Towards 3D Human Pose Construction Using WiFi

Wenjun Jiang[†], Hongfei Xue[†], Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, Lu Su[*]
State University of New York at Buffalo, Buffalo, NY USA
{wenjunji, hongfeix, cmiao, shiyangw, senlin, chongtia, smurali3, haochenh, zhisun,lusu}@buffalo.edu

ABSTRACT

This paper presents WiPose, the first 3D human pose construction framework using commercial WiFi devices. From the pervasive WiFi signals, WiPose can reconstruct 3D skeletons composed of the joints on both limbs and torso of the human body. By overcoming the technical challenges faced by traditional camera-based human perception solutions, such as lighting and occlusion, the proposed WiFi human sensing technique demonstrates the potential to enable a new generation of applications such as health care, assisted living, gaming, and virtual reality. WiPose is based on a novel deep learning model that addresses a series of technical challenges. First, WiPose can encode the prior knowledge of human skeleton into the posture construction process to ensure the estimated joints satisfy the skeletal structure of the human body. Second, to achieve cross environment generalization, WiPose takes as input a 3D velocity profile which can capture the movements of the whole 3D space, and thus separate posture-specific features from the static objects in the ambient environment. Finally, WiPose employs a recurrent neural network (RNN) and a smooth loss to enforce smooth movements of the generated skeletons. Our evaluation results on a real-world WiFi sensing testbed with distributed antennas show that WiPose can localize each joint on the human skeleton with an average error of 2.83*cm*, achieving a 35% improvement in accuracy over the state-of-the-art posture construction model designed for dedicated radar sensors.

CCS CONCEPTS

• **Networks** → *Wireless access points, base stations and infrastructure*; • **Human-centered computing** → *Interaction techniques.*

KEYWORDS

Human Pose Construction; WiFi Sensing; Deep Learning

ACM Reference Format:
Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, Lu Su. 2020. Towards 3D Human Pose Construction Using WiFi. In *The 26th Annual International Conference on Mobile Computing and Networking (MobiCom '20), September 21–25, 2020, London, United Kingdom.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3372224.3380900

[*]Lu Su is the corresponding author.
[†]The first two authors contributed equally to this work.

MobiCom '20, September 21–25, 2020, London, United Kingdom

ACM ISBN 978-1-4503-7085-1/20/09…$15.00
https://doi.org/10.1145/3372224.3380900

1 INTRODUCTION

In recent years, significant research efforts have been spent towards building intelligent wireless sensing systems, with the goal of leveraging pervasive wireless signals to perceive and understand the activities of humans. Thus far, various wireless sensing systems and algorithms have been proposed, mainly to track the position of the monitored human subject and recognize his/her activities through analyzing the signals reflected off the human body. With the localization and recognition accuracy progressively increased, a fundamental question rises: *how much information related to human body and activity is carried in the wireless signals?* and more importantly, *is it rich enough to image the human body like a camera?*

A recent pioneer study [59, 60] offers a preliminary answer to the above question. It is revealed that with the supervision of visual information, radio frequency (RF) signals can be used to generate 2D and even 3D skeletal representations of the human body. By overcoming the technical challenges faced by traditional camera-based human perception solutions, such as occlusion, poor lighting, clothing, as well as privacy issues, wireless human sensing technique demonstrates the potential to enable a new generation of applications capable of supporting more sophisticated interactions between humans and their physical surroundings. Despite the inspiring findings presented in [59, 60], the prohibitive requirements in both hardware (i.e., a carefully assembled and synchronized 16+4 T-shaped antenna array) and RF signals (i.e., Frequency Modulated Continuous Wave with a broad signal bandwidth of 1.78 GHz) severely limit the application scope of their system.

To tackle this problem, we propose to make use of the pervasive WiFi devices, and "image" 3D human postures from WiFi signals. More specifically, we aim to *reconstruct 3D skeletons composed of the joints on both limbs (i.e., arms and legs) and torso (e.g., shoulders, waist, hip) of the human body.* With the ubiquity of WiFi devices, such system could facilitate a wide spectrum of applications. Examples include health care and assisted living where the activities of the elder and patients need to be monitored without jeopardizing their privacy, gaming and virtual reality where human postures need to be transferred into the virtual world from an environment full of occlusions, and theft detection in groceries/shopping malls where thieves tend to cover their hand movements with clothes, bags, books, etc.

However, to unleash the power of WiFi-carried information, we have to address a series of challenges. First, the generated postures should be realistic-looking. It is undesired that the constructed

skeleton has, for example, unrealistically long or short limbs. Second, WiFi signals usually carry substantial information that is specific to the environment where the postures are recorded and the human subject who performs the postures. As a result, a posture construction model that is trained on a specific subject in a specific environment will typically not work well when being applied to generate another subject's skeleton from the WiFi signals that are recorded in a different environment. Third, the synthesized movement of skeletons should be continuous and smooth, in consistency with the real human activities.

To address the above challenges, we propose to adopt deep learning techniques, which have been proved to be effective on extracting useful knowledge from complicated data. In this paper, we develop a deep learning framework, named **WiPose**, to extract human postures from WiFi signals. First, WiPose can *encode the prior knowledge of human skeleton into the posture construction process* to ensure the estimated joints satisfy the skeletal structure of the human body. Second, to achieve cross environment generalization, WiPose takes as input *a 3D velocity profile which can capture the movements of the whole 3D space*, and thus separate posture-specific features from the static objects in the ambient environment. Finally, WiPose employs *a recurrent neural network (RNN) and a smooth loss* to enforce smooth movements of the generated skeletons.

In order to evaluate the proposed WiPose framework, we develop a WiFi sensing testbed with distributed antennas to collect WiFi data, and use the VICON motion capture system [2] to generate high precision 3D human skeletons as the ground truth to train our proposed deep learning model. Our evaluation results show that WiPose can localize each joint on the human skeleton with *an average error of* 2.83*cm*, achieving a 35% *improvement in accuracy* over the state-of-the-art posture construction model proposed in [60].

To the best of our knowledge, this is *the first investigation on 3D human pose construction using commercial WiFi devices*. With its superior effectiveness and generalizability, the proposed WiPose framework symbolizes a major step towards the practical deployment of wireless human sensing systems in real world.

2 SYSTEM OVERVIEW

In this paper, we consider a real-life scenario where the human subject is monitored by multiple WiFi devices whose signals are affected by the subject's activities. Our goal in this paper is to reconstruct the subject's 3D skeletons using the WiFi signals collected from these wireless devices. Figure 1 shows an overview of our proposed system, which contains three major components: data collection, data preprocessing and skeleton construction.

- **Data Collection.** The function of this component is to collect the WiFi signals that can be used to reconstruct the subject's 3D skeleton. During the data collection process, we use one transmitter and several distributed receiving antennas to capture the posture of the human subject. In addition to the collection of WiFi data, we also use a VICON motion capture system [2] to generate high precision 3D skeleton of the subject, which is used as the ground truth to train the proposed deep learning model in our system.
- **Data Preprocessing.** This component extracts the Channel State Information (CSI) from the collected WiFi signals, and then preprocess the CSI data so that they can be fed into the proposed deep learning model. Specifically, we first perform data denoising to remove the phase offset of the CSI signals. Then, we divide the denoised CSI data into non-overlapping small segments and transform each segment to a representation that can be fed into the deep learning model.
- **Skeleton Construction.** This component is designed to construct the subject's 3D skeleton using the representation generated from the CSI data. To achieve the goal, we propose a deep learning model that can encode the skeletal structure of human body in the training process to ensure the realisticness of the generated postures. The details of the proposed deep learning model are described in Section 3.

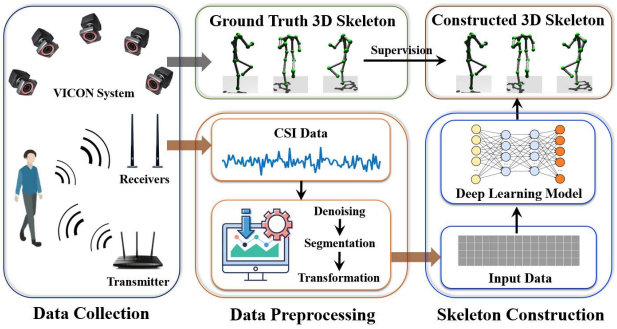


Figure 1: System Overview.

3 METHODOLOGY

3.1 Overview

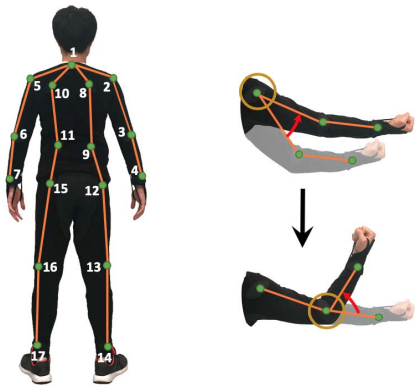
Our goal is to reconstruct 3D human posture from CSI data, specifically to generate the joints and body segments connecting the joints. There are some challenges towards this end. For example, the estimated positions of the joints should be close to the ground-truth, the generated posture should be realistic-looking, and the synthesized body movement should be smooth.

An intuitive solution is to directly and independently estimate the position of each joint from CSI data. However, due to the low spatial resolution and error-prone nature of WiFi signals, such estimation cannot be accurate. As a result, the constructed skeleton may not look real (e.g., unrealistically long or short limbs).

To address this challenge, we propose a deep learning framework, named WiPose, which *can encode the prior knowledge of human skeleton into the posture construction process to ensure the estimated joints satisfy the skeletal structure of the human body.*

The intuition is that in 3D space, we can treat the segments of human body as individual *rigid bodies* with fixed length. As shown in Figure 2a, we model human skeleton as a tree with the nodes being the joints and the edges being the body segments. On the skeleton tree, with the length of body segments being fixed, to infer the position of each joint (i.e., tree node), we only need to estimate the rotation of its associated body segment (i.e., tree edge) with regard to its parent joint.

In our model, the rotations of body segments are recursively estimated from the root joint to the leaf joints. For example, as shown in Figure 2b, if we only consider the movement of an arm, the shoulder is the parent joint of the elbow, which is also the parent of the hand. We can decompose the arm movement as first rotating the



(a) Skeleton (b) Rotation

Figure 2: Human skeleton and joint rotation.

elbow with respect to the shoulder and then rotating the hand with respect to the elbow. This process is called forward kinematics [3]. Through learning the rotation of the joints and applying them to a given skeletal structure, we can make the estimated joints naturally satisfy the constraints of the skeleton, and thus the reconstructed human posture will look realistic.

Moreover, we carefully choose the neural network structure as well as the loss functions to make our model adaptive to *unseen human subjects* and enforce the synthesized body movement to be *smooth*. The details will be presented in the next subsection.

3.2 Neural Network

The proposed WiPose deep learning framework is illustrated in Figure 3. After preprocessing, we transformed the raw CSI data extracted from M distributed antennas into a sequence of input data. We denote the t-th input data as x_t, and the whole sequence of input data as $x_{1:T}$.

After that, we employ convolutional neural networks (CNNs) to extract spatial features from x_t. In particular, we use stacked four-layer CNNs, and in each CNN layer, the dimension of the kernels (i.e., filters) is determined by the dimension of x_t, which can be 1D, 2D or 3D. After each convolutional layer, we add a batch norm layer to normalize the mean and variance, followed by a leaky rectified linear unit (Leaky ReLU) to add non-linearity to the model and a dropout layer to avoid the over-fitting of the model.

After the four-layer CNNs, we get a sequence of feature vector $z_{1:T}$ from the input $x_{1:T}$. Since a body movement usually spans multiple time slots, there are high temporal dependencies between the consecutive data samples. To learn the relationship between consecutive data samples, we further feed $z_{1:T}$ into a recurrent neural network (RNN), which is an ideal model for this task due to its ability of connecting the hidden states of temporally dependent data. To capture relatively long movement, we adopt Long Short-Term Memory (LSTM) [12], an effective and widely used RNN. In our model, we put three-layer LSTMs on top of the CNNs.

The last and ultimate task of our neural network is to apply the learned features to a given skeletal structure to construct the posture of the subject through recursively estimating the rotation of the body segments, a process called forward kinematics. Given there are N joints on the skeleton tree, the forward kinematics

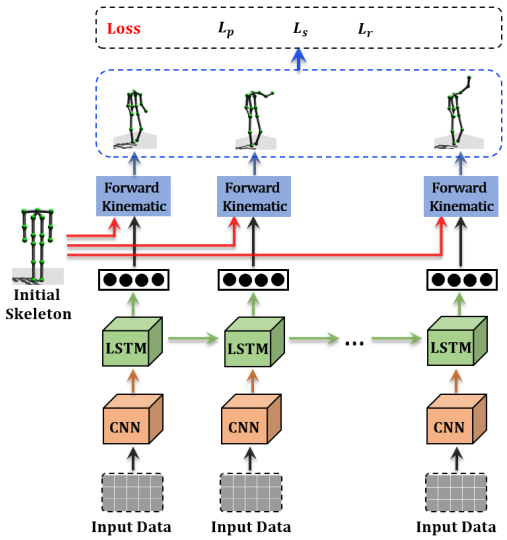


Figure 3: Model Overview.

process is mathematically defined as follows:

$$p^i = p^{parent(i)} + R^i(\bar{p}^i - \bar{p}^{parent(i)}), \tag{1}$$

where $p^i \in \mathbb{R}^3$ is the 3D coordinate of joint $i, i = 2, 3, \cdots, N$, $p^{parent(i)} \in \mathbb{R}^3$ is the parent joint of p^i on the skeleton tree, and $\bar{p}^i, \bar{p}^{parent(i)}$ are the initial position of $p^i, p^{parent(i)}$ respectively. R^i denotes the rotation of the joint p^i with respect to its parent. It is an orthogonal 3×3 matrix defined in the 3D rotation group, denoted as $\mathbb{SO}^3$. The 3D rotation group can be represented by unit quaternions. A quaternion is a *hyper complex number* with the form $a+b\mathbf{i}+c\mathbf{j}+d\mathbf{k}$, where a, b, c, d are real numbers and $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are quaternion units. The unit quaternion is a quaternion with norm one. It can be transformed from θ radians about a unit axis $e = (e_x, e_y, e_z)$, represented as quaternion $(\cos(\theta/2), e_x \sin(\theta/2), e_y \sin(\theta/2), e_z \sin(\theta/2))$. Given a unit quaternion $q^i = (q^i_a, q^i_b, q^i_c, q^i_d)$, the corresponding rotation matrix can be derived through:

$$R^i = \begin{bmatrix} 1-2({q^i_c}^2+{q^i_d}^2) & 2q^i_bq^i_c-2q^i_dq^i_a & 2q^i_bq^i_d+2q^i_cq^i_a \\ 2q^i_bq^i_c+2q^i_dq^i_a & 1-2({q^i_b}^2+{q^i_d}^2) & 2q^i_cq^i_d-2q^i_bq^i_a \\ 2q^i_bq^i_d-2q^i_cq^i_a & 2q^i_cq^i_d+2q^i_bq^i_a & 1-2({q^i_b}^2+{q^i_c}^2) \end{bmatrix}.$$

Compared with rotation matrices, unit quaternions are more compact, more numerically stable, and more efficient [1], which is ideal for us to represent the movement of the human body. Since directly calculating quaternions from the position of the joints (called inverse kinematics) is an ill-posed problem because the given joint positions can be fulfilled by multiple joint rotations, we learn it through back-propagation using the forward kinematics layer [40]. The inputs of the forward kinematics layer are the initial skeletal structure of the human subject and the learned features from LSTM. The forward kinematics layers will treat the features as the rotation of the joints and then apply them to the given skeletal structure to construct the actual joint positions. In this way, our neural network will focus on learning the skeleton independent movement features (i.e., the rotation of the joints) from the input data. Note that although the skeletal structure of each subject is necessary during the training process, it is optional during the testing process, because as long as we get the rotation of each

joint, the human motion will be the same no matter what skeletal structure we apply them to.

3.3 Loss Functions

Training the neural network comes down to minimizing the error between the predicted position of each joint i at each time slot t, denoted as $\hat{p}_t^i$, with the corresponding ground truth being p_t^i. In order to achieve this, given that there are N joints on the skeleton tree and the input CSI sequence contains T data samples, we first minimize the *position loss* L_p, which is defined as the L_2 norm between $\hat{p}_t^i$ and p_t^i:

$$L_p = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\|\hat{p}_t^i - p_t^i\|_2. \tag{2}$$

The position loss treats the posture at each time point independently. As a result, the movement of a joint over time may not be smooth, which will cause the estimated posture jitter. In order to solve this problem, we add a *smooth loss* to make the difference between consecutive postures during estimation similar to that of the ground truth:

$$L_s = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{N}\sum_{i=1}^{N}\|(\hat{p}_t^i - \hat{p}_{t-1}^i) - (p_t^i - p_{t-1}^i)\|_H. \tag{3}$$

Here $\|\cdot\|_H$ is the Huber norm. For $z = (z_1, z_2, \cdots, z_n)$, the Huber norm is defined as:

$$\|z\|_H = \frac{1}{n}\sum_{i=1}^{n} huber(z_i), \tag{4}$$

where

$$huber(z_i) = \begin{cases} 0.5z_i^2 & \text{if} |z_i| < 1 \\ |z_i| - 0.5 & \text{otherwise.} \end{cases} \tag{5}$$

Moreover, since the position of a joint is inferred through recursively rotating the joints from the root joint to that joint, the estimation error on the joint position may accumulate during this process. The position estimation for a joint may be misled if the learned position of its parent joint has already deviated from the ground truth. So it is necessary to introduce a loss to penalize the error in the relative position of a joint with respect to its parent, as follows:

$$L_r = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N-1}\sum_{i=2}^{N}\|(\hat{p}_t^i - \hat{p}_t^{parent(i)}) - (p_t^i - p_t^{parent(i)})\|_H. \tag{6}$$

Since the length of each body segment is fixed in our model, the relative position of a joint with respect to its parent is only determined by how much it rotated from an initial relative position. Therefore, we call this loss the *rotation loss*.

With all these three losses, we can finally give the overall objective function as follows:

$$J = L_p + \beta L_s + \gamma L_r, \tag{7}$$

where β, γ are the pre-defined hyper-parameters to balance the three losses. We optimize the above objective function through Adam [21].

3.4 Skeleton Construction

After our model is trained, we can apply it to construct the skeleton of a subject from a CSI sequence of arbitrary length. If the skeletal structure of the subject in terms of bone length is available, we can incorporate it in the model as the initial skeleton and apply the rotation of joints on it to improve the skeleton estimation. If such prior knowledge is not available, we can either roughly estimate the subject's skeletal structure based on, for example, a photo of the subject and/or his height information, or simply use a standard skeletal structure. As we mentioned before, since our model can precisely estimate the rotation of all the joints, even if the inputted skeletal structure is not exactly the same as that of the monitored subject, our model can still reconstruct the subject's postures.

4 CROSS-DOMAIN INPUTS

When deploying the skeleton reconstruction system in real world, one challenge we will face is how to make the system adapt to new environments and human subjects. WiFi signals usually carry substantial information that is specific to the environment where the postures are recorded and the human subject who performs the postures. On one hand, the WiFi signals, when being transmitted, may be penetrating, reflected, and diffracted by the media (e.g., air, glass) and objects (e.g., wall, furniture) in the ambient environment. On the other hand, different human subjects with different ages, genders, heights, weights, and body shapes affect the signals in different ways, even if they are taking the same posture. As a result, a posture construction model that is trained on a specific subject in a specific environment will typically not work well when being applied to generate another subject's skeleton from the WiFi signals that are recorded in a different environment.

In this paper, we refer to a pair of environment and human subject as *domain*. The domain where the model is trained is called the source domain, and the domain where the model is applied is called the target domain.

To achieve cross-domain generalization, one possible solution is to add an adversarial neural network to the proposed deep learning framework to extract domain independent features, as in existing work on domain independent activity recognition [20]. However, this solution will add significant overhead in data collection and model re-training. Instead, we look to move the domain generalization ability from the upper model level downward to the lower signal level,

Following this idea, we adopt body-coordinate velocity profile (BVP) [64] that describes power distribution over different velocities, at which body parts involved in the posture movements. BVP naturally can separate posture-specific features from the ambient objects whose velocities are 0. It is extracted directly from the CSI data through physical model, and can be fed as the input to our deep learning framework in a seamless manner. Although the BVP feature can capture the kinetic characteristics of human posture, it only models the horizontal velocity and ignores the vertical velocity. In order to construct 3D human skeleton, we extend the BVP to 3D velocity profile which can capture the movements of the whole 3D space. The details of this extension will be presented in the rest of this section.

As discussed previously in Section 3.4, our proposed neural network can achieve cross-subject generalization by incorporating

the subject's skeletal structure. We thus focus our discussion on cross-environment issue in this section.

4.1 CSI Preprocessing

The movement of the subject between a pair of transmitter and receiver will lead to Doppler effect, which shifts the frequency of the signal collected by the receiver. The Doppler frequency shift (DFS) $f_D(t)$ is defined as the change in the length of the signal propagation path $d(t)$ [33, 44]:

$$f_D(t) = -\frac{1}{\lambda}\frac{d}{dt}d(t), \tag{8}$$

where λ is the wave-length. Moreover, transmitted signal arrives at the receiver through multiple paths, so the CSI data collected by each receiver can be modeled as:

$$H(f,t) = (H_s(f) + \sum_{l=1}^{L}\alpha_l(f,t)e^{j2\pi\int_{-\infty}^{t}f_{Dl}(u)du})e^{j\epsilon(f,t)}, \tag{9}$$

where $H_s(f)$ represents the CSI from static paths (corresponding to the line-of-sight signal and the signals reflected by ambient objects), and we are interested in the CSI from L dynamic paths (corresponding to the signals reflected by the moving human body). α_l is the attenuation of the l-th path, and $\epsilon(f,t)$ is the phase offset caused by timing alignment offset, sampling frequency offset and carrier frequency offset. The power of $H_s(f)$ is much stronger than the signals reflected from the moving body parts because it contains the signal from the line-of-sight path.

In order to remove the phase offset, we first calculate the conjugate multiplication of CSI as [44]:

$$\begin{aligned} H(f,t)\overline{H(f,t)} = (H_s(f)H_s(f) + \\ \sum_{k=1}^{L}\sum_{l=1}^{L}\alpha_l(f,t)\alpha_k(f,t)e^{j2\pi\int_{-\infty}^{t}f_{Dl}(u)-f_{Dk}(u)du} + \\ H_s(f)\sum_{l=1}^{L}\alpha_l(f,t)e^{j2\pi\int_{-\infty}^{t}f_{Dl}(u)du} + \\ H_s(f)\sum_{l=1}^{L}\alpha_l(f,t)e^{j2\pi\int_{-\infty}^{t}-f_{Dl}(u)du}). \end{aligned}$$

The first term on the right hand side represents static signals and has the highest power. Since it does not contain the velocity information we care about, we remove it through subtracting the mean value from the conjugate multiplication. The second term has very small value so we can ignore it. The remaining third and fourth terms will be further used to extract 3D velocity profile.

4.2 3D Velocity Profile

After we deducted the static components from the conjugate multiplication $H(f,t)\overline{H(f,t)}$, we conduct short-time Fourier transform on the remaining dynamic components to extract DFS profile. For each time snapshot in the spectrograms, we treat it as a DFS profile $D \in R^{F\times M}$, where F is the length of frequency bins and M is the number of transceiver links. Here the number of transceiver links equals the number of receivers because we have only one transmitter in our system. Our DFS profile is symmetric to the zero frequency because it contains both $f_{Dl}(t)$ and $-f_{Dl}(t)$. The DFS profiles are still domain-dependent since they might be different for different wireless links. Next, we will derive the domain independent 3D velocity profile from DFS profiles.

We establish a coordinate system with the origin set to be the location of the subject with a fixed height, the x-axis to be the orientation of the subject, and the z-axis to be the upward vertical direction. We define the 3D velocity profile to be a 3D tensor V of size $K\times K\times K$, representing that there are K possible velocities on each of the x-axis, y-axis, and z-axis. We denote the location of the transmitter to be $\vec{l}_t = (x_t, y_t, z_t)$, and the location of the m-th receiver to be $\vec{l}_r^{(m)} = (x_r^{(m)}, y_r^{(m)}, z_r^{(m)})$. Given a velocity $\vec{v} = (v_x, v_y, v_z)$, the power of some frequency components will increase due to the contribution of this velocity component. Let the affected frequency component on the m-th link be $f^{(m)}(\vec{v})$, it can be derived as:

$$f^{(m)}(\vec{v}) = a_x^{(m)}v_x + a_y^{(m)}v_y + a_z^{(m)}v_z, \tag{10}$$

where $a_x^{(m)}$, $a_y^{(m)}$ and $a_z^{(m)}$ are coefficients which can be computed from the locations of the transmitter and the m-th receiver:

$$\begin{aligned} a_x^{(m)} &= \frac{1}{\lambda}(\frac{x_t}{\left\|\vec{l}_t\right\|_2} + \frac{x_r^{(m)}}{\left\|\vec{l}_r^{(m)}\right\|_2}), \\ a_y^{(m)} &= \frac{1}{\lambda}(\frac{y_t}{\left\|\vec{l}_t\right\|_2} + \frac{y_r^{(m)}}{\left\|\vec{l}_r^{(m)}\right\|_2}), \\ a_z^{(m)} &= \frac{1}{\lambda}(\frac{z_t}{\left\|\vec{l}_t\right\|_2} + \frac{z_r^{(m)}}{\left\|\vec{l}_r^{(m)}\right\|_2}), \end{aligned} \tag{11}$$

Similar to that in [64], we use an assignment matrix $A^{(m)} \in \{0,1\}^{F\times K^3}$ to represent the project of the 3D velocity profile V to the DFS profile of the m-th link $D^{(m)}$ as:

$$A_{j,k}^{(m)} = \begin{cases} 1 & f_j = f^{(m)}(\vec{v}_k) \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where f_j is the j-th frequency sampling point on the DFS profile, and $\vec{v}_k$ is the k-th element on the vectorized V. Then the relation between the $D^{(m)}$ and V can be modeled as:

$$D^{(m)} = c^{(m)}A^{(m)}V, \tag{13}$$

with $c^{(m)}$ to be a scaling factor to model the different propagation loss on different links. This equation can be solved through compressed sensing as in [64].

5 TESTBEDS

5.1 VICON System

In this paper, we use the VICON motion capture system [2] to generate 3D human skeletons and take them as the ground truth to train our proposed deep learning model. As shown in Figure 4, our VICON system consists of 21 VICON Vantage cameras which can emit and receive infrared light. When collecting the motion data, we place 17 high precision pearl markers on each subject representing the human skeleton, and the positions of them are shown in Figure 2a. These markers are covered with highly reflective materials, and thus are able to reflect back far more infrared light than the surface of the subject so that they can be easily captured

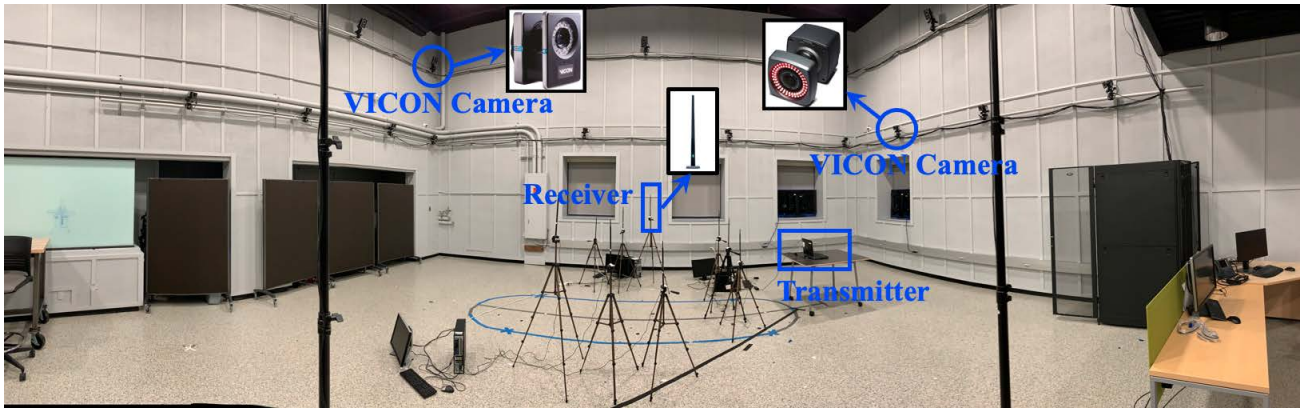


Figure 4: Testbeds and the basic scenario of posture construction.

by the VICON Vantage cameras. Our VICON system can estimate the position of the markers with errors less than 2*mm* [28]. The sampling rate of our VICON system is set as 10 Hz.

5.2 WiFi Testbed

Most of the existing wireless systems employed till date, both using dedicated hardware and COTS devices, to the best of knowledge, are built upon the receivers having multiple antennas assembled together, and sometimes even a perfectly designed antenna array. Such design makes it easy to extract parameters such as Angle of Arrival (AoA) and Time of Flight (ToF). However, in real world scenarios, although there are quite a few WiFi antennas available in devices such as smartphone, smart watches, laptops, these antennas are usually separate and not designed for sensing tasks. In order to simulate this scenario, we develop a WiFi sensing testbed with distributed antennas as shown in Figure 4.

Our WiFi testbed consists of one laptop and three desktops. Each computer is equipped with Intel 5300 wireless NIC connected with three antennas. We use the laptop as the transmitter and activate one antenna of it. In order to build a distributed wireless sensing system, we treat each antenna on the desktops as one receiver, and place them at different locations. Linux 802.11 CSI tools [13] are used on our testbed to log CSI data. We set the CSI tools in the monitor mode to enable the transmitter to broadcast UDP packets to the receivers. Thus, all the receivers can simultaneously receive packets from the transmitter. We set our testbed to transmit WiFi signals on channel 165 (5.825 GHz) where there is little interference from other devices. The packet rate is set at 1000 packets per second and the bandwidth is set at 20 MHz.

Ubuntu 14.04 is installed on both the transmitter and the receivers. We use a VICON motion capturing system installed on Windows 10 to collect the ground truth skeleton. We record the local timestamps of the CSI data and the skeleton data, and use the timestamps to align the data. To achieve this, it is important to synchronize the local clock on all the receivers and the VICON system. We use network time protocol (NTP) to ensure the synchronization and achieves average synchronization error of 7 ms.

6 EXPERIMENTS

6.1 Experimental Setting

6.1.1 Data Collection and Preprocessing. In our experiment, 10 volunteers (including both men and women) are employed as the subjects to collect the CSI data, and we consider 16 typical daily-life activities including: lifting left/right hand for 45/90/180 degree, lifting both hands for 90/180 degree, sweeping left/right hand for 45/90 degree, sweeping both hands for 90 degree, lifting leg, waving hand, walking on the spot. Each subject is asked to conduct each activity for one minute. While we are collecting the CSI data, we also use the VICON system to simultaneously collect the skeleton data for the purpose of model training and evaluation. The sampling rate of the skeleton data is 10 Hz, and the sampling rate of the CSI data is about 1000 Hz.

We first interpolate the CSI data to obtain uniform sampling periods through nearest-neighbor interpolation. Then we transform the raw CSI data into two different types of inputs for the proposed and baseline models. One is the *denoised CSI data*, from which we remove the phase offset through calculating the absolute value of the raw CSI data. The other is the *3D velocity profile*, which is calculated from the raw CSI data as discussed in Section 4.

In order to align the CSI sequence with the skeleton sequence, we use 100 CSI samples from 9 receiving antennas to estimate one skeleton. Specifically for the denoised CSI data, we concatenate 100 CSI samples (each of which is the channel state measurement of 30 subcarriers) to form a 9×30×100 matrix as the input. While for the 3D velocity profile, we calculate each profile from 100 CSI samples. In our experiment, the height of the 3D velocity profile origin is set as 1.3 m, the range of the velocities on all the axes are set as [-2.0, 2.0] m/s, and the number of possible velocities K is set as 9, so the 3D velocity profile input is of size 9×9×9.

Finally, since the makers are attached to the clothes of the subject, and thus may be shaking when the subject conducts activities. As a result, the obtained skeleton cannot always keep the bone length of the subject constant. In order to compensate such error, for each activity and each subject, we fix the marker on the root joint, and normalize the position of the other markers through adjusting the length of each body segment to the average length of that body segment. The normalized skeleton is used as ground truth for training both the proposed and baseline models.

6.1.2 Model Setting. When implementing the stacked four-layer CNNs in our model, we use 2D convolution operation for the CSI data and 3D convolution operation for the 3D velocity profile. The numbers of the convolutional filters in these layers are set as 64, 128, 64 and 1, respectively, for both types of input data. Some techniques

Table 1: Average joint localization errors (unit: *mm*) for the basic scenario.

Joint Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Overall
RFPose3D (CSI)	6	28	92	165	26	78	136	14	18	14	18	19	27	26	20	27	28	43.6
WiPose (CSI)	0	17	59	105	16	52	90	7	13	7	13	14	19	18	14	19	20	28.3
RFPose3D (3DVP)	37	59	134	222	61	129	210	44	45	45	45	40	46	46	42	50	51	76.7
WiPose (3DVP)	0	23	77	133	22	64	108	9	17	10	17	19	26	25	19	27	27	36.7

such as batch normalization [19] and dropout [37] (the dropout rate is set as 0.2) are also incorporated in our implementation of CNNs. The adopted activation function is Leaky ReLU [51] with the parameter α set as 0.02. For the implementation of LSTM, we set the hidden state number as 544 and the dropout rate as 0.1. The hyper-parameters β and γ in the loss function are both set as 1.0.

6.1.3 Baseline. RFPose3D is the state-of-the-art deep learning model for 3D human skeleton construction from FMCW-based RF signals. In this paper, we implement the deep learning model of RFPose3D [60] (except the input interface in order to fit WiFi data) as our baseline. RFPose3D regards human skeleton reconstruction problem as a joint classification problem. This model builds a 18-layer convolutional network with residual connections [15], and classifies each skeleton joint to a predefined voxel in 3D space. In our implementation, we use the 2D convolution on the denoised CSI inputs and the 3D convolution on the 3D velocity profile inputs.

6.2 Performance Evaluation for Basic Scenario

We first evaluate the performance of the proposed framework in the basic scenario that is shown in Figure 4. In this scenario, we place one transmitter and nine receiving antennas in an empty area of a room. The receiving antennas are equally divided into three groups. One group of the antennas are placed in front of the subject, and the other two groups are placed on the left and right hand sides of the subject, respectively. To generate the 3D velocity profile from the CSI data, we place the receiving antennas in different heights.

Posture Construction Using CSI Data. In this experiment, we first collect CSI data to generate the postures of 10 subjects. For each subject and each activity, the first 70% of the data samples are used for training, and the rest 30% are used for testing. As for the performance measure, we use the *average joint localization error*, which is defined as the average Euclidean distance between the predicted joint locations and the ground truths for all the subjects and activities.

Table 1 reports the average joint localization error for each joint that is indexed in Figure 2a. We also calculate the overall result by averaging the errors for all the joints. The results show that our proposed framework performs much better than the baseline. The overall prediction error of our framework is only 28.3*mm* while that of the baseline method is 43.6*mm*. For individual joints, the estimation errors for the 3rd (right elbow), 4th (right hand), 6th (left elbow), and 7th (left hand) joints are larger than those of other joints. One of the reasons is that the reflected signals from human arms are weaker than that from other parts of the human body, due to the small reflection area. Additionally, in our activity design, the arms of the subjects have a larger moving range, which makes it much harder to track the above four joints than other parts of the body. Next, we analyze the advantages of our proposed framework through visualizing the constructed skeleton from CSI data.

Figure 5 shows some examples of the constructed 3D skeletons. The results of RFPose3D and our proposed WiPose framework are plotted in the third and fourth row, respectively. The first row of this figure pictures the corresponding human postures, and the ground truth skeletons generated by our VICON system are shown in the second row. Note that the skeletons generated by the VICON system are not perfectly consistent with the postures shown in the video frames. This is caused by the deviation of the positions of the markers attached to the subject's clothes from the true locations of his skeleton joints. In practice, this problem can be fixed by adjusting the coordinates of these markers in the 3D space. In Figure 5, we color the skeleton joints with green and the body segments with grey. Also, we highlight the incorrectly predicted and distorted body segments with orange color. In this figure, columns (a), (b) and (c) correspond to three different postures of three different subjects. Columns (d)-(f) are consecutive postures captured when one subject perform the activity of walking on the spot.

The results in column (a) and (b) of Figure 5 show that the 3D skeletons constructed by our proposed WiPose framework are almost the same as the ground truths while those of the baseline have a distorted left arm in both column (a) and (b) and even a completely wrong gesture in column (b). For the posture in column (c), although both our framework and the baseline have an incorrect construction on the subject's right arm, the skeleton shape output by our framework is more realistic than that of the baseline. The results in columns (a)-(c) demonstrate that our WiPose framework has clear advantages in constructing realistic 3D skeletons with high accuracy compared with the baseline method. The reason is that our WiPose framework can encode the prior knowledge of human skeleton into the posture construction process to ensure that the estimated joints satisfy the skeletal structure of the human body. However, the baseline method independently estimates the positions of individual joints without explicitly enforcing the physical relationship among them, and thus it may generate distorted body segments due to the localization errors of individual joints.

In columns (d)-(h) of Figure 5, the subject is performing a continuous action of walking on the spot. We can see that some of the skeletons constructed by the baseline model are not correct and the baseline's synthesized body movements are not smooth, either. In contrast, our proposed framework reconstructs the activity truthfully and smoothly. This is mainly because our framework is capable of capturing the temporal relationship between consecutive postures and making their difference during estimation consistent with that of the ground truth.

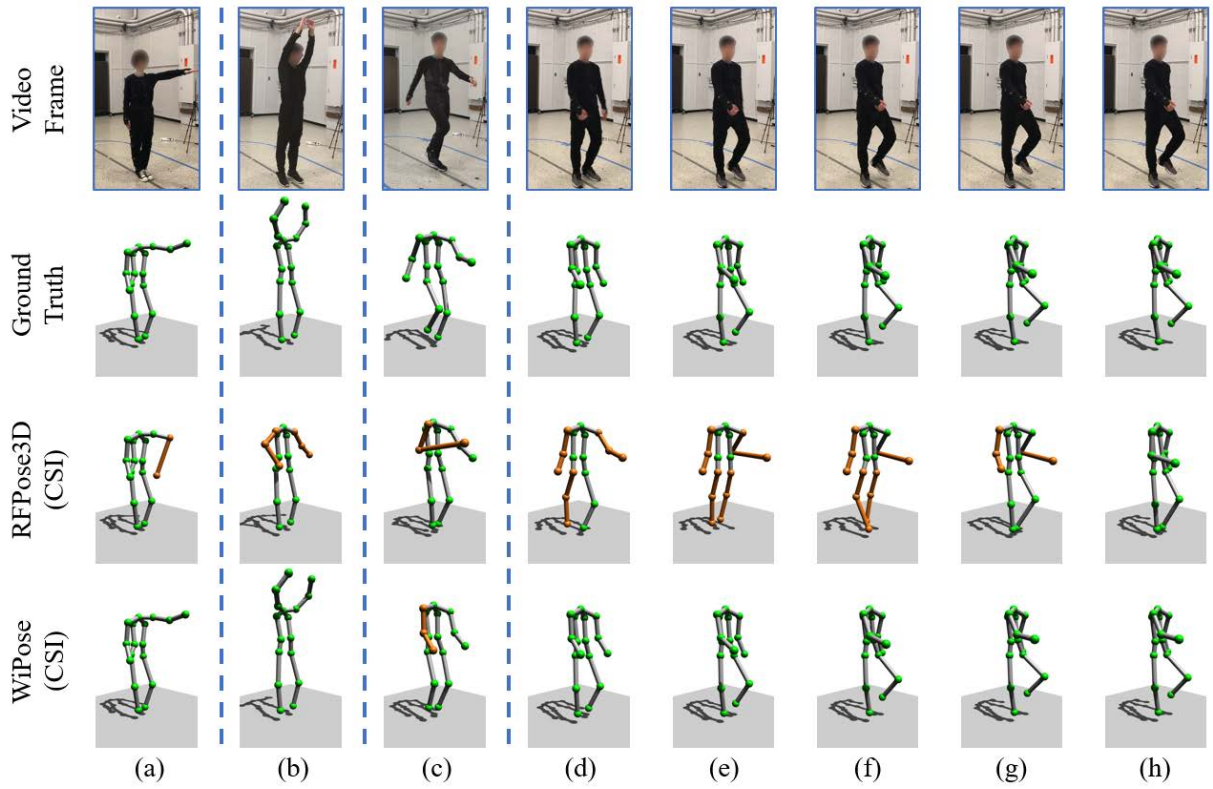


Figure 5: The examples of the constructed skeletons in the basic scenario.

Posture Construction Using 3D Velocity Profile. Besides the CSI data, we also use the 3D velocity profile (denoted as 3DVP) as the model input to evaluate the performance of our proposed framework. The results are also shown in Table 1, from which we can see that WiPose can still achieve high accuracy and perform much better than the baseline on the estimation of every single joint. These results demonstrate that our proposed deep learning model is flexible to learn useful information from different types of input data. Additionally, the results in Table 1 show that the performance of both our framework and the baseline on the velocity profile is worse than that on the CSI data. This is mainly because when the velocity profile is generated from the CSI data, though domain-specific information is removed, some information related to the body posture is also lost.

Bad Lighting Condition. We also evaluate the performance of our proposed framework under bad lighting condition, where the vision-based solutions usually have poor performance. Figure 6 shows an illustrative example. Here the model input is the denoised CSI data. The image on the left shows the posture of the subject, which is very blurry due to the bad lighting. The middle and right images are the ground truth and the skeleton constructed by WiPose, respectively. As we can see, our constructed skeleton is almost the same as the ground truth, which demonstrates the advantage of our proposed framework over the vision-based solutions.

Effect of the Training Rate. In the above experiments, we use the first 70% of the data samples as the training data. Next, we evaluate the performance of our WiPose framework on less training data. Specifically, we use the denoised CSI data as input and gradually decrease the data used for training from the first 70% to the first 40% (the amount of training samples are decreased from 112 minutes to 64 minutes). We report the results in Table 2, from which we can see the average joint localization errors of both WiPose and RFPose3D increase when the training rate decreases. However, our proposed WiPose framework still has better performance than RFPose3D in all cases, and it can reconstruct the human skeleton with an error of less than 40*mm* even when the training rate is only 40%.

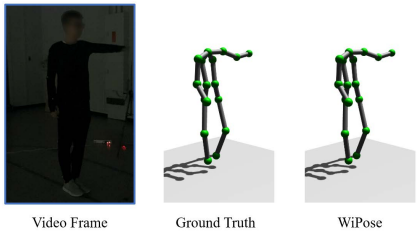
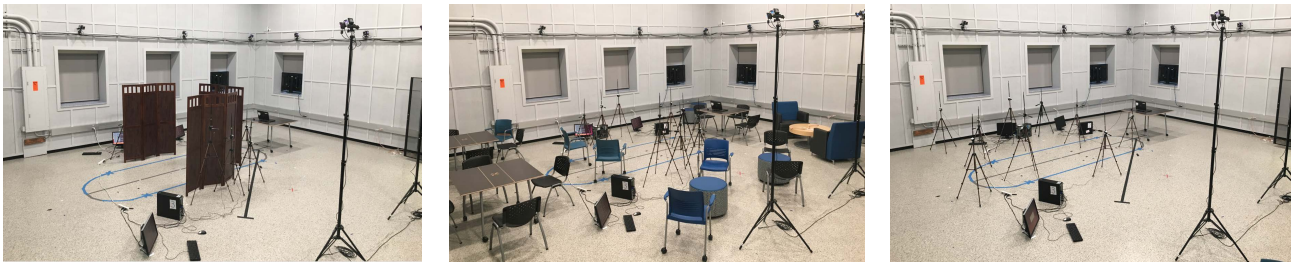


Figure 6: The example of the constructed skeleton in the basic scenario with bad lighting.

Table 2: Average joint localization errors (unit: *mm*) for different training rates.

Training Rate	40%	50%	60%	70%
RFPose3D (CSI)	57.1	52.4	48.6	43.6
WiPose (CSI)	39.1	34.6	32.1	28.3

Effect of the Packet Rate. In addition to the training rate, we also evaluate the effect of the packet rate on the performance of



(a) (b) (c)

Figure 7: Different scenarios of posture construction. (a) The occluded scenario. (b) The furnished scenario. (c) The scenario with randomly deployed antennas.

our model. In this experiment, we first collect CSI data with the transmitter sending packets at 1000 Hz, then we consider another two cases in which we down-sample the collected CSI data to 500 Hz and 250 Hz, respectively. For the two cases, the input size of our model is decreased to $9 \times 30 \times 50$ and $9 \times 30 \times 25$, respectively. We then adjust the kernel size of the CNN and feature size of the RNN accordingly to estimate the skeleton. The results in Table 3 show that our proposed WiPose framework performs much better than RFPose3D when the packet rate varies. We can also observe that the variation of the packet rate has little effect on the performance of WiPose. For example, even when the packet rate is reduced to a quarter (i.e., 250 Hz) of the full rate, the growth of the localization error is less than 5% when using our WiPose framework.

Table 3: Average joint localization errors (unit: *mm*) for different packet rates.

Packet Rate (Hz)	250	500	1000
RFPose3D (CSI)	46.6	45.2	43.6
WiPose (CSI)	29.7	29.0	28.3

Effect of the Number of Receiving Antennas. Finally, we evaluate the performance of our model when the number of receiving antennas reduces. Here we vary the number of receiving antennas from 6 to 9, and report the results in Table 4. As we can see, our proposed WiPose framework still outperforms the baseline method in all cases.

6.3 Performance Evaluation for Occluded Scenario

To investigate the effect of occlusion on the performance of our proposed WiPose framework, we place a wooden screen between the subject and each group of the receiving antennas. The occluded scenario is shown as Figure 7a. In this experiment, we collect the CSI data for 6 of the employed 10 subjects. For each subject and activity, we also use the first 70% of the data samples as the training data and the rest 30% as testing data.

Table 5 reports the average joint localization errors for the occluded scenario. As we can see, for both the CSI data and the 3D velocity profile, our proposed WiPose framework outperforms the baseline method on the estimation of each joint. Additionally, the results in Table 5 are close to those in Table 1, which means WiPose is robust and it can still have good performance even there exist occlusions when collecting the CSI data. This is mainly because the

Table 4: Average joint localization errors (unit: *mm*) for different number of receiving antennas.

Antenna Number	6	7	8	9
RFPose3D (CSI)	44.3	44.0	43.9	43.6
WiPose (CSI)	31.9	30.7	29.6	28.3

WiFi signal is capable of penetrating the wooden screen. Although the screen attenuates the signal power, most of the posture related information can still be retained in the signal. To further verify this point, we also visualize the 3D subject skeletons constructed by WiPose. Here we consider three different "antenna" views (i.e., the front, right and left views with respect of the subject, each of which aligns with the line-of-sight of a group of receiving antennas towards the subject) from which the subject is not visible due to the occlusion of the screen. Figure 8 shows two examples of the constructed skeletons when the CSI data are taken as the model input. The first row of this figure pictures the corresponding human postures. The second to the fourth rows show the ground truth skeletons and our constructed skeletons from the front, right and left views of the subject, respectively. We can see the constructed 3D skeletons based on WiPose are almost the same as the ground truths. These results further demonstrate that our proposed WiPose framework is able to construct 3D human skeletons with high precision in the occluded scenario, where the vision-based solutions usually do not work.

6.4 Performance Evaluation for Cross-domain Posture Construction

Different Settings of the Room. We first quantitatively evaluate our model on three unseen scenarios during the training process. In this experiment, we only use the CSI data collected in the basic scenario (as shown in Figure 4) to train our model. Since we have only one room that is equipped with the VICON system, the three unseen scenarios are simulated through changing the setting of the room. Specifically, we first simulate a furnished scenario, as shown in Figure 7b, in which we furnish the room with some desks and chairs that can change the transmission of the wireless signals. The second testing scenario is the occluded scenario that is described in Section 6.3. In this scenario, the screen can attenuate the power of the signals reflected from the subject. Additionally, we also consider a random antenna scenario where the receiving antennas are randomly placed, which can be seen in Figure 7c. The

Table 5: Average joint localization errors (unit: *mm*) for the occluded scenario.

Joint Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Overall
RFPose3D (CSI)	5	22	73	127	23	80	133	11	14	11	15	15	21	22	16	22	22	37.2
WiPose (CSI)	0	18	65	113	17	61	105	7	12	7	12	13	17	17	13	17	18	30.1
RFPose3D (3DVP)	28	56	142	237	53	138	226	35	37	35	37	36	41	42	37	48	46	74.9
WiPose (3DVP)	0	32	117	204	30	110	188	12	22	13	22	24	29	30	23	33	34	54.2

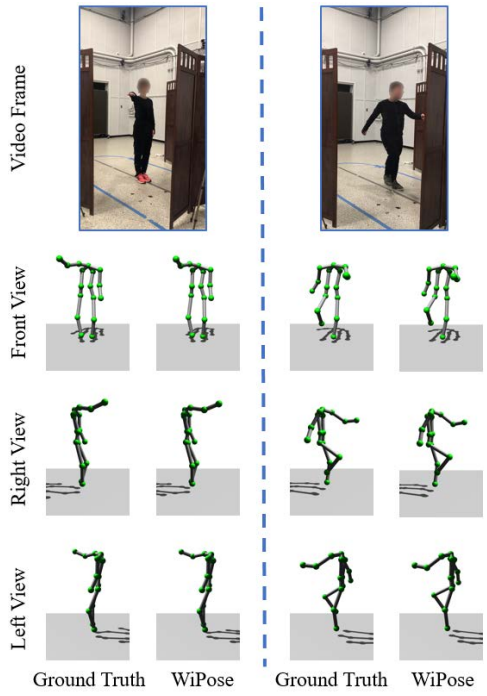


Figure 8: The examples of the constructed skeletons for the occluded scenario.

random deployment in this scenario certainly makes change to the signals collected by the receiving antennas. In all the above three scenarios, we collect the CSI data for 6 of the employed 10 subjects.

Table 6 shows the average joint localization errors for the above three scenarios. In this table, we only report the overall results that are calculated by averaging the errors for all the joints. The results show that our proposed WiPose framework performs much better than the baseline method in all cases, which demonstrates the advantage of WiPose over the baseline when testing the model in different settings of the room.

When comparing Table 6 with Table 1, we find that after the testing domain is changed from the basic scenario (i.e., the same domain) to other scenarios (i.e., different domains), the performance of our framework and the baseline on CSI data drops much more than that on the 3D velocity profile. This is mainly because the 3D velocity profile is not only able to separate the subject's movements from surrounding static objects but also independent of the deployment of antennas.

Moreover, the results in Table 6 show that our proposed WiPose framework can make better use of the cross-domain input (i.e., the 3D velocity profile) than the baseline method. After the input data

Table 6: Average joint localization errors (unit: *mm*) for posture construction in different settings of the room.

Scenarios	Input	Model	Overall
Furnished	CSI	RFPose3D	135.3
		WiPose	95.3
	3DVP	RFPose3D	119.9
		WiPose	61.2
Occluded	CSI	RFPose3D	147.3
		WiPose	121.6
	3DVP	RFPose3D	125.5
		WiPose	79.1
Random Antenna	CSI	RFPose3D	133.9
		WiPose	95.2
	3DVP	RFPose3D	126.9
		WiPose	87.7

is changed from CSI to 3D velocity profile, our framework can achieve 25.6*mm* improvement, as opposed to the marginal 8.1*mm* improvement made by the baseline. The reason is that not all the body segments are moving at any given time. The 3D velocity profile, which is designed to remove the information specific to static objects, would inevitably incur a loss in the information related to the body posture. Such loss makes it difficult for the baseline method to achieve a truthful construction of the skeleton. In contrast, our proposed model explicitly incorporates prior knowledge of skeletal structure, and thus is more tolerant of the loss of posture information caused by the velocity profile.

Different Subjects. In addition to changing the setting of the room, we also evaluate the performance of our proposed WiPose framework across different subjects in the basic scenario. Specifically, we pick up two subjects who are shorter than the others as the testing subjects, and use the others' data to train the models. This makes the testing set quite biased. We report the overall average joint localization errors in Table 7, from which we can see that the baseline method performs much worse than our framework in this biased setting. This is mainly because the baseline method needs to simultaneously estimate the skeleton structures and the movements of the testing subjects, and thus is more vulnerable to the errors in either of these two aspects. However, our proposed framework encodes the skeletal structures of the subjects in the model, so it still has good performance even when the testing subjects are unseen in the training data.

Strong Reflectors and Different Rooms. In order to demonstrate the feasibility of our model working in real practice, we further evaluate it in more challenging scenarios. We first study the effect of strong reflectors on the performance of the proposed WiPose framework. Specifically, we train the model with CSI data of 5 subjects and 9 activities (i.e., lifting left/right hand for 90/180

(a) (b) (c)

Figure 9: Different scenarios of posture construction. (a) The furnished scenario with strong reflectors. (b) The conference room. (c) The lobby.

Table 7: Average joint localization errors (unit: *mm*) for cross-subject posture construction.

Input	Model	Overall	Input	Model	Overall
CSI	RFPose3D	141.7	3DVP	RFPose3D	138.8
	WiPose	103.6		WiPose	85.8

degree, lifting both hands for 90 degree, sweeping left/right hand for 90 degree, sweeping both hands for 90 degree, walking on the spot) collected from an environment similar to the basic scenario, and then test the model in the room with two metal whiteboards near the antennas. The setting of this scenario is shown in Figure 9a. We report the results in Table 8, from which we can see the WiPose with 3DVP can achieve the best performance.

Table 8: Average joint localization errors (unit: *mm*) for the scenario with strong reflectors.

Input	Model	Overall	Input	Model	Overall
CSI	RFPose3D	130.2	3DVP	RFPose3D	114.8
	WiPose	107.4		WiPose	58.6

We also directly apply the trained model on 3DVP to two new scenarios in a different building: a conference room and the lobby, where we do not have any training data. The settings for the two scenarios are shown in Figure 9b and Figure 9c, respectively. Since it is difficult to move the VICON system to these two scenarios to provide the ground truth skeletons, here we only qualitatively analyze the performance of our proposed model. In Figure 10 and Figure 11, we show the generated skeletons of some postures and the corresponding snapshots. Based on our observation, for most of the subjects and most of the activities (e.g., Figure 10a, Figure 10b, Figure 11a, and Figure 11b), our WiPose framework can correctly reconstruct human postures from WiFi signals, even though the model is trained in a different building. This confirms the cross-environment generalization ability of our model.

There are still some cases where WiPose fail to reconstruct some body segments. For example, as shown in Figure 10c, WiPose can correctly capture the movement of the left arm but fails on the right arm. In our experiments, we also observe that some postures are more difficult to be reconstructed than others. For example, the activity of lifting left hand for 180 degree is apt to be mistaken as lifting left hand for 90 degree (Figure 11c). We speculate that it is because the transmitter and receiving antennas are placed at the

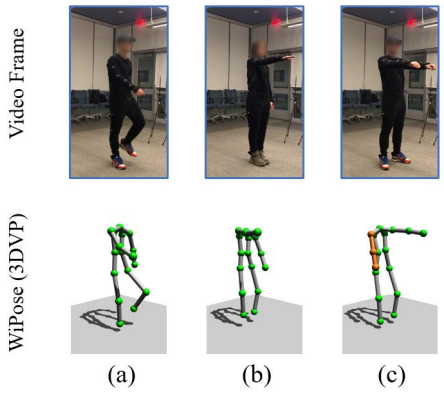


Figure 10: Posture construction in the conference room.

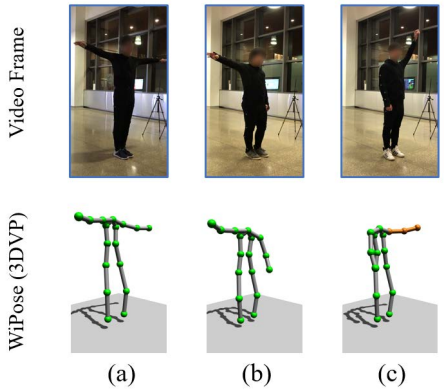


Figure 11: Posture construction in the lobby.

heights ranging from 0.74*m* to 1.28*m*, and thus the WiFi signals from overhead may be weaker than the other parts of the body.

6.5 Running Time Analysis

Here we provide a quantitative analysis of the efficiency of our proposed WiPose framework. Specifically, we measure the inference runtime of all 3 components of our model: CNN, RNN, and FK. The results on a single NVIDIA Titan XP GPU for every 1 second WiFi data is summarized in Table 9. From the table, we can see that it takes our model only 0.016*s* to reconstruct 10 frames of human posture from 1 second WiFi data. As a comparison, RFPose3D spends 0.029*s* to do the same work because it uses a much deeper neural network and many skip layers.

Table 9: The inference time of the models on a single NVIDIA Titan XP GPU for every 1 second of WiFi data.

Model	CNN	RNN	FK	Total	RFPose3D
Time (s)	0.001	0.001	0.014	0.016	0.029

7 DISCUSSION

Although our proposed WiPose framework can reconstruct human postures with high accuracy in many different scenarios, it still has some limitations.

For example, in our current design, we assume that the subject is performing activities on the spot. However, many activities in real-life involves the change of locations. To construct the postures of moving subject, we can make use of the 3D velocity profile derived from the CSI data. 3DVP captures the power distribution over absolute velocities, and thus can be utilized to track the location of the subject. After we estimate the location of the subject, we can construct the subject's skeleton at his current location. Since 3D velocity profile is robust with regard to the location change of the subject, our proposed WiPose can still truthfully construct the skeleton of the subject.

Moreover, our model currently can construct the postures of only a single subject. Actually, it can be extended to the scenes with multiple subjects through jointly estimating 3D velocity profile for each subject. Specifically, we can first estimate the number of subjects and their locations. Assume that there are n subjects at different locations, the frequency shift at each antenna is contributed by the movements of all these subjects. We use one vector of length K^3 (K is the number of possible velocities on one axis) to parameterize the vectorized 3D velocity profile of one subject, and concatenate all the vectors to form a vector of length nK^3. Using the method proposed in Section 4.2, we can jointly estimate the 3D velocity profile for each subject. A WiPose model can then be applied on each subject's velocity profile to construct his skeleton.

8 RELATED WORK

WiFi Sensing: Recently, a considerable number of investigations are conducted to make use of WiFi devices to enable various sensing tasks, including indoor localization and tracking [5, 22, 25, 33, 39, 42, 48–50, 52, 53, 65], walking speed estimation [17, 43], breathing and heartbeats inference [6, 55, 58, 61], human identification [16, 43, 54], human or human movement detection [32, 34, 47, 66], and gesture recognition [27, 31, 35, 56, 62, 63]. But all the above methods are considered as coarse-grained person perception. In contrast, our work aims to achieve a much finer grained person perception, i.e., human pose estimation.

Human Pose Estimation: In computer vision area, human pose estimation is a well-studied problem. With the development of deep learning algorithms and the emergence of abundant annotated 2D human poses datasets, lots of work [7–11, 14, 18, 29, 30, 38, 46] are focused on 2D human pose estimation. In addition to using conventional RGB cameras on the 2D pose estimation, some researchers also make efforts on estimating 3D human pose with VICON system [36] and RGB-Depth cameras [57]. Despite the great success achieved by vision-based approaches, the performance of these methods can be severely impaired by bad illumination, occlusion and blurry. Most importantly, privacy issues occur when cameras are deployed to monitor the human subjects. Recent light-based approaches [23, 24] protect user privacy through using photodiodes instead of cameras, but they still cannot work in dark or occluded scenarios. In contrast, our WiFi-based approach both avoids the privacy issue, and is immune to the lighting and occlusion conditions.

There are also LiDAR based solutions, which can achieve fine-grained person perception from 3D point clouds. For example, the systems proposed in [26, 45] can be used to detect and recognize a human subject. However, the LiDAR sensors are usually expensive and power-consuming, therefore are difficult to apply for daily and household use.

To overcome the limitations of above solutions, recent efforts are made to explore RF signal based solutions. Among them, RFCapture [4] demonstrates that by using RF signal, a coarse description of human body can be generated even through a wall. Through a teacher-student deep network, RFPose[59] is able to extract 2D human skeletons from RF signals with the supervision of visual information. The later version, RFPose3D [60], extends this framework to achieve 3D skeleton construction. Despite their impressive performance, the prohibitive requirements of these approaches in both hardware (i.e., a carefully assembled and synchronized 16 + 4 T-shaped antenna array) and RF signals (i.e., Frequency Modulated Continuous Wave with a broad signal bandwidth of 1.78 GHz) severely limit the application scope of their system. Most recently, Wang et.al. [41] propose a human pose estimation system which can construct 2D skeleton from WiFi data. However, the success of this work is highly dependent on some features, e.g., Part Affinity Fields (PAFs) [7] and Segmentation Masks(SM) [14], which are suitable for only 2D scenes, making it hard to be extended to construct 3D skeletons. Compared with the aforementioned RF signal based solutions, our work is the first one that can accurately construct fine-grained 3D human skeletons from different forms of WiFi signals.

9 CONCLUSIONS

In this paper, we investigate the possibility of using pervasive WiFi signals to image the human body like a camera. Specifically, we propose a deep learning framework, named WiPose, that can construct 3D human pose using commercial WiFi devices. The proposed framework is able to encode the prior knowledge of human skeleton into the posture construction process to ensure the realisticness of the generated postures. It is also promising to achieve cross environment generalization by taking as input a 3D velocity profile which can capture the movements of the whole 3D space, and thus separate posture-specific features from the static objects in the ambient environment. Additionally, WiPose employs a recurrent neural network (RNN) and a smooth loss to enforce smooth movements of the generated skeletons. The experimental results based on a real-world WiFi sensing testbed demonstrate that our proposed WiPose framework can construct 3D human pose with high precision.

ACKNOWLEDGMENTS

We thank our anonymous shepherd and reviewers for their insightful comments and suggestions on this paper. This work was supported in part by the US National Science Foundation under grant CNS-1652503.

REFERENCES

[1] [n.d.]. Quaternions and spatial rotation. https://en.wikipedia.org/wiki/Quaternions_and_spatial_rotation.

[2] [n.d.]. VICON Motion Systems. https://www.vicon.com.

[3] Karim Abdel-Malek and Jasbir Singh Arora. 2013. *Human Motion Simulation: Predictive Dynamics*. Academic Press.

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

[5] Fadel Adib and Dina Katabi. 2013. *See through walls with WiFi!* Vol. 43. ACM.

[6] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2015. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 837–846.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.

[8] Xianjie Chen and Alan L Yuille. 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in neural information processing systems*. 1736–1744.

[9] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. 2015. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1347–1355.

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2334–2343.

[11] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. 2014. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3582–3589.

[12] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* 28, 10 (2016), 2222–2232.

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

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Feng Hong, Xiang Wang, Yanni Yang, Yuan Zong, Yuliang Zhang, and Zhongwen Guo. 2016. 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, 47–56.

[17] Chen-Yu Hsu, Yuchen Liu, Zachary Kabelac, Rumen Hristov, Dina Katabi, and Christine Liu. 2017. Extracting gait velocity and stride length from surrounding radio signals. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2116–2126.

[18] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*. Springer, 34–50.

[19] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[20] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 289–304.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. Spotfi: Decimeter level localization using wifi. In *ACM SIGCOMM computer communication review*, Vol. 45. ACM, 269–282.

[23] Tianxing Li, Chuankai An, Zhao Tian, Andrew T Campbell, and Xia Zhou. 2015. Human sensing using visible light communication. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 331–344.

[24] Tianxing Li, Qiang Liu, and Xia Zhou. 2016. Practical human sensing in the light. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. 71–84.

[25] Xiang Li, Daqing Zhang, Qin Lv, Jie Xiong, Shengjie Li, Yue Zhang, and Hong Mei. 2017. IndoTrack: Device-free indoor human tracking with commodity Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 72.

[26] Daniel Maturana and Sebastian Scherer. 2015. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 922–928.

[27] Pedro Melgarejo, Xinyu Zhang, Parameswaran Ramanathan, and David Chu. 2014. Leveraging directional antenna capabilities for fine-grained gesture recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 541–551.

[28] Pierre Merriaux, Yohan Dupuis, Rémi Boutteau, Pascal Vasseur, and Xavier Savatier. 2017. A study of vicon system positioning performance. *Sensors* 17, 7 (2017), 1591.

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

[30] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4929–4937.

[31] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. ACM, 27–38.

[32] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, Fugui He, and Tianzhang Xing. 2018. Enabling contactless detection of moving humans with dynamic speeds using CSI. *ACM Transactions on Embedded Computing Systems (TECS)* 17, 2 (2018), 52.

[33] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, and Kyle Jamieson. 2017. Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 6.

[34] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, and Zimu Zhou. 2014. PADS: Passive detection of moving targets with dynamic speed using PHY layer information. In *2014 20th IEEE international conference on parallel and distributed systems (ICPADS)*. IEEE, 1–8.

[35] Kun Qian, Chenshu Wu, Zimu Zhou, Yue Zheng, Zheng Yang, and Yunhao Liu. 2017. Inferring motion direction using commodity wi-fi for interactive exergames. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1961–1972.

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

[37] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[38] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*. 1799–1807.

[39] Deepak Vasisht, Swarun Kumar, and Dina Katabi. 2016. Decimeter-level localization with a single WiFi access point. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*. 165–178.

[40] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. 2018. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8639–8648.

[41] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained Person Perception using WiFi. In *Proceedings of the IEEE International Conference on Computer Vision*.

[42] Ju Wang, Hongbo Jiang, Jie Xiong, Kyle Jamieson, Xiaojiang Chen, Dingyi Fang, and Binbin Xie. 2016. LiFS: low human-effort, device-free localization with fine-grained subcarrier information. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 243–256.

[43] Wei Wang, Alex X Liu, and Muhammad Shahzad. 2016. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 363–373.

[44] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*. ACM, 65–76.

[45] Zhe Wang, Yang Liu, Qinghai Liao, Haoyang Ye, Ming Liu, and Lujia Wang. 2018. Characterization of a RS-LiDAR for 3D Perception. In *2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE, 564–569.

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

[47] Chenshu Wu, Zheng Yang, Zimu Zhou, Xuefeng Liu, Yunhao Liu, and Jiannong Cao. 2015. Non-invasive detection of moving and stationary human with wifi. *IEEE Journal on Selected Areas in Communications* 33, 11 (2015), 2329–2342.

[48] Dan Wu, Daqing Zhang, Chenren Xu, Yasha Wang, and Hao Wang. 2016. WiDir: walking direction estimation using wireless signals. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 351–362.

[49] Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. 2019. mD-Track: Leveraging Multi-Dimensionality for Passive Indoor Wi-Fi Tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*. ACM, 1–16.

[50] Jie Xiong and Kyle Jamieson. 2013. Arraytrack: A fine-grained indoor location system. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*. 71–84.

[51] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015).

[52] Zheng Yang, Zimu Zhou, and Yunhao Liu. 2013. From RSSI to CSI: Indoor localization via channel response. *ACM Computing Surveys (CSUR)* 46, 2 (2013), 25.

[53] Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a mobile device into a mouse in the air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 15–29.

[54] Yunze Zeng, Parth H Pathak, and Prasant Mohapatra. 2016. WiWho: wifi-based person identification in smart spaces. In *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*. IEEE Press, 4.

[55] Fusang Zhang, Daqing Zhang, Jie Xiong, Hao Wang, Kai Niu, Beihong Jin, and Yuxiang Wang. 2018. From fresnel diffraction model to fine-grained human respiration sensing with commodity wi-fi devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 53.

[56] Ouyang Zhang and Kannan Srinivasan. 2016. Mudra: User-friendly Fine-grained Gesture Recognition using WiFi Signals. In *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*. ACM, 83–96.

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

[58] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 95–108.

[59] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.

[60] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. ACM, 267–281.

[61] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*. 4100–4109.

[62] Xiaolong Zheng, Jiliang Wang, Longfei Shangguan, Zimu Zhou, and Yunhao Liu. 2016. Smokey: Ubiquitous smoking detection with commercial wifi infrastructures. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 1–9.

[63] Xiaolong Zheng, Jiliang Wang, Longfei Shangguan, Zimu Zhou, and Yunhao Liu. 2017. Design and implementation of a CSI-based ubiquitous smoking detection system. *IEEE/ACM Transactions on Networking* 25, 6 (2017), 3781–3793.

[64] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 313–325.

[65] Rui Zhou, Xiang Lu, Pengbiao Zhao, and Jiesong Chen. 2017. Device-free presence detection and localization with SVM and CSI fingerprinting. *IEEE Sensors Journal* 17, 23 (2017), 7990–7999.

[66] Zimu Zhou, Zheng Yang, Chenshu Wu, Longfei Shangguan, and Yunhao Liu. 2013. Omnidirectional coverage for device-free passive human detection. *IEEE Transactions on Parallel and Distributed Systems* 25, 7 (2013), 1819–1829.