Person-in-WiFi 3D: End-to-End Multi-Person 3D Pose Estimation with Wi-Fi

Kangwei Yan[1], Fei Wang[1]*, Bo Qian[1], Han Ding[1], Jinsong Han[2], Xing Wei[1]
[1]Xi'an Jiaotong University, Xi'an 710049, China
[2]Zhejiang University, Hangzhou 310058, China
{yankangwei,qb90531}@stu.xjtu.edu.cn {feynmanw,dinghan,weixing}@xjtu.edu.cn,hanjinsong@zju.edu.cn

Abstract

Wi-Fi signals, in contrast to cameras, offer privacy protection and occlusion resilience for some practical scenarios such as smart homes, elderly care, and virtual reality. Recent years have seen remarkable progress in the estimation of single-person 2D pose, single-person 3D pose, and multi-person 2D pose. This paper takes a step forward by introducing Person-in-WiFi 3D, a pioneering Wi-Fi system that accomplishes multi-person 3D pose estimation. Person-in-WiFi 3D has two main updates. Firstly, it has a greater number of Wi-Fi devices to enhance the capability for capturing spatial reflections from multiple individuals. Secondly, it leverages the Transformer for end-to-end estimation. Compared to its predecessor, Person-in-WiFi 3D is storage-efficient and fast. We deployed a proof-of-concept system in 4m × 3.5m areas and collected a dataset of over 97K frames with seven volunteers. Person-in-WiFi 3D attains 3D joint localization errors of 91.7mm (1-person), 108.1mm (2-person), and 125.3mm (3-person), comparable to cameras and millimeter-wave radars. The project page is at https://aiotgroup.github.io/Person-in-WiFi-3D.

1. Introduction

Human pose estimation is a critical technology with broad applications in areas like elderly care, virtual reality, and smart homes. To achieve precise pose estimation, researchers have explored various methods, including cameras [2, 6, 18, 26, 33], radars [1, 13, 15, 23, 39], and Wi-Fi signals [10, 21, 22, 28, 29, 41]. Among these, camera-based solutions are most extensively studied, supported by a large research community and a wealth of labeled and unlabeled data. This has led to the development of well-known frameworks like convolutional pose machines [33], OpenPose [2], AlphaPose [6], Hourglasses network [18], HRNet [26], and more. However, camera solutions are not always applicable due to their dependence on proper lighting conditions and field of view. They also struggle in severe occlusion scenarios. Additionally, cameras capturing sensitive information such as identity and appearance can lead to privacy concerns in scenarios where privacy is paramount.

Unlike camera-based solutions, Wi-Fi methods are resilient to occlusions and do not capture sensitive personal details, making them well-suited for indoor scenarios. Current Wi-Fi solutions have advanced in estimating single-person 2D/3D poses. This process involves a regression problem, mapping Wi-Fi signal variations, caused by an individual's movements and presence, to their corresponding 2D/3D pose coordinates. For instance, WiSPPN [28] predicts 2D keypoint coordinates using pose adjacent matrix similarity loss. Similarly, MetaFi++ [41] estimates 2D coordinates employing mean squared error loss. In single-person 3D pose estimation, solutions like WiPose [10], Winect[21], and GoPose [22] also utilize mean squared error to learn 3D coordinates. In the case of Wi-Fi-based multi-person pose estimation, it's challenging to distinctly segment individuals from 1-dimensional (1D) Wi-Fi signals. Addressing this, Person-in-WiFi [29] adopts techniques from OpenPose [2], initially regressing keypoint heatmaps and part affinity fields, and then associating these with individuals. One alternative approach is Densepose from Wi-Fi [7], which transforms 1D Wi-Fi signals to 1280×720×3 image-like tensors under the supervision of synchronized images. This method may favor overfitting colors in training scenes, such as the color of subjects' clothing and surrounding objects, as Wi-Fi signals do not inherently capture color information.

Up to now, multi-person 3D pose estimation using Wi-Fi signals remains an unsolved challenge. In our initial attempt to evolve Person-in-WiFi into a 3D version for 14 keypoints, we represented multi-person poses with 3D keypoint heatmaps $\in 14 \times 64 \times 64 \times 64$ and 3D part affinity fields $\in 42 \times 64 \times 64 \times 64$. We replaced 2D operations, like convolutions in Person-in-WiFi, with 3D counterparts, and modified the pose-processing algorithms to produce 3D coordinates from the 3D heatmaps and fields. However, this deep network failed to converge. We identified six major

*Corresponding author.

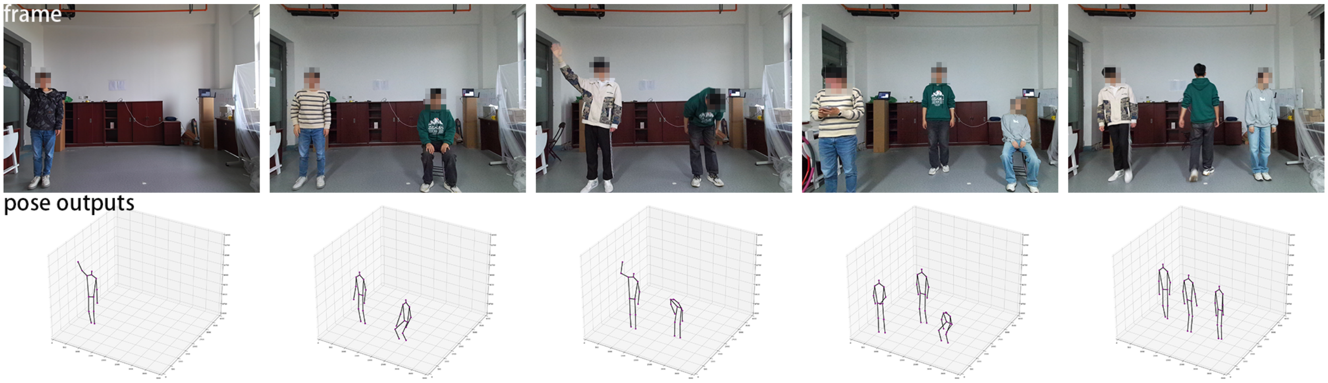


Figure 1. This paper presents Person-in-WiFi 3D, the first multi-person 3D pose estimation system with Wi-Fi signals.

shortcomings in this approach: (1) the 3D network's excessive size; (2) slow training of the 3D network; (3) sluggish 3D post-processing; (4) coarse spatial resolution, for example, e.g., $4000\text{mm}/64 = 62.5\text{mm}$ in a $4\text{m}\times 4\text{m}\times 4\text{m}$ space; (5) the lack of an end-to-end process, leading to error accumulation during network prediction and post-processing; (6) inefficient use of storage and memory, as large matrices for keypoint heatmaps and part affinity fields are primarily used to store 3D pose coordinates $\in p\times 14\times 3$ for p persons.

In this paper, we introduce Person-in-WiFi 3D. Compared to its 2D predecessor, Person-in-WiFi 3D features two significant updates. Firstly, it incorporates three Wi-Fi receivers placed at the corners of the sensing area, capturing more signal reflections from different human body parts. Secondly, Person-in-WiFi 3D is based on Transformer [27] and DETR [3], enabling direct estimation of multi-person 3D poses from Wi-Fi signals. Person-in-WiFi 3D consists of three modules: a Wi-Fi encoder, a pose decoder, and a refine decoder. The Wi-Fi encoder is designed to extract global context from tokenized Wi-Fi signals. In the pose decoder, multiple queries, initialized randomly, interact with the globally extracted information to predict a set of poses. The refine decoder then further refines these predictions. We approach multi-person pose estimation as a set prediction task and utilize a set-based Hungarian loss, ensuring each individual's ground-truth pose is uniquely predicted. Our system achieves 3D joint localization errors of 91.7mm in single-person scenarios, 108.1mm in two-person scenarios, and 125.3mm in three-person scenarios, as proven on a self-collected dataset with over 97,000 Wi-Fi samples. Our contributions in this work are outlined below:

(1) We introduce Person-in-WiFi 3D, a pioneering system designed for estimating multi-person 3D poses using Wi-Fi signals, marking a first in this area.

(2) We develop the Wi-Fi Pose Transformer, an innovative algorithm that converts Wi-Fi signals into multi-person 3D poses in an end-to-end manner.

(3) We comprehensively evaluate Person-in-WiFi 3D in our self-collected dataset, over 97,000 samples. Our dataset is ethically cleared for public release. This makes it the first Wi-Fi pose estimation dataset available for open access, promising to expedite future research in this domain.

2. Related work

2.1. Human Pose from Images

Multi-person pose estimation from images is a task of detecting joint keypoints of every person in images. It basically has two main paradigms, i.e., top-down and bottom-up. In the top-down paradigm, a person detector like Faster RCNN [8] and Yolo [20] first crops regions of every person from an image, then a single-person pose estimator regresses the person keypoints region by region. This paradigm is more accurate and has many record-breaking work, such as Hourglass networks [18], AlphaPose [6], and HR-Net [26]. In contrast, the bottom-up paradigm firstly regresses all keypoints of all persons in an image, then performs keypoint grouping methods, e.g., part affinity fields [2], associative embedding [19], and graph clustering [11], to align the regressed keypoints to every person. This paradigm has the advantage of reducing the time cost of pose estimation as the number of people largely increases. Recently, following the above two paradigms, multi-person 3D pose estimation can be achieved with RGB/D images [34, 43]. Since Wi-Fi signals carry the combined changes caused by all persons in the surroundings, we cannot crop information belonging to every person from Wi-Fi signals. Thus, the bottom-up paradigm is more suitable for multi-person pose estimation from Wi-Fi signals.

2.2. Human Pose from Wi-Fi

Human pose estimation from WiFi signals is a relatively new and emerging research area that targets privacy-preserving and occlusion use scenarios. Current Wi-Fi work have made notable advancements in single-person 2D pose estimation [28, 41], single-person 3D pose estimation [10, 21, 22], and multi-person 2D pose estimation [29]. However, the multi-person 3D pose estimation is still unsolved. In the hardware aspect, WiSPPN [28] and MetaFi++ [41] use 1 transmitter and 1 receiver for single-person 2D pose estimation. Several work achieve single-person 3D pose estimation with more Wi-Fi transceivers to capture more reflection information from human body. For example, WiPose [10] uses 1 transmitter and 3 receivers; GoPose [22] and Winect [21] both apply 1 transmitter and 4 receivers. Unfortunately, these single-person pose estimation approaches cannot extend to multi-person pose estimation because their algorithms have no person detection strategy or keypoint grouping method. Person-in-WiFi [29] is the first multi-person 2D pose estimation work, which leverages part affinity fields [2] for keypoint grouping. However, multi-person 3D pose estimation from Wi-Fi signals is still an open problem.

3. Wi-Fi Signals

Channel State Information.

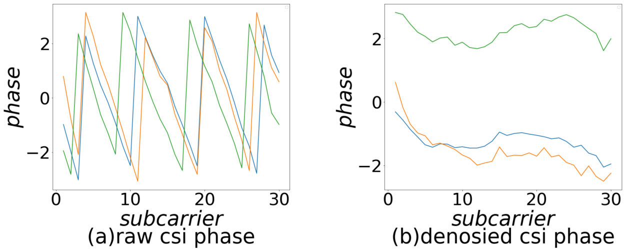
Wi-Fi devices communicate via multiple subcarriers in orthogonal frequency. The propagation process of the Wi-Fi signals between the transmitter and the receiver can be formalized as Eq. 1.

$$Y_i = H_i X_i + n_i, i \in [1, n] \tag{1}$$

where i is the subcarrier index; X_i is the bits sent by the transmitter via the i^{th} subcarrier; Y_i is the bits received at the receiver via the i^{th} subcarrier; n_i represents the noise. During the propagation, Wi-Fi signals undergo various distortions including the influence of propagation distance, scattering, power fading, etc. H_i symbolizes the signal distortion summary, technically known as the Channel State Information (CSI). It is represented as a complex-value number, from which we can derive both amplitude and phase. In Wi-Fi systems, CSI phases are often disrupted by noise and other interferences like time lag, random phase offsets due to device imperfections, and measurement process noise [30, 35]. We employ a phase denoising method based on linear transformation, as proposed in PhaseFi [30], to mitigate these interferences in CSI phases. An example of this denoising result is illustrated in Fig. 2.

Recording CSI over a period captures time-series Wi-Fi distortions triggered by environmental changes, including the presence and movements of people nearby. This is the underlying principle enabling Wi-Fi signals to estimate human poses.

Hardware Configurations. Our data collection setup includes four ThinkPad X201 laptops, one as a transmitter and the remaining three as receivers, all equipped with Intel 5300 network cards. The deployment is arranged as illustrated in Fig. 3. The transmitter is configured to broadcast Wi-Fi in channel 128 (5.64GHz) with 30 subcarriers and one antenna at a rate of 300 packets per second. The three receivers are set up to monitor the channel, each utilizing three antennas. Concurrently, an Azure Kinect camera captures RGB-D videos at 15 frames per second. We manually synchronize Kinect with the three receivers before the recording begins. Consequently, CSI samples, sized $1 \times 3 \times 3 \times 30 \times 20$, are synchronized with one video frame. The five dimensions of CSI samples respectively represent the meanings of
(#transmitter, #receiver, #antenna, #subcarrier, #time).



(a)raw csi phase

(b)denosied csi phase

Figure 2. CSI Phase denoising, where the left is the original phase, and the right is the phase after denoising.

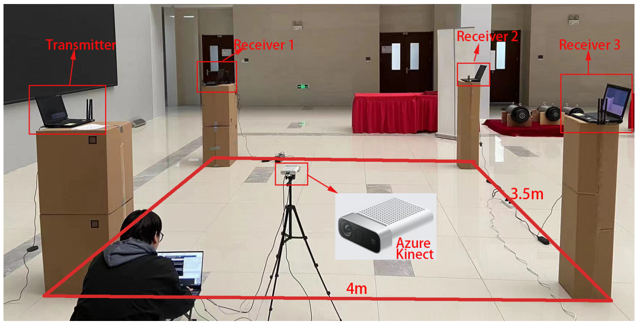


Figure 3. One of our experiment areas, a rectangle of $4m \times 3.5m$, where the camera is in the center of the horizontal axis facing the experimental area.

4. Transformer for Person-in-WiFi 3D

- **Tokenization.** The first issue we address is the effective tokenization of CSI samples. In language models, tokens are typically discrete words or characters organized temporally. In Vision Transformers (ViT) [4], tokens are spatially ordered image patches. Our aim is to sequence CSI tokens, considering CSI samples are inherently time-series data. Besides, the distinct locations of the transmitter, receivers, and antennas incorporate spatial information. We, therefore, flatten these to form a tensor, whose dimensions are interpreted as
#transmitter×#receiver×#antenna×#time, #subcarrier.
with the first dimension encompassing spatial-temporal elements. For our specific CSI samples, this results in 180 tokens ($1 \times 3 \times 3 \times 20$), where each token is a vector $\in 1 \times 30$,

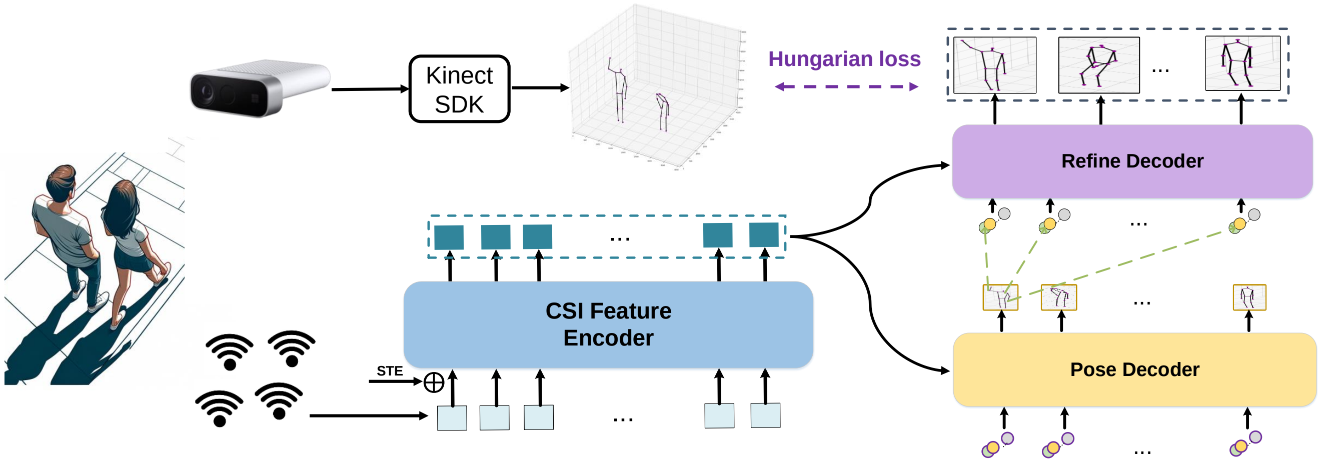


Figure 4. Person-in-WiFi 3D employs a teacher-student framework, where the Azure Kinect serves as the supervisory element. The primary function of the Wi-Fi encoder and decoder within this framework is to generate 3D poses. Subsequently, the refine decoder is tasked with refining these initial outputs. To facilitate training the network in an end-to-end manner, a set-based Hungarian loss is utilized.

representing the subcarriers. We then combine the amplitudes and denoised phases along the subcarrier dimension, leading to each token being $\in 1 \times 60$. A fully-connected layer is subsequently used to upscale each token to a dimension of $\in 1 \times 256$.

• **Spatial-temporal Embedding.** CSI sample tokens are organized in a spatial-temporal sequence. To aid in distinguishing these tokens, we introduce randomly initialized and learnable spatial-temporal embeddings (STE), each of dimension $\in 1 \times 256$. These embeddings are added with each token. After that, we have an input of $\in 180 \times 256$, which forms the foundation for the Wi-Fi encoder to effectively learn pose representation from the CSI samples.

• **CSI Encoder.** Our architecture incorporates six encoder layers to process CSI features, each comprising a multi-head self-attention module and a feed-forward network, which are fundamental components in the Transformer architecture [27]. After the encoder stage, we utilize a Multi-Layer Perceptron to produce an initial set of 3D poses and corresponding scores, denoted as $\mathcal{P}_{init} \in 180 \times (3K)$ and $\mathcal{S}_{init} \in 180 \times 1$, respectively. In our experiments, we select K as 14, which represents the 3D coordinates of 14 distinct keypoints.

• **Pose Decoder.** Following the approach of DETR [3], our Pose decoder, equipped with 100 randomly initialized and learnable queries, regresses a set of 3D body coordinates $\mathcal{P} \in 100 \times (3K)$ along with corresponding confidence scores. The decoder layers are structured as basic blocks of the DETR decoder [3], and we stack three such layers.

Additionally, drawing inspiration from Deformable-DETR [42], we are not to predict a full pose at the final layer, but rather to predict an offset at each layer. The output poses from each layer are calculated as the cumulative sum of the coordinates from the previous layer and the predicted offsets for that layer. In other words, given the pose $\mathcal{P}_{d-1}$ predicted by the $(d-1)^{th}$ decoder layer, the d^{th} layer outputs the pose as per the equation:

$$\mathcal{P}_d = \mathcal{P}_{d-1} + \Delta\mathcal{P}_d \tag{2}$$

where $\Delta\mathcal{P}_d$ represents the offsets predicted at the d^{th} layer. The initialized poses $\mathcal{P}_0 \in 100 \times (3K)$ are sampled from P_{init} based on their confidence score ranking. This method of offset prediction ensures that each layer of the decoder is adequately constrained.

• **Refine Decoder.** The concept of refinement in our system is inspired by PETR [24]. Our refine decoder is specifically crafted to fine-tune the keypoint coordinates for more precise predictions. The layers of the refine decoder are modeled after the basic blocks of the DETR decoder [3], and we employ three such layers in the stack. Mirroring the pose decoder, each layer in the refine decoder produces a relative offset $\Delta\mathcal{J} = (\Delta x, \Delta y, \Delta z)$. The joint coordinates output by the d^{th} layer adhere to the equation:

$$\mathcal{J}_d = \mathcal{J}_{d-1} + \Delta\mathcal{J}_d \tag{3}$$

Differing from the pose decoder, the refine decoder selects only G bodies for refinement. In the training phase, we use set-based Hungarian bipartite matching [14] between the pose decoder outputs $\mathcal{P} \in 100 \times (3K)$ and the ground truth $P_{gt} \in G \times (3K)$ to choose G bodies for refinement. During inference, in the absence of ground truth, this selection is based on confidence scores.

• **Loss Function.** In our system, we implement a set-based Hungarian loss [14] to ensure each person's ground-truth pose receives a unique prediction. The focal loss function [16] is used as the classification loss to determine the accuracy of predicting a body as a person. Additionally, we employ Mean Squared Error (MSE) loss for keypoint regression. The total loss function is expressed as:

$$L = L_{cls} + \lambda L_{kpt} \tag{4}$$

where λ is a balancing weight between classification and regression. The formula for L_{kpt} is:

$$L_{kpt} = \text{mean}(||\hat{P} - P_{gt}||_2). \tag{5}$$

In this equation, $\hat{P}$ and P_{gt} denote the predicted 3D joint coordinates and the ground truth, respectively.

The formula for L_{cls} is given by:

$$L_{cls} = \begin{cases} -\alpha(1-\hat{y})^{\gamma}\log(\hat{y}) & \text{if } y = 1 \\ -\alpha\hat{y}^{\gamma}\log(1-\hat{y}) & \text{otherwise.} \end{cases} \tag{6}$$

Here, α and γ are factors adjusting the impact of positive vs. negative samples and easy vs. hard samples, while y and $\hat{y}$ represent the class labels and confidence scores, respectively.

• **Training Details.** We optimize Person-in-WiFi 3D using the Adam optimizer [12], set with a momentum of 0.9 and a weight decay of 10^{-4}. The batch size for training is set at 32. The training process spans 500 epochs, starting with an initial learning rate of 2×10^{-5}. This rate is reduced by a factor of 0.1 at the 450th epoch. In Eq.6, we set an α value of 0.25 and a γ value of 2. To balance the classification and keypoint localization losses, the weight λ is set at 35 in Eq. 4.

5. Experiment

5.1. Dataset Acquisition

Dataset diversity. We recruited 7 volunteers, with heights of 160-177cm and weights of 55-70kg, to perform 8 daily actions in 3 locations with different degrees of multipath effects: an office, a classroom, and a corridor. The actions were reaching out, raising hands, bending over, stretching, sitting down, lifting legs, standing, and walking. In each location, 1-4 volunteers freely performed these actions for 40 seconds, and we recorded 152 clips of 40 seconds each, including 32 single-person, 48 two-person, 48 three-person, and 24 four-person scenarios. This approach resulted in a total of 456 CSI and RGB-D clips, ensuring the dataset captured a wide variety of volunteers, actions, places, and combinations of these elements across different time.

Dataset statistics. We possess 456 RGB-D clips, each 40 seconds long, with a total of 600 frames per clip (first 550 for training, last 50 for testing) and 270,000 frames overall. Multi-person 3D keypoint coordinates are generated from these video clips using the Kinect Body Tracking SDK, then act as CSI annotations for network training and evaluation. However, the SDK's performance on our clips was not entirely satisfactory. We meticulously cleaned the dataset, frame by frame, discarding samples that the SDK failed, e.g., the 4-person case. As a result, we have a dataset with over 97,000 samples, the details of which are outlined in Table. 1.

	1-person	2-person	3-person	all
training	28121	36242	25583	89946
test	2586	3184	2054	7824

Table 1. Dataset statistics.

5.2. Evaluation Metrics

(1) Mean Per Joint Dimension Location Error (MPJDLE). Person-in-WiFi 3D estimates multi-person pose in 3D dimension: horizontal, vertical, and depth. We can compute the location error of the horizontal dimension as reported in mmPose-NLP [23] with Eq. 7 and Eq. 8.

$$\text{PJDLE(k, h)} = \frac{1}{F}\sum_{f=1}^{F}\frac{1}{P_f}\sum_{p=1}^{P_f}|pre(f,p,k,h) - gt(f,p,k,h)|_1 \tag{7}$$

$$\text{MPJDLE(h)} = \frac{1}{K}\sum_{k=1}^{K} E_{PJDLE}(k,h) \tag{8}$$

where $pre(f,p,k,h)$ and $gt(f,p,k,h)$ denote the predicted and ground truth positions, respectively, for the horizontal dimension of the k^{th} joint of the p^{th} person in the f^{th} frame. The $||_1$ notation is used to calculate the L1 distance. Therefore, PJDLE(k,h) represents the Per Joint Dimension Localization Error (PJDLE) for the k^{th} joint in the horizontal dimension. MPJDLE(h) is the Mean PJDLE calculated across all joints in the horizontal dimension. Similarly, we can compute MPJDLE(v) for vertical MPJDLE, and MPJDLE(d) for depth MPJDLE.

(2) Mean Per Joint Position Error (MPJPE). If we compute L2 distances between 3D predictions and 3D ground truth as Eq. 9, we have joint position errors.

$$\text{PJPE(k)} = \frac{1}{F}\sum_{f=1}^{F}\frac{1}{P_f}\sum_{p=1}^{P_f}||pre(f,p,k) - gt(f,p,k)||_2 \tag{9}$$

$$\text{MPJPE} = \frac{1}{K}\sum_{k=1}^{K} E_{PJPE}(k) \tag{10}$$

PJPE(k) is for the PJPE of the kth joint. MPJPE is the mean of all PJPE(k), which is widely used to evaluate 3D human pose estimation [9].

5.3. Results

Table. 2 presents PJPE and PJDLEs across horizontal, vertical, and depth dimensions, with units in millimeters (mm).

Joint	PJPE	PJDLE(h)	PJDLE(v)	PJDLE(d)
neck	81.6	41.0	39.6	47.5
head	88.4	43.3	43.2	52.2
left shoulder	90.5	47.5	41.2	52.6
right shoulder	92.5	50.3	40.9	53.8
left elbow	124.1	65.2	55.3	71.6
left hip	72.9	38.1	30.2	46.3
right elbow	132.9	71.2	58.3	76.1
right hip	74.7	40.5	30.2	46.8
left hand	182.1	91.0	98.2	87.9
left knee	85.1	38.9	33.2	58.3
right hand	202.9	97.8	111.3	97.9
right knee	86.9	40.4	33.0	59.4
left ankle	92.1	41.2	38.4	61.4
right ankle	94.1	43.2	39.2	62.4
Mean	107.2	53.5	49.4	62.4

Table 2. MPJPE and MPJDLEs, measured in millimeters (mm).

Notably, the largest prediction errors occur in the left and right hands, with 182.1mm and 202.9mm, respectively. This higher error rate is attributed to the greater diversity in hand movements compared to overall body movements. The final MPJPE is calculated at 107.2mm, a level of precision suitable for various applications like indoor person localization and tracking, intrusion detection, and coarse-grained action recognition. Similarly, PJDLEs for the left and right hands also exhibit the largest errors, for reasons analogous to those for MPJPE. The mean joint dimension location errors, MPJDLEs, are 53.5mm, 49.4mm, and 62.4mm in the horizontal, vertical, and depth dimensions, respectively.

Table. 3 displays the estimation errors in scenarios with one, two, and three persons. We observe that the prediction errors escalate with an increasing number of people. For instance, considering MPJPE, there is an increase of 16.4mm from 1-person scenarios (91.7mm) to 2-person scenarios (108.1mm), and a further increase of 17.2mm from 2-person to 3-person scenarios, closely aligning with the 16.4mm increment. We infer that transitioning from single-person to multi-person 3D pose estimation is not inherently problematic for the Wi-Fi system. Rather, the current limitations in achieving multi-person 3D pose estimation are algorithmic, not due to inherent constraints of the capabilities of Wi-Fi system.

Metric	1-person	2-person	3-person
MPJPE	91.7	108.1	125.3
MPJDLE(h)	42.4	51.2	60.6
MPJDLE(v)	43.3	47.7	53.8
MPJDLE(d)	50.0	60.4	69.7

Table 3. As the number of people in the scene increases, the prediction errors increase.

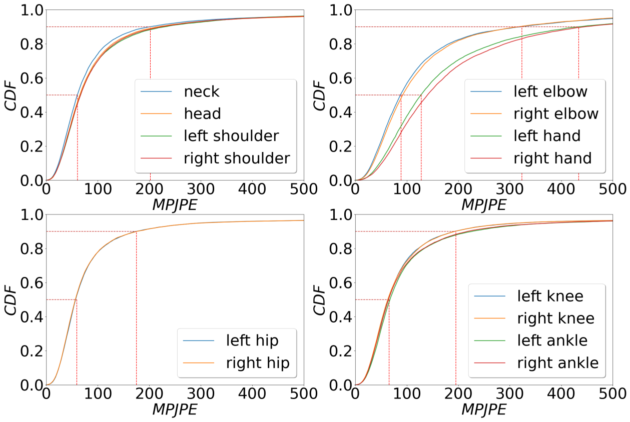


Figure 5. Cumulative distribution of MPJPE in four groups i.e., upper body, middle body, legs, and arms.

To gain a clearer understanding of prediction error distribution, we analyzed the cumulative distribution of MPJPE, categorizing it into four segments: upper body, arms, middle body, and legs. This categorization, shown in Fig. 5, helps prevent overlapping of the curves. From the figure, it is evident that 50% of the MPJPEs for pose estimation in the upper body, middle body, and legs are under 80mm, while 90% are under 170mm. However, for the arms, particularly the left and right hands, the MPJPEs are significantly higher, with 50% of estimations falling within 140mm and 90% within 380mm. Both Table. 2 and Fig. 5 underscore the increased difficulty in accurately estimating arm joints compared to others. We think this problem could be relieved by enhancing the dataset with more diverse arm positions in future research.

5.4. Visualization and Failure cases

We visualize multiple instances of multi-person 3D pose estimation predictions across three different rooms, as shown in Fig.1. It is important to note that these images are used solely for visualization purposes and are not included in the deep model training. The figure demonstrates that Person-in-WiFi 3D effectively localizes people, which could be leveraged for tracking purposes. Additionally, it can estimate a variety of actions, such as sitting, raising hands, and walking. More crucially, even though the MPJPEs for arm joints are higher compared to other joints, the predictions are sufficiently accurate to identify the actions being performed. Therefore, Person-in-WiFi 3D is also capable of supporting action and interaction recognition applications.

In Fig.6, we present some typical failure cases of Person-in-WiFi 3D. The leftmost figure illustrates the system is challenging to accurately estimate the positions of spare hand gestures. The middle figure depicts a scenario where Person-in-WiFi 3D incorrectly locates a person about one

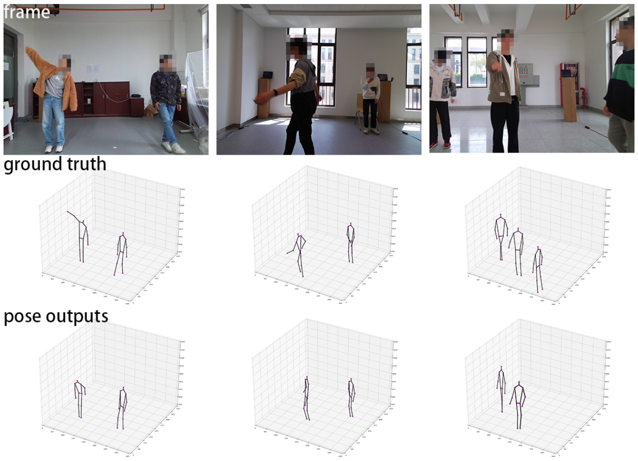


Figure 6. Failure cases, (1) spare gestures (2) location prediction error, (3) out of sensing area.

meter away from their actual position. The last figure highlights an instance where Person-in-WiFi 3D fails to detect a person situated out of the system's coverage area. Addressing these shortcomings will likely require collecting more data, deploying additional devices, etc.

Metric	Person-in-WiFi 3D	WiPose*
MPJPE	91.7	101.8
MPJPE(h)	42.4	42.1
MPJPE(v)	43.3	47.8
MPJPE(d)	50.0	58.0

Table 4. Person-in-WiFi 3D is more accurate than WiPose. (WiPose* is not officially released yet. The results here originate from our independent reproduction of WiPose.)

5.5. Comparison with existing work

- **Comparison with WiPose.** All current Wi-Fi-based pose estimation research has not been open-sourced in terms of signal processing, network frameworks, or datasets. We independently reproduced WiPose [10] and tested it using our dataset for single-person scenarios. As shown in Table. 4, Person-in-WiFi 3D demonstrates better performance in single-person 3D pose estimation compared to WiPose.
- **Comparison with Person-in-WiFi.** Person-in-WiFi [29] focuses on multi-person 2D pose estimation. We dedicated nearly half a year to adapting it for 3D pose estimation, but unfortunately, these efforts were unsuccessful. Person-in-WiFi faces several challenges in the realm of multi-person 3D pose estimation, challenges which our Person-in-WiFi 3D effectively overcomes.

(1) inefficient use of storage and memory. Extending Person-in-WiFi for 3D poses necessitates the use of 3D Joint Heatmaps (JHMs) and 3D Part Affinity Fields (PAFs) for pose grouping, leading to significant storage and memory requirements. Specifically. The JHMs $\in 14 \times 64 \times 64 \times 64 \times 64$ for 14 keypoints, and the PAFs are sized at $42 \times 64 \times 64 \times 64 \times 64$ for 42 limbs across 3D axes. Storing JHMs and PAFs for a single frame requires about 7 to 8MB, amounting to roughly 2TB for all frames. In contrast, our method uses a ground truth tensor sized as the number of persons $p \times 14 \times 3$ axes per frame, which totals to approximately 200MB for all frames, significantly reducing the storage demand.

(2) slow training. We extend Person-in-WiFi to a two-stream 3D UNet. Since UNet has to store all middle feature maps, JHMs, and PAFs, the minibatch size is limited to 4 on an Nvidia 3090 GPU. On our dataset, to train one epoch, the extended Person-in-WiFi takes $\sim$6 hours on one Nvidia 3090 GPU. In contrast, our work supports the minibatch size of 64, only taking 20mins to train one epoch.

(3) sluggish pose grouping. We also extended the pose grouping from 2D to 3D. It takes about 10s to conduct pose grouping for 1 frame using 3D JHMs and PAFs. In contrast, our method can produce 3D human poses at 54fps.

- **Comparison to cameras and millimeter-wave radars.** Recent camera-based solutions for multi-person 3D pose estimation have reported MPJPEs in the range of 54-70mm [5, 32, 40]. Zhang et al. [37] reported MPJPE of 29.9mm on self-collected data using millimeter-wave radars. The MPJPE of Person-in-WiFi 3D stands at 91.7mm in single-person scenarios and 107.2mm in multi-person scenarios, comparable to cameras and millimeter-wave radars. We also tried to apply PETR [24] directly applied to our task, and the MPJPE is 219.0, showcasing the effectiveness of our enhancements. Notably, Person-in-WiFi 3D is trained with annotations automatically generated by the Azure Kinect SDK, whose performance limits the upper bound of Person-in-WiFi 3D. Nevertheless, Person-in-WiFi 3D is still the first work that achieves multi-person 3D pose estimation with Wi-Fi signals.

5.6. Ablation Study

- **Number of receivers.** Our configuration includes a diagonally opposed transmitter, two transmitter, and three receivers, arranged at the corners of a rectangular area to establish the sensing zone. With three pairs of transmitters and receivers, Person-in-WiFi 3D is adept at capturing Wi-Fi signals reflected from the human body at different angles. While the reduction of receivers will lead to a considerable drop in system performance, as shown in Table 5.

	3 receivers	R1+R2	R1+R3	R2+R3	R2
MPJPE(mm)	107.2	148.5	142.8	144.2	173.0

Table 5. Performance on the number of receivers.

	without Refine decoder	with Refine decoder
MPJPE	116.1	107.2
MPJPE(h)	54.7	53.5
MPJPE(v)	46.9	49.4
MPJPE(d)	65.8	62.4

Table 6. Refine decoder improves pose estimation performance.

Metric	amplitude	amplitude+phase.	A. and P.(de.)
MPJPE	192.4	137.7	107.2
MPJPE(h)	105.6	61.6	53.5
MPJPE(v)	103.1	71.1	49.4
MPJPE(d)	71.9	70.0	62.4

Table 7. Incorporating phase information and applying denoising significantly improves the system outcomes. 'A. and P.(de.)' indicates the use of amplitude with denoised phase.

Metric	Cross-Person	Cross-Environment
MPJPE	131.6	626.4
MPJPE(h)	76.1	414.5
MPJPE(v)	57.4	304.8
MPJPE(d)	62.2	133.0

Table 8. Experimental results of the Cross-Person and Cross-Environment settings.

- **Refine decoder.** We conduct an ablation experiment on the Refine decoder, with results detailed in Table. 6. These results demonstrate that the Refine decoder improves pose estimation performance.
- **Phase and Phase Denoising.** Table. 7 presents the ablation study results on phase denoising. We explore three scenarios: using only amplitude, combining amplitude with non-denoised phase, and using amplitude with denoised phase. The experimental findings indicate that incorporating phase information and applying denoising significantly improves outcomes of Person-in-WiFi 3D.
- **Occlusion.** To assess the adaptability of Person-in-WiFi 3D to occlusion scenarios, we placed a white screen obstructing the front perspective of the Kinect, while a side-view camera captured images, as shown in Fig. 7. The figure demonstrates Person-in-WiFi 3D remains functional in detecting persons even when their view is obstructed. Due to the absence of ground truth data from the occluded Kinect perspective, we are unable to provide quantitative results for these scenarios.
- **Cross-domain ability.** Person-in-WiFi 3D undergoes evaluation in both leave-one-person-out and leave-one-environment-out scenarios. The results, displayed in Table.8, show that in cross-person cases, Person-in-WiFi 3D effectively recognizes poses despite individual differences in body shape and movement habits. However, Person-in-WiFi 3D does not implement any cross-environment adaptation strategies, leading to a large decline in performance during cross-environment testing.

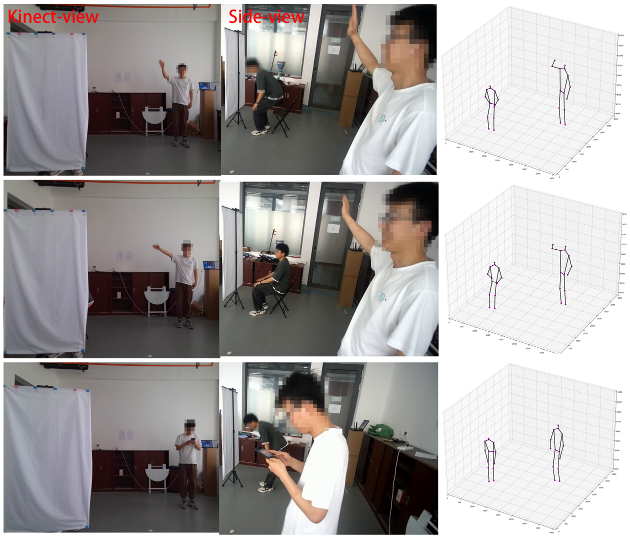


Figure 7. Person-in-WiFi 3D remains functional in detecting persons when their view is obstructed.

6. Conclusion

We presented Person-in-WiFi 3D, a fully end-to-end approach that takes the multi-person 3D pose estimation task as a set-based prediction problem and achieves it elegantly with a Transformer framework. However, the current system still has several limitations.

(1) The upper-bound performance Person-in-WiFi 3D is severely limited by the Azuere Kinect Body Tracking SDK. However, the current SDK works badly in cases where there are more than 3 persons in the view. Future work can introduce VICON system [25], IMUs [36], and optical markers [17] for more accurate annotations.

(2) Person-in-WiFi 3D has 4 distributed Wi-Fi transceivers. Accurately adjusting their relative spatial configurations, including positions, elevations, and antenna orientations, to different places is challenging to ensure cross-location generalization. Future efforts could focus on increasing the tolerance of spatial configurations, perhaps by using environment-independent features like BVP [38] and propagation models such as Fresnel Zone [31] in Wi-Fi signal preprocessing and deep networks, or focus on large model pre-training and efficient cross-environment finetuning.

Acknowledgements – This work was supported by the National Natural Science Foundation of China under grant 62102307, U21A20462, 62372400, and 62372365, "Pioneer" and "Leading Goose" R&D Program of Zhejiang under grant No.2024C03287, Key Research and Development Program of Shaanxi (ProgramNo.2021GXLH-Z-021), China Postdoctoral Science Foundation under grant 2023T160511 and 2021M692562, and CAAI-MindSpore Open Fund; and developed on Openl Community.

References

[1] Sizhe An and Umit Y Ogras. Fast and scalable human pose estimation using mmwave point cloud. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pages 889–894, 2022. 1

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1, 2, 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2, 4

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[5] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7204–7213, 2020. 7

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1, 2

[7] Jiaqi Geng, Dong Huang, and Fernando De la Torre. Densepose from wifi. *arXiv preprint arXiv:2301.00250*, 2022. 1

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 5

[10] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. Towards 3d human pose construction using wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2020. 1, 3, 7

[11] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 718–734. Springer, 2020. 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[13] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. m3track: mmwave-based multi-user 3d posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pages 491–503, 2022. 1

[14] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[15] Guangzheng Li, Ze Zhang, Hanmei Yang, Jin Pan, Dayin Chen, and Jin Zhang. Capturing human pose using mmwave radar. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2020. 1

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[17] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 8

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 1, 2

[19] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[21] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. Winect: 3d human pose tracking for free-form activity using commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):1–29, 2021. 1, 3

[22] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. Gopose: 3d human pose estimation using wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–25, 2022. 1, 3

[23] Arindam Sengupta and Siyang Cao. mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1, 5

[24] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022. 4, 7

[25] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 8

[26] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 1, 2

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[28] Fei Wang, Stanislav Panev, Ziyi Dai, Jinsong Han, and Dong Huang. Can wifi estimate person pose? *arXiv preprint arXiv:1904.00277*, 2019. 1, 2, 3

[29] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. Person-in-wifi: Fine-grained person perception using wifi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5452–5461, 2019. 1, 3, 7

[30] Xuyu Wang, Lingjun Gao, and Shiwen Mao. Phasefi: Phase fingerprinting for indoor localization with a deep learning approach. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015. 3

[31] Xuanzhi Wang, Kai Niu, Anlan Yu, Jie Xiong, Zhiyun Yao, Junzhe Wang, Wenwei Li, and Daqing Zhang. Wimeasure: Millimeter-level object size measurement with commodity wifi devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(2):1–26, 2023. 8

[32] Zitian Wang, Xuecheng Nie, Xiaochao Qu, Yunpeng Chen, and Si Liu. Distribution-aware single-stage models for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13096–13105, 2022. 7

[33] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 1

[34] Qingqiang Wu, Guanghua Xu, Sicong Zhang, Yu Li, and Fan Wei. Human 3d pose estimation in a lying position by rgb-d images for medical diagnosis and rehabilitation. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5802–5805. IEEE, 2020. 2

[35] Yaxiong Xie, Zhenjiang Li, and Mo Li. Precise power delay profiling with commodity wifi. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 53–64, 2015. 3

[36] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13167–13178, 2022. 8

[37] Jianxiong Zhang, Zhongping Cao, Wen Ding, Rihui Cheng, Xuemei Guo, and Guoli Wang. Multi-spectrum fusion towards 3d human pose estimation using mmwave radar. In *Proceedings of 2022 Chinese Intelligent Systems Conference: Volume I*, pages 220–232. Springer, 2022. 7

[38] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8671–8688, 2021. 8

[39] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018. 1

[40] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 550–566. Springer, 2020. 7

[41] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal*, 2023. 1, 2, 3

[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4

[43] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. 3d human pose estimation in rgbd images for robotic task learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1986–1992. IEEE, 2018. 2