In-Home Daily-Life Captioning Using Radio Signals

Lijie Fan*, Tianhong Li*, Yuan Yuan, and Dina Katabi

MIT CSAIL

Abstract. This paper aims to caption daily life –i.e., to create a textual description of people's activities and interactions with objects in their homes. Addressing this problem requires novel methods beyond traditional video captioning, as most people would have privacy concerns about deploying cameras throughout their homes. We introduce RF-Diary, a new model for captioning daily life by analyzing the privacy-preserving radio signal in the home with the home's floormap. RF-Diary can further observe and caption people's life through walls and occlusions and in dark settings. In designing RF-Diary, we exploit the ability of radio signals to capture people's 3D dynamics, and use the floormap to help the model learn people's interactions with objects. We also use a multi-modal feature alignment training scheme that leverages existing video-based captioning datasets to improve the performance of our radio-based captioning model. Extensive experimental results demonstrate that RF-Diary generates accurate captions under visible conditions. It also sustains its good performance in dark or occluded settings, where video-based captioning approaches fail to generate meaningful captions.[1]

1 Introduction

Captioning is an important task in computer vision and natural language processing; it typically generates language descriptions of visual inputs such as images or videos [33,37,10,3,38,23,29,36,17,27,25,35]. This paper focuses on *in-home daily-life captioning*, that is, creating a system that observes people at home, and automatically generates a transcript of their everyday life. Such a system would help older people to age-in-place. Older people may have memory problems and some of them suffer from Alzheimer's. They may forget whether they took their medications, brushed their teeth, slept enough, woke up at night, ate their meals, etc. Daily life captioning enables a family caregiver, e.g., a daughter or son, to receive text updates about their parent's daily life, allowing them to care for mom or dad even if they live away, and providing them peace of mind about the wellness and safety of their elderly parents. More generally, daily-life captioning can help people track and analyze their habits and routine at home, which can empower them to change bad habits and improve their life-style.

But how do we caption people's daily life? One option would be to deploy cameras at home, and run existing video-captioning models on the recorded videos. However, most people would have privacy concerns about deploying cameras at home, particularly in the bedroom and bathroom. Also, a single camera usually has a limited field of view; thus, users would need to deploy multiple cameras covering different rooms, which

* Indicates equal contribution. Correspondence to Tianhong Li <tianhong@mit.edu>.

[1] For more information, please visit our project webpage: http://rf-diary.csail.mit.edu

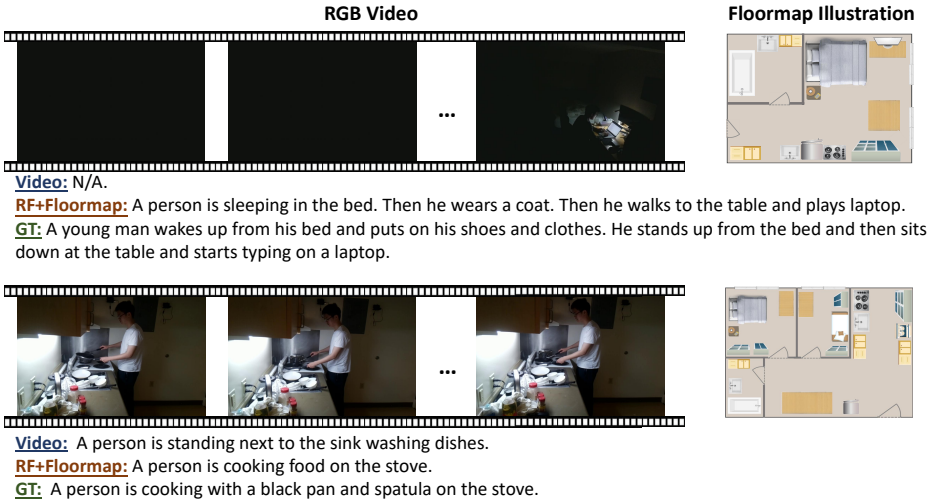


Fig. 1. Event descriptions generated from videos and RF+Floormap. The description generated from video shows its vulnerability to poor lighting and confusing images, while RF-Diary is robust to both. The visualization of floormap shown here is for illustration. The representation used by our model is person-centric and is described in detail in section 4.2.

would introduce a significant overhead. Moreover, cameras do not work well in dark settings and occlusions, which are common scenarios at home.

To address these limitations, we propose to use radio frequency (RF) signals for daily-life captioning. RF signals are more privacy-preserving than cameras since they are difficult to interpret by humans. Signals from a single RF device can traverse walls and occlusions and cover most of the home. Also, RF signals work in both bright and dark settings without performance degradation. Furthermore, the literature has shown that one can analyze the radio signals that bounce off people's bodies to capture people's movements [1,2], and track their 3D skeletons [43].

However, using RF signals also introduces new challenges, as described below:

– **Missing objects information:** RF signals do not have enough information to differentiate objects, since many objects are partially or fully transparent to radio signals. Their wavelength is on the order of a few centimeters, whereas the wavelength of visible light is hundreds nanometer [4]. Thus, it is also hard to capture the exact shape of objects using RF signals.
– **Limited training data:** Currently, there is no training dataset that contains RF signals from people's homes with the corresponding captions. Training a captioning system typically requires tens of thousands of labeled examples. However, collecting a new large captioning dataset with RF in people's homes would be a daunting task.

In this paper, we develop RF-Diary, an RF-based in-home daily-life captioning model that addresses both challenges. To capture objects information, besides RF signals, RF-Diary also takes as input the home floormap marked with the size and location of static objects like bed, sofa, TV, fridge, etc. Floormaps provide information about the surrounding environment, thus enabling the model to infer human interactions with

objects. Moreover, floormaps are easy to measure with a cheap laser-meter in less than 10 minutes (Section 4.2). Once measured, the floormap remains unchanged for potentially years, and can be used for all future daily-life captioning from that home.

RF-Diary proposes an effective representation to integrate the information in the floormap with that in RF signals. It encodes the floormap from the perspective of the person in the scene. It first extracts the 3D human skeletons from RF signals as in [43] and then at each time step, it shifts the reference system of floormap to the location of the extracted skeleton, and encodes the location and orientation of each object with respect to the person in the scene. This representation allows the model, at each time step, to focus on various objects depending on their proximity to the person.

To deal with the limited availability of training data, we propose a multi-modal feature alignment training scheme to leverage existing video-captioning datasets for training RF-Diary. To transfer visual knowledge of event captioning to our model, we align the features generated from RF-Diary to the same space of features extracted from a video-captioning model trained on existing large video-captioning datasets. Once the features are aligned, we use a language model to generate text descriptions.

Figure 1 shows the performance of our model in two scenarios. In the first scenario, a person wakes up from bed, puts on his shoes and clothes, and goes to his desk to work on his laptop. RF-Diary generates a correct description of the events, while video-captioning fails due to poor lighting conditions. The second scenario shows a person cooking on the stove. Video-captioning confuses the person's activity as washing dishes because, in the camera view, the person looks as if he were near a sink full of dishes. In contrast, RF-Diary generates a correct caption because it can extract 3D information from RF signals, and hence can tell that the person is near the stove not the sink.

To evaluate RF-Diary, we collect a captioning dataset of RF signals and floormaps, as well as the synchronized RGB videos. Our experimental results demonstrate that: 1) RF-Diary can obtain comparable results to video-captioning in visible scenarios. Specifically, on our test set, RF-Diary achieves 41.5 average BLEU and 26.7 CIDEr, while RGB-based video-captioning achieves 41.1 average BLEU and 27.0 CIDEr. 2) RF-Diary continues to work effectively in dark and occluded conditions, where video-captioning methods fail. 3) Finally, our ablation study shows that the integration of the floormaps into the model and the multi-modal feature alignment both contribute significantly to improving performance.

Finally, we summarize our contributions as follows:

- We are the first to caption people's daily-life at home, in the presence of bad lighting and occlusions.
- We also introduce new modalities: the combination of RF and floormap, as well as new representations for both modalities that better tie them together.
- We further introduce a multi-modal feature alignment training strategy for knowledge transfer from a video-captioning model to RF-Diary.
- We evaluate our RF-based model and compare its performance to past work on video-captioning. Our results provide new insights into the strengths and weaknesses of these two types of inputs.

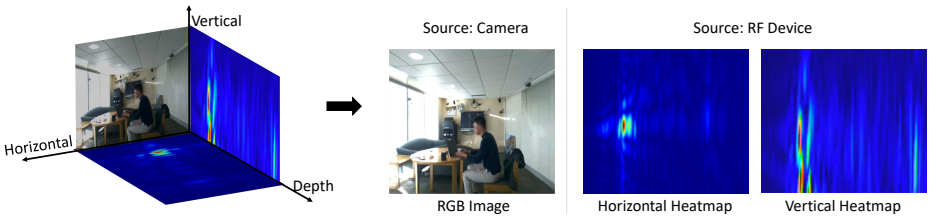


Fig. 2. RF heatmaps and an RGB image recorded at the same time.

2 Related Work

(a) RGB-Based Video Captioning. Early works on video-captioning are direct extensions of image captioning. They pool features from individual frames across time, and apply image captioning models to video-captioning [34]. Such models cannot capture temporal dependencies in videos. Recent approaches, e.g., sequence-to-sequence video-to-text (S2VT), address this limitation by adopting recurrent neural networks (RNNs) [33]. In particular, S2VT customizes LSTM for video-captioning and generates natural language descriptions using an encoder-decoder architecture. Follow-up papers improve this model by introducing an attention mechanism [37,10,22], leveraging hierarchical architectures [3,38,17,29,12,13], or proposing new ways to improve feature extraction from video inputs, such as C3D features [37] or trajectories [36]. There have also been attempts to use reinforcement learning to generate descriptions from videos [27,25,35], in which they use the REINFORCE algorithm to optimize captioning scores.

(b) Human Behavior Analysis with Wireless Signals. Recently, there has been a significant interest in analyzing the radio signals that bounce off people's bodies to understand human movements and behavior. Past papers have used radio signals to track a person's location [1], extract 3D skeletons of nearby people [43], or do action classification [19]. To the best of our knowledge, we are the first to generate natural language descriptions of continuous and complex in-home activities using RF signals. Moreover, we introduce a new combined modality based on RF+Floormap and a novel representation that highlights the interaction between these two modalities, as well as a multi-modal feature alignment training scheme to allow RF-based captioning to learn from existing video captioning datasets.

3 RF Signal Preliminary

In this work, we use a radio commonly used in prior works on RF-based human sensing [43,14,15,20,26,31,7,40,42,39,44,11,16]. The radio has two antenna arrays organized vertically and horizontally, each equipped with 12 antennas. The antennas transmit a waveform called FMCW [30] and sweep the frequencies from 5.4 to 7.2 GHz. Intuitively, the antenna arrays provide angular resolution and the FMCW provides depth resolution.

Our input RF signal takes the form of two-dimensional heatmaps, one from the horizontal array and the other from the vertical array. As shown in Figure 2, the horizontal heatmap is similar to a depth heatmap projected on a plane parallel to the ground, and

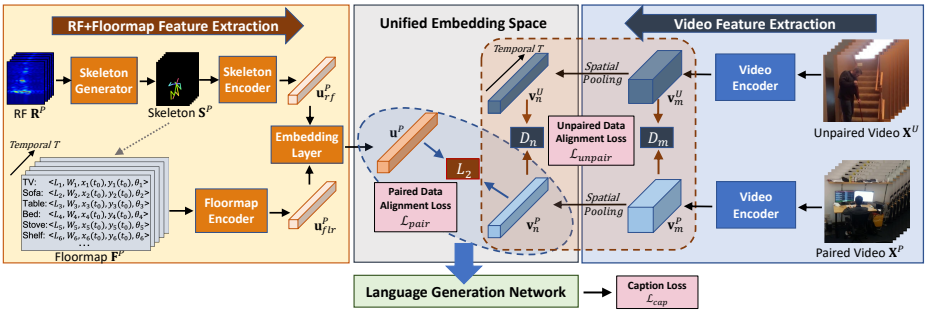


Fig. 3. Model architecture. It contains four parts: RF+Floormap feature extraction (the left yellow box), video feature extraction (the right blue box), unified embedding space for feature alignment (the center grey box), and language generation network (the bottom green box). RF-Diary extracts features from RF signals and floormaps and combines them into a unified human-environment feature map. The features are then taken by the language generation network to generate captions. RF-Diary also extracts features from both paired videos (synchronized videos with RF+Floormap) and unpaired videos (an existing video captioning dataset), and gets the video representation. These features are used to distill knowledge from existing video dataset to RF-Diary. During training, RF-Diary uses the caption loss and the feature alignment loss to train the network. During testing, RF-Diary takes only the RF+Floormap without videos as input and generates captions.

the vertical heatmap is similar to a depth heatmap projected on a plane perpendicular to the ground. Red parts in the figure correspond to large RF power, while blue parts correspond to small RF power. The radio generates 30 pairs of heatmaps per second.

RF signals are different from vision data. They contain much less information than RGB images. This is because the wave-length of RF signals is few centimeters making it hard to capture objects' shape using RF signals; they may even totally miss small objects such as a pen or cellphone. However, the FMCW radio enables us to get a relatively high resolution on depth (∼8cm), making it much easier to locate a person. We harness this property to better associate RF signals and floormaps in the same coordinate system.

4 RF-Diary

RF-Diary generates event captions using RF signals and floormaps. As shown in Figure 3, our model first performs feature extraction from RF signals and floormaps, then combine them into a unified feature (the left yellow box). The combined feature is taken by a language generation network to generate captions (the bottom green box). Below, we describe the model in detail. We also provide implementation details in Appendix A.

4.1 RF Signal Encoding

RF signals have properties totally different from visible light, which are usually not interpretable by human. Therefore, it can be hard to directly generate captions from RF signals. However, recent works demonstrate that it is possible to generate accurate human skeletons from radio signals [43], and that the human skeleton is a succinct yet informative representation for human behavior analysis [18,9]. In this work, we first

generate 3D human skeletons from RF signals, then extract the feature representations of the 3D skeletons.

Thus, the first stage in RF-Diary is a skeleton generator network, which has an architecture similar to the one in [43], with 90-frame RF heatmaps (3 seconds) as input; we refer to these 90 frames as an RF segment. The skeleton generator first extracts information from the RF segment with a feature extraction network (12-layer ResNet). This is followed by a region proposal network (6-layer ResNet) to generate region proposals for potential human bounding boxes and a pose estimation network (2-layer ResNet) to generate the final 3D skeleton estimations based on the feature maps and the proposals. Note that these are dynamic skeletons similar to skeletons extracted from video segment.

After we obtain the 3D skeletons from RF signals, we extract the feature representation through a skeleton encoder from each skeleton segment **S**. The skeleton encoder is a Hierarchical Co-occurrence Network (HCN) [18], which is a CNN-based network architecture for skeleton-based action recognition. We use the features from the last convolutional layer of HCN, denoted as $\mathbf{u}_{rf}$, as the encoded features for RF signals.

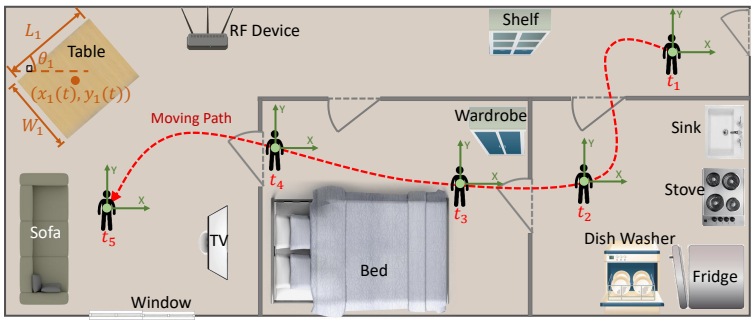


Fig. 4. Illustration of floormap representation. Noted that this figure is not the input to our model but only a visualization. Red dotted line in the apartment denotes the moving path of a person from time t_1 to t_5. Green axes X-Y centered at the person illustrate our person-centric coordinate system, where the origin of the coordinate system is changed through time along with people's location. Under this person-centric coordinate system, at t^{th} time step, we describe each object using a 5-element tuple: (length L, width W, center coordinates $x(t), y(t)$, and rotation θ), as exemplified using the **Table** in the figure.

4.2 Floormap Encoding

Many objects are transparent or semi-transparent to RF signals and act in a manner similar to air [2]. Even objects that are not transparent to RF signals, they can be partially invisible; this is because they can deflect the incident RF signal away from the emitting radio, preventing the radio from sensing them [41,43]. Thus, to allow the model to understand interactions with objects, we must provide additional information about the surrounding environment. But we do not need to have every aspect of the home environment since most of the background information, e.g. the color or the texture

of furniture, is irrelevant to captioning daily life. Therefore, we represent the in-home environment using the locations and sizes of objects – the floormap. The floormap is easily measured with a laser meter. Once measured, the floormap tends to remain valid for a long time. In our model, we only consider static objects relevant to people's in-home activities, e,g., bed, sofa, stove, or TV. To demonstrate the ease of collecting such measurements, we have asked 5 volunteers to measure the floormap for each of our test environments. The average time to measure one environment is about 7 mins.

Let M be the number of objects, N be the maximum number of instances of an object, then the floormap can be represented by a tensor $f \in \mathbb{R}^{M\times N\times O}$, where O denotes the dimension for the location and size of each object, which is typically 5, i.e., length L, width W, the coordinate of the center $x(t), y(t)$, and its rotation θ.

Since people are more likely to interact with objects close to them, we set the origin point of the floormap reference system to the location of the 3D skeleton extracted from RF. Specifically, we use a person-centric coordinate system as shown in the green X-Y coordinates in Figure 4. A person is moving around at home in a red-dotted path from time t_1 to t_5. At each time step t_i, we set the origin of the 3D coordinate system to be the center of the 3D human skeleton and the X-Y plane to be parallel to the floor plane. The orientation of the X-axis is parallel to the RF device and Y-axis is perpendicular to the device. Each object is then projected onto the X-Y plane. For example, at time t_i, the center coordinates of the **Table** is $(x_1(t_i), y_1(t_i))$, while its width, length and rotation (L_1, W_1, θ_1) are independent of time. The floormap at time t_i is thus generated by describing each object k using a 5-element tuple: $(L_k, W_k, x_k(t_i), y_k(t_i), \theta_k)$, as shown in the left yellow box in Figure 3. In this way, each object's location is encoded w.r.t. the person's location, allowing our model to pay different attention to objects depending on their proximity to the person at that time instance.

To extract features of the floormaps **F**, we use a floormap encoder which is a two-layer fully-connected network which generates the encoded features for floormaps $\mathbf{u}_{flr}$.

Using the encoded RF signal features $\mathbf{u}_{rf}$ and floormap features $\mathbf{u}_{flr}$, we generate a unified human-environment feature:

$$\mathbf{u} = \psi(\mathbf{u}_{rf} \oplus \mathbf{u}_{flr}),$$

where $\oplus$ denotes the concatenation of vectors, and ψ denotes an encoder to map the concatenated features to a joint embedding space. Here we use another two-layer fully-connected sub-network for ψ.

4.3 Caption Generation

To generate language descriptions based on the extracted features, we use an attention-based sequence-to-sequence LSTM model similar to the one in [33,35]. During the encoding stage, given the unified human-environment feature $\mathbf{u} = \{u_t\}_1^T$ with time dimension T, the encoder LSTM operates on its time dimension to generate hidden states $\{h_t\}_1^T$ and outputs $\{o_t\}_1^T$. During the decoding stage, the decoder LSTM uses h_T as an initialization for hidden states and take inputs of previous ground-truth words with an attention module related to $\{u_t\}_1^T$, to output language sequence with m words $\{w_1, w_2, ..., w_m\}$. The event captioning loss $\mathcal{L}_{cap}(\mathbf{u})$ is then given by a negative-log-likelihood loss between the generated and the ground truth caption similar to [33,35].

5 Multi-Modal Feature Alignment Training

Training RF-Diary requires a large labeled RF captioning dataset. Collecting such a dataset would be a daunting task. To make use of the existing large video-captioning dataset (e.g., Charades), we use a multi-model feature alignment strategy in the training process to transfer knowledge from video-captioning to RF-Diary. However, RGB videos from Charades and RF signals have totally different distributions both in terms of semantics and modality. To mitigate the gap between them and make the knowledge distillation possible, we collect a small dataset where we record synchronized videos and RF from people performing activities at home. We also collect the floormaps and provide the corresponding natural language descriptions (for dataset details see section 6(a)). The videos in the small dataset are called paired videos, since they are in the same semantic space as their corresponding RF signal, while the videos in large existing datasets are unpaired videos with the RF signals. Both the paired and unpaired videos are in the same modality. Since the paired videos share the same semantics with the RF data, and the same modality with the unpaired videos, they can work as a bridge between video-captioning and RF-Diary, and distill knowledge from the video data to RF data.

Our multi-modal feature alignment training scheme operates by aligning the features from RF+Floormaps and those from RGB videos. During training, our model extracts features from not only RF+Floormaps, but also paired and unpaired RGB videos, as shown in Figure 3 (the right blue box). Besides the captioning losses, we add additional paired-alignment loss between paired videos and RF+Floormaps and unpaired-alignment loss between paired and unpaired videos. This ensures features from the two modalities are mapped to a unified embedding space (the center grey box). Below, we describe the video encoder and the feature alignment method in detail.

5.1 Video Encoding

We use the I3D model [5] pre-trained on Kinetics dataset to extract the video features. For each 64-frame video segment, we extract the Mixed_5c features from I3D model, denoted as v_m. We then apply a spatial average pooling on top of the Mixed_5c feature and get the spatial pooled video-segment feature v_n. For a video containing T non-overlapping video-segment, its Mixed_5c features and the spatial-pooled features are denoted as $\mathbf{v}_m = \{v_m(t)\}_1^T$ and $\mathbf{v}_n = \{v_n(t)\}_1^T$. Therefore, the extracted features of paired videos $\mathbf{X}^P$ and unpaired videos $\mathbf{X}^U$ are denoted as $\mathbf{v}_m^P$, $\mathbf{v}_n^P$ and $\mathbf{v}_m^U$, $\mathbf{v}_n^U$. We use the spatial-pooled features to generate captions through the language generation model. The corresponding captioning loss is $\mathcal{L}_{cap}(\mathbf{v}_n^P)$ and $\mathcal{L}_{cap}(\mathbf{v}_n^U)$.

5.2 Alignment of Paired Data

Since the synchronized video and RF+Floormap correspond to the exact same event, we use L_2 loss to align the features from paired video $\mathbf{v}_n^P$ in Sec 5.1 and RF+Floormap $\mathbf{u}^P$ in Sec 4.2 (we deonte a P here to indicate the paired data) to be consistent with each other in a unified embedding space, i.e., the paired data alignment loss $\mathcal{L}_{pair}(\mathbf{u}^P, \mathbf{v}_n^P) = \left\|\mathbf{u}^P - \mathbf{v}_n^P\right\|_2$.

5.3 Alignment of Unpaired Data

Existing large video-captioning datasets have neither synchronized RF signal nor the corresponding floormaps, so we cannot use the L_2-norm for alignment. Since we collect a small paired dataset, we can first train a video-captioning model on both paired and unpaired datasets, and then use the paired dataset to transfer knowledge to RF-Diary. However, the problem is that since the paired feature alignment is only applied on the paired dataset, the video-captioning model may overfit to the paired dataset and cause inconsistency between the distribution of features from paired and unpaired datasets. To solve this problem, we align the paired and unpaired datasets by making the two feature distributions similar. We achieve this goal by applying discriminators on different layers of video features that enforces the video encoder to generate indistinguishable features given $\mathbf{X}^P$ and $\mathbf{X}^U$. Specifically, we use two different layers of video features, i.e., $\mathbf{v}_m$ and $\mathbf{v}_n$ in Sec 5.1, to calculate the discriminator losses $\mathcal{L}_{unpair}(\mathbf{v}_m^P, \mathbf{v}_m^U)$ and $\mathcal{L}_{unpair}(\mathbf{v}_n^P, \mathbf{v}_n^U)$. Since features from the paired videos are also aligned with the RF+Floormap features, this strategy effectively aligns the feature distribution of the unpaired video with the feature distribution of RF+Floormaps. The total loss of training process is shown as below:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{cap}(\mathbf{u}^P) + \mathcal{L}_{cap}(\mathbf{v}_n^P) + \mathcal{L}_{cap}(\mathbf{v}_n^U) \\ +\mathcal{L}_{pair}(\mathbf{u}^P, \mathbf{v}_n^P) \\ +\mathcal{L}_{unpair}(\mathbf{v}_n^P, \mathbf{v}_n^U) + \mathcal{L}_{unpair}(\mathbf{v}_m^P, \mathbf{v}_m^U).\end{aligned}$$

6 Experiments

(a) Datasets: We collect a new dataset named RF Captioning Dataset (RCD). It provides synchronized RF signals, RGB videos, floormaps, and human-labeled captions to describe each event. We generate floormaps using a commercial laser meter. The floormaps are marked with the following objects: cabinet, table, bed, wardrobe, shelf, drawer, stove, fridge, sink, sofa, television, door, window, air conditioner, bathtub, dishwasher, oven, bedside table. We use a radio device to collect RF signals, and a multi-camera system to collect multi-view videos, as the volunteers perform the activities. The synchronized radio signals and videos have a maximum synchronization error of 10 ms. The multi-camera system has 12 viewpoints to allow for accurate captioning even in cases where some viewpoints are occluded or the volunteers walk from one room to another room.

To generate captions, we follow the method used to create the Charades dataset [28] –i.e., we first generate instructions similar to those used in Charades, ask the volunteers to perform activities according to the instructions, and record the synchronized RF signals and multi-view RGB videos. We then provide each set of multi-view RGB videos to Amazon Mechanical Turk (AMT) workers and ask them to label 2-4 sentences as the ground-truth language descriptions.

We summarize our dataset statistics in Table 1. In total, we collect 1,035 clips in 10 different in-door environments, including bedroom, kitchen, living room, lounge, office, etc. Each clip on average spans 31.3 seconds. The RCD dataset exhibits two types of

#environments	#clips	avg len	#action types	#object types	#sentences	#words	vocab	len (hrs)
10	1,035	31.3s	157	38	3,610	77,762	6,910	8.99

Table 1. Statistics of our RCD dataset.

diversity. *1. Diversity of actions and objects:* Our RCD dataset contains 157 different actions and 38 different objects to interact with. The actions and objects are the same as the Charades dataset to ensure a similar action diversity. The same action is performed at different locations by different people, and different actions are performed at the same location. For example, all of the following actions are performed in the bathroom next to the sink: brushing teeth, washing hands, dressing, brushing hair, opening/closing a cabinet, putting something on a shelf, taking something off a shelf, washing something, etc. *2. Diversity of environments:* Environments in our dataset differ in their floormap, position of furniture, and the viewpoint of the RF device. Further, each environment and all actions performed in that environment are included either in testing or training, but not both.

(b) Train-test Protocol: To evaluate RF-Diary under visible scenarios, we do a 10-fold cross-validation on our RCD Dataset. Each time 9 environments are used for training, and the other 1 environment is used for testing. We report the average performance of 10 experiments. To show the performance of RF-Diary under invisible scenarios, e.g., with occlusions or poor lighting conditions, we randomly choose 3 environments (with 175 clips) and collect corresponding clips under invisible conditions. Specifically, in these 3 environments, we ask the volunteers to perform the same series of activities twice under the visible and invisible conditions (with the light on and off, or with an occlusion in front of the RF device and cameras), respectively. Later we provide the same ground truth language descriptions for the clips under invisible conditions as the corresponding ones under visible conditions. During testing, clips under invisible scenarios in these 3 environments are used for testing, and clips in the other 7 environments are used to train the model.

During training, we only use RGB videos from 3 cameras with good views instead of all 12 views in the multi-modal feature alignment between the video-captioning model and RF-Diary model. Using multi-view videos will provide more training samples and help the feature space to be oblivious to the viewpoint. When testing the video-captioning model, we use the video from the master camera as it covers most of the scenes. The master camera is positioned atop of the RF device for a fair comparison.

We leverage the Charades caption dataset [28,35] as the unpaired dataset to train the video-captioning model. This dataset provides captions for different in-door human activities. It contains 6,963 training videos, 500 validation videos, and 1,760 test videos. Each video clip is annotated with 2-5 language descriptions by AMT workers.

(c) Evaluation Metrics: We adopt 4 caption evaluation scores widely used in video-captioning: BLEU [24], METEOR [8], ROUGE-L [21] and CIDEr [32]. BLEU-n analyzes the co-occurrences of n words between the predicted and ground truth sentences. METEOR compares exact token matches, stemmed tokens, paraphrase matches, as well as semantically similar matches using WordNet synonyms. ROUGE-L measures the longest common subsequence of two sentences. CIDEr measures consensus in captions

by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n words. According to [28], CIDEr offers the highest resolution and most similarity with human judgment on the captioning task for in-home events. We compute these scores using the standard evaluation code provided by the Microsoft COCO Evaluation Server [6]. All results are obtained as an average of 5 trials. We denote B@n, M, R, C short for BLEU-n, METEOR, ROUGE-L, CIDEr.

6.1 Quantitative Results

We compare RF-Diary with state-of-the-art video-captioning baselines [17,33,37,10]. The video-based models are trained on RGB data from both the Charades and RCD training sets, and tested on the RGB data of the RCD test set. RF-Diary is trained on RF and floormap data from the RCD training sets. It also uses the RGB data from Charades and RCD training sets in multi-modal feature alignment training. It is then tested on the RF and floormap data of the RCD test set.

The results on the left side of Table 2 show that RF-Diary achieves comparable performance to state-of-the-art video captioning baselines in visible scenarios. The right side of Table 2 shows that RF-Diary also generates accurate language descriptions when the environment is dark or with occlusion, where video-captioning fails completely. The little reduction in RF-Diary's performance from the visible scenario is likely due to that occlusions attenuate RF signals and therefore introduce more noise in the RF heatmaps.

Methods	Visible Scenario							Dark/Occlusion Scenario						
	B@1	B@2	B@3	B@4	M	R	C	B@1	B@2	B@3	B@4	M	R	C
S2VT [33]	57.3	40.4	27.2	19.3	19.8	27.3	18.9	-	-	-	-	-	-	–
SA [37]	56.8	39.2	26.7	19.0	18.1	25.9	22.1	-	-	-	-	-	-	–
MAAM [10]	57.8	41.9	28.2	19.3	20.7	27.1	21.2	-	-	-	-	-	-	–
HTM [17]	61.3	44.6	32.2	22.1	21.3	28.3	26.5	-	-	-	-	-	-	–
HRL [35]	**62.5**	45.3	32.9	**23.8**	**21.7**	28.5	**27.0**	-	-	-	-	-	-	–
RF-Diary	62.3	**45.9**	**33.9**	23.5	21.1	**28.9**	26.7	**61.5**	**45.5**	**33.1**	**22.6**	**21.1**	**28.3**	**25.9**

Table 2. Quantitative results for RF-Diary and video-based captioning models. All models are trained on Charades and RCD training set, and tested on the RCD test set. The left side of the Table shows the results under visible scenarios, and the right side of the Table shows the results under scenarios with occlusions or without light.

6.2 Ablation Study

We conduct several ablation studies to demonstrate the necessity of each component in RF-Diary. All experiments here are evaluated on the visible test set of RCD.

3D Skeleton vs. Locations: One may wonder whether simply knowing the location of the person is enough to generate a good caption. This could happen if the RCD dataset has low diversity, i.e., each action is done in a specific location. This is however not the case in the RCD dataset, where each action is done in multiple locations, and each location may have different actions. To test this point empirically, we compare our model which extracts 3D skeletons from RF signals with a model that extracts only people locations from RF. We also compare with a model that extracts 2D skeletons with no locations (in this case the floormap's coordinate system is centered on the RF device).

Method	B@1	B@2	B@3	B@4	M	R	C
Locations	52.0	37.4	24.3	17.2	15.7	22.1	19.1
2D Skeletons	56.5	39.8	26.9	18.8	18.0	24.1	22.3
3D Skeletons	**62.3**	**45.9**	**33.9**	**23.5**	**21.1**	**28.9**	**26.7**

Table 3. Comparison between using different human representations.

Table 3 shows that replacing *3D skeletons* with *locations* or *2D skeletons* yields poor performance. This is because *locations* do not contain enough information of the actions performed by the person, and *2D skeletons* do not contain information of the person's position with respect to the objects on the floormaps. These results show that: 1) our dataset is diverse and hence locations are not enough to generate correct captioning, and 2) our choice of representation, i.e., *3D skeletons*, which combines information about both the people's locations and poses provides the right abstraction to learn meaningful features for proper captioning.

Person-Centric Floormap Representation: In this work, we use a person-centric coordinate representation for the floormap and its objects, as described in subsection 4.2. What if we simply use the image of the floormap with the objects, and mark the map with the person's location at each time instance? We compare this *image-based floormap* representation to our person-centric representation in Table 4. We use ResNet-18 pre-trained on ImageNet to extract features from the floormap image. The result shows that the image representation of floormap can achieve better performance than not having the floormap, but still worse than our person-centric representation. This is because it is much harder for the network to interpret and extract features from an image representation, since the information is far less explicit than our person-centric coordinate-based representation.

Method	B@1	B@2	B@3	B@4	M	R	C
w/o floormap	56.3	40.8	27.7	18.5	18.1	24.0	22.1
image-based floormap	60.1	43.9	31.5	21.6	20.1	26.7	24.4
person-centric floormap	**62.3**	**45.9**	**33.9**	**23.5**	**21.1**	**28.9**	**26.7**
person-centric floormap+noise	61.6	45.8	33.7	23.4	21.0	28.7	26.5

Table 4. Performance of RF-Diary with or without using floormap, with different floormap representations, and with gaussian noise.

Measurement Errors: We analyze the influence of floormap measurement errors on our model's performance. We add a random gaussian noise with a 20cm standard deviation on location, 10cm on size and 30 degrees on object rotation. The results in the last row of Table 4 show that the noise has very little effect on performance. This demonstrates that our model is robust to measurement errors.

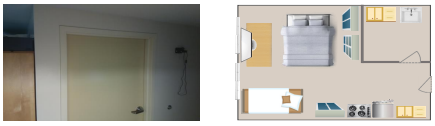
Feature Alignment: Our feature alignment framework consists of two parts: the L_2-norm between paired dataset, and the discriminator between unpaired datasets. Table 5 quantifies the contribution of each of these alignment mechanisms to RF-Diary's performance. The results demonstrate that our multi-modal feature alignment training scheme helps RF-Diary utilize the knowledge of the video-captioning model learned from the large video-captioning dataset to generate accurate descriptions, while training only on a rather small RCD dataset. We show a visualization of the features before and after alignment in the Appendix.

Video: A person walks into the room cleans the table. He sits down on chair.
RF+Floormap: A person walks to the table. Then he tidies the table and sits down at the table.
GT: A person enters the room and starts cleaning up the stuffs on the table. He then sits at the table when done.

(a)

Video: A person is texting on his phone. He then starts calling someone.
RF+Floormap: A person is working on his phone. Then he talks on his phone.
GT: A person plays with his phone. He then makes a phone call.

(b)

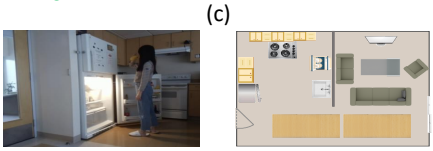


Video: N/A.
RF+Floormap: A person is brushing his teeth in the bathroom. Then he brushes hair.
GT: A person is in the bathroom. He takes out his tooth brush and cup, and then brushes his teeth. Then he starts brushing his hair.

(c)

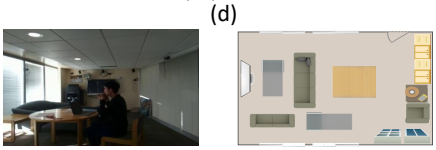
Video: N/A.
RF+Floormap: A person removes his clothes and puts it in wardrobe. Then he sits at a table and starts typing.
GT: A person walks into the room. He takes off his coat and hangs it in the wardrobe. Then he walks to the table, sits down and works on his laptop.

(d)



Video: A woman opens a fridge. She takes some juice and drink it and walks away.
RF+Floormap: A person opens the fridge, takes out something and then drinks water. Then he walks to the sink.
GT: In the kitchen, a woman takes some juice from fridge and drink it. She then closes the fridge door and goes to the sink.

(e)



Video: A person sits at a table. He opens snacks from the table and eats.
RF+Floormap: A person sits at a table. He then starts eating something.
GT: A person sits down at a dinning table. Then he takes a chocolate bar from the table and eats it.

(f)

Fig. 5. Examples from our RCD test set. Green words indicate actions. Blue words indicate objects included in floormap. Brown words indicate small objects not covered by floormap. Red words indicate the misprediction of small objects from RF-Diary. The first row shows RF-Diary can generate accurate captions compared to the video-based captioning model under visible scenarios. The second row shows that RF-Diary can still generate accurate captions when the video-based model does not work because of poor lighting conditions or occlusions. The third row shows the limitation of RF-Diary that it may miss object color and detailed descriptions of small objects.

6.3 Qualitative Result

In Figure 5, we show six examples from the RCD test set. The first row under each image is the caption generated by state of the art video-based captioning model [35], the second row is the caption generated by RF-Diary, and the third row is the ground truth caption labeled by a human.

The result shows that RF-Diary can generate accurate descriptions of the person's activities (green) and interaction with the surrounding environment (blue), and continue to work well even in the presence of occlusions (Figure 5 (c)), and poor lighting (Figure 5 (d)). Video-based captioning is limited by bad lighting, occlusions and the camera's field of view. So if the person exits the field of view, video captioning can miss some of the events (Figure 5 (e)).

Method	B@1	B@2	B@3	B@4	M	R	C
w/o L_2	52.5	38.0	25.7	18.5	16.6	23.1	20.3
w/o discrim	59.4	44.1	31.4	21.0	19.8	26.3	24.6
RF-Diary	**62.3**	**45.9**	**33.9**	**23.5**	**21.1**	**28.9**	**26.7**

Table 5. Performance of RF-Diary network on RCD with or without L_2 loss and discriminator. Note that without adding the L_2 loss, RF-Diary will not be affected by the video-captioning model. So if without the L_2 loss, then adding the discriminator loss on video-captioning model or not will not affect the RF-Diary's performance.

Besides poor lighting conditions, occlusions and field of view, video-captioning is also faced with privacy problems. For example, in Figure 5 (b), the person just took a bath and is not well-dressed. The video will record this content which is quite privacy-invasive. However, RF signal can protect privacy since it is not interpretable by a human, and it does not contain detailed information because of the relatively low resolution.

We also observe that RF-Diary has certain limitations. Since RF signals cannot capture details of objects such as color, texture, and shape, the model can mispredict those features. It can also mistake small objects. For example, in Figure 5 (e), the person is actually drinking orange juice, but RF-Diary predicts he is drinking water. Similarly, in Figure 5 (f), our model reports that the person is eating but cannot tell that he is eating a chocolate bar. The model also cannot distinguish the person's gender, so it always predicts "he" as shown in Figure 5 (e).

6.4 Additional Notes on Privacy

In comparison to images, RF signal is privacy-preserving because it is difficult to interpret by humans. However, one may also argue that since RF signals can track people though walls, they could create privacy concerns. This issue can be addressed through a challenge-response authentication protocol that prevent people from maliciously using RF signals to see areas that they are not authorized to access. More specifically, previous work [1] demonstrates that RF signals can sense human trajectories and locate them in space. Thus, whenever the user sets up the system to monitor an area, the system first challenges the user to execute certain moves (e.g., take two steps to the right, or move one meter forward), to ensure that the monitored person is the user. The system also asks the user to walk around the area to be monitored, and only monitors that area. Hence, the system would not monitor an area which the user does not have access to.

7 Conclusion

In this paper, we introduce RF-Diary, a system that enables in-home daily-life captioning using RF signals and floormaps. We also introduce the combination of RF signal and floormap as new complementary input modalities, and propose a feature alignment training scheme to transfer the knowledge from large video-captioning dataset to RF-Diary. Extensive experimental results demonstrate that RF-Diary can generate accurate descriptions of in-home events even when the environment is under poor lighting conditions or has occlusions. We believe this work paves the way for many new applications in health monitoring and smart homes.

References

1. Adib, F., Kabelac, Z., Katabi, D., Miller, R.C.: 3d tracking via body radio reflections. In: 11th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 14). pp. 317–329 (2014)
2. Adib, F., Katabi, D.: See through walls with WiFi!, vol. 43. ACM (2013)
3. Baraldi, L., Grana, C., Cucchiara, R.: Hierarchical boundary-aware neural encoder for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1657–1666 (2017)
4. Barbrow, L.: International lighting vocabulary. Journal of the SMPTE **73**(4), 331–332 (1964)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
7. Chetty, K., Chen, Q., Ritchie, M., Woodbridge, K.: A low-cost through-the-wall fmcw radar for stand-off operation and activity detection. In: Radar Sensor Technology XXI. vol. 10188, p. 1018808. International Society for Optics and Photonics (2017)
8. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. pp. 376–380 (2014)
9. Du, Y., Fu, Y., Wang, L.: Skeleton based action recognition with convolutional neural network. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). pp. 579–583. IEEE (2015)
10. Fakoor, R., Mohamed, A.r., Mitchell, M., Kang, S.B., Kohli, P.: Memory-augmented attention modelling for videos. arXiv preprint arXiv:1611.02261 (2016)
11. Fan, L., Li, T., Fang, R., Hristov, R., Yuan, Y., Katabi, D.: Learning longterm representations for person re-identification using radio signals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10699–10709 (2020)
12. Gan, C., Gan, Z., He, X., Gao, J., Deng, L.: Stylenet: Generating attractive visual captions with styles. In: CVPR. pp. 3137–3146 (2017)
13. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: CVPR. pp. 5630–5639 (2017)
14. Hsu, C.Y., Ahuja, A., Yue, S., Hristov, R., Kabelac, Z., Katabi, D.: Zero-effort in-home sleep and insomnia monitoring using radio signals. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(3), 1–18 (2017)
15. Hsu, C.Y., Hristov, R., Lee, G.H., Zhao, M., Katabi, D.: Enabling identification and behavioral sensing in homes using radio reflections. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. p. 548. ACM (2019)
16. Hsu, C.Y., Liu, Y., Kabelac, Z., Hristov, R., Katabi, D., Liu, C.: Extracting gait velocity and stride length from surrounding radio signals. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 2116–2126 (2017)
17. Hu, Y., Chen, Z., Zha, Z.J., Wu, F.: Hierarchical global-local temporal modeling for video captioning. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 774–783 (2019)
18. Li, C., Zhong, Q., Xie, D., Pu, S.: Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 786–792 (2018)
19. Li, T., Fan, L., Zhao, M., Liu, Y., Katabi, D.: Making the invisible visible: Action recognition through walls and occlusions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 872–881 (2019)

20. Lien, J., Gillian, N., Karagozler, M.E., Amihood, P., Schwesig, C., Olson, E., Raja, H., Poupyrev, I.: Soli: Ubiquitous gesture sensing with millimeter wave radar. ACM Transactions on Graphics (TOG) **35**(4), 142 (2016)
21. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
22. Long, X., Gan, C., de Melo, G.: Video captioning with multi-faceted attention. Transactions of the Association for Computational Linguistics **6**, 173–184 (2018)
23. Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y.: Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1029–1038 (2016)
24. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
25. Pasunuru, R., Bansal, M.: Reinforced video captioning with entailment rewards. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 979–985 (2017)
26. Peng, Z., Muñoz-Ferreras, J.M., Gómez-García, R., Li, C.: Fmcw radar fall detection based on isar processing utilizing the properties of rcs, range, and doppler. In: 2016 IEEE MTT-S International Microwave Symposium (IMS). pp. 1–3. IEEE (2016)
27. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732 (2015)
28. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526. Springer (2016)
29. Song, J., Guo, Z., Gao, L., Liu, W., Zhang, D., Shen, H.T.: Hierarchical lstm with adjusted temporal attention for video captioning. arXiv preprint arXiv:1706.01231 (2017)
30. Stove, A.G.: Linear fmcw radar techniques. In: IEE Proceedings F (Radar and Signal Processing). vol. 139, pp. 343–350. IET (1992)
31. Tian, Y., Lee, G.H., He, H., Hsu, C.Y., Katabi, D.: Rf-based fall monitoring using convolutional neural networks. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(3), 137 (2018)
32. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
33. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE international conference on computer vision. pp. 4534–4542 (2015)
34. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1494–1504 (2015)
35. Wang, X., Chen, W., Wu, J., Wang, Y.F., Yang Wang, W.: Video captioning via hierarchical reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4213–4222 (2018)
36. Wu, X., Li, G., Cao, Q., Ji, Q., Lin, L.: Interpretable video captioning via trajectory structured localization. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6829–6837 (2018)
37. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE international conference on computer vision. pp. 4507–4515 (2015)

38. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4584–4593 (2016)
39. Zhang, Z., Tian, Z., Zhou, M.: Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor. IEEE Sensors Journal **18**(8), 3278–3289 (2018)
40. Zhao, M., Adib, F., Katabi, D.: Emotion recognition using wireless signals. In: Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking. pp. 95–108. ACM (2016)
41. Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7356–7365 (2018)
42. Zhao, M., Liu, Y., Raghu, A., Li, T., Zhao, H., Torralba, A., Katabi, D.: Through-wall human mesh recovery using radio signals. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10113–10122 (2019)
43. Zhao, M., Tian, Y., Zhao, H., Alsheikh, M.A., Li, T., Hristov, R., Kabelac, Z., Katabi, D., Torralba, A.: Rf-based 3d skeletons. In: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. pp. 267–281. ACM (2018)
44. Zhao, M., Yue, S., Katabi, D., Jaakkola, T.S., Bianchi, M.T.: Learning sleep stages from radio signals: A conditional adversarial architecture. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 4100–4109. JMLR. org (2017)