% reduction in MAE and 35.8% reduction in RMSE compared with the best results from RF-only models. These results confirm the complementarity between the RGB and RF modalities and demonstrate the efficacy of fusion-based vital sensing for human physiological estimation.

Measurement in Challenging Scenarios. In addition to the overall performance enhancement achieved under the previous general scenarios, multimodal fusion for remote physiology also has the potential to enable modality-agnostic prediction when one of the sensors is missing or unavailable. Specifically, our model leverages CA mechanisms to assign adaptive weights to each modality depending on the surrounding conditions, which enables stable estimation even when one sensor is inapplicable. Table 2 summarizes the vital estimation results on the MMD-rPPG dataset under challenging scenarios where one of the sensors is unavailable for remote physiology due to either dark conditions for RGB or occluded settings for RF. It is evident that such corrupted input streams make previous RGB- and RF-only models fail completely under dark and occluded conditions, respectively. In contrast, the proposed model maintains great robustness in both cases. Note that the model shows comparable performance with its single-sensor counterparts, even when the corrupted input from another sensor is fed into the network.

Ablation Study

To further evaluate the effectiveness of each element in the proposed Fusion-Vital model, we performed ablation experiments on the MMD-rPPG dataset.

Input Modality of RGB and RF. Our Fusion-Vital model

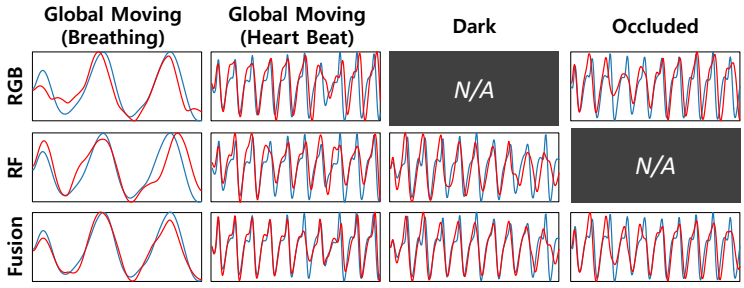

Figure 3: Qualitative examples of 10-s vital wave estimations obtained from RGB-only (Liu et al. 2020), RF-only (Choi, Kang, and Kim 2022), and Fusion-Vital models. The gold-standard signal is represented with blue, and the estimated one is with red line in the figure. The first and second columns show the cases of respiration and heart beat monitoring when a person has global swinging movements, respectively. The third and fourth columns show the results under more challenging scenarios, where the data were recorded under dark lighting or occlusion, respectively. Note that each signal is normalized to [-1, 1].

Fuion Module	MAE↓	RMSE↓	ρ↑	Std↓
(Wang et al. 2021)	1.99	4.21	0.92	3.85
(Prakash, Chitta, and Geiger 2021)	1.87	3.74	0.92	3.81
(Li et al. 2022)	1.71	3.22	0.95	3.20
Ours w/o Sensor-Wise Projection	1.66	3.23	0.96	3.18
Ours	**1.61**	**3.05**	**0.97**	**3.02**

Table 4: Heart rate estimation performance applying another fusion strategies or variants of the proposed fusion module.

is characterized by aligning RGB and RF inputs on the time-difference domain (i.e., motional input for RGB and Doppler input for RF) to ensure the temporal equivalence of each modality as well as improve robustness on external spuriousness. To investigate the effectiveness of these time-difference-based input pairs, we compared the estimation performance using different input combinations of RGB and RF modalities, as summarized in Table 3. The results demonstrate that the use of time-difference inputs is significantly more effective for vital estimation than the direct application of conventional time domain inputs (i.e., spatial video for RGB and unwrapped range input for RF). Furthermore, the outcomes of non-equivalent temporal orders between RGB and RF (second and third rows in Table 3) clearly highlight the significance of temporal alignment for each sensor input.

Fusion Components. To examine the effectiveness of the proposed transformer-based temporal fusion module for physiological measurements, we conducted experiments by comparing its performance against those of alternative multimodal fusion strategies (Wang et al. 2021; Prakash, Chitta, and Geiger 2021; Li et al. 2022), developed mainly for spatial fusion, and a variant of ours (temporal fusion module without sensor-wise projection). The results, presented in Table 4, demonstrate that the proposed time-difference-centric attention outperforms the conventional fusion strategies in all metrics. Moreover, we observed that the sensor-wise projection in our module leads to an additional reduction of 3.01% in MAE and 5.57% in RMSE, indicating the significance of considering sensor-wise characteristics through their independent projection during fusion.

Qualitative Results

Fig. 3 presents a qualitative comparison of the baseline and proposed models under various scenarios. From the results, we can observe the following: 1) when the subject exhibits unwanted global motion, our fusion-based approach can more robustly measure the respiration or heart beat signal, implying the superiority of the fusion-based approach over the single-sensor approaches, especially in the presence of global motion; and 2) in dark or occluded settings where either sensor is inherently unavailable, only the Fusion-Vital model can reconstruct stable outputs.

Meanwhile, in terms of overhead, the respective inference times of RGB-only, RF-only, and our RGB-RF model, for a single frame were represented as 18.8 ms, 33.6 ms, and 29.3 ms, respectively, all of which allow near-real-time processing under workstation settings.

Conclusion

In this study, we presented a novel Fusion-Vital model, which represents the first remote physiological reconstruction approach based on the deep multimodal fusion of RGB and RF sequences. To enable effective alignment between the disparate video and RF dimensions, as well as the straightforward reflection of minute vital signatures, we introduced a new RGB-RF pairwise format based on time-difference signatures. Furthermore, the Fusion-Vital model features a CA-based fusion transformer, which enables feature-level adaptive fusion between multisensor streams. We evaluated the performance of the proposed Fusion-Vital model using both the public RRM-static dataset and a newly constructed MMD-rPPG dataset. The experimental results demonstrate that the RGB and RF modalities can complement each other's information for vital monitoring tasks, which is what enables the proposed model to significantly outperform current state-of-the-art models.

References

Balakrishnan, G.; Durand, F.; and Guttag, J. 2013. Detecting Pulse from Head Motions in Video. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 3430–3437. Portland, OR, USA.

Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; and Heide, F. 2020. Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 11682–11692. Virtual.

Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; and Dubois, J. 2019. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.*, 124(1): 82–90.

Boyer, R. 2011. Performance Bounds and Angular Resolution Limit for the Moving Colocated MIMO Radar. *IEEE Trans. Signal Process.*, 59(4): 1539–1552.

Chen, W.; and McDuff, D. 2018. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In *Eur. Conf. Comput. Vis. (ECCV)*, 349–365.

Cheng, Y.; Xu, H.; and Liu, Y. 2021. Robust Small Object Detection on the Water Surface through Fusion of Camera and Millimeter Wave Radar. In *Int. Conf. Comput. Vis. (ICCV)*, 15243–15252.

Choi, I.-O.; Kim, M.; Choi, J.-H.; Park, J.-K.; Park, S.-H.; and Kim, K.-T. 2021. Robust Cardiac Rate Estimation of an Individual. *IEEE Sensors J.*, 21(13): 15053–15064.

Choi, J.-H.; Kang, K.-B.; and Kim, K.-T. 2022. Remote Respiration Monitoring of Moving Person Using Radio Signals. In *Eur. Conf. Comput. Vis. (ECCV)*, 253–270. Tel Aviv, Israel.

Choi, J.-H.; Kim, J.-E.; Jeong, N.-H.; Kim, K.-T.; and Jin, S.-H. 2020. Accurate People Counting Based on Radar: Deep Learning Approach. In *IEEE Radar Conf. (RadarConf)*, 1–5. Florence, Italy.

Choi, J.-H.; Kim, J.-E.; and Kim, K.-T. 2021. People Counting Using IR-UWB Radar Sensor in a Wide Area. *IEEE Internet Things J.*, 8(7): 5806–5821.

Choi, J.-H.; Kim, J.-E.; and Kim, K.-T. 2022. Deep Learning Approach for Radar-based People Counting. *IEEE Internet Things J.*, 9(10): 7715–7730.

Dong, X.; Zhuang, B.; Mao, Y.; and Liu, L. 2021. Radar Camera Fusion via Representation Learning in Autonomous Driving. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (CVPRW)*, 1672–1681. Virtual.

Estepp, J. R.; Blackford, E. B.; and Meier, C. M. 2014. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *IEEE Conf. Syst., Man, Cybern. (SMC)*, 1462–1469.

Fogle, O. R.; and Rigling, B. D. 2012. Micro-Range/Micro-Doppler Decomposition of Human Radar Signatures. *IEEE Trans. Aerosp. Electron. Syst.*, 48(4): 3058–3072.

Ha, U.; Assana, S.; and Adib, F. 2020. Contactless Seismocardiography via Deep Learning Radars. In *ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 1–14.

Hwang, J.-J.; Kretzschmar, H.; Manela, J.; Rafferty, S.; Armstrong-Crews, N.; Chen, T.; and Anguelov, D. 2022. CramNet: Camera-Radar Fusion with Ray-Constrained Cross-Attention for Robust 3D Object Detection. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Eur. Conf. Comput. Vis. (ECCV)*, 388–405. Tel Aviv, Israel.

Janda, F.; Pangerl, S.; Lang, E.; and Fuchs, E. 2013. Road boundary detection for run-off road prevention based on the fusion of video and radar. In *IEEE Intell. Veh. Symp. (IV)*, 1173–1178. Gold Coast, QLD, Australia.

Ji, Z.; and Prokhorov, D. 2008. Radar-vision fusion for object classification. In *Int. Conf. Inform. Fusion*, 1–7. Cologne, Germany.

Jiang, C.; Guo, J.; He, Y.; Jin, M.; Li, S.; and Liu, Y. 2020. mmVib: Micrometer-Level Vibration Measurement with mmWave Radar. In *ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 1–13. Virtual.

Li, C.; and Lin, J. 2008. Random Body Movement Cancellation in Doppler Radar Vital Sign Detection. *IEEE Trans. Microw. Theory Techn.*, 56(12): 3143–3152.

Li, J.; and Stoica, P. 2008. *MIMO radar signal processing*. Hoboken, New Jersey, USA: John Wiley & Sons.

Li, M.; and Lin, J. 2018. Wavelet-Transform-Based Data-Length-Variation Technique for Fast Heart Rate Detection Using 5.8-GHz CW Doppler Radar. *IEEE Trans. Microw. Theory Techn.*, 66(1): 568–576.

Li, Y.; Yu, A. W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Lu, Y.; Zhou, D.; Le, Q. V.; Yuille, A.; and Tan, M. 2022. DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 17182–17191. New Orleans, Louisiana, USA.

Lin, J.; Gan, C.; Wang, K.; and Han, S. 2022. TSM: Temporal Shift Module for Efficient and Scalable Video Understanding on Edge Devices. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5): 2760–2774.

Liu, X.; Fromm, J.; Patel, S.; and McDuff, D. 2020. Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 1–23. Virtual.

Long, Y.; Morris, D.; Liu, X.; Castro, M.; Chakravarty, P.; and Narayanan, P. 2021a. Full-Velocity Radar Returns by Radar-Camera Fusion. In *Int. Conf. Comput. Vis. (ICCV)*, 16198–16207. Virtual.

Long, Y.; Morris, D.; Liu, X.; Castro, M.; Chakravarty, P.; and Narayanan, P. 2021b. Radar-Camera Pixel Depth Association for Depth Completion. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 12507–12516. Virtual.

Lu, H.; Han, H.; and Zhou, S. K. 2021. Dual-GAN: Joint BVP and Noise Modeling for Remote Physiological Measurement. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 12404–12413. Virtual.

McDuff, D. J.; Hernandez, J.; Gontarek, S.; and Picard, R. W. 2016. COGCAM: Contact-Free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera.

In *ACM CHI Conf. Hum. Factors Comput. Syst.*, 4000–4004. New York, NY, USA.

Mercuri, M.; Liu, Y.-H.; Lorato, I.; Torfs, T.; Wieringa, F.; Bourdoux, A.; and Van Hoof, C. 2018. A Direct Phase-Tracking Doppler Radar Using Wavelet Independent Component Analysis for Non-Contact Respiratory and Heart Rate Monitoring. *IEEE Trans. Biomed. Circuits Syst.*, 12(3): 632–643.

Mercuri, M.; Lorato, I.; Liu, Y.-H.; P. Wieringa, F.; Van Hoof, C.; and Torfs, T. 2019. Vital-sign monitoring and spatial tracking of multiple people using a contactless radar-based sensor. *Nature Electron.*, 2: 252–262.

Monkaresi, H.; Calvo, R. A.; and Yan, H. 2014. A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam. *IEEE J. Biomed. Health Inform.*, 18(4): 1153–1160.

Nabati, R.; and Qi, H. 2021. CenterFusion: Center-Based Radar and Camera Fusion for 3D Object Detection. In *IEEE Winter Conf. Applic. Comput. Vis. (WACV)*, 1527–1536. Virtual.

NeuLog. 2017. Heart Rate & Pulse logger sensor NUL-208. https://neulog.com/heart-rate-pulse/. Accessed: 2023-12-31.

Nowara, E. M.; McDuff, D.; and Veeraraghavan, A. 2021. The Benefit of Distraction: Denoising Camera-Based Physiological Measurements Using Inverse Attention. In *Int. Conf. Comput. Vis. (ICCV)*, 4955–4964.

Obadi, A. B.; Zeghid, M.; Kan, P. L. E.; Soh, P. J.; Mercuri, M.; and Aldayel, O. 2022. Optimized Continuous Wavelet Transform Algorithm Architecture and Implementation on FPGA for Motion Artifact Rejection in Radar-Based Vital Signs Monitoring. *IEEE Access*, 1–21.

Park, J.-K.; Hong, Y.; Lee, H.; Jang, C.; Yun, G.-H.; Lee, H.-J.; and Yook, J.-G. 2019. Noncontact RF Vital Sign Sensor for Continuous Monitoring of Driver Status. *IEEE Trans. Biomed. Circuits Syst.*, 13(3): 493–502.

Poh, M.-Z.; McDuff, D. J.; and Picard, R. W. 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10): 10762–10774.

Prakash, A.; Chitta, K.; and Geiger, A. 2021. Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 7077–7087. Virtual.

Qian, K.; Zhu, S.; Zhang, X.; and Li, L. E. 2021. Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary Lidar and Radar Signals. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 444–453. Virtual.

Revanur, A.; Li, Z.; Ciftci, U. A.; Yin, L.; and Jeni, L. A. 2021. The First Vision for Vitals (V4V) Challenge for Non-Contact Video-Based Physiological Estimation. In *Int. Conf. Comput. Vis. Worksh. (ICCVW)*, 2760–2767. Virtual.

Spetlik, R.; Franc, V.; Cech, J.; and Matas, J. 2018. Visual Heart Rate Estimation with Convolutional Neural Network. In *Brit. Mach. Vis. Conf. (BMVC)*, 3–6. Newcastle, UK.

Tu, J.; Hwang, T.; and Lin, J. 2016. Respiration Rate Measurement Under 1-D Body Motion Using Single Continuous-Wave Doppler Radar Vital Sign Detection System. *IEEE Trans. Microw. Theory Techn.*, 64(6): 1937–1946.

Verkruysse, W.; Othar Svaasand, L.; and Stuart Nelson, J. 2008. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26): 21434–21445.

Wang, C.; Ma, C.; Zhu, M.; and Yang, X. 2021. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 11794–11803. Virtual.

Wang, W.; den Brinker, A. C.; Stuijk, S.; and de Haan, G. 2017. Algorithmic Principles of Remote PPG. *IEEE Trans. Biomed. Eng.*, 64(7): 1479–1491.

Wang, X.; Xu, L.; Sun, H.; Xin, J.; and Zheng, N. 2016. On-Road Vehicle Detection and Tracking Using MMW Radar and Monovision Fusion. *IEEE Trans. Intell. Transp. Syst.*, 17(7): 2075–2084.

Wang, Z.-K.; Kao, Y.; and Hsu, C.-T. 2019. Vision-Based Heart Rate Estimation Via A Two-Stream CNN. In *IEEE Int. Conf. Image Process. (ICIP)*, 3327–3331. Taipei, Taiwan.

Xu, S.; Sun, L.; and Rohde, G. K. 2014. Robust efficient estimation of heart rate pulse from video. *Biomed. Opt. Express*, 5(4): 1124–1135.

Yu, Z.; Peng, W.; Li, X.; Hong, X.; and Zhao, G. 2019. Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement. In *Int. Conf. Comput. Vis. (ICCV)*, 151–160.

Zhang, J.; Tang, H.; Chen, D.; and Zhang, Q. 2012. deStress: Mobile and remote stress monitoring, alleviation, and management platform. In *IEEE Glob. Commun. Conf. (GLOBECOM)*, 2036–2041. Anaheim, CA, USA.

Zhang, S.; Zheng, T.; Chen, Z.; and Luo, J. 2022. Can We Obtain Fine-grained Heartbeat Waveform via Contact-free RF-sensing? In *IEEE Conf. Comput. Commun. (INFOCOM)*, 1759–1768. London, UK.

Zhao, M.; Yue, S.; Katabi, D.; Jaakkola, T. S.; and Bianchi, M. T. 2017. Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture. In *Int. Conf. Mach. Learn. (ICML)*, volume 70, 4100–4109. Sydney, Australia.

Zheng, T.; Chen, Z.; Cai, C.; Luo, J.; and Zhang, X. 2020. V2iFi: In-Vehicle Vital Sign Monitoring via Compact RF Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, 4(2): 1–27.

Zheng, T.; Chen, Z.; Zhang, S.; Cai, C.; and Luo, J. 2021. MoRe-Fi: Motion-Robust and Fine-Grained Respiration Monitoring via Deep-Learning UWB Radar. In *ACM Conf. Embedded Netw. Sens. Syst. (SenSys)*, 111–124. New York, NY, USA.