Person-in-WiFi: Fine-grained Person Perception using WiFi

Fei Wang[†¶*] Sanping Zhou[‡¶] Stanislav Panev[¶] Jinsong Han[†§] Dong Huang[¶]

[†]School of Cyber Science and Technology, Zhejiang University

[¶]The Robotics Institute, Carnegie Mellon University

[‡]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

[§]Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies

{feiwang,spanev,donghuang}@cmu.edu sanpingzhou@stu.xjtu.edu.cn hanjinsong@zju.edu.cn

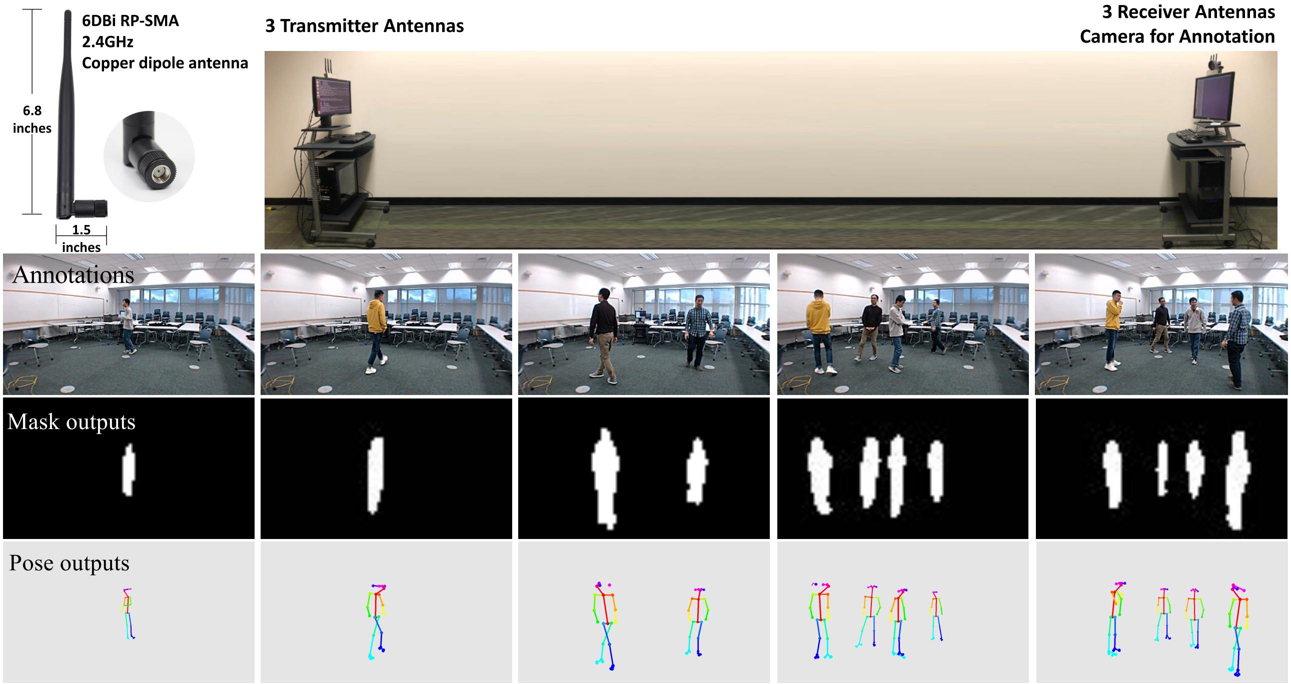

Figure 1. Person-in-WiFi. Top: WiFi antennas as sensors for person perception. Receiver antennas record WiFi signals as inputs to Person-in-WiFi. The rest rows are, images used to annotate WiFi signals, and two outputs: person segmentation masks and body poses.

Abstract

Fine-grained person perception such as body segmentation and pose estimation has been achieved with many 2D and 3D sensors such as RGB/depth cameras, radars (e.g. RF-Pose), and LiDARs. These solutions require 2D images, depth maps or 3D point clouds of person bodies as input. In this paper, we take one step forward to show that fine-grained person perception is possible even with 1D sensors: WiFi antennas. Specifically, we used two sets of WiFi antennas to acquire signals, i.e., one transmitter set and one receiver set. Each set contains three antennas horizontally lined-up as a regular household WiFi router. The WiFi signal generated by a transmitter antenna, penetrates through and reflects on human bodies, furniture, and walls, and then superposes at a receiver antenna as 1D signal samples. We developed a deep learning approach that uses annotations on 2D images, takes the received 1D WiFi signals as input, and performs body segmentation and pose estimation in an end-to-end manner. To our knowledge, our solution is the first work based on off-the-shelf WiFi antennas and standard IEEE 802.11n WiFi signals. Demonstrating comparable results to image-based solutions, our WiFi-based person perception solution is cheaper and more ubiquitous than radars and LiDARs, while invariant to illumination and has little privacy concern comparing to cameras.

[*] Work done when Fei Wang was a visiting PhD student at CMU.

1. Introduction

To conduct fine-grained person perception like human body segmentation and pose estimation, three main categories of sensors have been used: cameras (2D images), radars (depth maps), and LiDARs (3D point clouds). These approaches require a minimal spatial resolution of sensor outputs. For instance, pixel resolution of 300×300 pixels from cameras [30], depth resolution of 2 cm for radars [59], or angular resolution comparable to 32-beam LiDARs [54, 33]. Moreover, camera-based solutions are limited by technical challenges such as variety of clothing, background, lighting and occlusion, and social limitations such as privacy concerns. Radar sensors require dedicated hardware, *e.g.*, RF-Pose [59] and RF-Capture [1] produce depth maps by the Frequency Modulated Continuous Wave (FMCW) technology, which requires carefully assembled and synchronized $16 + 4$ T-shaped antenna array with very broad bandwidth (1.78 GHz). High-definition LiDARs are very expensive and power-consuming, therefore are difficult to apply to daily and household use.

In this paper, we propose a fine-grained person perception solution using prevalent WiFi antennas and standard IEEE 802.11n WiFi signals. Such WiFi devices is wildly available in warehouse, hospital, office, home where the low illumination, blind spots, privacy issues make cameras not applicable, while radars and LiDARs are too expensive and power-consuming to install. The challenge is that a WiFi antenna can only receive signal as the amplitude/phase of Electromagnetic (EM) waves. The received amplitude is an one dimensional summary of the 3D space. Reconstructing fine-grained spatial information from the 1D summary is a severely ill-posed problem. It is even more challenging for person perception: (1) Joint interference on WiFi signal by the human body and environment via the multiple propagation path effect [57]. (2) Variety of EM properties among bodies due to bone, muscle and fat distribution [51]. (3) Temporal physical changes due to breath and heartbeats [52]. Due to these challenges, WiFi antennas have only been explored preliminarily on detecting the presence or a rough body mass even with a large antenna array [23, 22]. To the best of our survey, using WiFi devices on fine-grained person perception has never been addressed.

To solve above ill-posed problem, our solution learns from many 1D samples of the environment and human bodies. Specifically, we used two sets of off-the-shelf WiFi devices, one as transmitter set (T) and the other as receiver set (R). Three antennas were lined up in each set similar to a standard WiFi router (shown in Figure 1). WiFi signals were recorded at 30 frequencies centered at 2.4 GHz (IEEE 802.11n WiFi communication standard). We recorded RGB videos and computed body segmentation masks and body joints to annotate the signals. This setting provides 9 propagating pairs among T and R antennas, 30 1D superposing patterns per antenna pairs, and multiple 2D spatial annotations of human bodies. We developed a deep learning approach that uses annotations from RGB videos, WiFi samples as input, and reconstructs 2D body segmentation mask and body joint coordinates. Experiments showed that our approach has a comparable ability of person perception as what computer vision approaches can achieve on 2D images. Figure 1 shows examples of our Person-in-WiFi approach. To our knowledge, this is the first work that demonstrates:

1. Fine-grained person perception can be achieved using pervasive WiFi antennas.
2. To sense the human body in 2D, the physical spatial layout of sensors can be as low as one dimension.
3. A deep learning solution to map WiFi signals to human body segmentation mask and joint coordinates.

2. Related Work on Person Perception

Camera-based. Deep learning has significantly advanced human pose estimation [48, 47, 10, 14, 37, 55, 55, 9] on images captured by monocular cameras, as well as those with optical flow and motion captures [24, 16, 36, 61]. Recent prevalent approaches [20, 11, 15, 35, 56] use a powerful person detector such as Faster R-CNN [42], SSD [30] Yolo [41], FPN [29] to crop Region-of-Interest of each person from image feature maps. Then, body-wise pose estimation is done independently on the cropped feature maps. This two-stage schema gains higher performance than previous approaches those are based on global joint heat maps such as OpenPose [9].

Unfortunately, we cannot benefit from this two-stage schema because it is not possible to crop 2D pixels of the human body from WiFi signals. Inspired by [9], we developed a deep learning approach to generate Joint Heat Maps (JHMs) and Part Affinity Fields (PAFs) directly from WiFi signals. Each JHM encodes one type of joint of all persons, and each PAF encodes the direction and length of person limbs. Then person-wise poses are computed from the JHMs and PAFs similar to [9].

Radar-based. Adib et.al. [2] introduced a Frequency Modulated Continuous Wave (FMCW) radar system with broad bandwidth from 5.56 GHz to 7.25 GHz for indoor human localization, obtaining a locating resolution of 8.8 cm. This system is built with the Software-Defined Radar (SDR) toolkit and T-shaped antenna arrays. Besides, this system is well-synchronized to enable computation on Time-of-Flight (ToF) of EM wave undergoing transmission, refraction, and reflection, before being received. The ToFs are then used to generate depth maps of the environment. In [1], they promoted the system by focusing on moving

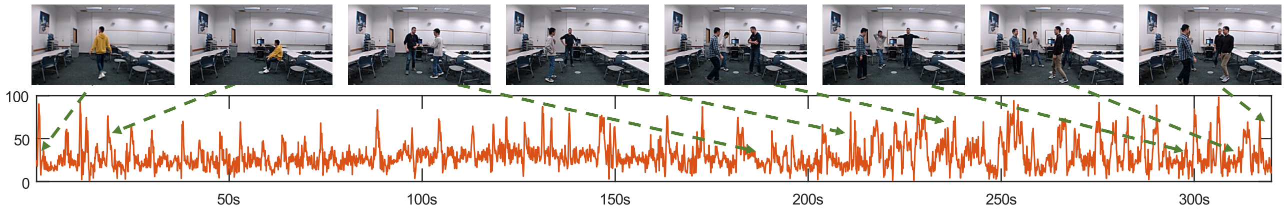

Figure 2. WiFi CSI samples recorded during single person moving and multiple person interaction around 320 seconds. The orange curve contains CSI samples of one WiFi signal frequency between one transmitter antenna and one receiver antenna.

person, and generate a rough single person outline with sequential depth maps. Recently, they applied deep learning approaches to do fine-grained human pose estimation using a similar system, termed RF-Pose [59].

LiDAR-based. LiDAR captures 3D point clouds and has been widely used in autonomous robots for Simultaneous Localization and Mapping (SLAM) [21, 13], person detection [54, 33], tracking [45, 28] and surveillance [7, 8, 44]. LiDAR sensors provide less spatial resolution than cameras. For instance, a Full HD camera with 90° diagonal field-of-view provides an angular resolution of $\approx 0.03°$, whereas the most advanced LiDARs on the market can provide up to $\approx 0.08°$ resolution . Affordable LiDARs usually have at least one magnitude lower angular resolution than the much more affordable cameras. Moreover, LiDARs have sampling rate in the range of 5-20 Hz, which is much lower than other sensors such as cameras (20-60 Hz) or WiFi adapters (100 Hz). To increase robustness, many researchers combine LiDAR with RGB cameras [38, 32, 19] or with motion sensors [12] for pedestrian detection.

WiFi-based. WiFi has been only explored for coarse-grained perception such as indoor localization with EM propagating models [3, 27] and classifying a closed-set of activities, such as opening a door [39], keystroke [4] , and hand control [50]. Wision [23] generated a bubble-like 2D heatmap to image single static person using a 8×8 WiFi antenna array. [22] generated the hologram of static objects by sweeping a WiFi antenna in 2D space and recording signals, which virtually simulates a 2D antenna array.

Till now, fine-grained person perception with WiFi signal, such as body segmentation and pose estimation, has not been well-explored. In this paper, we take one step forward to make this happen.

3. Person Perception with WiFi Signals

3.1. Methodology

We first consider the simplest setting $\mathcal{W}(\cdot)$ of a WiFi sensing system (Figure 3 (a)): one transmitting antenna, one receiving antenna and one EM frequency. A person stands still between two antennas, and one pulse signal broadcasts from the transmitting antenna. Due to the different EM properties of the human body from the floor, ceiling, furniture, etc., the signal penetrates, refracts and reflects at countless points and directions on the body. This process may probe rich spatial information of both human body (P) and environment (E) for person perception.

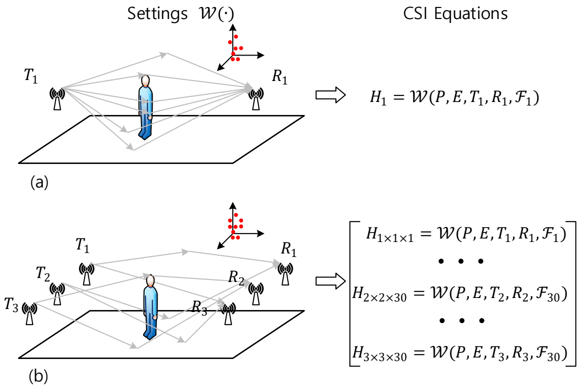
Settings $\mathcal{W}(\cdot)$
CSI Equations
T_1
R_1
$H_1 = \mathcal{W}(P, E, T_1, R_1, \mathcal{F}_1)$
(a)
T_1
T_2
T_3
R_1
R_2
R_3
$H_{1\times1\times1} = \mathcal{W}(P, E, T_1, R_1, \mathcal{F}_1)$
$H_{2\times2\times30} = \mathcal{W}(P, E, T_2, R_2, \mathcal{F}_{30})$
$H_{3\times3\times30} = \mathcal{W}(P, E, T_3, R_3, \mathcal{F}_{30})$
(b)

Figure 3. WiFi sensing system. H: CSI sample, P: person body, E: environment, T: transmitter antenna, R: receiver antenna, F: EM frequency.

Unfortunately, when the penetrated, refracted and reflected signals arrive at the receiving antenna, they superpose as a single signal sample, which is then extracted as Channel State Information (CSI) [18]. As a result, the spatial information probed by WiFi signals is collapsed to a single CSI numeric, from which reconstructing the fine-grained spatial information of human body is an ill-posed problem. For instance, if we want to perceive human body in a 100×100 px image coordinate (denoted by $\mathbf{I}(P)$) from one CSI signal (denoted by H), we have to solve 10^4 unknowns given one $\mathbf{I}(P) = f(H)$ equation.

We alleviate this problem by using the following two solutions: (1) Increasing the number of equations. In our person perception equipment, as shown in Figure 3 (b), we use 3 transmitting antennas (T), 3 receiving antennas (R) and 30 EM frequencies (F). As a reward, the $3 \times 3 = 9$ propagation pairs between antennas can capture the signals from different paths. The 30 EM frequencies generate 30 different superposing patterns at receiver antennas. This is because signals of different wavelengths can perceive objects at different scales. Moreover, we record $\mathbf{I}$ as video frames at 20 FPS and the CSI signals $\mathbf{H}$ at 100 Hz, such that each

$\mathbf{I}$ corresponds to 5 sequential CSI samples. As a result, the system in Figure 3 (b) generates $3 \times 3 \times 30 \times 5 = 1350$ equations of $\mathbf{H}$ for one setting $\mathcal{W}(\cdot)$ of person (P) and environment (E). Our problem is reduced to learn a less ill-posed function $\mathbf{I}(P) = f(\mathbf{H})$, with 1350 equations and 10^4 unknowns. Note that the number of antennas, EM frequencies and CSI sampling rate are subject to IEEE 802.11n/ac WiFi communication standard and cannot be increased indefinitely. (2) Constraining the mapping complexity. We generate multiple spatial representations of person body from $\mathbf{I}(P)$ and learn to map CSI to them using a multi-task DNNs. All these representations share the same spatial layout while highlight different body structures such as body mask, joints and limbs. This approach basically augments the data labels and further relieves the ill-posed problem.

3.2. WiFi Signal, CSI, and Hardware

In the prevalent IEEE 802.11n/ac WiFi communication system, digital packages are carried in parallel by EM waves with multiple frequencies, called orthogonal frequency division multiplexing (OFDM) technology. These packages are transmitted between multiple antenna pairs, called multiple-input-multiple-output (MIMO). CSI is computed from signals between each pair of antennas at each frequency. A CSI sample, c_i, is computed as $c_i = y_i/x_i$, where x_i and y_i are the transmitted and received digital packages. Because of this, c_i is irrelevant to the digital content of packages, but a measure of signal changes due to the reflection, refraction, absorption of EM wave with the person body and environment. Using CSI of WiFi, person perception is fundamentally possible.

To record CSI samples, we used Intel 5300 WiFi NICs and leveraged an open source tool [18], recorded CSI of 30 EM waves with a bandwidth of 20 MHz centering at the standard 2.4 GHz WiFi. The 2.4 GHz EM signal has a wavelength of around 12.5 cm. Similar to standard household WiFi routers, we uniformly spaced three receiver antennas within a wavelength, 12.5 cm. This setting maximizes the difference of CSI captured at different receiver antennas. Figure 2 shows CSI samples corresponding to different person poses and locations under the same scene.

4. Deep Learning for Person-in-WiFi

4.1. Data and Annotations

We recorded CSI at 100 Hz from receiver antennas and videos at 20 FPS from an RGB camera attached with receiver antennas. The videos are only used for annotating CSI. We synchronized CSI samples and video frames according to time stamps. In order to reduce the correlation between person body and environment, we collected data under 6 scenes in a laboratory office and 10 scenes in a classroom, shown in Figure 4. Eight volunteers were asked to perform daily activities while the number of concurrent persons in the video varied from 1 to 5 (See Table 1).

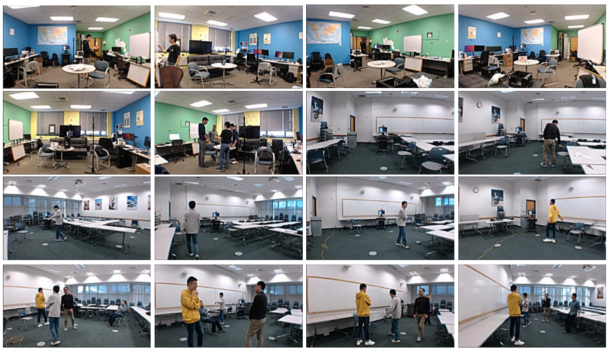

Figure 4. Data collection under 16 indoor scenes.

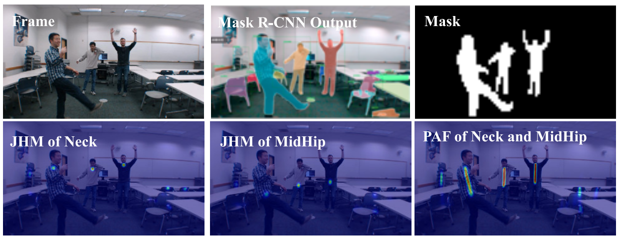

Figure 5. Example of annotations from a video frame: body segmentation mask computed by Mask R-CNN [20], JHMs and PAFs computed by OpenPose [9].

#P	1	2	3	4	5	Total
#F	99,366	13,030	20,476	20,214	1,541	154,627

Table 1. Statistics of data: Number of concurrent persons (#P) and number of video frames (#F).

From each video frame, we generated ground truth annotation for CSI as follows. For body segmentation, we used Mask R-CNN [20] to produce Segmentation Masks (SM) of persons, a $1 \times 46 \times 82$ tensor, where 46 and 82 are height and width, respectively. For pose estimation, as explained in Section 2, we cannot use a person detector like Faster R-CNN [42], SSD [30] or Yolo [41] to crop a person from the input CSI. We used the latest Body-25 model of OpenPose [9] to output body Joint Heat Maps (JHMs) and Part Affinity Fields (PAFs). For each frame, JHMs is a $26 \times 46 \times 82$ tensor, where the 26 corresponds to 25 joints and 1 background. The PAFs is a $52 \times 46 \times 82$ tensor where 52 is for x and y coordinates of 26 limbs. Figure 5 shows an example of annotations on a video frame.

4.2. Networks

Our deep neural networks (Figure 6) maps a CSI tensor to three output tensors: SM, JHMs and PAFs, where JHMs and PAFs are used later for the joint association as in [9].

The input tensor ($150 \times 3 \times 3$) contains 5 CSI samples cor-

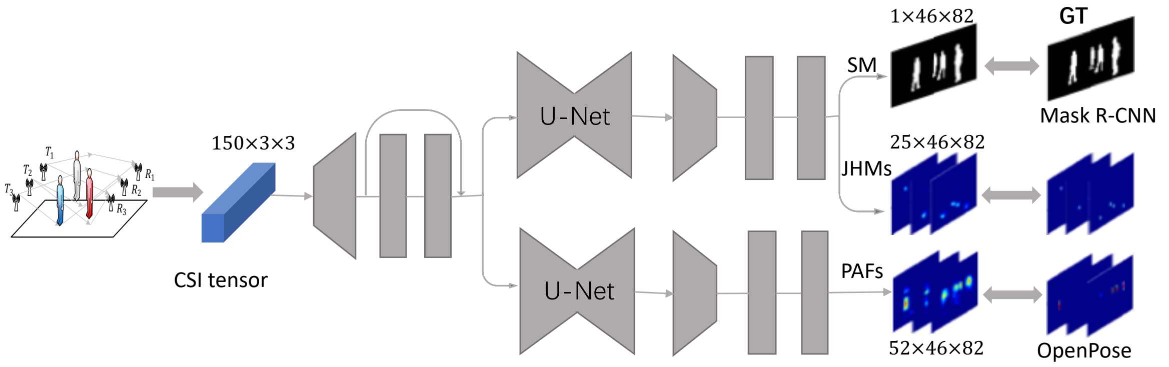

Figure 6. Deep Neural Network for Person-in-WiFi: mapping from CSI to the body Segmentation Mask (SM), Joint Heatmaps (JHMs) and Part Affinity Fields (PAFs).

responding to one video frame. The outputs are SM, JHMs and PAFs, all resized to $c \times 46 \times 82$. The input tensor is first upsampled to $150 \times 96 \times 96$, feed to a residual convolution block, and U-Nets [43]. U-Nets outputs are then downsampled to match ground truth using kernels with stride 2 on height and stride 1 on width. We found that SM (full body heatmaps) and JHMs (local joints/limbs heatmaps) are highly complementary, and one U-Net for SM and JHMs produced similar results as two independent U-Nets.

We here go deeper and discuss how the spatial information embedded in CSI is reconstructed and mapped to SM, JHMs and PAFs. We interpret in the view of Receptive Field (RF) of convolutional operation [46]. Observe that dimensions of stacked CSI represent temporal information (5), EM frequency (30), and transmitting pairs among antennas (3×3), respectively. Because of the different relative distances and angles among transmitter and receiver antennas, the 3×3 transmitting pairs capture 9 different 1D summaries of the same scene. Although the difference is subtle due to the small intervals comparing to distances to the human body, these 1D summaries are directly induced by the spatial layout of sensors. By reorganizing and reweighing, these 9 numbers can potentially be to reconstruct 2D information of the scene. This is the reason we perform 2D convolution along the 3×3 dimension of the input tensor. Observe that, the feature map after downsampling part of U-Nets has an RF size of 140, which is larger than the height and width of the up-sampled $150 \times 96 \times 96$ tensor. This ensures that the feature maps in U-Nets observed all 9 views among transmitter and receiver antennas. With supervision from annotations, the feature maps in U-Nets are forced to match the 2D spatial layout of the SM, JHMs and PAFs.

4.3. Loss and Matthew Weight

The network is trained over the sum of multiple losses

$$\mathcal{L} = \lambda_1 L_{\text{SM}} + \lambda_2 L_{\text{JHM}} + \lambda_3 L_{\text{PAF}} \tag{1}$$

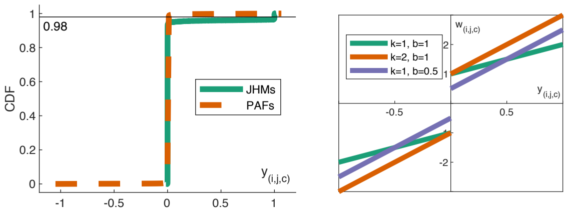

Figure 7. Left: CDF of values of JHMs and PAFs. Right: examples of three Matthew Weight functions.

where L_{SM}, L_{JHM} and L_{PAF} are losses on body SM, JHMs and PAFs, respectively. $\lambda_i, i \in 1, 2, 3$ are scalar weights to balance for these three losses. We use Binary Cross Entropy Loss to compute L_{SM} as in [20, 43, 31]. Following [9], we set λ_2 and λ_3 as 1. λ_1 is empirically set as 0.1 to balance L_{SM} with L_{JHM} and L_{PAF}. Next, we go details about the problem we faced when optimizing L_{JHM} and L_{PAF}, and the approach we proposed to tackle it.

Taking the JHMs loss as an example, directly using the popular L2 loss [11, 15, 35, 56] fails to generate good JHMs, see the middle of Figure 8. This is because the body joints only occupy very few pixels in the image, while L2 loss tends to average the regression error over all pixels. Figure 7 shows the Cumulative Distribution Function (CDF) of one JHMs tensor ($26 \times 46 \times 82 = 98072$ scalars), showing that 98% of the pixels are occupied by background, only less than 2% are for joints. This problem could be partially relieved by multiple cascaded regression stages like OpenPose or Stacked Hourglass Networks [34]. Both solutions make networks much heavier. Leading top-down approaches focus on cropped person-wise features. But one cannot directly crop persons from CSI tensors. We use a simple but efficient loss to make networks pay more attention to body joints than the background:

$$L_{\text{JHM}}^{(i,j,c)} = w_{(i,j,c)} \cdot \left\| \hat{y}_{(i,j,c)} - y_{(i,j,c)} \right\|_2^2, \tag{2}$$

where $w_{(i,j,c)}$ is the element-wise weight at index

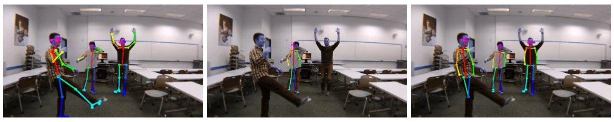

Figure 8. Matthew Weight (MW) improves pose estimation. Left: ground-truth by OpenPose [9]; Middle: results with L2 loss; Right: results with L2 loss plus MW.

(i, j, c), which is used to adjust optimizing attentions on JHMs; $\hat{y}_{(i,j,c)}$ and $y_{(i,j,c)}$ are the prediction and annotation of JHMs at (i, j, c). We propose to use the Matthew Weight (MW) to achieve the attention mechanism.

$$w_{(i,j,c)} = k \cdot y_{(i,j,c)} + b \cdot \mathbb{I}(y_{(i,j,c)}), \tag{3}$$

where $\mathbb{I}(\cdot)$ outputs $+1$ when $y_{(i,j,c)} \geq 0$, otherwise -1. Figure 7 are three MW examples. Note that MW is higher on larger elements (the body joints) in JHMs. Similarly, we applied MW in computing PAFs loss, L_{PAF}. Figure 8 shows an example that MW significantly improves pose estimation comparing to directly using L2 loss.

4.4. Implementation Details

We implemented the networks in PyTorch. The batch size was 32, and the initial learning rate is 0.001. An Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ was used in training. We used a $k = 1, b = 1$ MW in computing L_{JHM} and a $k = 1, b = 0.3$ MW in computing L_{PAF}. The networks were trained for 20 epochs.

We used an OpenPose Python API to conduct multi-person joint association given JHMs and PAFs. The output tensor is $p \times 25 \times 3$, where p represents the number of persons that networks detected, 25×3 denotes the x axis, y axis, and confidences of 25 body joints.

5. Experiments

Data were collected by groups of subjects (1-5 persons per group). Each group was asked to perform a continuous motion in the scene. We used automatic annotations for segmentation (Mask R-CNN) and poses (OpenPose) on mono-camera images that were synchronized with CSI samples. Note that this is a proof-of-concept experiments and can be further improved by high-quality manual annotations and multi-camera images (for occlusions or behind a wall).

Body Segmentation Metrics: Mean Intersection over Union (mIoU) and mAP (over AP@50 to AP@95) as used in the COCO challenge, where:

$$\text{AP@}a = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(100 \cdot IOU_n \geq a) \tag{4}$$

where N is the number of test frames, and $\mathbb{I}$ is a logical operation which outputs 1 if True and outputs 0 if False. All metrics are the higher the better.

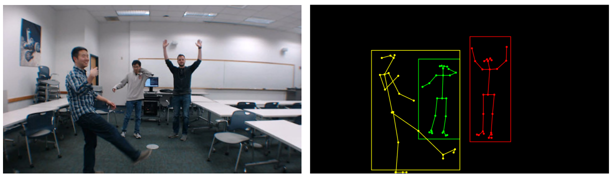

Figure 9. Aligning body joints and person bounding-boxes for computing the PCK metric.

Pose Estimation Metrics: Percentage of Correct Keypoint (PCK) [6, 58, 34]. We made a slight modification as Equation 5 considering annotations we have.

$$\text{PCK}_{\text{i}}\text{@}a = \frac{1}{P} \sum_{p=1}^{P} \mathbb{I}\left(\frac{\|pd_i^p - gt_i^p\|_2^2}{\sqrt[2]{w^{p2} + h^{p2}}} \leq a\right), \tag{5}$$

where $\mathbb{I}$ are the same as Equation 4. P is the amount of persons in test frames. i denotes the index of body joint and $i \in \{1, 2, ..., 25\}$. $\|pd_i^p - gt_i^p\|_2^2$ is the Euclidean pixel distance between the prediction and ground-truth, which is normalized by the diagonal length of the person bounding box, $\sqrt[2]{w^{p2} + h^{p2}}$. To get person bounding boxes, we aligned body joint coordinates from OpenPose [9] with the bounding box from Mask R-CNN [20] (see Figure 9).

We did not use the Object Keypoint Similarity (OKS) AP@a of the COCO Keypoint Detection challenge for two reasons: (1) Our 25 body joints requires 25 hyper-parameters to compute OKS, but the COCO dataset only provides 18; (2) The COCO dataset hyper-parameters are based on statistics of COCO data and may introduce bias in evaluating our dataset.

In the first experiment, the first 80% of samples of each subject group were used for training and the later 20% for testing. The training and testing samples are different in locomotion and body poses, but share the person identities and environments. The amount of training/test samples are 123631 and 30996, respectively.

5.1. Performance of Body Segmentation

The mAP over AP@50-AP@95 of body segmentation is 0.38 (see Table 2). High values of AP@50-AP@70 mean that person profiles can be properly detected from WiFi signals. Low values of AP@80-AP@95 indicate that subtle body masks are not well-detected. Figure 10 qualitatively show masks from WiFi comparing to the annotations by Mask R-CNN [20]. Most body locations, torsos, legs can be well-segmented, which is good enough for safety applications such as detecting falling of elderly [53] and physical conflicts among people.

5.2. Performance of Pose Estimation

Since we used Body-25 model of OpenPose to annotate the poses, 25 PCKs were computed for the 25 body

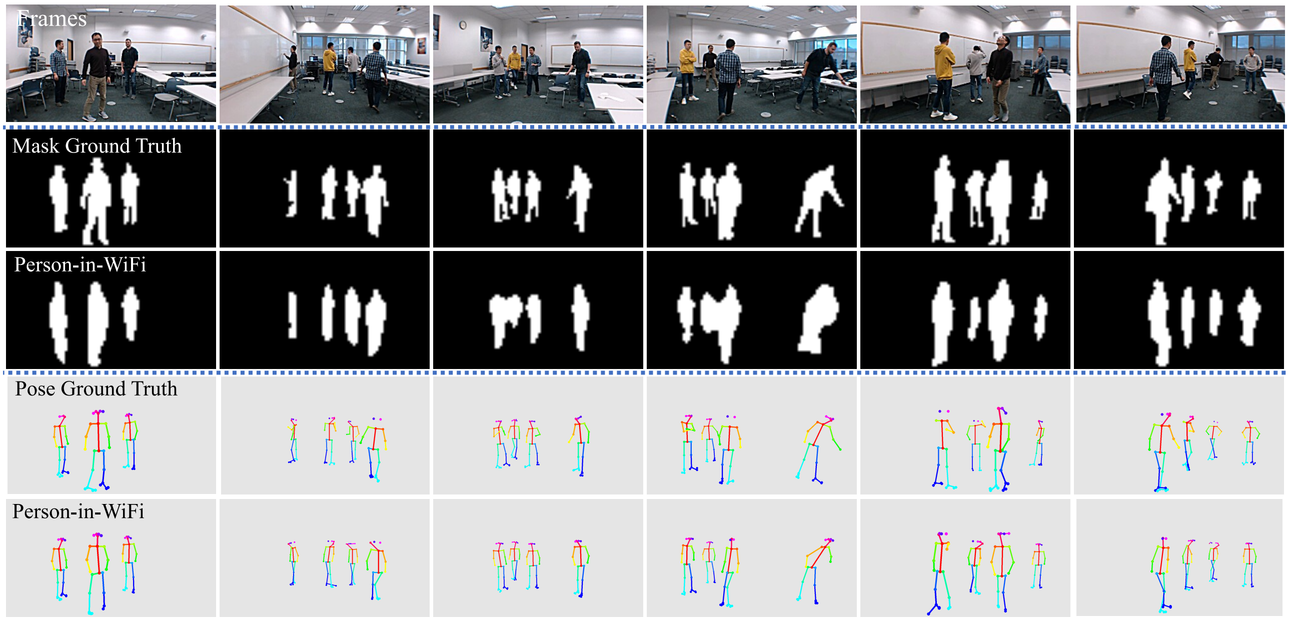

Figure 10. Peson-in-WiFi results of body segmentation and pose estimation comparing to annotations by Mask R-CNN and and OpenPose.

mIoU	**mAP**	AP@50	AP@55	AP@60	AP@65	AP@70	AP@75	AP@80	AP@85	AP@90	AP@95
0.65	0.38	0.91	0.85	0.75	0.59	0.40	0.20	0.07	0.01	0	0

Table 2. mIoU, mAP and APs of body segmentation. All metrics are the higher the better.

joints. We plot PCKs in 4 groups in Figure 11 and analyze the performance of pose estimation. The 4 groups of joints are Head {Nose, REye, LEye, REar, LEar}, Torso&Arms {Neck, Rshoulder, RElbow, RWrist, LShoulder, LElbow, LWrist}, Legs {MidHip, RHip, RKnee, LHip, LKnee} and Feet {RAnkle, LAnkle, LBigToe, LSmallToe, LHeel, RBigToe, RSmallToe, RHeel}.

As shown in Figure 11, the estimation of most joints produced high PCKs (vertical axis) at low (0.1) normalized distance error (horizontal axis). In other word, most joints were located within less than 0.1 of diagonal length of the person bounding box. Generally, joints of large body parts like in group Torso&Arms and group Legs have higher PCKs, while joints in group Head or group Feet tend to have lower PCKs. We will analyze the failure cases in the next subsection. Figure 10 show pose estimation achieved using WiFi comparing to annotations from OpenPose.

5.3. Failure cases

Several failure cases exist in our current results (see Figure 12) (1) Lack of spatial resolution (See Figure 12 (a-b)). Small limbs may be bypassed or mixed in WiFi EM waves due to the diffraction effect. For instance, WiFi signals at 2.4 GHz have a wavelength of around 12.5 cm, and may miss an object of less than 12.5 cm along its direct propagation path. However, multiple propagation paths by 3 receiver antennas and countless reflection paths of signals can capture the trace of small limbs. Figure 10 showed many successful cases. The failures could be improved by higher weights on regression errors of small limbs, more data and temporal smoothing. (2) Rare poses (Figure 12 (c-d)). More diversity in data and hard-example mining can improve the results. (3) Incomplete annotations: Camera has narrower field-of-view (70° horizontally) than the WiFi antennas that broadcast signals in 360°. Annotations from a single camera is incomplete on occluded body parts (Figure 12 (e-f)). Annotation with multiple cameras could address the issues.

	Mask-RCNN	OpenPose	Ours
mIOU	0.83	-	0.66
mPCK@0.20	-	89.48	78.75

Table 3. Gaps between Person-in-WiFi (Trained on annotations of camera-based approaches) and camera-based approaches.

5.4. Gaps with Camera-based Approaches

Above Person-in-WiFi models were trained on, therefore bounded by the annotations produced using Mask R-CNN and OpenPose. It is still possible to evaluate the gaps between two perception approaches. Table. 3 compares the results on 160 samples that were uniformly selected from above test set and manually annotated [49]. The quantitative gaps are noticeable, but could be reduced with more data and high-quality annotations, considering that Mask R-

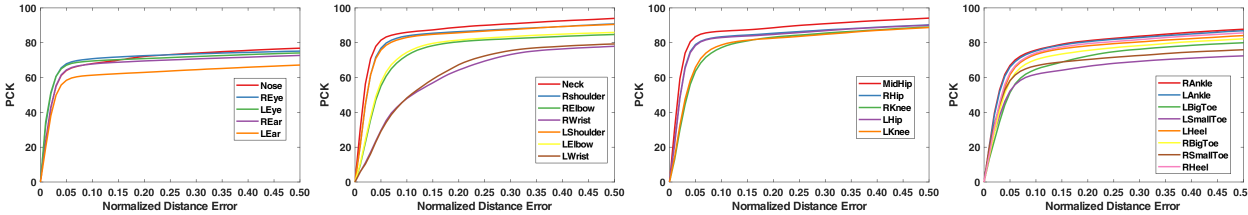

Figure 11. PCKs of pose estimation. Horizontal axis: normalized distance error of joints (see Figure 9). Vertical axis: PCKs of 25 body joints plot in four groups: (1) Head, (2) Torso&Arms, (3) Legs, (4) Feet. PCKs are the higher the better.

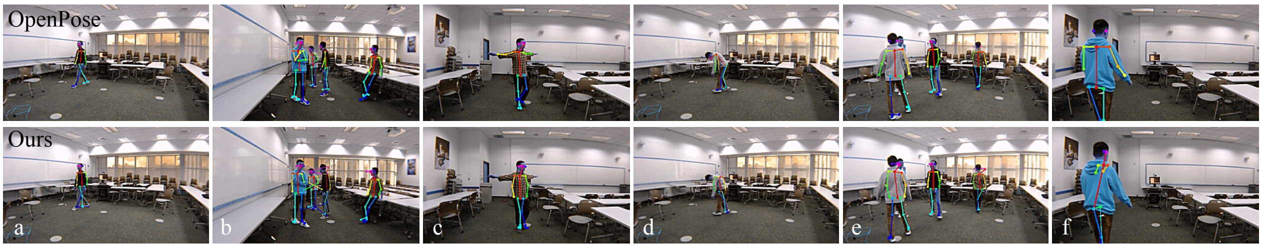

Figure 12. Example of failure cases: (a-b) Lack of spatial resolution; (c-d) Rare poses; (e-f) Incomplete annotations from camera view.

CNN and OpenPose were trained with abundant data.

5.5. Deployment in Untrained Environment

WiFi contains coupled scattering patterns of human bodies and scenes. To make the system scene-irrelevant, we hope to suppress the scene information and keep the body information in the networks. Some work we found to address this issue was for activity classification [25, 60], a much simpler task than Person-in-WiFi. As a preliminary attempt to deploy Person-in-WiFi to untrained environment, we build a GAN to transform original CSI tensors into new tensors that cannot be differentiated by their scenes.

Step 1: pre-training a binary environment discriminator (D) which takes a random pair of CSI tensors as inputs, and produces 1 if the paired tensors are from a same environment, and 0 otherwise; **Step 2:** training the network in Fig. 13; Fixing discriminator (D) in Step 1; updating a Unet generator network (G), such that any pairs of generator outputs (GCSI) produce 1s (same environment). Meanwhile, GCSI tensors are used as input tensors of the Person-in-WiFi network (see Fig. 6). The generator and Person-in-WiFi network are updated simultaneously.

We conducted preliminary experiments on 14 training scenes and 2 testing scenes. Above training approach improved segmentation mIoU from 0.12 to 0.24, improved pose estimation mPCK@0.20 from 19.34 to 31.06. Nevertheless, further improvement on untrained environment requires more data and annotations.

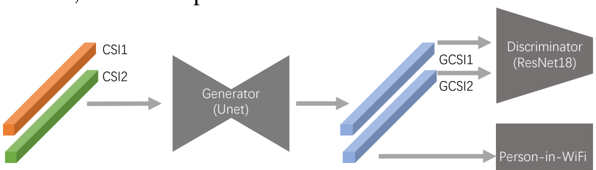

Figure 13. Adversarial training for environment invariance.

5.6. Potential Extensions

- Temporal Extension: Temporal stability and speed-up could be introduced following [56, 17, 40] from the PoseTrack challenge [5].
- 3D and Total Capture [26] Extension: Using multiple well-calibrated cameras to provide 3D annotations, our system could produce 3D poses and voxel segmentation.

6. Conclusion

WiFi devices as perception sensors are invariant to illumination and privacy-friendly comparing to cameras, while are cheaper, smaller, and more power efficient than radars and LiDARs. In this paper, we present the first work that given 1D data received at WiFi antennas, it is possible to reconstruct 2D fine-grained spatial information of human bodies. Our Person-in-WiFi approach is based on off-the-shell WiFi antennas lined-up as regular house-hold WiFi routers, making it very easy to develop perception applications in any indoor environments such as warehouse, hospital, office and home.

Acknowledgements: This work was supported in part by the National Science Foundation under grant RI-1617953, National Natural Science Foundation of China under grant 61872285 and 61572396, and the Fundamental Research Funds for the Central Universities. Fei Wang and Sanping Zhou were supported by the China Scholarship Council. Jinsong Han and Dong Huang were corresponding authors. The content is solely the responsibility of the authors and does not necessarily represent the official views of the supporting agencies.

References

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *TOG*, 34(6):219, 2015. 2

[2] Fadel Adib, Zachary Kabelac, Dina Katabi, and Robert C Miller. 3d tracking via body radio reflections. In *NSDI*, volume 14, pages 317–329, 2014. 2

[3] Fadel Adib and Dina Katabi. See through walls with wifi! *SIGCOMM Comput. Commun. Rev.*, 43(4):75–86, Aug. 2013. 3

[4] Kamran Ali, Alex X Liu, Wei Wang, and Muhammad Shahzad. Keystroke recognition using wifi signals. In *MobiCom*, pages 90–102. ACM, 2015. 3

[5] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018. 8

[6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 6

[7] Csaba Benedek. 3d people surveillance on range data sequences of a rotating lidar. *Pattern Recognition Letters*, 50:149–158, 2014. 3

[8] Csaba Benedek, Bence Gálai, Balázs Nagy, and Zsolt Jankó. Lidar-based gait analysis and activity recognition in a 4d surveillance system. *TCSVT*, 28(1):101–113, 2018. 3

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 4, 5, 6

[10] Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, pages 1736–1744, 2014. 2

[11] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *arXiv preprint*, 2018. 2, 5

[12] Arthur Daniel Costea, Robert Varga, and Sergiu Nedevschi. Fast boosting based detection using scale invariant multimodal multiresolution filtered features. In *CVPR*, pages 6674–6683, 2017. 3

[13] David Droeschel and Sven Behnke. Efficient continuous-time slam for 3d lidar-based online mapping. *ICRA*, pages 1–9, 2018. 3

[14] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*, pages 1347–1355, 2015. 2

[15] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, volume 2, 2017. 2, 5

[16] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354, 2015. 2

[17] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 350–359, 2018. 8

[18] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review*, 41(1):53–53, 2011. 3, 4

[19] Xiaofeng Han, Jianfeng Lu, Ying Tai, and Chunxia Zhao. A real-time lidar and vision based pedestrian detection system for unmanned ground vehicles. In *ACPR*, pages 635–639. IEEE, 2015. 3

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988. IEEE, 2017. 2, 4, 5, 6

[21] Wolfgang Hess, Damon Kohler, Holger Rapp, and Daniel Andor. Real-time loop closure in 2d lidar slam. In *ICRA*, pages 1271–1278. IEEE, 2016. 3

[22] Philipp M Holl and Friedemann Reinhard. Holography of wi-fi radiation. *Physical review letters*, 118(18):183901, 2017. 2, 3

[23] Donny Huang, Rajalakshmi Nandakumar, and Shyamnath Gollakota. Feasibility and limits of wi-fi imaging. In *SenSys*, pages 266–279. ACM, 2014. 2, 3

[24] Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. Modeep: A deep learning framework using motion features for human pose estimation. In *ACCV*, pages 302–315. Springer, 2014. 2

[25] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. Towards environment independent device free human activity recognition. In *MobiCom*, pages 289–304. ACM, 2018. 8

[26] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 8

[27] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. Spotfi: Decimeter level localization using wifi. *SIGCOMM Comput. Commun. Rev.*, 45(4):269–282, Aug. 2015. 3

[28] Angus Leigh, Joelle Pineau, Nicolas Olmedo, and Hong Zhang. Person tracking and following with 2d laser scanners. In *ICRA*, pages 726–733. IEEE, 2015. 3

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 2, 4

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 5

[32] Damien Matti, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Combining lidar space clustering and convolu-

tional neural networks for pedestrian detection. *CoRR*, abs/1710.06160, 2017. 3

[33] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, pages 922–928. IEEE, 2015. 2, 3

[34] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 5, 6

[35] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017. 2, 5

[36] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, pages 1913–1921, 2015. 2

[37] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, pages 4929–4937, 2016. 2

[38] Cristiano Premebida, Joao Carreira, Jorge Batista, and Urbano Nunes. Pedestrian detection combining rgb and dense lidar data. In *IROS*, pages 4112–4117. IEEE, 2014. 3

[39] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. Whole-home gesture recognition using wireless signals. In *MobiCom*, pages 27–38. ACM, 2013. 3

[40] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4620–4628, 2019. 8

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 2, 4

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 2, 4

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[44] John Shackleton, Brian VanVoorst, and Joel Hesch. Tracking people with a 360-degree lidar. In *AVSS*, pages 420–426. IEEE, 2010. 3

[45] Samir Shaker, Jean J Saade, and Daniel Asmar. Fuzzy inference-based person-following robot. *International Journal of Systems Applications, Engineering and Development*, 2(1):29–34, 2008. 3

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[47] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, pages 1799–1807, 2014. 2

[48] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 2

[49] Ketaro Wada. labelme: Image Polygonal Annotation with Python. `https://github.com/wkentaro/labelme`, 2016. 7

[50] Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han. Joint activity recognition and indoor localization with wifi fingerprints. *IEEE Access*, 7:80058–80068, 2019. 3

[51] Fei Wang, Jinsong Han, Shiyuan Zhang, Xu He, and Dong Huang. Csi-net: Unified human body characterization and action recognition. *arXiv preprint arXiv:1810.03064*, 2018. 2

[52] Hao Wang, Daqing Zhang, Junyi Ma, Yasha Wang, Yuxiang Wang, Dan Wu, Tao Gu, and Bing Xie. Human respiration detection with commodity wifi devices: do user location and body orientation matter? In *Ubicomp*, pages 25–36. ACM, 2016. 2

[53] Yuxi Wang, Kaishun Wu, and Lionel M Ni. Wifall: Device-free fall detection by wireless networks. *TMC*, 16(2):581–594, 2017. 6

[54] Zhe Wang, Yang Liu, Qinghai Liao, Haoyang Ye, Ming Liu, and Lujia Wang. Characterization of a rs-lidar for 3d perception. *arXiv preprint arXiv:1709.07641*, 2017. 2, 3

[55] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *ICCV*, pages 4724–4732, 2016. 2

[56] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 466–481, 2018. 2, 5, 8

[57] Chouchang Yang and Huai-Rong Shao. Wifi-based indoor positioning. *IEEE Communications Magazine*, 53(3):150–157, 2015. 2

[58] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *TPAMI*, 35(12):2878–2890, 2013. 6

[59] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *CVPR*, pages 7356–7365, 2018. 2, 3

[60] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pages 313–325. ACM, 2019. 8

[61] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, pages 4966–4975, 2016. 2