Through-Wall Human Pose Estimation Using Radio Signals

Through-Wall Human Pose Estimation Using Radio Signals

Mingmin Zhao Tianhong Li Mohammad Abu Alsheikh Yonglong Tian Hang Zhao
Antonio Torralba Dina Katabi
MIT CSAIL

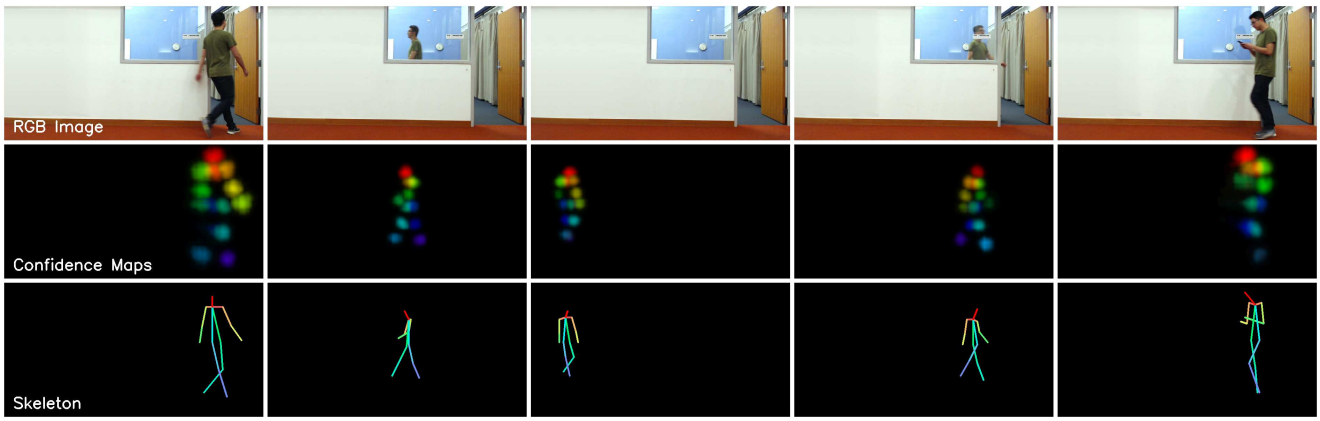


Figure 1: The figure shows a test example with a single person. It demonstrates that our system tracks the pose as the person enters the room and even when he is fully occluded behind the wall. **Top:** Images captured by a camera colocated with the radio sensor, and presented here for visual reference. **Middle:** Keypoint confidence maps extracted from RF signals *alone*, without any visual input. **Bottom:** Skeleton parsed from keypoint confidence maps showing that we can use RF signals to estimate the human pose even in the presence of full occlusion.

Abstract

This paper demonstrates accurate human pose estimation through walls and occlusions. We leverage the fact that wireless signals in the WiFi frequencies traverse walls and reflect off the human body. We introduce a deep neural network approach that parses such radio signals to estimate 2D poses. Since humans cannot annotate radio signals, we use state-of-the-art vision model to provide cross-modal supervision. Specifically, during training the system uses synchronized wireless and visual inputs, extracts pose information from the visual stream, and uses it to guide the training process. Once trained, the network uses only the wireless signal for pose estimation. We show that, when tested on visible scenes, the radio-based system is almost as accurate as the vision-based system used to train it. Yet, unlike vision-based pose estimation, the radio-based system can estimate 2D poses through walls despite never trained on such scenarios. Demo videos are available at our website.

1. Introduction

Estimating the human pose is an important task in computer vision with applications in surveillance, activity recognition, gaming, etc. The problem is defined as generating 2D skeletal representations of the joints on the arms and legs, and keypoints on the torso and head. It has recently witnessed major advances and significant performance improvements [30, 27, 28, 46, 31, 20, 10, 16, 33, 12, 47, 37, 45, 13]. However, as in any camera-based recognition task, occlusion remains a fundamental challenge. Past work deals with occlusion by hallucinating the occluded body parts based on the visible ones. Yet, since the human body is deformable, such hallucinations are prone to errors. Further, this approach becomes infeasible when the person is fully occluded, behind a wall or in a different room.

This paper presents a fundamentally different approach to deal with occlusions in pose estimation, and potentially other visual recognition tasks. While visible light is easily blocked by walls and opaque objects, radio frequency (RF) signals in the WiFi range can traverse such occlusions. Further, they reflect off the human body, providing an opportunity to track people through walls. Recent advances in wireless systems have leveraged those properties to detect people [5] and track their walking speed through occlusions [19]. Past systems however are quite coarse: they either track only one limb at any time [5, 4], or generate a static and coarse description of the body, where body-parts observed at different time are collapsed into one frame [4]. Use of wireless signals to produce a detailed and accurate

description of the pose, similar to that achieved by a state-of-the-art computer vision system, has remained intractable.

In this paper, we introduce RF-Pose, a neural network system that parses wireless signals and extracts accurate 2D human poses, even when the people are occluded or behind a wall. RF-Pose transmits a low power wireless signal (1000 times lower power than WiFi) and observes its reflections from the environment. Using only the radio reflections as input, it estimates the human skeleton. Fig. 1 shows an example output of RF-Pose tracking a person as he enters a room, becomes partially visible through a window, and then walks behind the wall. The RGB images in the top row show the sequence of events and the occlusions the person goes through; the middle row shows the confidence maps of the human keypoints extracted by RF-Pose; and the third row shows the resulting skeletons. Note how our pose estimator tracks the person even when he is fully occluded behind a wall. While this example shows a single person, RF-Pose works with multiple people in the scene, just as a state-of-art vision system would.

The design and training of our network present different challenges from vision-based pose estimation. In particular, there is no labeled data for this task. It is also infeasible for humans to annotate radio signals with keypoints. To address this problem, we use cross-modal supervision. During training, we attach a web camera to our wireless sensor, and synchronize the the wireless and visual streams. We extract pose information from the visual stream and use it as a supervisory signal for the wireless stream. Once the system is trained, it only uses the radio signal as input. The result is a system that is capable of estimating human pose using wireless signals only, without requiring human annotation as supervision. Interestingly, the RF-based model learns to perform pose estimation even when the people are fully occluded or in a different room. It does so despite it has never seen such examples during training.

Beyond cross-modal supervision, the design of RF-Pose accounts for the intrinsic features of RF signals including low spatial resolution, specularity of the human body at RF frequencies that traverse walls, and differences in representation and perspective between RF signals and the supervisory visual stream.

We train and test RF-Pose using data collected in public environments around our campus. The dataset has hundreds of different people performing diverse indoor activities: walking, sitting, taking stairs, waiting for elevators, opening doors, talking to friends, etc. We test and train on different environments to ensure the network generalizes to new scenes. We manually label 2000 RGB images and use them to test both the vision system and RF-Pose. The results show that on visible scenes, RF-Pose has an average precision (AP) of 62.4 whereas the vision-based system used to train it has an AP of 68.8. For through-wall scenes, RF-Pose has an AP of 58.1 whereas the vision-based system fails completely.

We also show that the skeleton learned from RF signals extracts identifying features of the people and their style of moving. We run an experiment where we have 100 people perform free walking, and train a vanilla-CNN classifier to identify each person using a 2-second clip of the RF-based skeleton. By simply observing how the RF-based skeleton moves, the classifier can identify the person with an accuracy over 83% in both visible and through wall scenarios.

2. Related Work

(a) Computer Vision: Human pose estimation from RGB images generally falls into two main categories: Top-down and bottom-up methods. Top-down methods [16, 14, 29, 15] first detect each people in the image, and then apply a single-person pose estimator to each people to extract keypoints. Bottom-up methods [10, 31, 20], on the other hand, first detect all keypoints in the image, then use post-processing to associate the keypoints belonging to the same person. We build on this literature and adopt a bottom-up approach, but differ in that we learn poses from RF signals. While some prior papers use sensors other than conventional cameras, such as RGB-D sensors [50] and Vicon [35], unlike RF signals, those data inputs still suffer from occlusions by walls and other opaque structures.

In terms of modeling, our work is related to cross-modal and multi-modal learning that explores matching different modalities or delivering complementary information across modalities [8, 11, 36, 34]. In particular, our approach falls under cross-modal teacher-student networks [8], which transfer knowledge learned in one data modality to another. While past work only transfers category-level discriminative knowledge, our network transfers richer knowledge on dense keypoint confidence maps.

(b) Wireless Systems: Recent years have witnessed much interest in localizing people and tracking their motion using wireless signals. The literature can be classified into two categories. The first category operates at very high frequencies (e.g., millimeter wave or terahertz) [3]. These can accurately image the surface of the human body (as in airport security scanners), but do not penetrate walls and furniture.

The second category uses lower frequencies, around a few GHz, and hence can track people through walls and occlusions. Such through-wall tracking systems can be divided into: device-based and device-free. Device-based tracking systems localize people using the signal generated by some wireless device they carry. For example, one can track a person using the WiFi signal from their cellphone [44, 24, 40]. Since the tracking is performed on the device not the person, one can track different body-parts by attaching different radio devices to each of them.

On the other hand, device-free wireless tracking systems do not require the tracked person to wear sensors on their body. They work by analyzing the radio signal reflected off the person's body. However, device-free systems typically have low spatial resolution and cannot localize multiple body parts simultaneously. Different papers either localize the whole body [5, 23], monitor the person's walking speed [43, 19], track the chest motion to extract breathing and heartbeats [6, 51, 52], or track the arm motion to identify a particular gesture [32, 26]. The closest to our work is a system called RF-Capture which creates a coarse description of the human body behind a wall by collapsing multiple body parts detected at different points in time [4]. None of the past work however is capable of estimating the human pose or simultaneously localizing its various keypoints.

Finally, some prior papers have explored human identification using wireless signals [49, 43, 18]. Past work, however, is highly restrictive in how the person has to move, and cannot identify people from free-form walking.

3. RF Signals Acquisition and Properties

Our RF-based pose estimation relies on transmitting a low power RF signal and receiving its reflections. To separate RF reflections from different objects, it is common to use techniques like FMCW (Frequency Modulated Continuous Wave) and antenna arrays [4]. FMCW separates RF reflections based on the distance of the reflecting object, whereas antenna arrays separate reflections based on their spatial direction. In this paper, we introduce a radio similar to [4], which generates an FMCW signal and has two antenna arrays: vertical and horizontal (other radios are also available [1, 2]). Thus, our input data takes the form of two-dimensional heatmaps, one for each of the horizontal and vertical antenna arrays. As shown in Fig. 2, the horizontal heatmap is a projection of the signal reflections on a plane parallel to the ground, whereas the vertical heatmap is a projection of the reflected signals on a plane perpendicular to the ground (red refers to large values while blue refers to small values). Note that since RF signals are complex numbers, each pixel in this map has a real and imaginary components. Our radio generates 30 pairs of heatmaps per second.

It is important to note that RF signals have intrinsically different properties than visual data, i.e., camera pixels.

- First, RF signals in the frequencies that traverse walls have low spatial resolution, much lower than vision data. The resolution is typically tens of centimeters [5, 2, 4], and is defined by the bandwidth of the FMCW signal and the aperture of the antenna array. In particular, our radio has a depth resolution about 10 cm, and its antenna arrays have vertical and horizontal angular resolution of 15 degrees.

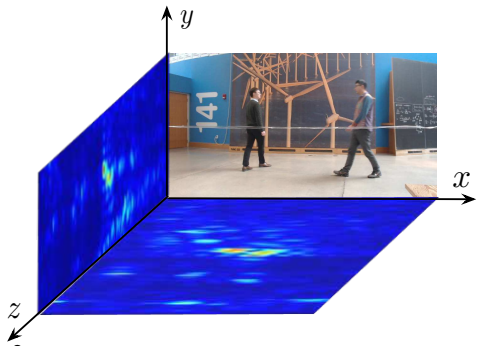


Figure 2: RF heatmaps and an RGB image recorded at the same time.

- Second, the human body is specular in the frequency range that traverse walls [9]. RF specularity is a physical phenomenon that occurs when the wavelength is larger than the roughness of the surface. In this case, the object acts like a reflector - i.e., a mirror - as opposed to a scatterer. The wavelength of our radio is about 5cm and hence humans act as reflectors. Depending on the orientation of the surface of each limb, the signal may be reflected towards our sensor or away from it. Thus, in contrast to camera systems where any snapshot shows all unoccluded key-points, in radio systems, a single snapshot has information about a subset of the limbs and misses limbs and body parts whose orientation at that time deflects the signal away from the sensor.
- Third, the wireless data has a different representation (complex numbers) and different perspectives (horizontal and vertical projections) from a camera.

The above properties have implications for pose estimation, and need to be taken into account in designing a neural network to extract poses from RF signals.

4. Method

Our model, illustrated in Fig. 3, follows a teacher-student design. The top pipeline in the figure shows the teacher network, which provides cross-modal supervision; the bottom pipeline shows the student network, which performs RF-based pose estimation.

4.1. Cross-Modal Supervision

One challenge of estimating human pose from RF signals is the the lack of labelled data. Annotating human pose by looking at RF signals (e.g., Fig. 2) is almost impossible. We address this challenge by leveraging the presence of well established vision models that are trained to predict human pose in images [25, 7].

We design a cross-modal teacher-student network that transfers the visual knowledge of human pose using synchronized images and RF signals as a bridge. Consider a synchronized pair of image and RF signals $(\mathbf{I}, \mathbf{R})$, where

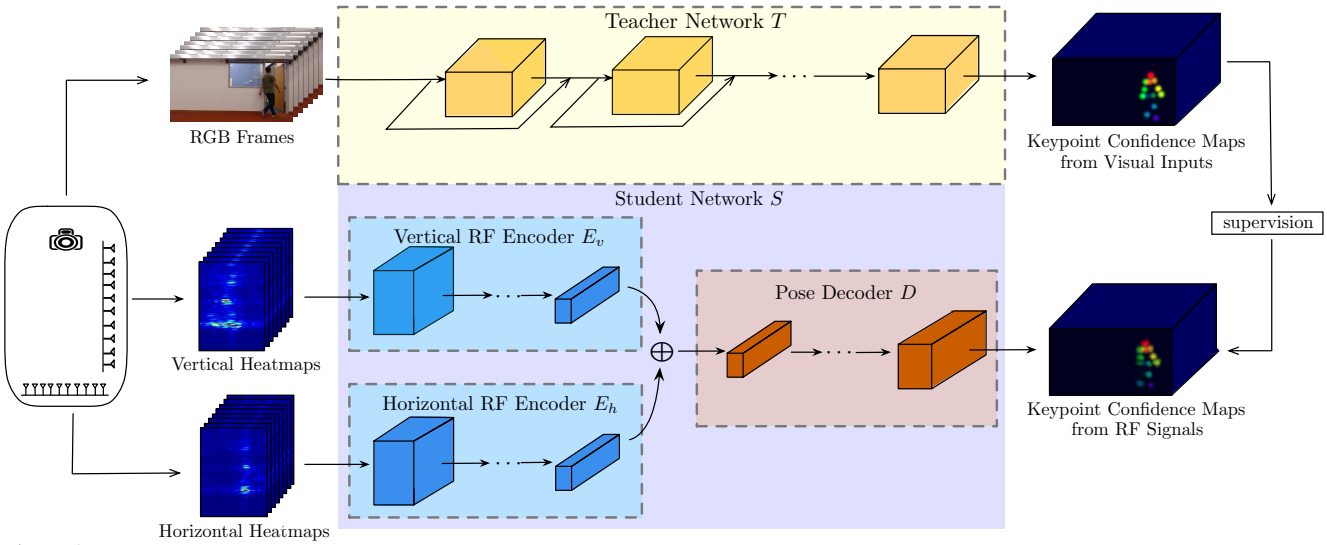


Figure 3: Our teacher-student network model used in RF-Pose. The upper pipeline provides training supervision, whereas the bottom pipeline learns to extract human pose using only RF heatmaps.

$\mathbf{R}$ denotes the combination of the vertical and horizontal heatmaps, and $\mathbf{I}$ the corresponding image in Fig. 2. The teacher network $\mathbf{T}(\cdot)$ takes the images $\mathbf{I}$ as input and predicts keypoint confidence maps as $\mathbf{T}(\mathbf{I})$. These predicted maps $\mathbf{T}(\mathbf{I})$ provide cross-modal supervision for the student network $\mathbf{S}(\cdot)$, which learns to predict keypoint confidence maps from the RF signals. In this paper, we adopt the 2D pose estimation network in [10] as the teacher network. The student network learns to predict 14 keypoint confidence maps corresponding to the following anatomical parts of the human body: head, neck, shoulders, elbows, wrists, hips, knees and ankles.

The training objective of the student network $\mathbf{S}(\cdot)$ is to minimize the difference between its prediction $\mathbf{S}(\mathbf{R})$ and the teacher network's prediction $\mathbf{T}(\mathbf{I})$:

$$\min_{\mathbf{S}} \sum_{(\mathbf{I},\mathbf{R})} L(\mathbf{T}(\mathbf{I}), \mathbf{S}(\mathbf{R})) \tag{1}$$

We define the loss as the summation of binary cross entropy loss for each pixel in the confidence maps:

$$L(\mathbf{T}, \mathbf{S}) = -\sum_{c} \sum_{i,j} \mathbf{S}_{ij}^{c} \log \mathbf{T}_{ij}^{c} + (1 - \mathbf{S}_{ij}^{c}) \log (1 - \mathbf{T}_{ij}^{c}),$$

where $\mathbf{T}_{ij}^{c}$ and $\mathbf{S}_{ij}^{c}$ are the confidence scores for the (i, j)-th pixel on the confidence map c.

4.2. Keypoint Detection from RF Signals

The design of our student network has to take into account the properties of RF signals. As mentioned earlier, the human body is specular in the RF range of interest. Hence, we cannot estimate the human pose from a single RF frame (a single pair of horizontal and vertical heatmaps) because the frame may be missing certain limbs tough they are not occluded. Further, RF signals have low spatial resolution. Hence, it will be difficult to pinpoint the location of a keypoint using a single RF frame. To deal with these issues, we make the network learn to aggregate information from multiple snapshots of RF heatmaps so that it can capture different limbs and model the dynamics of body movement. Thus, instead of taking a single frame as input, we make the network look at sequences of frames. For each sequence, the network outputs keypoint confidence maps as the number of frames in the input – i.e., while the network looks at a clip of multiple RF frames at a time, it still outputs a pose estimate for every frame in the input.

We also want the network to be invariant to translations in both space and time so that it can generalize from visible scenes to through-wall scenarios. Therefore, we use spatio-temoral convolutions [22, 39, 42] as basic building blocks for the student networks.

Finally, the student network needs to transform the information from the views of RF heatmaps to the view of the camera in the teacher network (see Fig. 2). To do so, the model has to first learn a representation of the information in the RF signal that is not encoded in original spatial space, then decode that representation into keypoints in the view of the camera. Thus, as shown in Fig. 3, our student network has: 1) two RF encoding networks $E_h(\cdot)$ and $E_v(\cdot)$ for horizontal and vertical heatmap streams, and 2) a pose decoding network $D(\cdot)$ that takes a channel-wise concatenation of horizontal and vertical RF encodings as input and predicts keypoint confidence maps. The RF encoding networks uses strided convolutional networks to remove spatial dimensions [48, 41] in order to summarize information from the original views. The pose decoding network then uses fractionally strided convolutional networks to decode keypoints in the camera's view.

4.3. Implementation and Training

RF encoding network. Each encoding network takes 100 frames (3.3 seconds) of RF heatmap as input. The RF encoding network uses 10 layers of $9 \times 5 \times 5$ spatio-temporal convolutions with $1 \times 2 \times 2$ strides on spatial dimensions every other layer. We use batch normalization [21] followed by the ReLU activation functions after every layer.

Pose decoding network. We combine spatio-temporal convolutions with fractionally strided convolutions to decode the pose. The decoding network has 4 layers of $3 \times 6 \times 6$ with fractionally stride of $1 \times \frac{1}{2} \times \frac{1}{2}$, except the last layer has one of $1 \times \frac{1}{4} \times \frac{1}{4}$. We use Parametric ReLu [17] after each layer, except for the output layer, where we use sigmoid.

Training Details. We represent a complex-valued RF heatmap by two real-valued channels that store the real and imaginary parts. We use a batch size of 24. Our networks are implemented in PyTorch.

4.4. Keypoint Association

The student network generates confidence maps for all keypoints of all people in the scene. We map the keypoints to skeletons as follows. We first perform non-maximum suppression on the keypoint confidence maps to obtain discrete peaks of keypoint candidates. To associate keypoints of different persons, we use the relaxation method proposed by Cao *et al*. [10] and we use Euclidean distance for the weight of two candidates. Note that we perform association on a frame-by-frame basis based on the learned keypoint confidence maps. More advanced association methods are possible, but outside the scope of this paper.

5. Dataset

We collected synchronized wireless and vision data. We attached a web camera to our RF sensor and synchronized the images and the RF data with an average synchronization error of 7 milliseconds.

We conducted more than 50 hours of data collection experiments from 50 different environments (see Fig. 4), including different buildings around our campus. The environments span offices, cafeteria, lecture and seminar rooms, stairs, and walking corridors. People performed natural everyday activities without any interference from our side. Their activities include walking, jogging, sitting, reading, using mobile phones and laptops, eating, etc. Our data includes hundreds of different people of varying ages. The maximum and average number of people in a single frame are 14 and 1.64, respectively. A data frame can also be empty, i.e., it does not include any person. Partial occlusions, where parts of the human body are hidden due to furniture and building amenities, are also present. Legs and arms are the most occluded parts.

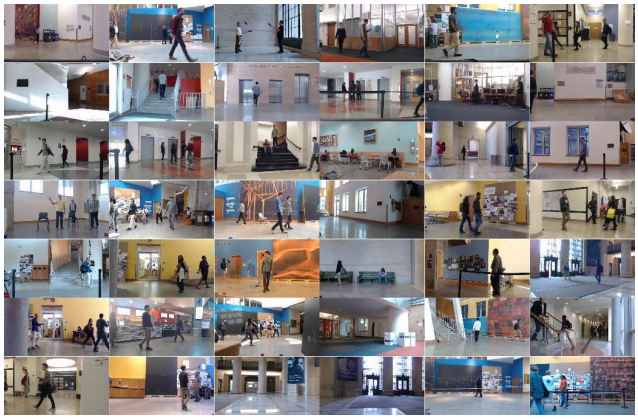


Figure 4: Different environments in the dataset.

To evaluate the performance of our model on through-wall scenes, we build a mobile camera system that has 8 cameras to provide ground truth when the people are fully occluded. After calibrating the camera system, we construct 3D poses of people and project them on the view of the camera colocated with RF sensor. The maximum and average number of people in each frame in the through-wall testing set are 3 and 1.41, respectively. This through-wall data was *only for testing* and was not used to train the model.

6. Experiments

RF-Pose is trained with 70% of the data from visible scenes, and tested with the remaining 30% of the data from visible scenes and all the data from through-wall scenarios. We make sure that the training data and test data are from different environments.

6.1. Setup

Evaluation Metrics: Motivated by the COCO keypoints evaluation [25] and as is common in past work [10, 29, 16], we evaluate the performance of our model using the average precision over different object keypoint similarity (OKS). We also report AP^{50} and AP^{75}, which denote the average precision when OKS is 0.5 and 0.75, and are treated as loose and strict match of human pose, respectively. We also report AP, which is the mean average precision over 10 different OKS thresholds ranging from 0.5 to 0.95.

Baseline: For visible and partially occluded scenes, we compare RF-Pose with OpenPose [10], a state-of-the-art vision-based model, that also acts as the teacher network.

Ground Truth: For visible scenes, we manually annotate human poses using the images captured by the camera colocated with our RF sensor. For through-wall scenarios where the colocated camera cannot see people in the other room, we use the eight-camera system described in 5 to provide ground truth. We annotate the images captured by all eight cameras to build 3D human poses and project them on the

Methods	Visible scenes			Through-walls		
	AP	AP^{50}	AP^{75}	**AP**	AP^{50}	AP^{75}
RF-Pose	62.4	**93.3**	70.7	**58.1**	**85.0**	**66.1**
OpenPose[10]	**68.8**	77.8	**72.6**	-	-	-

Table 1: Average precision in visible and through-wall scenarios.

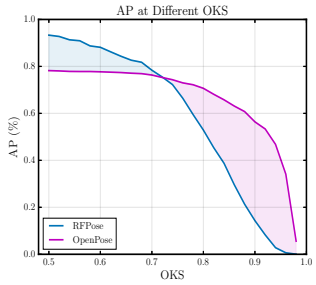


Figure 5: Average precision at different OKS values.

Methods	Hea	Nec	Sho	Elb	Wri	Hip	Kne	Ank
RF-Pose	**75.5**	**68.2**	62.2	56.1	51.9	**74.2**	63.4	54.7
OpenPose[10]	73.0	67.1	**70.8**	**64.5**	**61.5**	71.4	**68.4**	**68.3**

Table 2: Average precision of different keypoints in visible scenes.

view of the camera colocated with the radio. We annotate 1000 randomly sampled images from the visible-scene test set and another 1000 examples from the through-wall data.

6.2. Multi-Person Pose Estimation Results

We compare human poses obtained via RF signals with the corresponding poses obtained using vision data. Table 1 shows the performance of RF-Pose and the baseline when tested on both visible scenes and through-wall scenarios. The table shows that, when tested on visible scenes, RF-Pose is almost as good as the vision-based OpenPose that was used to train it. Further, when tested on through-wall scenarios, RF-Pose can achieve good pose estimation while the vision-based baseline completely fail due to occlusion.

The performance of RF-Pose on through-wall scenarios can be surprising because the system did not see such examples during training. However, from the perspective of radio signals, a wall simply attenuates the signal power, but maintains the signal structure. Since our model is space invariant, it is able to identify a person behind a wall as similar to the examples it has seen in the space in front of a wall.

An interesting aspect in Table 1 is that RF-Pose outperforms OpenPose for AP^{50}, and becomes worse at AP^{75}. To further explore this aspect, we plot in Fig. 5 the average precision as a function of OKS values. The figure shows that at low OKS values (< 0.7), our model outperforms the vision baseline. This is because RF-Pose predicts less false alarm than the vision-based solution, which can generate fictitious skeletons if the scene has a poster of a person, or a human reflection in a glass window or mirror. In contrast, at high OKS values (> 0.75), the performance of RF-Pose degrades fast, and becomes worse than vision-based approaches. This is due to the intrinsic low spatial resolution of RF signals which prevents them from pinpointing the exact location of the keypoints. The ability of RF-Pose to exactly locate the keypoints is further hampered by imperfect synchronization between the RF heatmaps and the ground truth images.

Next, we zoom in on the various keypoints and compare their performance. Table 2 shows the average precision of RF-Pose and the baseline in localizing different body parts including head, right and left shoulders, elbows, wrists, hips, knees, and ankles. The results indicate that RF signals are highly accurate at localizing the head and torso (neck and hips) but less accurate in localizing limbs. This is expected because the amount of RF reflections depends on the size of the body part. Thus, RF-Pose is better at capturing the head and torso, which have large reflective areas and relatively slow motion in comparison to the limbs. As for why RF-Pose outperforms OpenPose on some of the keypoints, this is due to the RF-based model operating over a clip of a few seconds, whereas the OpenPose baseline operates on individual images.

Finally, we show a few test skeletons to provide a qualitative perspective. Fig. 6 shows sample RF-based skeletons from our test dataset, and compares them to the corresponding RBG images and OpenPose skeletons. The figure demonstrates RF-Pose performs well in different environments with different people doing a variety of everyday activities. Fig. 7 illustrates the difference in errors between RF-Pose and vision-based solutions. It shows that the errors in vision-based systems are typically due to partial occlusions, bad lighting [1], or confusing a poster or wall-picture as a person. In contrast, errors in RF-Pose happen when a person is occluded by a metallic structure (e.g., a metallic cabinet in Fig. 7(b)) which blocks RF signals, or when people are too close and hence the low resolution RF signal fails to track all of them.

6.3. Model Analysis

We use guided back-propagation [38] to visualize the gradient with respect to the input RF signal, and leverage the information to provide insight into our model.

Which part of the RF heatmap does RF-Pose focus on? Fig. 8 presents an example where one person is walking in front of the wall while another person is hidden behind it. Fig. 8(c) shows the raw horizontal heatmap. The two large boxes are the rescaled versions of the smaller boxes and zoom in on the two people in the figure. The red patch indicated by the marker is the wall, and the other patches are multipath effects and other objects. The gradient in Fig. 8(d) shows that RF-Pose has learned to focus its at-

[1] Images with bad lighting are excluded during training and testing.

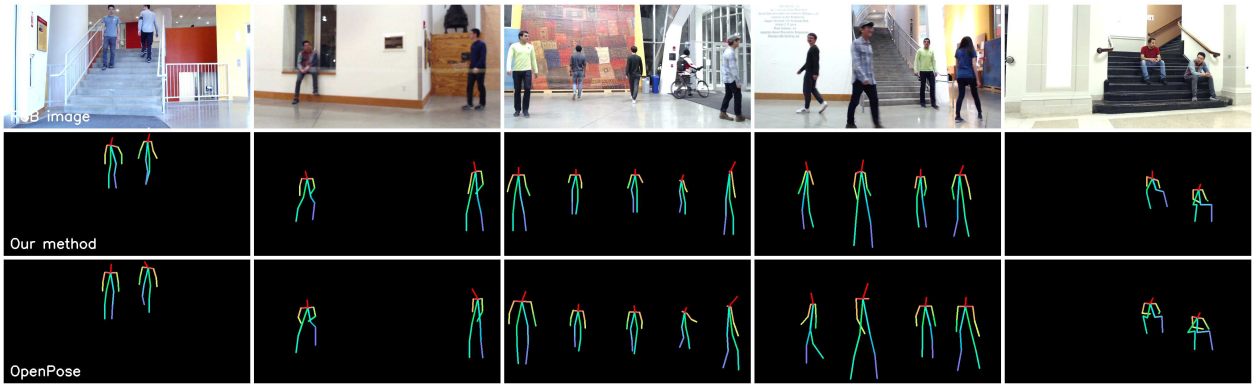
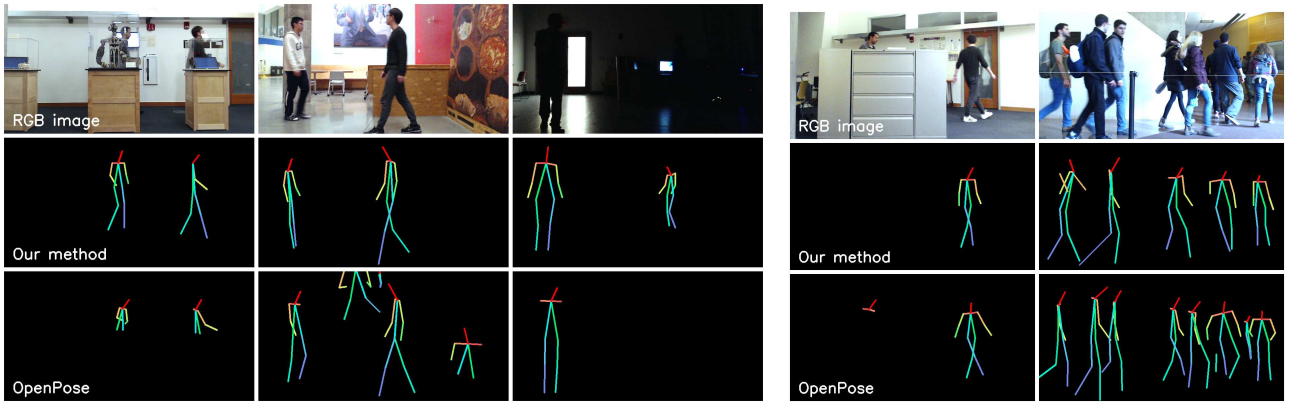


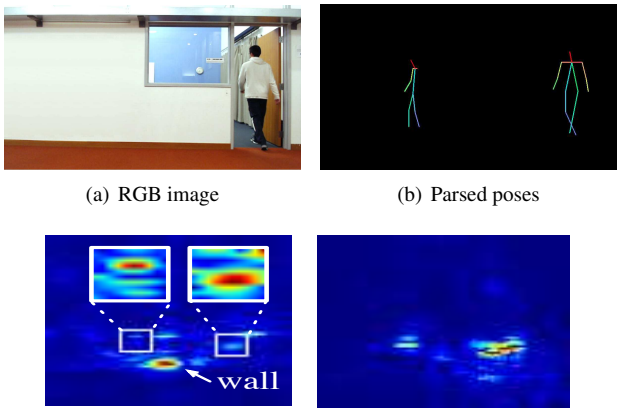
Figure 6: Pose estimation on different activities and environments. **First row:** Images captured by a web camera (shown as a visual reference). **Second row:** Pose estimation by our model using RF signals *only* and without any visual input. **Third row:** Pose estimation using OpenPose based on images from the first row.



(a) Failure examples of OpenPose due to occlusioin, posters, and bad lighting.

(b) Failure examples of ours due to metal and crowd.

Figure 7: Common failure examples. **First row:** Images captured by a web camera (shown as a visual reference). **Second row:** Pose estimation by our model using RF signals *only* and without any visual input. **Third row:** Pose estimation using OpenPose based on images from the first row.



(a) RGB image

(b) Parsed poses

(c) Horizontal Heatmap

(d) Gradients

Figure 8: Attention of the model across space

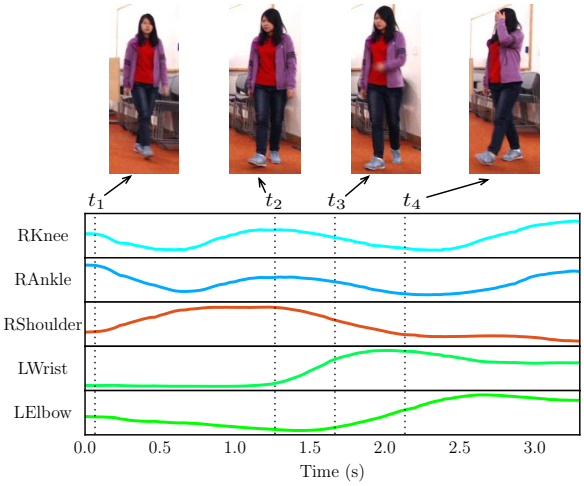


Figure 9: Activation of different keypoints over time.

tention on the two people in the scene and ignore the wall, other objects, and multipath.

How does RF-Pose deal with specularity? Due to the specularity of the human body, some body parts may not reflect much RF signals towards our sensor, and hence may be de-emphasized or missing in some heatmaps, even though they are not occluded. RF-Pose deals with this issue by taking as input a sequences of RF frames (i.e., a video clip RF heatmaps). To show the benefit of processing sequences of

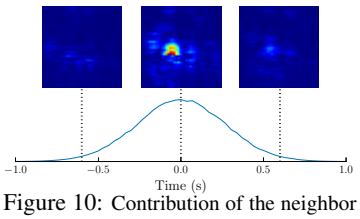


Figure 10: Contribution of the neighbor to the current frame.

# RF frames	AP
6	30.8
20	50.8
50	59.1
100	**62.4**

Table 3: Average precision of pose estimation trained on varying lengths of input frames.

RF frames, we sum up the input gradient in all pixels in the heatmaps to obtain activation per RF frame. We then plot in Fig. 9 the activation as a function of time to visualize the contribution of each frame to the estimation of various keypoints. The figure shows: that the activations of the right knee (RKnee) and right ankle (RAnkle) are highly correlated, and have peaks at time t_1 and t_2 when the person is taking a step with her right leg. In contrast, her left wrist (LWrist) gets activated after she raises her forearm at t_3, whereas her left elbow (LElbow) remains silent until t_4 when she raises her backarm.

Fig. 9 shows that, for a single output frame, different RF frames in the input sequence contribute differently to the output keypoints. This emphasizes the need for using a sequence of RF frames at the input. But how many frames should one use? Table 3 compares the model's performance for different sequence length at the input. The average precision is poor when the inout uses only 6 RF frames and increases as the sequence length increases.

But how much temporal information does RF-Pose need? Given a particular output frame, i, we compute the contributions of each of the input frames to it as a function of their time difference from i. To do so, we back-propagate the loss of a single frame w.r.t. to the RF heatmaps before it and after it, and sum up the spatial dimensions. Fig. 10 shows the results, suggesting that RF-Pose leverages RF heatmaps up to 1 second away to estimate the current pose.

6.4. Identification Using RF-Based Skeleton

We would like to show that the skeleton generated by RF-Pose captures personalized features of the individuals in the scene, and can be used by various recognition tasks. Thus, we experiment with using the RF-based skeleton for person identification.

We conduct person identification experiment with 100 people in two settings: visible environment, where the subject and RF device are in the same room, and through-wall environment, where the RF device captures the person's reflections through a wall. In each setting, every person walks naturally and randomly inside the area covered by our RF device, and we collect 8 and 2 minutes data separately for training and testing. The skeleton heatmaps are extracted by the model trained on our pose estimation dataset, which never overlaps with the identification dataset. For each setting, we train a 10-layer vanilla CNN to identify people based on 50 consecutive frames of skeleton heatmaps.

Method	Visible scenes		Through-walls	
	Top1	Top3	Top1	Top3
RF-Pose	83.4	96.1	84.4	96.3

Table 4: Top1 and top3 identification percent accuracy in visible and through-wall settings.

Table 4 shows that RF-based skeleton identification can reach 83.4% top1 accuracy in visiable scenes. Interestingly, even when a wall blocks the device and our pose extractor never sees these people and such environments during training, the extracted skeletons can still achieve 84.4% top1 accuracy, showing its robustness and generalizability regardless of the wall. As for top3 accuracy, we achieve more than 96% in both settings, demonstrating that the extracted skeleton can preserve most of the discriminative information for identification even though the pose extractor is never trained or fine-tuned on the identification task.

7. Scope & Limitations

RF-Pose leverages RF signals to infer the human pose through occlusions. However, RF signals and the solution that we present herein have some limitations: First, the human body is opaque at the frequencies of interest – i.e., frequencies that traverse walls. Hence, inter-person occlusion is a limitation of the current system. Second, the operating distance of a radio is dependent on its transmission power. The radio we use in this paper works up to 40 feet. Finally, we have demonstrated that our extracted pose captures identifying features of the human body. However, our identification experiments consider only one activity: walking. Exploring more sophisticated models and identifying people in the wild while performing daily activities other than walking is left for future work.

8. Conclusion

Occlusion is a fundamental problem in human pose estimation and many other vision tasks. Instead of hallucinating missing body parts based on visible ones, we demonstrate a solution that leverages radio signals to accurately track the 2D human pose through walls and obstructions. We believe this work opens up exciting research opportunities to transfer visual knowledge about people and environments to RF signals, providing a new sensing modality that is intrinsically different from visible light and can augment vision systems with powerful capabilities.

Acknowledgments: We are grateful to all the human subjects for their contribution to our dataset. We thank the NETMIT members and the anonymous reviewers for their insightful comments.

References

[1] Texas instruments. http://www.ti.com/. 3

[2] Walabot. https://walabot.com/. 3

[3] Reusing 60 ghz radios for mobile radar imaging. In *In Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015. 2

[4] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand. Capturing the human figure through a wall. *ACM Transactions on Graphics*, 34(6):219, November 2015. 1, 3

[5] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller. 3D tracking via body radio reflections. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, 2014. 1, 3

[6] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015. 3

[7] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 3

[8] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, NIPS, 2016. 2

[9] P. Beckmann and A. Spizzichino. The scattering of electromagnetic waves from rough surfaces. *Norwood, MA, Artech House, Inc.*, 1987. 3

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2017. 1, 2, 4, 5, 6

[11] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2016. 2

[12] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *CVPR*, 2016. 1

[13] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017. 1

[14] H. Fang, S. Xie, and C. Lu. Rmpe: Regional multi-person pose estimation. *arXiv preprint arXiv:1612.00137*, 2016. 2

[15] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2014. 2

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, ICCV, 2017. 1, 2, 5

[17] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, ICCV, 2015. 5

[18] F. Hong, X. Wang, Y. Yang, Y. Zong, Y. Zhang, and Z. Guo. Wfid: Passive device-free human identification using WiFi signal. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, 2016. 3

[19] C.-Y. Hsu, Y. Liu, Z. Kabelac, R. Hristov, D. Katabi, and C. Liu. Extracting gait velocity and stride length from surrounding radio signals. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI, 2017. 1, 3

[20] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Proceedings of the European Conference on Computer Vision*, ECCV, 2016. 1, 2

[21] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML, 2015. 5

[22] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, January 2013. 4

[23] K. R. Joshi, D. Bharadia, M. Kotaru, and S. Katti. Wideo: Fine-grained device-free motion tracing using rf backscatter. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, 2015. 3

[24] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti. Spotfi: Decimeter level localization using wifi. In *ACM SIGCOMM Computer Communication Review*, 2015. 2

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European conference on computer vision*, ECCV, 2014. 3, 5

[26] P. Melgarejo, X. Zhang, P. Ramanathan, and D. Chu. Leveraging directional antenna capabilities for fine-grained gesture recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014. 3

[27] A. Newell and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*, 2016. 1

[28] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, ECCV, 2016. 1

[29] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards Accurate Multi-person Pose Estimation in the Wild. In *CVPR*, 2017. 2, 5

[30] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV, 2015. 1

[31] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2016. 1, 2

[32] Q. Pu, S. Gupta, S. Gollakota, and S. Patel. Whole-home gesture recognition using wireless signals. In *Proceedings*

of the 19th Annual International Conference on Mobile computing & Networking, MobiCom. ACM, 2013. 3

[33] M. R. Ronchi and P. Perona. Benchmarking and error diagnosis in multi-instance pose estimation. 2017. 1

[34] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba. Learning cross-modal embeddings for cooking recipes and food images. *Training*, 720:619–508, 2017. 2

[35] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, March 2010. 2

[36] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013. 2

[37] J. Song, L. Wang, L. Van Gool, and O. Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[38] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 6

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, ICCV, 2015. 4

[40] D. Vasisht, S. Kumar, and D. Katabi. Decimeter-level localization with a single wifi access point. In *NSDI*, 2016. 2

[41] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, NIPS, 2016. 4

[42] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 4

[43] W. Wang, A. X. Liu, and M. Shahzad. Gait recognition using WiFi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016. 3

[44] J. Xiong and K. Jamieson. Arraytrack: A fine-grained indoor location system. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, 2013. 2

[45] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 1

[46] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016. 1

[47] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[48] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2010. 4

[49] Y. Zeng, P. H. Pathak, and P. Mohapatra. Wiwho: wifi-based person identification in smart spaces. In *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*, page 4, 2016. 3

[50] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 2

[51] M. Zhao, F. Adib, and D. Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, MobiCom, 2016. 3

[52] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*, ICML, 2017. 3