Remote Respiration Monitoring of Moving Person Using Radio Signals

Jae-Ho Choi[1], Ki-Bong Kang[1,2], and Kyung-Tae Kim[1]

[1] POSTECH, Republic of Korea
[2] Samsung Electronics, Republic of Korea
{jhchoi93,kkb131,kkt}@postech.ac.kr

Abstract. Non-contact respiration rate measurement (nRRM), which aims to monitor one's breathing status without any contact with the skin, can be utilized in various remote applications (e.g., telehealth or emergency detection). The existing nRRM approaches mainly analyze fine details from videos to extract minute respiration signals; however, they have practical limitations in that the head or body of a subject must be quasi-stationary. In this study, we examine the task of estimating the respiration signal of a non-stationary subject (a person with large body movements or even walking around) based on radio signals. The key idea is that the received radio signals retain both the reflections from human global motion (GM) and respiration in a mixed form, while preserving the GM-only components at the same time. During training, our model leverages a novel multi-task adversarial learning (MTAL) framework to capture the mapping from radio signals to respiration while excluding the GM components in a self-supervised manner. We test the proposed model based on the newly collected and released datasets under real-world conditions. This study is the first realization of the nRRM task for moving/occluded scenarios, and also outperforms the state-of-the-art baselines even when the person sits still.

Keywords: Non-Contact Respiration Rate Measurement, Radio Signal, Multi-Task Adversarial Learning.

1 Introduction

Respiration rate (RR) is an important clinical indicator directly reflecting the status of the human ventilation system. In this respect, continuous monitoring of one's RR is helpful for general health care, especially for telehealth or emergency detection in patients with breathing disorders such as chronic obstructive pulmonary disease and SARS-CoV-2 (COVID-19) [1,40]. Traditional measurements for RR are typically based on contact devices such as chest belts, contact photoplethysmography (PPG), and airflow sensing, which require direct contact with the skin of the subject, hence induces significant discomfort and measurement discontinuities. As alternative to the contact solutions, non-contact RR measurement (nRRM) approaches have recently attracted scholarly attention,

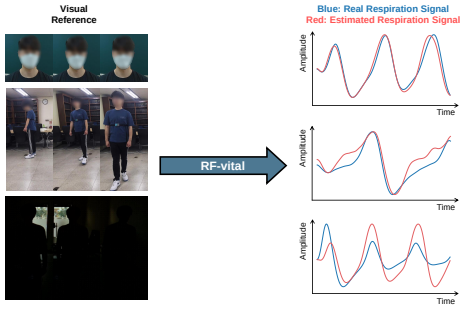

Fig. 1. We propose a RF-vital model that learns the mapping from radio reflections to human respiration signal based on a novel MTAL framework. Several test examples of our RF-vital model demonstrate the feasibility of recovering the fine respiration signs even under occluded, dark, and moving scenarios

most of which leverage the physiological signatures extracted from facial videos [27,16,15,47,38,30,37,22,49,24,3,52,29,32,33,51,23,34,28,35].

However, the skin color changes originating from human breathing cycles are significantly marginal and easily contaminated by head movements of the subject, struggling outside controlled settings (e.g., a scenario where a person must sit approximately still while facing forward) [2,10,40]. Moreover, a single camera view cannot cope effectively with misaligned/occluded faces as well as dark settings, which are quite common scenarios in daily life. Consequently, nRRM for a non-stationary subject (a person with large random body movements or even walking around) has rarely been explored.

To realize robust nRRM systems even against such challenging scenarios, we propose to use radio frequency (RF) signals reflected from radar as an input modality. Radar is an electromagnetic sensor capable of measuring radial depth changes for its targets of interest with high sensitivity. Accordingly, it can capture the horizontal displacements around the chest modulated from human vital signs, while maintaining stable measurements in the presence of head movements, face occlusion, and even large motions. Furthermore, the RF sensor typically operates in GHz band, making it intrinsically unaffected by the surrounding illumination (THz band) or dark conditions as well as completely free from privacy issues.

In fact, there have been several attempts to achieve nRRM based on RF signals previously [45,9,39,26,17,31,44,5,14]. RF-based nRRM methods usually first estimate the radial distance of the human body from the raw reflected signals. Considering that the extracted radial distances with respect to time directly reflects the physiological signals modulated from the body-depth variations, the RR can be recovered via several signal decomposition techniques such as advanced filtering [9,39,26,31] and deep learning (DL) [14,56]. However, these approaches still have limitations in overcoming the large motion scenarios. Such vigorous movements of each individual force a dynamic range of the signal to be significantly enlarged, greatly inflating the distance estimation errors. Particu-

larly, the radial distance of a person changes both along the global motion (GM) induced from the stagger/gait and along the respiratory motion (RM) from the inhalation-exhalation cycle, whereas the RM components maintain much smaller displacements than the GM; therefore, they are likely to be obscured in the radio reflection data.

To tackle this problem and achieve nRRM even for a moving subject, we propose a novel RF-vital model, characterized by newly-introduced input formats for radio reflections and a multi-task adversarial learning (MTAL) framework. Specifically, our U-Net style network [42] takes a radio joint time-frequency (RJTF) map as input (which is completely free from the distance estimation issues), then attempts to reconstruct the subject's respiration signal (i.e., RM) and spatial trajectory (i.e., GM). During training, the decoder for the GM is co-trained with the feature encoder in an adversarial manner, thereby facilitating the latent representation to be irrelevant to the GM of a person and reflect only the desired RM. Such adversarial mapping on GM can be accomplished based on our key observations that the reflected RF signals not only provide the RM-GM mixture, but also preserve GM-only self-supervision simultaneously. Meanwhile, to prevent the model from learning identity (ID)-dependent short-cuts, we add an auxiliary identification task, which is also trained in an adversarial manner.

This study is the first to report the realization of an nRRM over a randomly moving person. We evaluate our RF-vital model on two nRRM datasets consisting of synchronized RF signals, respiratory signals, and RGB videos, which were collected from different base scenarios. The first dataset was obtained in ideal situations, where a person sits nearly still with her/his head facing forward. The second dataset was collected from much more challenging scenarios, where the subject was allowed to stand and even move around freely in various directions. We release our datasets to further advance the RF-based nRRM research. The experimental results show that our RF-vital model outperforms the state-of-the-art video- and RF-based nRRM approaches in static scenarios. Moreover, as shown in Fig. 1, it continues to work properly in large motion scenarios, where the current methods fail completely. Furthermore, our methods can provide robust estimations, even in dark-light conditions and occlusions, enabling more realistic implementations of nRRM. We believe that our approach is also applicable for detecting various vital signs in humans, such as heart rate. Nonetheless, in this study, we only focus on estimating respiratory signals.

2 Related Work

2.1 Video-Based Physiological Measurements

Because the diffuse reflectance spectra of the skin (typically facial region) change along with the human physiological movements, remote prediction of one's vital signals can be achieved by capturing the subtle light reflections using a camera [46,48,40]. The problem is that such diffuse components reflected back from the camera are substantially marginal and easily affected by nuisance factors owing to head motions and light changes. The traditional methods exploit combinations

of different color profiles [27,16,15,47] to retrieve illumination-invariant signatures or exploit signal decomposition techniques, such as independent component analysis (ICA) [38,30,37,22] and principal component analysis (PCA) [49,24] to enhance the signal-to-noise ratio (SNR) of the physiological signals. With the advent of DL in the pattern recognition field, there have also been attempts to employ its powerful nonlinear fitting capability to video-based physiological monitoring, achieving substantial performance improvements [3,52,29,32,33,51,23,34]. The recent approaches further advanced the robustness of the network on head motions by introducing multi-task temporal shift or inverse attention [28,35].

2.2 RF-Based Physiological Measurements

RF signal involves human physiology mainly based on changes in body depth instead of the reflectance in the facial area, so it is less influenced by head motions. Based on the signal property that the received phase components linearly indicate the subject's radial depth with microscopic sensitivity, most RF-based physiological measurements rely on the estimated phase information. Tu *et al.* [45] demonstrated the feasibility of RF-based vital monitoring in a controlled setting. Regarding the generic applications in the presence of small 1-D body movements, several motion compensation methods have been proposed using signal decomposition techniques [9,39,26], wavelet transform [17,31,44], and fuzzy logic [5]. Recently, Ha *et al.* [14] devised an approach to recover the original physiological waveforms from the radio reflections by leveraging a deep supervised encoder-decoder framework. However, these methods fundamentally assume accurate phase estimations (i.e., distance estimations) as *priori*, which are likely to fail under large body movements. Therefore, they can still be applicable to only limited scenarios (e.g., situations where a person sits and shakes her/his body back and forth). Our study aims at more general settings, where a person can stand and even walk around by introducing a new image-like input modality for RF signals and a MTAL strategy.

2.3 Indoor Sensing with RF Signals

The RF system employs wireless reflections for surrounding detection, enabling illuminance-invariant and privacy-preserving sensing. The past wireless systems for indoor environments tend to be biased towards localization and tracking [4,36,21,50,8,7,6]; nonetheless, recent advances in RF hardware and DL-based analysis techniques have facilitated the implementation of more sophisticated tasks based on radio signals. For example, Zhao *et al.* [53,55,54] developed RF-based 2D/3D pose estimation systems, which have been proven to work even through walls. Fan *et al.* [12,11] extended the results for wireless captioning and person re-identification tasks.

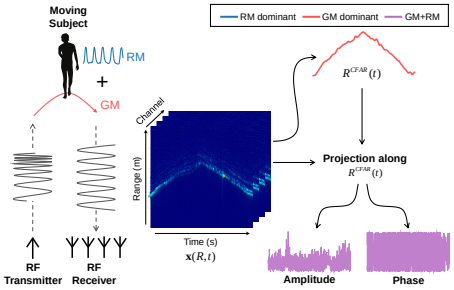

Fig. 2. Pre-processing pipeline for radio-projected profiles. We leverage the CFAR thresholding technique to obtain global trajectories from the channel-wise range-time RF heatmaps. Projecting along the CFAR-output, radio-projected profiles can be extracted, whose magnitude and phase values (purple line) retain both the GM and RM components for a moving person

3 RF Signal Preliminary

3.1 Depth Estimation from RF Signal

The RF sensor periodically transmits an radio signal and receives reflections from its surroundings. Large bodies of RF sensing systems use a frequency-modulated continuous-wave (FMCW) technique for signal modulation [12,53,55,54,13,43], which has also been adopted in our work. After the basic pre-processing from RF raw reflections (see supplementary material for details), we can obtain channel-wise 2D complex range-time heatmaps $\mathbf{x}(R,t) = \{x_m(R,t)\}_{m=1}^{4}$ (see Fig. 2), where R is the radial distance from the transmitter, t is the time, and m represents the receiver index from the distributed array antennas. The magnitude of each RF heatmap (i.e., $|x(R,t)|$) directly indicates the reflected energy level at each distance. Therefore, it is possible to estimate the radial depth of an individual by detecting only the high absolute energy values from $|x(R,t)|$.

Meanwhile, the range resolution of a RF system is determined solely by its transmitted signal bandwidth as [18]:

$$\Delta R = \frac{c}{2BW} = 0.1\,\mathrm{m}, \tag{1}$$

where c is the speed of light and BW is the signal bandwidth, which is set to 1.5 GHz in our RF system. This implies that the general range detections from $|x(R,t)|$ cannot fundamentally involve the microscopic displacement variations originating from human respiration (with displacements of $\sim$ 1 mm in typical [31]). Thus, instead of exploiting range detections from the RF heatmaps as in most RF-based indoor applications [7,53,54,11,12,50,43], we leverage the detected profile itself to retrieve the respiratory signatures beyond the resolution limit.

Namely, as shown in Fig. 2, projection along the detected trajectory of a subject can convert $x(R,t)$ into a 1D temporal signal (hereinafter referred to as

the radio-projected profile) with a complex format [5]:

$$x\left(R^{CFAR}\left(t\right),t\right)=I\left(t\right)+jQ\left(t\right)=\alpha\left(t\right)\exp\left(j\theta\left(t\right)\right), \tag{2}$$

where $R^{CFAR}\left(t\right)$ denotes the coarse distance of a person obtained from a direct detection on $\left|x\left(R,t\right)\right|$ via constant false alarm rate (CFAR) thresholding [41]. It should be noted that the magnitude and phase of the projected signal profiles are further decomposed as [25]:

$$\sqrt{I\left(t\right)^{2}+Q\left(t\right)^{2}}=\alpha\left(t\right)\approx\sqrt{\frac{P_{t}G\sigma\lambda^{2}}{\left(4\pi\right)^{3}\bar{R}\left(t\right)^{4}}}, \tag{3}$$

$$\tan^{-1}\left(\frac{Q\left(t\right)}{I\left(t\right)}\right)+2\pi k=\theta\left(t\right)=\frac{4\pi}{\lambda}\bar{R}\left(t\right), \tag{4}$$

where P_t, G, σ, and λ represent the transmit power, antenna gain, electromagnetic reflectivity, and signal wavelength, respectively, all of which are approximately constant over time. $\bar{R}\left(t\right)$ refers to the radial depth of a subject from the transmitter, and $k\left(=\pm0,1,\cdots\right)$ is the ambiguity factor in estimating the phase. From Eq. (3) and (4), it can clearly be noticed that the magnitude and phase components of the projected signal also reflect the radial depth of the subject. Particularly, contrary to $R^{CFAR}\left(t\right)$ estimated from the coarse range detection on $\left|x\left(R,t\right)\right|$, $\bar{R}\left(t\right)$ is not confined by the range resolution limit, and thus, it retains exquisite sensitivity such that the vital signals with marginal displacements can even be captured [19].

3.2 Motivation for RF-vital Model

Let us consider the radial distance over time for a person with large motion. Because the radial distance for a moving person changes along both the GM of the body and the fluctuating depth owing to RM, $\bar{R}\left(t\right)$ can be expressed as a linear summation of the GM and RM components: $\bar{R}\left(t\right)=\bar{R}_{GM}\left(t\right)+\bar{R}_{RM}\left(t\right)$, where $\bar{R}_{GM}\left(t\right)$ and $\bar{R}_{RM}\left(t\right)$ denote the distance variations induced from the GM (i.e., body movements such as swinging, staggering, and walking) and RM (i.e., the body depth changes owing to breathing), respectively. The RM components, which are quasi-isotropic in any part of the torso and exactly coincide with the inhalation/exhalation cycles, allow human respiration to be recovered depending on the information of body depth variations, from any azimuth angle. Moreover, such chest-based sensing does not suffer from the prerequisite for continuous face tracking, and even maintains an enhanced SNR compared to the extraction from RGB face pixels.

The problem is that while the displacement of RM oscillates at the microscopic level, the radial distance caused by GM (i.e., $\bar{R}_{GM}\left(t\right)$) changes rapidly; hence, the signal strength of $\bar{R}_{RM}\left(t\right)$ present within $\alpha\left(t\right)$ and $\theta\left(t\right)$ becomes substantially trivial. For robust extraction of the RM component, it is essential to highlight the dominance of $\bar{R}_{RM}\left(t\right)$ in $\bar{R}\left(t\right)$, while suppressing the influence of $\bar{R}_{GM}\left(t\right)$. This is not a simple task because the GM and RM components are

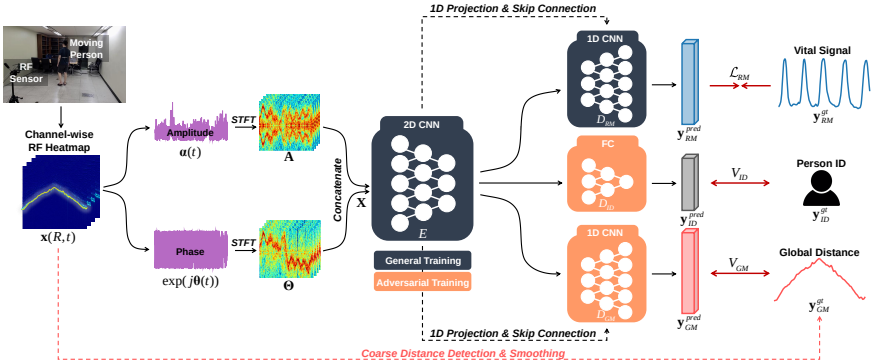

Fig. 3. Overall RF-vital model architecture. It first transforms the radio-projected profiles into RJTF map format, which is subsequently fed in to the U-Net style network [42] composed of one 2D encoder, two 1D decoders, and one discriminator. The network is trained based on the MTAL strategy. During training, the RM decoder attempts to reduce the discrepancy between the real and predicted respiration (black box), whereas the GM decoder and the ID discriminator are co-trained with the encoder in an adversarial manner (orange box), such that the features cannot preserve the signatures regarding the GM and ID of a person

entangled in $\bar{R}(t)$ for every interval as well as the explicit separation of $\bar{R}_{GM}(t)$ requires additional utilization of motion/localization sensors, which makes the overall system extremely bulky.

To tackle these challenges, we propose a novel MTAL framework (Fig. 3), which leverages the domain properties of the RF signals mentioned above: the range detections $R^{CFAR}(t)$ from the RF heatmap cannot fundamentally involve the minute displacements of RM owing to its resolution limits in hardware, but is able to coarsely track the GM of the human body. This implies that $R^{CFAR}(t)$ can act as a powerful model of $\bar{R}_{GM}(t)$. Based on this insight, we devise a network that learns the mapping from the RF inputs composed of $\boldsymbol{\alpha}(t) = \{\alpha_m(t)\}_{m=1}^{4}$ and $\boldsymbol{\theta}(t) = \{\theta_m(t)\}_{m=1}^{4}$ to the desired RM component, while simultaneously pushing out the GM component modeled from $\mathbf{R}^{CFAR}(t) = \{R_m^{CFAR}(t)\}_{m=1}^{4}$ in a self-supervised manner.

4 Methodology

RF-vital is a model for estimating the human respiration signal, given the channel-wise radio reflections as input. As illustrated in Fig. 3, our RF-vital pipeline consists of three main stages: 1) input transformation to convert the radio received signals into the newly proposed input modality, named RJTF maps; 2) representation encoding from the RJTF maps by leveraging 2D convolution modules; 3) decoding branches composed of two 1D convolutional decoders (for RM and GM), and an ID discriminator. During training, these modules are

guided in an end-to-end manner based on the MTAL strategy such that the RM signatures prevail among the latent features, whereas the shortcuts provided by a person's ID and GM are suppressed.

4.1 RJTF Mapping

Although $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\theta}(t)$ intrinsically encompass the RM of a person, adopting them directly as input modalities for our RF-vital model causes two complications. First, their 1D signal formats project not only the RM components but also all the undesired GM and noise-induced elements in a single dimension. Therefore, they are not suitable for distinguishing the fine $\bar{R}_{RM}(t)$. Second, the presence of GM makes it infeasible to achieve correct estimation of $\bar{R}(t)$ from the radio-projected profiles due to the significantly enlarged dynamic range. Considering these problems, we introduce the RJTF map as the input modality for our RF-vital model to further clarify $\bar{R}_{RM}(t)$ and to avoid the estimation ambiguity problem.

The RJTF map takes advantage of the Doppler characteristics of RF signals (see supplementary material for details), which can entail the information of instantaneous distance changes in radial direction [18]. Namely, instead of directly estimating $\bar{R}(t)$ from the radio-projected profiles, we rather perform short-time Fourier transform (STFT) on $\boldsymbol{\alpha}(t)$ and $\exp(j\boldsymbol{\theta}(t))$ to obtain joint time-frequency images $\mathbf{A}$ and $\boldsymbol{\Theta}$ (Fig. 3). The additional Doppler frequency dimension in the RJTF map spans the instantaneous change in the radial distance, so is capable of tracking the human vital signs as well. Particularly, it simply scatters the distance changes of all body parts with respect to time on the 2D domain, being free from the burden for accurate distance estimation. Finally, we aggregate all the channel-wise spectrogram images in concatenated forms, resulting in the final RJTF map $\mathbf{X} \in \mathbb{R}^{8\times T\times F}$, where T and F denote the dimensions in time and Doppler frequency, respectively.

4.2 RF-vital Model Based on MTAL

Overall Architecture. As shown in Fig. 3, our RF-vital model adopts a 2D convolutional encoder to convert the input RJTF map $\mathbf{X}$ into high-level representations. These are subsequently fed into two parallel 1D convolutional decoders responsible for predicting the subject's respiratory signal and global body motion, respectively, and a fully-connected network responsible for identifying her/his ID. Regarding the overall encoder-decoder architecture, we leverage the modified form of a U-Net [42] architecture, where the multi-scale features produced from a series of 2D convolution layers within the encoder are averaged along the frequency dimension and skip-connected to the corresponding 1D convolution layers in each decoder network (see supplementary material for fully detailed architecture).

RM Decoder. The RM decoder $D_{RM}(\cdot)$ aims at extending the correlation with the real respiration signal from the low-dimensional RF feature. We devise the

RM decoder based on series of 1D convolution and up-convolution layers matching the temporal dimensions of the encoder to reconstruct a T-length respiration signal from the representations. In addition, a tanh layer is added after the last convolutional module to bound the predicted values to $[-1, 1]$. During training, the network becomes optimized based on the L_1 distance between the predicted and real respiration signals:

$$\mathcal{L}_{RM} = \left\| \mathbf{y}_{RM}^{pred} - \mathbf{y}_{RM}^{gt} \right\|_1, \tag{5}$$

where $\mathbf{y}_{RM}^{pred}$ denotes the output signals from D_{RM}, and $\mathbf{y}_{RM}^{gt}$ refers to the ground-truth respiration signals measured with the contact chest belt.

GM Decoder. A major complication in accomplishing nRRM for a moving person is the entanglement of human GM and RM within the input, which, in turn, precludes the model from the high-fidelity separation of RM. Moreover, it is impossible to acquire a GM-dominant data (i.e., data affected only by the subject's GM without the RM component at all) paired with the network input, further complicating the disentanglement of the RM features. We address this challenge through a novel adversarial training strategy guided by range-detection-based self-supervision.

The GM decoder consists of 1D convolution and up-convolution layers identical to those of the RM decoder; however, it performs a completely different role. The encoder and GM decoder are trained in an adversarial manner such that the model is encouraged to exclude the GM-dependent features. Let the encoding network be denoted as $E(\cdot)$ and the decoding network for GM as $D_{GM}(\cdot)$. Then, the optimization target can be defined as:

$$\min_{E} \max_{D_{GM}} V_{GM} = - \left\| \mathbf{y}_{GM}^{pred} - \mathbf{y}_{GM}^{gt} \right\|_1, \tag{6}$$

where $\mathbf{y}_{GM}^{pred}$ is the estimated GM component from the decoder, i.e., $\mathbf{y}_{GM}^{pred} = D_{GM}(E(\mathbf{X}))$, and $\mathbf{y}_{GM}^{gt}$ is the ground-truth GM-dominant data. Recall that the coarse range detection $R^{CFAR}(t)$, which is obtained from the direct detection in the RF time-range heatmap $x(R, t)$, can predominantly reflect only the human GM component owing to its range resolution limit in hardware. Based on this, we hypothesize that $R^{CFAR}(t)$ has a great potential to serve as a self-supervision for GM-dominant signals. We average $\mathbf{R}^{CFAR}(t)$ along the receiver channel domain, which is subsequently passed through the linear interpolation and smoothing filter to mitigate the influence of false detections and noise, resulting in the final T-length $\mathbf{y}_{GM}^{gt}$. From the adversarial learning between the RF encoder and GM decoder, the encoding network E can further focus on the RM-dominant signals, while eliminating the GM-dominant features.

Discriminator for Person ID. Because the input radio reflection contains unexpected subject-dependent signatures (such as gait patterns or average staying positions) besides the vital signs, the person ID may provide strong shortcuts

for predicting breath signals. For example, the network may learn the person ID through the gait pattern of each individual to reconstruct the subject-dependent respiration signal. Such shortcuts not only degrade the generalizability of the model for unseen subjects but also contradict our intention for the RF-based nRRM task.

To address this problem, we devise an ID discriminator that operates also in an adversarial manner during the training, similar to the case of the GM decoder. We first construct a network for ID discrimination, which consists of three fully-connected and soft-max layers to take the flattened features extracted from the encoder E as input and classify the person ID as output. Denoting this discriminator as $D_{ID}(\cdot)$, the adversarial training between D_{ID} and encoder F can be achieved using a cross-entropy loss function:

$$\min_{E} \max_{D_{ID}} V_{ID} = \sum_{n=1}^{N} \left(\mathbf{y}_{ID}^{gt}\right)_n \cdot \log\left(\left(\mathbf{y}_{ID}^{pred}\right)_n\right), \tag{7}$$

where N denotes the total number of subjects in the training data, $\mathbf{y}_{ID}^{pred}$ is the N-length output vector representing the probability for the person ID, and $\mathbf{y}_{ID}^{gt}$ is the one-hot encoded ground-truth vector. $(\cdot)_n$ represents the n-th element of an arbitrary vector.

In summary, two decoders for RM and GM, and one ID discriminator are trained together in an end-to-end manner based on MTAL. Therefore, the overall loss can be defined as:

$$\begin{aligned}\min_{\{E,D_{RM}\}} \max_{\{D_{GM},D_{ID}\}} V = &\left\|\mathbf{y}_{RM}^{pred} - \mathbf{y}_{RM}^{gt}\right\|_1 \\ &- \eta_1 \left\|\mathbf{y}_{GM}^{pred} - \mathbf{y}_{GM}^{gt}\right\|_1 + \eta_2 \sum_{n=1}^{N} \left(\mathbf{y}_{ID}^{gt}\right)_n \cdot \log\left(\left(\mathbf{y}_{ID}^{pred}\right)_n\right),\end{aligned} \tag{8}$$

where $\eta_1 = 0.3$ and $\eta_2 = 0.2$ are the balancing factors, which have been selected empirically in our experiments. Note that the proposed MTAL strategy aims at developing a RM decomposition model in the presence of large body motions which can similarly be applied based on other input modalities (e.g. video), but it is worthwhile to adopt RF signal given that the GM-dominant self-supervision can intrinsically be provided.

5 Experimental Results

5.1 Datasets and Experimental Setup

Since there is no public dataset for the RF-based nRRM tasks, we collected two datasets for the static/moving settings. For acquiring RF data, we utilized a commercial FMCW radar (IWR1443BOOST, Texas Instruments Inc.) operating in the 77 GHz frequency band with a 1000 pulse repetition frequency. The

following details the collected datasets.

RRM-static RRM-static dataset contains 2.4 h of synchronized RF reflected signals, uncompressed RGB videos captured at 1280×720 resolution and 30 fps through a Razer Kiyo Pro webcam, and ground-truth respiration signals recorded from the contact chest belt. The measurements were collected from 13 subjects in an indoor room, in which each individual was requested to sit in a chair and face forward, ensuring quasi-stationary settings. The participants were also asked to hold their breath periodically during the experiments to generate negative data samples.

RRM-moving RRM-moving dataset is obtained under conditions similar to RRM-static; however, this case was based on non-stationary, i.e., moving settings, where 13 participants were able to stand and even walk around, reflecting more challenging and realistic scenarios such as staggering, looking backward, and turning around. This dataset spans 7 h of random movements and includes some negative samples regarding walking around while holding one's breath.

Implementation Details. The overall algorithm for the RF-vital model was implemented based on 10 s of sequential frame data with a sliding window of 2.5-s intervals, resulting in 3527/10171 RF frames for RRM-static and RRM-moving, respectively. The received RF signals were transformed into RJTF maps using STFT based on a Hann window of 300 ms duration, hop length of 60 ms, and FFT size of 256. To train the network, we adopted ADAM [20] optimizer with a learning rate of 0.0001 and a batch size of 64.

Regarding the quantitative evaluation of the nRRM algorithm, we followed the protocols in [28]. That is, we measure the RRs of a person by post-processing the output signals through a band pass filter with a [0.08 Hz, 0.6 Hz] passband range, which are then compared with real RR measurements in beats per minute unit (BPM) using several standard metrics: mean absolute error (MAE), root mean square error (RMSE), standard deviation (Std), and Pearson's correlation coefficient (ρ). For train-test split, the datasets were divided into 13 folds corresponding to each participant so that the network model could be trained and tested through subject-independent 13-fold cross-validation.

5.2 Quantitative results

We compare the proposed RF-vital model with seven state-of-the-art non-contact vital monitoring baselines (three video- [3,35,28] and four RF-based methods [45,31,14,56]). Regarding the video-based methods, we used 10-s video clips corresponding to the RF data, and center-cropped them to 400×400 pixels to focus only on the facial areas.

Considering the left side of Table 1, our RF-vital model outperforms the previous baselines under static conditions, achieving a 51.8% reduction in MAE and 57.1% in RMSE. Furthermore, the right side of Table 1 demonstrates the feasibility of realizing nRRM even in moving conditions, where the previous models

Table 1. Quantitative comparison of the RF-vital and seven baseline methods based on the RRM-static and RRM-moving datasets

Method	**Input**	**RRM-static (BPM)**				**RRM-moving (BPM)**			
		MAE↓	RMSE↓	ρ↑	Std↓	MAE↓	RMSE↓	ρ↑	Std↓
CAN [3]	RGB	3.16	5.83	0.57	5.21	Not applicable			
Nowara *et al.* [35]	RGB	2.51	4.58	0.67	4.25				
MTTS-CAN [28]	RGB	2.65	4.13	0.69	4.04				
Tu *et al.* [45]	RF (1D)	5.46	7.31	0.19	4.86				
Mercuri *et al.* [31]	RF (1D)	2.52	5.64	0.54	5.47				
Zheng *et al.* [56]	RF (1D)	1.68	3.82	0.72	3.45				
Ha *et al.* [14]	RF (1D)	1.37	3.36	0.75	3.21				
RF-vital	RF (2D)	**0.66**	**1.44**	**0.88**	**1.43**	**3.67**	**7.02**	**0.32**	**6.39**

Table 2. Comparison between different input RF formats

Model Input	MAE↓	RMSE↓	ρ↑	Std↓
unwrapped phase signal	-	-	-	-
RJTFmap (phase only)	4.92	7.20	0.23	6.97
RJTFmap	**3.67**	**7.02**	**0.32**	**6.39**

Table 3. Estimation performance for different combinations of the decoding branches

Use of Decoder	MAE↓	RMSE↓	ρ↑	Std↓
RM only	5.04	7.96	0.26	7.15
RM + ID	4.76	7.64	0.30	6.85
RM + GM	3.85	7.10	**0.36**	6.98
RM + GM + ID	**3.67**	**7.02**	0.32	**6.39**

completely fail because of inconsistent facial tracking induced from the erratic and occluded face regions (for video-based approaches), or significant ambiguity for prerequisite distance estimation (for RF-based approaches). Particularly, it is remarkable that the RR estimation results of our model under moving conditions are comparable to those of Tu *et al.* [45] under static cases.

5.3 Ablation Study

For further in-depth analysis of the effectiveness of each component in the RF-vital model, we conduct ablation studies based on the RRM-moving dataset.

RJTF Map. To analyze the potential utility of the proposed RJTF map, we investigated the numerical performance by changing the input modality for training our RF-vital model. As candidates for the model input, we adopt unwrapped phase signals (i.e., $\theta(t)$ in Eq. (4)) widely utilized in RF-based nRRM methods [31,14,5,9,39,17,44], four-channel RJTF maps based only on RF phase, and eight-channel RJTF maps based on both amplitude and phase components. As shown

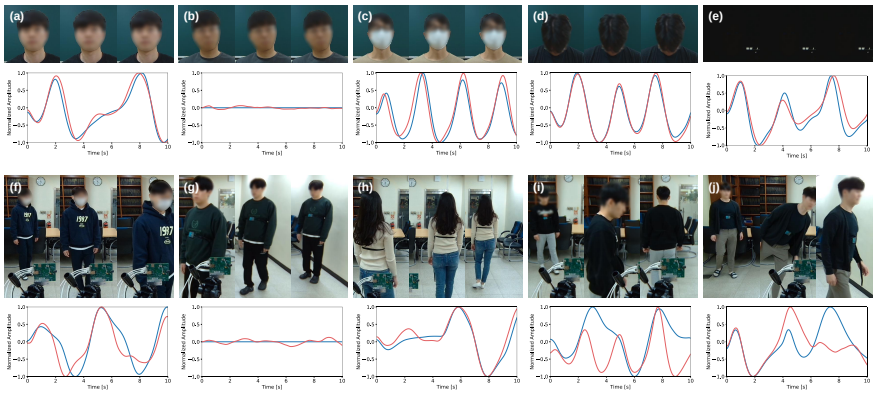

Fig. 4. Qualitative results of our RF-vital model for various realistic scenarios. The first and second rows show the reference video samples and corresponding estimation results under stationary cases. The third and fourth rows show the results under more challenging moving conditions. The ground-truth and the predicted respiration signals are indicated by blue and red lines, respectively. Note that each signal is determined to be the case of holding breath if the average absolute amplitude is less than 0.2

in Table 2, the model with the unwrapped phase signal fails entirely owing to huge unwrapping errors (i.e., distance estimation errors), generating only random jitters for the network output. This implies the inability of the conventional direct distance estimations for involving respiration signatures in the presence of GM. On the contrary, we observe that the proposed RJTF map can present a solution to resolve the estimation ambiguity problem using Doppler effect. In particular, exploiting the amplitude-based spectrograms as well with the phase spectrograms can further improve the performance, reducing the MAE by 25.4% and RMSE by 2.5% compared to the phase-only RJTF maps.

MTAL Strategy. We explore the effectiveness of the proposed MTAL strategy. Specifically, we trained the network with respect to three different combinations of decoding pipelines and evaluated the measurement performance of each model (Table 3). The comparison between the cases with and without the GM decoder clearly verifies the efficacy of the adversarial training on the GM components (23.6% and 10.8% reduction in MAE and RMSE, respectively) in encouraging the network to focus more on the desired RM components. Furthermore, considering the last row of Table 3, we observe that the proposed ID discriminator can provide an additional reduction in MAE by 4.7% and RMSE by 1.1%, demonstrating its potential to improve the model generalizability for an unseen person.

5.4 Qualitative Results

Fig. 4 visualizes the qualitative outcomes of our RF-vital model under stationary/moving conditions. Also, we measure the robustness on occlusion and poor

light conditions based on additional test samples collected from occluded faces or dark illuminance. Each example in the figure represents the real/estimated respiration signal and its corresponding RGB scene.

Results for Stationary Cases. The results for a static person (the first and second rows of the figure) show that the proposed RF-vital model provides distinct outputs between a person with regular breathing [Fig. 4(a)] and a person holding breath [Fig. 4(b)], implying that our model can serve as a promising non-contact solution for people with respiratory disorders. Particularly, because radio reflections convey respiratory signs through the depth of the body instead of the exposed skin surface, it is possible to conduct stable and privacy-preserving predictions even when a person wears a mask [Fig. 4(c)] or bows her/his head [Fig. 4(d)]. Furthermore, we observe that our RF-vital model maintains robustness in a dark setting [Fig. 4(e)], in which the video-based approaches are likely to suffer from significant performance degradation.

Results for Moving Cases and Limitations. The third and fourth rows in the figure demonstrate that our model still works for a moving subject. It can be noticed that the predicted outcomes reflect the respiratory signs of each individual walking toward the sensor [Fig. 4(f)] or with her/his back [Fig. 4(h)], under various spatial angles. In addition, the model can certainly factor out the unusual cases of walking around while holding one's breath [Fig. 4(g)].

However, we observed some failure cases in the RF-vital model under the scenarios with rapid movements. For example, when a person suddenly changes direction, the model generates erroneous signals as shown in [Fig. 4(i)]. Moreover, the RF-vital model tends to show vulnerability to large motions in the vertical direction, such as a large faltering or falling [Fig. 4(j)]. Such failures may have been affected by misalignment with the chest caused from the poor vertical resolution in our sensing system.

6 Conclusion

In this study, we present a novel RF-vital model, the first approach for implementing the nRRM task over a randomly moving individual. We propose the use of radio reflections as an input modality for the RF-vital model, based on its domain property that can capture the microscopic changes in human body depth, while preserving GM-only signals simultaneously. By leveraging the GM-dominant signals as self-supervision, we can devise a MTAL strategy that induces the network to focus more on the desired RM components, while pushing out GM components. The extensive experimental results show that the proposed RF-vital model can provide robust estimations in the presence of a moving person, occluded face, and poor illumination, demonstrating its potentiality for realizing practical vital monitoring solutions.

References

1. Ali, M., Elsayed, A., Mendez, A., Savaria, Y., Sawan, M.: Contact and remote breathing rate monitoring techniques: A review. IEEE Sensors J. **21**(13), 14569–14586 (2021)
2. Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J.: Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognit. Lett. **124**(1), 82–90 (2019)
3. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Eur. Conf. Comput. Vis. (ECCV). pp. 349–365 (September 2018)
4. Chintalapudi, K., Padmanabha Iyer, A., Padmanabhan, V.N.: Indoor localization without the pain. In: ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom). p. 173–184 (2010)
5. Choi, I.O., Kim, M., Choi, J.H., Park, J.K., Park, S.H., Kim, K.T.: Robust cardiac rate estimation of an individual. IEEE Sensors J. **21**(13), 15053–15064 (2021)
6. Choi, J.H., Kim, J.E., Jeong, N.H., Kim, K.T., Jin, S.H.: Accurate people counting based on radar: Deep learning approach. In: IEEE Radar Conf. (RadarConf). pp. 1–5 (2020)
7. Choi, J.H., Kim, J.E., Kim, K.T.: Deep learning approach for radar-based people counting. IEEE Internet Things J. pp. 1–16 (2021)
8. Choi, J.H., Kim, J.E., Kim, K.T.: People counting using IR-UWB radar sensor in a wide area. IEEE Internet Things J. **8**(7), 5806–5821 (2021)
9. Ding, C., Yan, J., Zhang, L., Zhao, H., Hong, H., Zhu, X.: Noncontact multiple targets vital sign detection based on VMD algorithm. In: IEEE Radar Conf. (RadarConf). pp. 0727–0730 (2017)
10. Estepp, J.R., Blackford, E.B., Meier, C.M.: Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In: IEEE Conf. Syst., Man, Cybern. (SMC). pp. 1462–1469 (2014)
11. Fan, L., Li, T., Fang, R., Hristov, R., Yuan, Y., Katabi, D.: Learning longterm representations for person re-identification using radio signals. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 10699–10709 (2020)
12. Fan, L., Li, T., Yuan, Y., Katabi, D.: In-home daily-life captioning using radio signals. In: Eur. Conf. Comput. Vis. (ECCV). pp. 105–123 (2020)
13. Guan, J., Madani, S., Jog, S., Gupta, S., Hassanieh, H.: Through fog high-resolution imaging using Millimeter wave radar. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 11461–11470 (2020)
14. Ha, U., Assana, S., Adib, F.: Contactless seismocardiography via deep learning radars. In: ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom). pp. 1–14 (2020)
15. de Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rPPG. IEEE Trans. Biomed. Eng. **60**(10), 2878–2886 (2013)
16. de Haan, G., Van Leest, A.: A. improved motion robustness of remote-PPG by using the blood volume pulse signature. Physiol. Meas. **35**(9), 1913–1926 (2014)
17. He, M., Nian, Y., Liu, B.: Noncontact heart beat signal extraction based on wavelet transform. In: Int. Conf. Biomed. Eng. Informat. (BMEI). pp. 209–213 (2015)
18. Iovescu, C., Rao, S.: The fundamentals of millimeter wave sensors. Texas Instrum. pp. 1–8 (2017)
19. Jiang, C., Guo, J., He, Y., Jin, M., Li, S., Liu, Y.: mmVib: Micrometer-level vibration measurement with mmWave radar. In: ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom). pp. 1–13 (2020)

20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Int. Conf. Learn. Represent. (ICLR). pp. 1–15 (2015)
21. Kumar, S., Gil, S., Katabi, D., Rus, D.: Accurate indoor localization with zero start-up cost. In: ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom). p. 483–494 (2014)
22. Lam, A., Kuno, Y.: Robust heart rate measurement from video using select random patches. In: Int. Conf. Comput. Vis. (ICCV). pp. 3640–3648 (2015)
23. Lee, E., Chen, E., Lee, C.Y.: Meta-rPPG: Remote heart rate estimation using a transductive meta-learner. In: Eur. Conf. Comput. Vis. (ECCV). pp. 392–409 (2020)
24. Lewandowska, M., Rumiński, J., Kocejko, T., Nowak, J.: Measuring pulse rate with a webcam — A non-contact method for evaluating cardiac activity. In: Fed. Conf. Comput. Sci. Inf. Syst. (FedCSIS). pp. 405–410 (2011)
25. Li, J., Stoica, P.: MIMO radar signal processing. John Wiley & Sons, Hoboken, New Jersey, USA (2008)
26. Li, J., Liu, L., Zeng, Z., Liu, F.: Advanced signal processing for vital sign extraction with applications in uwb radar detection of trapped victims in complex environments. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **7**(3), 783–791 (2014)
27. Li, X., Chen, J., Zhao, G., Pietikäinen, M.: Remote heart rate measurement from face videos under realistic situations. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 4264–4271 (2014)
28. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. In: Adv. Neural Inform. Process. Syst. (NIPS). pp. 1–23 (2020)
29. McDuff, D.: Deep super resolution for recovering physiological information from videos. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (CVPRW). pp. 1480–1487 (2018)
30. McDuff, D.J., Sarah, G., Picard, R.W.: Improvements in remote cardiopulmonary measurement using a five band digital camera. IEEE Trans. Biomed. Eng. **61**(10), 2593–2601 (2014)
31. Mercuri, M., Lorato, I., Liu, Y.H., P. Wieringa, F., Van Hoof, C., Torfs, T.: Vital-sign monitoring and spatial tracking of multiple people using a contactless radar-based sensor. Nature Electron. **2**, 252–262 (2019)
32. Niu, X., Han, H., Shan, S., Chen, X.: SynRhythm: Learning a deep heart rate estimator from general to specific. In: Int. Conf. Pattern Recog. (ICPR). pp. 3580–3585 (2018)
33. Niu, X., Han, H., Shan, S., Chen, X.: VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In: Asian Conf. Comput. Vis. (ACCV). pp. 562–576 (2018)
34. Niu, X., Shan, S., Han, H., Chen, X.: RhythmNet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Trans. Image Process. **29**, 2409–2423 (2020)
35. Nowara, E.M., McDuff, D., Veeraraghavan, A.: The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In: Int. Conf. Comput. Vis. (ICCV). pp. 4955–4964 (2021)
36. Pan, J.J., Pan, S.J., Yin, J., Ni, L.M., Yang, Q.: Tracking mobile users in wireless networks via semi-supervised colocalization. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 587–600 (2012)
37. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt. Express **18**(10), 10762–10774 (2010)

38. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Eng. **58**(1), 7–11 (2011)
39. Ren, W., Qi, F., Foroughian, F., Kvelashvili, T., Liu, Q., Kilic, O., Long, T., Fathy, A.E.: Vital sign detection in any orientation using a distributed radar network via modified independent component analysis. IEEE Trans. Microw. Theory Techn. **69**(11), 4774–4790 (2021)
40. Revanur, A., Li, Z., Ciftci, U.A., Yin, L., Jeni, L.A.: The first vision for vitals (V4V) challenge for non-contact video-based physiological estimation. In: Int. Conf. Comput. Vis. Worksh. (ICCVW). pp. 2760–2767 (2021)
41. Rohling, H.: Radar CFAR thresholding in clutter and multiple target situations. IEEE Trans. Aerosp. Electron. Syst. **AES-19**(4), 608–621 (1983)
42. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Int. Conf. Med. Image Comput. Computer-Assist. Interven. (MICCAI). pp. 234–241 (2015)
43. Scheiner, N., Kraus, F., Wei, F., Phan, B., Mannan, F., Appenrodt, N., Ritter, W., Dickmann, J., Dietmayer, K., Sick, B., Heide, F.: Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using Doppler radar. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 2068–2077 (2020)
44. Tariq, A., Ghafouri-Shiraz, H.: Vital signs detection using Doppler radar and continuous wavelet transform. In: Eur. Conf. Antennas Propag. (EUCAP). pp. 285–288 (2011)
45. Tu, J., Hwang, T., Lin, J.: 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 (2016)
46. Verkruysse, W., Othar Svaasand, L., Stuart Nelson, J.: Remote plethysmographic imaging using ambient light. Opt. Express **16**(26), 21434–21445 (2008)
47. Wang, W., C. den Brinker, A., Stuijk, S., de Haan, G.: Amplitude-selective filtering for remote-PPG. Biomed. Opt. Express **8**(3), 1965–1980 (2017)
48. Wang, W., den Brinker, A.C., Stuijk, S., de Haan, G.: Algorithmic principles of remote PPG. IEEE Trans. Biomed. Eng. **64**(7), 1479–1491 (2017)
49. Wang, W., Stuijk, S., de Haan, G.: Exploiting spatial redundancy of image sensor for motion robust rPPG. IEEE Trans. Biomed. Eng. **62**(2), 415–425 (2015)
50. Xiong, J., Sundaresan, K., Jamieson, K.: ToneTrack: Leveraging frequency-agile radios for time-based indoor wireless localization. In: ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom). p. 537–549 (2015)
51. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: 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). pp. 151–160 (2019)
52. Zhan, Q., Wang, W., de Haan, G.: Analysis of CNN-based remote-PPG to understand limitations and sensitivities. Biomed. Opt. Express **11**(3), 1268–1283 (2020)
53. Zhao, M., Li, T., Alsheikh, M.A., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 7356–7365 (2018)
54. Zhao, M., Liu, Y., Raghu, A., Zhao, H., Li, T., Torralba, A., Katabi, D.: Through-wall human mesh recovery using radio signals. In: Int. Conf. Comput. Vis. (ICCV). pp. 10112–10121 (2019)
55. Zhao, M., Tian, Y., Zhao, H., Alsheikh, M.A., Li, T., Hristov, R., Kabelac, Z., Katabi, D., Torralba, A.: RF-based 3D skeletons. In: Conf. ACM Special Interest Group Data Commun. (SIGCOMM). p. 267–281 (2018)

56. Zheng, T., Chen, Z., Zhang, S., Cai, C., Luo, J.: MoRe-Fi: Motion-robust and fine-grained respiration monitoring via deep-learning UWB radar. In: ACM Conf. Embedded Netw. Sens. Syst. (SenSys). p. 111–124. New York, NY, USA (2021)