mPose: Environment- and Subject-Agnostic 3D Skeleton Posture Reconstruction Leveraging a Single mmWave Device

Cong Shi*, Li Lu†, Jian Liu‡, Yan Wang§, Yingying Chen*, Jiadi Yu†
*WINLAB, Rutgers University, New Brunswick, NJ, USA
†Zhejiang University, Zhejiang, China
‡University of Tennessee, Knoxville, TN, USA
§Temple University, Philadelphia, PA, USA
¶Shanghai Jiao Tong University, Shanghai, China

***Abstract*—Human skeleton posture reconstruction is an essential component for human-computer interactions (HCI) in various application domains. Traditional approaches usually rely on either cameras or on-body sensors, which induce privacy concerns or inconvenient practical setups. To address these practical concerns, this paper proposes a low-cost contactless skeleton posture reconstruction system, *mPose*, which can reconstruct a user's 3D skeleton postures using a single mmWave device. *mPose* does not require the user to wear any sensors and can enable a broad range of emerging mobile applications (e.g., VR gaming and pervasive user input) via mmWave-5G ready Internet of Things (IoT) devices. Particularly, the system extracts multi-dimensional spatial information from mmWave signals which characterizes the skeleton postures in a 3D space. To mitigate the impacts of environmental changes, *mPose* dynamically detects the user location and extracts spatial features from the mmWave signals reflected only from the user. Furthermore, we develop a deep regression method with a domain discriminator to learn a mapping between the spatial features and the joint coordinates of human body while removing subject-specific characteristics, realizing robust posture reconstruction across users. Extensive experiments, involving 17 representative body postures, 7 subjects, and 3 indoor environments, show that *mPose* outperforms contemporary state-of-the-art RF-based solutions with a lower average joint error of only ∼30mm, while achieving transferability across environments and subjects at the same time.**

***Index Terms*—3D Skeleton Posture Reconstruction, mmWave**

I. Introduction

Recent years have witnessed an upsurge of research in human movement tracking and recognition systems. Theses systems significantly facilitate human-computer interactions (HCI) and many emerging applications, such as virtual reality (VR), augmented reality (AR), fitness tracking, smart healthcare, and smart home control. Traditional approaches utilize vision-based techniques [1]–[3] or body attached sensors [4]–[6] to reconstruct skeleton postures and infer body movements. However, these approaches may suffer from illumination interference, incur privacy concerns, or introduce intrusive user experience. Therefore, a robust, privacy-preserving, and non-intrusive skeleton posture reconstruction method is highly demanded.

Non-intrusive approaches using radio frequency (RF) signals have been exploited to detect coarse-grained activities [7], monitor vital signs [8], and reconstruct body postures [9]–[11]. Among these studies, E-eyes [7] is a pioneer work to recognize indoor activities leveraging commodity WiFi devices (e.g., WiFi access points), while Liu *et al.* [8] first propose leveraging commodity WiFi to monitor vital signals including breathing rate and heart rate. WiTrack [9] can provide coarse tracking of body parts using customized antennas; RF-Pose [10] has been developed to estimate 2D human poses using multiple universal software radio peripheral (USRP) units mounted on the walls. However, these systems require multiple dedicated devices (e.g., customized antennas, USRP devices), which are prohibitively expensive and hard to deploy in practice. WiPose [11] demonstrates the feasibility of reconstructing 3D human posture using WiFi signals and deep learning, but it requires to deploy multiple antennas around the user, which has limited application scenarios.

As approaching the 5G era, low-cost Commercial Off-The-Shelf (COTS) mmWave hardware has been integrated into mobile devices (e.g., Soli on Google Pixel4 [12], [13]). Compared to WiFi, mmWave has a much shorter wavelength (i.e., about 1/16 of 5GHz WiFi wavelength), which renders mm-level wireless ranging and enables high-precision skeleton reconstruction beyond traditional RF-based methods. Additionally, mmWave devices are usually equipped with miniature antenna arrays for highly directional transmissions, and such arrays can be used to derive the angles of arriving signals, which facilitates fine-grained spatial sensing. As such, researchers have utilized mmWave to recognize hand gestures [13], vital signs [14], and human identity [15]. Sengupta *et al.* [16] show the initial success of using mmWave radar to detect and track human skeletal postures. However, the proposed solution needs multiple mmWave devices and does not guarantee the generalizability to different environments (e.g., room layouts, furniture placement) and users (e.g., with different heights, lengths of arms and legs).

In this paper, we propose a low-cost contactless skeleton posture reconstruction system, *mPose*, which tracks a user's

full-body 3D skeleton postures leveraging mmWave signals. Different from existing work, *mPose* can achieve fine-grained skeleton reconstruction with a single COTS mmWave device. By utilizing advanced signal processing and deep learning technologies, our system can dynamically remove posture-irrelevant information (e.g., reflections from static room objects and the user's location) and accurately track skeletal joints (e.g., wrists, elbows, knees, and ankles) in a 3D space across different domains (i.e., with different users or in different environments.) *mPose* can be conveniently deployed in a 5G or 802.11ad-enabled IoT device (e.g., a smartphone or a smart TV) without additional costs. It provides privacy-preserving tracking of fine-grained human poses, facilitating various applications, such as contactless smart appliance control, augmented reality and virtual reality games, fitness tracking, wellbeing monitoring, and smart city surveillance.

Realizing such a skeleton posture reconstruction system has a number of challenges in practice. First, it is challenging to reconstruct 3D full-body skeleton posture using a single device. Though mmWave signal can provide fine-grained information, existing approaches usually require at least two devices to achieve it [16]. Second, skeleton reconstruction usually occurs at different locations in a room or in different rooms with distinct layouts and furniture placement. Such heterogeneous environmental factors may cause different reflections that could significantly impact the skeleton reconstruction performance. Third, it is essential to have a general skeleton reconstruction model so that individual users do not need to suffer tedious training processes. However, due to differences in individual body shapes, it is difficult to train such a model to achieve high skeleton reconstruction accuracy for different users.

To address these challenges, we propose to extract fine-grained spatial information from mmWave signals reflected by human bodies to perform spatial tracking of the users' body parts. Specifically, Frequency-Modulated Continuous Wave (FMCW) [17] is used to obtain the range of the user's body parts. Capon beamforming is used with an antenna array to derive angles of arriving signals, which are integrated with the range information to enable spatial tracking of human body parts.

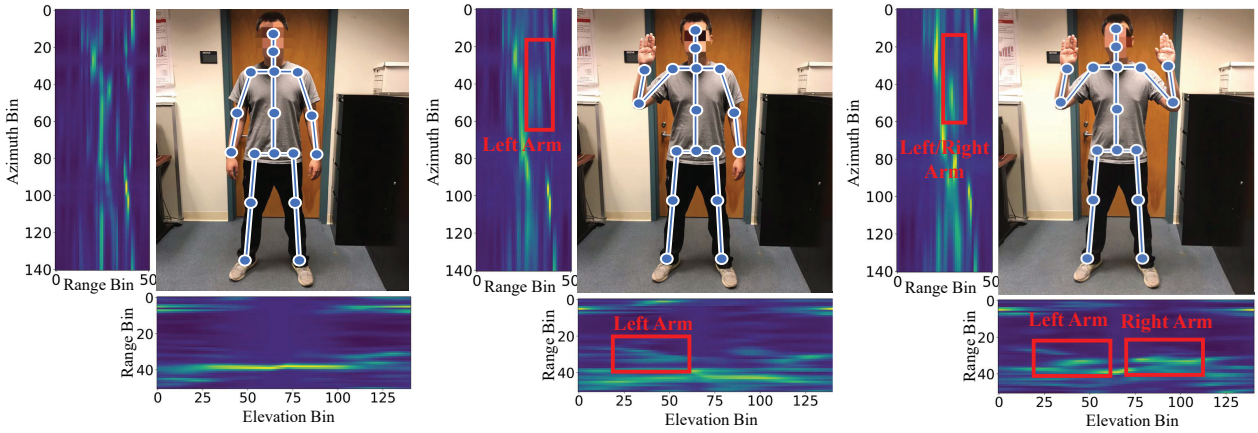
To enhance the resolution of spatial tracking, we exploit the virtual antenna technology [18] to develop a virtual antenna array with a much larger size than the physical antenna array. In addition, we dynamically estimate the user's location and extract the spatial features from the signals reflected from the user in a 3D contour, which mitigates the impacts from different environmental setups and user locations. A three-layer convolutional neural network (CNN) is designed to model human skeleton postures based on the extracted spatial features. Furthermore, *mPose* employs domain adversarial learning [19] to remove user-specific characteristics embedded in the spatial features and ensure reliable skeleton reconstruction across users. The major contributions of this work are summarized as follows:

- We show that it is possible to accurately reconstruct 3D full-body skeleton postures by using a single COTS mmWave device. We develop a low-cost 3D skeleton posture reconstruction system, *mPose*, which can precisely localize skeletal joints in a 3D space.
- We develop a target detection method that allows our system to extract spatial features from the signals reflected only by the user and mitigate the impacts from the changes of environments and user locations.
- We design a domain discriminator that removes user-specific characteristics embedded in the mmWave signals so as to achieve robust skeleton reconstruction across users with reasonable training efforts.
- We conduct extensive experiments involving 17 body postures, 7 subjects, and 3 different environments to evaluate the performance of *mPose*. The results demonstrate that our system can achieve a low average joint estimation error of 30mm while achieving environment- and user-independency.

II. Related Work

Exiting skeleton posture reconstruction studies usually rely on either attached sensors or cameras [2], [4], [20]–[23], [23]–[28]. For instance, Quwaider *et al.* [20] exploit inertial sensors attached to the human body and Hidden Markov Model (HMM) to reconstruct body postures. Shen *et al.* [23] develop a system that can track 3D postures of an entire arm (i.e., both wrist and elbow) using motion and magnetic sensors on smartwatches. While these approaches are intuitive, they all require users to wear additional sensors, which is not always feasible in practice. There are also a number of studies that utilize cameras for 2D/3D skeleton posture reconstruction [2], [22]. For example, Tompson *et al.* [22] design a spatial-model based on a deep convolutional neural network to map human postures in a video frame into 2D joint positions. By utilizing a camera array, Mehrizi *et al.* [2] demonstrate the potential to recover 3D human postures through analyzing image features of joint kinematics. However, these vision-based approaches usually incur privacy concerns, and they require line-of-sight and sufficient illumination, which largely limits their application scenarios.

To overcome the weaknesses of body-attached sensor- and vision-based solutions, researchers have been exploring WiFi-based approaches to recognize human activities [7], [29], monitor vital signs [8], or reconstruct human body postures [9]–[11], [30]–[32]. For example, E-eyes [7] is one of the first studies to recognize various daily activities (e.g., cooking, eating, and watching TV) using commodity WiFi devices. Liu *et al.* [8] developed the first non-intrusive vital sign tracking system that can monitor users' breathing and heart rates during sleep using WiFi signals. To reconstruct body postures, Zhao *et al.* [10] develop a system that uses customized antenna arrays and FMCW signals to estimate 2D postures. In addition, Adib *et al.* [9] propose to track the human arm motions leveraging FMCW radio operating on a universal software radio peripheral (USRP) device. However, these systems require dedicated RF settings (e.g., USRP devices and antenna arrays), which are costly and require to



(a) Standing (b) Lifting right arm (c) Lifting both arms

Fig. 1: mmWave device and the received signals impacted by different postures.

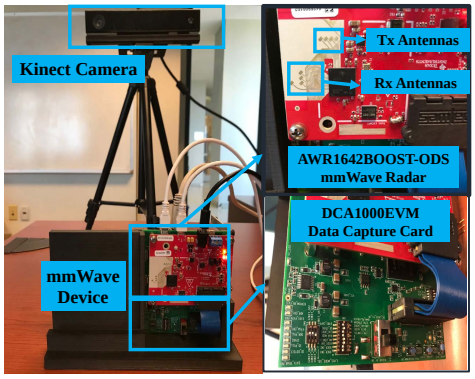


Fig. 2: Experimental Setup.

be installed in controlled environments. A recent work shows that channel state information (CSI) of WiFi signals could be used to reconstruct 3D human postures [11], but the system requires a large number of antennas distributed around the human body, making it hard to deploy in real-world scenarios. Additionally, due to the open-propagation nature, WiFi signals are very sensitive to various environmental factors (e.g., the room layout and moving objects), which inevitably cause reconstruction errors.

Compared to existing work, $mPose$ leverages highly directional mmWave signals to track skeletal joints and thus allows more reliable posture construction. The shorter wavelengths of mmWave signals also enable posture tracking with higher precision. A more recent study [16] shows the initial success of using mmWave signals to track human skeleton postures, but it requires two distributed mmWave devices to achieve 3D posture tracking and the performance of the system is moderate. In contrast, *mPose* could achieve high accuracy in reconstructing 3D postures by using a single mmWave transceiver. In addition, our system dynamically extracts spatial features from the target user to remove the impacts of user location and can be adapted to reconstruct postures across users with a reasonable amount of training efforts.

III. Preliminaries

In this section, we first introduce the fundamentals of mmWave sensing. Then, we investigate the feasibility of using mmWave signals for skeleton posture reconstruction. We also present some potential applications that can be benefited from mPose.

A. Fundamentals of mmWave Sensing

Recently, mmWave-based sensing techniques have been widely investigated in academia and industry. Due to the mmWave signals' high frequency (i.e., $> 30GHz$) and short wavelength (i.e., $< 10mm$), mmWave techniques can realize high-resolution measurements for fine-grained sensing. Moreover, mmWave transceivers can be implemented in a compact form that is small enough to fit into mobile devices (e.g., Google Pixel series [12]). As such, we propose to employ a single COTS mmWave device (i.e., TI AWR1642BOOST-ODS mmWave device), which is low-cost, portable, and highly-integrated, for 3D human posture reconstruction in this work. The mmWave device supports multiple-input and multiple-output (MIMO) with an antenna array (e.g., 2×4 antennas as shown in Figure 2) that can provide multiple dimensions of temporal and spatial information with one single measurement.

To reconstruct a user's skeleton posture, *mPose* needs to estimate the range (distance) and angle of different human body parts to the mmWave device. Particularly, our system utilizes the mmWave device to transmit a continuous chirp signal sweeping across a bandwidth within a fixed duration. Upon receiving the reflected mmWave signals, the mmWave device performs a dechirp operation on both the transmitted and the received signals to derive an intermediate frequency (IF) signal, which can be formulated as:

$$IF = sin[(f_{Tx} - f_{Rx})t + (\phi_{Tx} - \phi_{Rx})], \quad (1)$$

where f_{Tx} and f_{Rx} represent the instantaneous frequencies of the transmitted and received mmWave signals, respectively. While ϕ_{Tx} and ϕ_{Rx} denote the instantaneous phases of the two signals respectively. By integrating IF signals from multiple antennas of the mmWave device, *mPose* can extract spatial

information (i.e., range and angle) on a user's skeletal posture. We introduce how to use the IF signal to derive spatial features in Section V.

B. Feasibility of Reconstructing Skeleton Postures via a Single mmWave Device

To demonstrate the feasibility of utilizing mmWave signals to reconstruct 3D human postures, we conduct a preliminary experiment by examining the spatial information derived from a single portable mmWave device when a volunteer is performing a set of postures (i.e., standing, waving right arm, and waving both arms). Specifically, we use TI AWR1642BOOST-ODS shown in Figure 1(a) as the mmWave transceiver. A Microsoft Kinect [33] adjacent to the mmWave device is used to record the video as the ground truth 3D skeleton joint positions. Figure 1 (b) (c), and (d) show the captured video frames and the corresponding range-azimuth and range-elevation heatmaps of the three postures. The heatmap value represents the frequency response of mmWave at different angles (i.e., azimuths and elevations) and ranges (i.e., distances), with higher values representing stronger reflections. Across the three postures, we can find the contour of torso and limb across the range bin $25 \sim 40$ (i.e., $1.1m \sim 1.8m$), which is consistent with our ground truth. Furthermore, as the volunteer lifts the right arm as shown in Figure 1(b), intensity changes at the range bin 25 and azimuth bin 30 can be observed. Interestingly, in Figure 1(c), we can observe symmetric intensity changes in the range-elevation heatmap, representing the posture of lifting both arms. The preliminary study shows the feasibility of using the mmWave signals from a single device to capture the spatial information of different human body parts.

C. Potential Applications

mPose can reveal human full body skeleton postures in a fine-grained manner, which can be utilized to support an abroad array of applications.

Alerting acts of violence based on human posture: some existing studies demonstrated the potential of interpreting human activities based on the posture [25], [34]. As public safety raises increasing concern recently, inferring the activities of violence (e.g., sexual violence, physical fighting) by embedding mPose into public infrastructures (e.g., street lamps) is highly desirable. Upon detecting the violence, mPose can alert the perpetrator and contact the police to prevent injury, psychological harm, or even death. Different from vision-based approaches, our system will not disclose the privacy of human subjects since the personally identifiable information (e.g., facial information) is not embedded in mmWave signals. Furthermore, our mmWave-based system will not suffer from performance degradation due to complicated ambient lighting conditions.

Full-body AR/VR gaming: the unique immersive experience provided by AR/VR gaming has attracted millions of users around the world. The AR/VR games create a 3D virtual world and allow users to use hand-held controllers to interact with 3D objects in the virtual world [35], [36]. To improve user experience, mPose can facilitate AR/VR gaming by enabling a controller-free paradigm. The miniature mmWave sensors can be easily embedded into the AR/VR devices, such as AR/VR headsets, game consoles, or even smartphones.

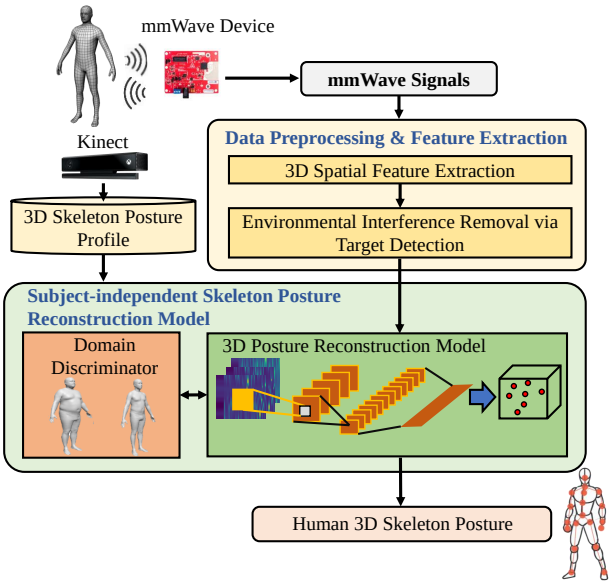


Fig. 3: System flow of *mPose*.

IV. Overview of mPose

A. Challenges

3D Skeleton Posture Reconstruction Using a Single mmWave Device. Compared to the existing solutions that use multiple devices or spatial-distributed antenna arrays, our single-device approach is challenging as we only have one device equipped with an on-board antenna array for sensing, which provides limited spatial diversity and a lower sensing resolution. Thus, it is crucial to extract in-depth features that can capture the unique impacts of different postures from mmWave signals.

Location and Environment Changes. In real-world scenarios, people may perform postures at different locations of a room or even in different rooms. Although mmWave signals reflected from the human body can be used to infer their relative distance and angle to the mmWave device, such estimation is prone to the varying location (e.g., causing different distances and angles between the subject and the mmWave device). Furthermore, the mmWave signals reflected from the objects in the room (e.g., furniture and home appliances) or walls can introduce interference to the received mmWave signals, leading to joint position estimation error. Our system needs to derive robust feature representations for accurate posture reconstruction.

Body Characteristic Differences Across Subjects. The mmWave signals also carry substantial information specific to individual subjects (e.g., body characteristics such as heights, length of arm and leg). Such body characteristics variations make it difficult to train a general skeleton reconstruction model applicable to different subjects, since the mapping

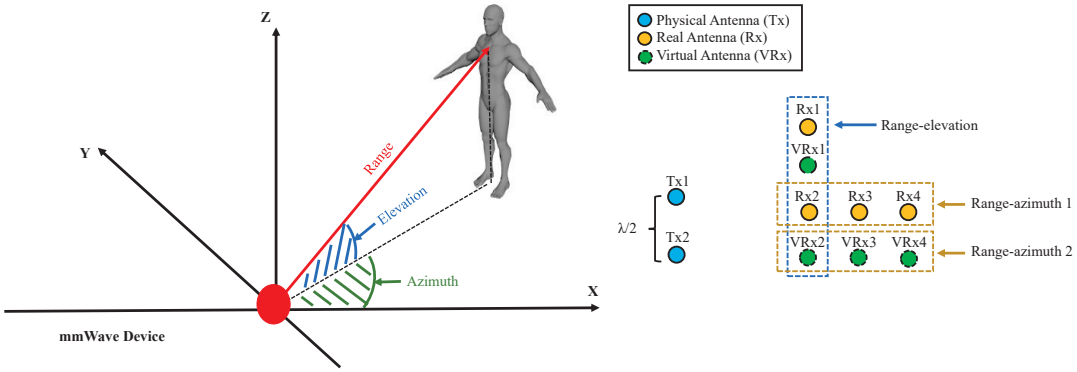


Fig. 4: Illustration of the antenna arrays used in *mPose*.

between the spatial features and the joint positions will not be consistent. To reduce training efforts during practical deployment, it is highly desired to enable accurate posture reconstruction across users.

B. System Architecture

In this work, we address the above challenges and develop a 3D skeleton posture reconstruction system, *mPose*, which can continuously track a user's joints using a single COTS mmWave device. *mPose* is designed to examine the mmWave signals to determine the spatial positions of joints for posture reconstruction. The system flow is illustrated in Figure 3. Specifically, in our system, the *Data Preprocessing* & *Feature Extraction* module first demodulates the reflected mmWave signals to obtain the ranges of various reflectors (e.g., the user's body parts, room objects), and then integrates the range information across antennas on the mmWave device to derive the angle information (i.e., azimuth and elevation). A virtual antenna array is constructed to enhance the resolution of the derived angle information, which enables fine-grained skeleton posture tracking on a single mmWave device. To ensure the robustness to user location and environment changes, the system dynamically tracks the user's location and extracts spatial features from the mmWave signals reflected from the user, which removes the impacts of the user's locations (i.e., the relative distances and angles between the user and the mmWave device) and environmental factors (e.g., reflections from walls and room objects).

After obtaining the spatial features, we develop a *Subject-independent Skeleton Posture Reconstruction Model* to map the spatial features to 3D coordinates of a user's skeletal joints (e.g., wrists, elbows, knees, and ankles) while removing user-specific body characteristics embedded in the spatial features. Instead of directly inferring the joint positions based on the spatial features, we resort to a deep-learning-based approach to learn a more reliable mapping between the joint positions and the spatial features. Particularly, we use the joint positions captured with Kinect [33] to train a deep regression model, which is referred to as a 3D posture model. The 3D posture model consists of a representation extractor that derives skeleton representations from spatial features and a 3D posture estimator that predicts 3D coordinates of 17 skeletal joints. In addition, *mPose* also employs a domain discriminator to remove subject-specific characteristics in the spatial features via domain adversarial training and adjust the 3D posture model to enable general skeleton reconstruction across users.

V. Design of mPose

A. mmWave Signal Design

We need to design the FMCW signal (i.e., a chirp signal) in a way that maximums the sensing power of *mPose*. In other words, we want to simultaneously achieve high range and angle resolutions for fine-grained skeleton reconstruction. The range resolution d_{res} of the FMCW signal is defined as [37]:

$$d_{res} = \frac{c}{2 \times B}, \tag{2}$$

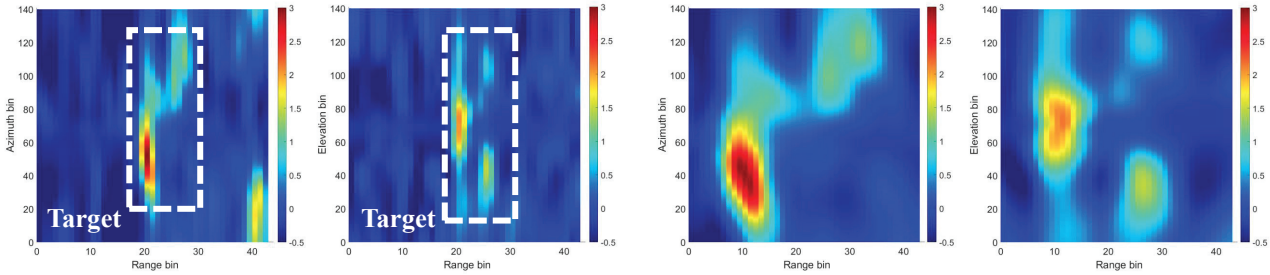
where c is the speed of light; B is the bandwidth of a mmWave chirp signal. According to the formulation, the range estimation resolution is determined by the bandwidth (i.e., B). Hence, the bandwidth of the mmWave chirp signal should be set as large as possible. Since the frequency band used by the mmWave device lies in the range of $77 \sim 81GHz$, the chirp bandwidth of *mPose* is selected as $81-77 = 4GHz$, with a range estimation resolution of $3.75cm$. Additionally, we use a short chirp duration $\tau = 33\mu s$ to maintain the capability of *mPose* on continuous sensing. The angle resolution of mmWave signals can be formulated as:

$$\theta_{res} = \frac{\lambda}{N \times d}, \tag{3}$$

where λ denotes the wavelength; N and d are the number of receiving antennas and the distance between two neighboring antennas in the antenna array, respectively. Since the mmWave device used in *mPose* is equipped an antenna array with $N = 4$ and $d = \frac{\lambda}{2}$, the angle resolution is $0.5°$. To enable fine-grained posture reconstruction, we synthesize a virtual antenna array to double the angle resolution, with the details elaborated in Section V-B.

B. 3D Spatial Feature Extraction

To reconstruct the skeleton posture from mmWave signals, *mPose* first estimates the range (i.e., distance) between each of the human body parts and the mmWave device. The principle of extracting such range information is to measure the frequencies of IF signals, which are linearly correlated



(a) Original heatmaps (b) Denoised heatmaps

Fig. 5: Illustration of the environmental interference removal through detecting the target user in a 3D contour.

with the distances to reflectors as mentioned in Section III-A. Specifically, we apply range-FFT on the IF signal to obtain the frequencies associated with strong reflections (i.e., f_{IF}), such as those from the subject's body parts and room objects, and then map the frequencies to distances:

$$d = \frac{f_{IF} \times c \times \tau}{2B}, \tag{4}$$

where c denotes the speed of light; B and τ are the bandwidth and the propagation time of the chirp signal, respectively. To reconstruct the user's posture in a 3D space, we further estimate the phase shifts (i.e., ω) associated with the Angle-of-Arrival (AoA) of the reflectors. Particularly, we apply angle-FFT upon the range information derived from all receiving antennas to obtain ω, which can be mapped to angles (e.g., azimuths or elevations):

$$\theta = sin^{-1}(\frac{\lambda \times \omega}{2\pi \times \Delta d}), \tag{5}$$

where Δd denotes the distance between two receiving antennas and λ represents the wavelength. The angle and range information together serve as the basis of posture reconstruction in *mPose*.

To enhance the resolution of angle estimation, *mPose* exploits a 1×8 antenna array with the diagram shown in Figure 4. Besides 4 physical antennas (i.e., yellow circles), *mPose* also simulates 4 virtual antennas (i.e., green circles), which doubles the angle resolution of the antenna array. Particularly, our system pairs $Tx1$ with $Rx1 \sim Rx4$ to form a 1×4 physical array. In addition, by pairing the other physical transmitting antenna (i.e., $Tx2$) with the same 4 physical receiving antennas, we can synthesize a 1×4 virtual array. The virtual array can operate together with the physical one by using time-division multiplexing, which alternatively transmits mmWave chirps signals using the physical transmitting antennas (i.e., $Tx1$ and $Tx2$) and receives the signals using the 4 physical receiving antennas (i.e., $Rx1 \sim Rx4$). Given the 8 receiving antennas, our system can achieve an angle resolution of $0.25°$.

Furthermore, to track skeletal joints in a Cartesian coordinate system, the antennas are organized into orthogonal subarrays, with two subarrays placed horizontally to capture range-azimuth and one subarray put vertically to derive range-elevation. Figure 4 illustrates such a spatial relationship (i.e., range, azimuth, and elevation) between the mmWave device and the torso of a subject (i.e., performing a standing posture). By sequentially performing range-FFT and angle-FFT on mmWave signals received with an antenna array, we derive a 2D heatmap that captures the frequency response of signals reflected from different angles at different distances:

$$F(d, \theta) = AngleFFT(RangeFFT(IF, d), \theta), \tag{6}$$

where $RangeFFT(\cdot, d)$ and $AngleFFT(\cdot, \theta)$ represent the FFT operations at range d and angle θ, respectively. We then combine multiple such 2D heat-maps (i.e., 2 range-azimuth heatmaps and 1 range-elevation heatmap) to derive 3D spatial features for posture reconstruction.

C. Environmental Interference Removal via Target Detection

Since the mmWave signals capture the spatial information of all reflectors within the field of view, the static objects in a room (e.g., furniture, walls) can introduce a significant amount of posture-irrelevant noises which interfere with our posture reconstruction system. In addition, the variations of the subject's location can alter the relative distance and angle between the subject and the mmWave device, introducing uncertainty to the derived spatial information of user posture. To mitigate these two types of interferences, *mPose* dynamically detects the 3D location of the target user's torso and creates a 3D contour to segment the spatial information (i.e., range-azimuth and range-elevation heatmaps), which can be used to remove the impacts of location and environment factors.

Particularly, *mPose* tracks the torso' 3D location through examining the reflection energy in the range and angle bins, since the torso normally leads to the strongest energy due to its larger reflection area compared to arms/legs. We denote the range, azimuth, and elevation of the torso as (r, θ, ψ). After obtaining the torso's location, *mPose* removes the location impacts by extracting the spatial information across ranges: $[r - \Delta r, r + \Delta r]$, where Δr is determined by the maximum arm length of the user. To remove the environmental impacts, we dynamically calculate the angles where the human body resides based on r to further segment the spatial information. Specifically, we extract the heatmaps across azimuth bins: $[\theta - \Delta\theta, \theta + \Delta\theta]$. The variable $\Delta\theta$ is calculated through:

$$\Delta\theta = arctan^{-1}\frac{l_{arm}}{2r}, \tag{7}$$

where l_{arm} is the user's arm span. Similarly, we calculate the elevation variable $\Delta\psi$ based on the user's height and extract the spatial information across the derived elevation

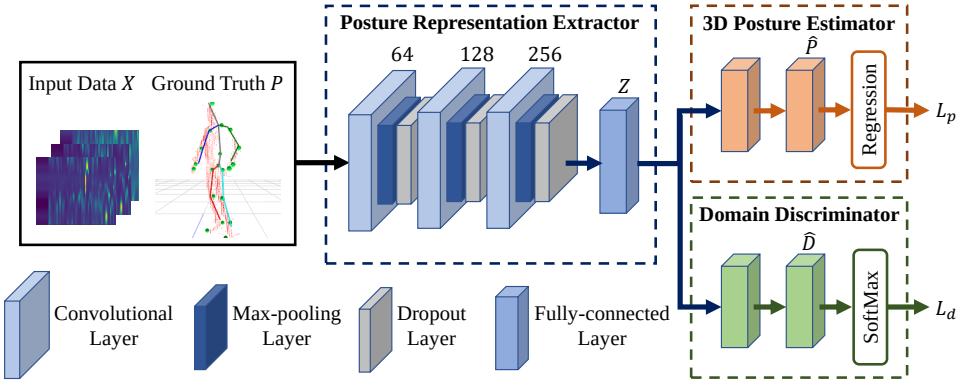


Fig. 6: Deep learning architecture used in *mPose* for subject-independent posture reconstruction.

bins. Since the varying distance between the subject and the mmWave device could change the size of the extracted range-azimuth and range-elevation heatmaps, we interpolate the heatmaps to a fixed dimension to mitigate the impacts of such variations, which ensures consistent heatmap sizes under the distance variations. With the approach, we can eliminate the interferences from the location and the environment in the heatmaps. Figure 5 shows a heatmap example before and after applying the proposed approach. We can find that compared with Figure 5 (a), the noises are significantly mitigated in Figure 5 (b), which confirms the effectiveness of our approach.

D. Subject-independent 3D Skeleton Posture Reconstruction

Model Overview. In order to construct 3D human posture, we build a deep learning model to map the extracted range-azimuth and range-elevation heatmaps to the coordinates of joints in the 3D space. Figure 6 shows the architecture of the proposed deep learning model. The deep learning model takes data X (i.e., extracted spatial information) and the 3D posture ground truth P derived by Microsoft Kinect as input. The ground truth includes the 3D positions of 17 skeletal joints (e.g., wrists, elbows, knees) [33]. The input data are first converted into a set of low-rank posture representations Z using the posture representation extractor implemented by a 3-layer CNN model. Based on the posture representations, a 3D posture estimator can predict the 3D joint coordinates (i.e., $\hat{P}$). In addition, a domain discriminator that can predict the subject label D is used to assist in training the posture representation extractor and posture estimator. By optimizing the posture representation extractor with the domain discriminator to achieve indistinguishable subject labels (i.e., maximizing the domain loss), our model can remove the subject-specific characteristics for general skeleton reconstruction across users.

3D Posture Reconstruction Model. The posture representation extractor is a 3-layer CNN model. Particularly, we use a convolutional layer with 2D filters in each CNN layer to calculate the feature maps. By integrating the 2D features maps of range-azimuth and range-elevation information, we can derive posture representations (i.e., Z) that characterize skeleton posture in the 3D space. In addition, a max-pooling layer is attached to the convolutional layer. Max-pooling can enhance the representation transferability by integrating multiple 2D feature points over a small neighborhood. To prevent overfitting, we use a dropout layer to randomly remove network parameters during training. The 2D feature maps are then flattened and compressed with a fully-connected layer. Given input data X, the posture representation extractor produces 3D posture representations as follows:

$$Z = F(X, \Phi), \tag{8}$$

where $F(\cdot)$ denotes the CNN model and Φ represents the trainable parameters (i.e., model weights) of the posture representation extractor. The model implementation of *mPose* employs 64, 128, and 256 filters in the 3 convolutional layers, respectively. The dropout rate of the dropout layer is set to 50%. We use Leaky ReLU with the parameter $\alpha = 0.01$ for both the convolutional and the fully-connected layers.

Based on the derived representations Z, *mPose* employs a 2-layer fully-connected neural network (i.e., posture estimator) to estimate the 3D position of each skeletal joint. The neural network further extracts non-linear abstractions that characterize the 3D skeleton posture. Based on the abstractions, a regression layer produces the joint coordinates $\hat{P}$. We define the mapping function as:

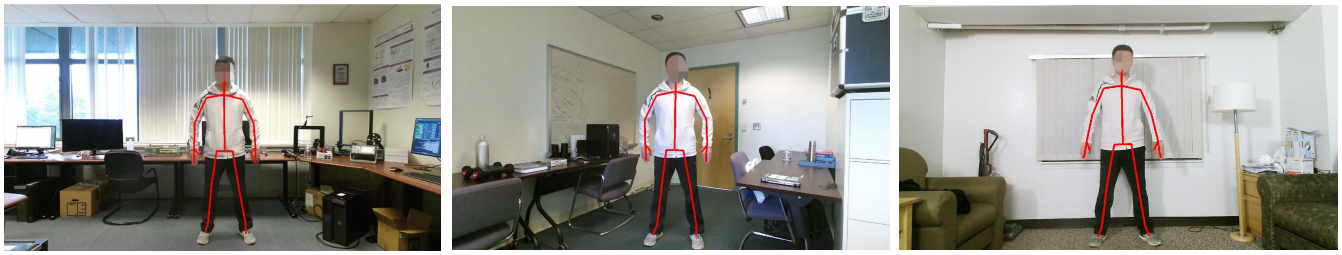
$$\hat{P} = G_p(Z, \Theta), \tag{9}$$

where $G_p(\cdot)$ represents the neural network and Θ denotes its trainable parameters. To optimize the posture estimator for predicting the 3D skeletal joint positions, we use Huber loss [38] to quantify the estimation cost. The joint estimation loss is defined as:

$$L_p = \begin{cases} \frac{1}{2}(P - \hat{P})^2, & |P - \hat{P}| \leq \delta \\ \delta \cdot (|P - \hat{P}| - \frac{1}{2}\delta), & |P - \hat{P}| > \delta \end{cases} \tag{10}$$

where P is the ground truth of the joint coordinates, and δ is the threshold for outlier detection.

Domain Discriminator. To achieve robust posture reconstruction across subjects, we use domain adversarial training [19] to optimize the posture representation extractor. The core component is a domain discriminator used in the training process for removing the domain-specific characteristics entangled in posture representations. Specifically, the domain



(a) Lab (b) Office (c) Apartment

Fig. 7: Illustration of the three environments used for evaluation.

discriminator is a 2-layer fully-connected neural network taking the posture representations Z as input. It infers the subject label as:

$$\hat{D} = G_d(Z, \Gamma), \tag{11}$$

where $G_d(\cdot)$ denotes the neural network and Γ represents the corresponding trainable parameters. We optimize the neural network with the cross-entropy loss as:

$$L_d = H(D, \hat{D}), \tag{12}$$

where $H(\cdot)$ represents the cross-entropy loss function, and D denotes the ground truth of the domain label.

The domain discriminator seems to contradict with our objective of domain-independent posture reconstruction. However, by using an adversarial loss, we can optimize the posture representation extractor to fool the domain discriminator so as to make the derived representations domain-independent. Specifically, during training, we apply a negative factor $-\lambda$ to the domain loss to force the posture representation extractor to maximize the domain loss. We define the adversarial loss to optimize the posture representation extractor as follows:

$$L_{adv} = L_p - \lambda L_d, \tag{13}$$

where L_p and L_d are the joint estimation loss and the domain loss, respectively. The factor λ is used to balance the performance of domain adaptation and posture reconstruction. In the adversarial training process, we iteratively optimize $\{\Phi, \Theta\}$ and Γ with Adam optimizer.

VI. Evaluation

A. Experimental Setup & Methodology

Hardware Components. We implement *mPose* with a TI AWR1642-ODS mmWave device as the sensing front end and a ThinkPad X1 Carbon laptop as the data processing backend. The mmWave device has an integrated antenna array with 2 transmitting antennas and 4 receiving antennas, which sends and receives chirp signals with a frequency range of $77 \sim 81GHz$ and a chirp duration of $33\mu s$. A TI DCA1000EVM data capture card is used to collect data from the mmWave device and forward them to the laptop. While we are collecting mmWave data, we use a Microsoft Kinect [33] to record the ground truth 3D coordinates of 17 skeletal joints (i.e., the same set of joints evaluated in WiPose [11]).

Data Collection. Experiments are conducted in three environments of different sizes and room objects as shown in Figure 7 to demonstrate the robustness of *mPose* across environments. The larger room (i.e., lab) is a public place and has a size of $28ft \times 25ft$ with desks, chairs, and many lab devices (e.g., desktops, 3D printers). The two smaller rooms (i.e., office and apartment) have sizes of $24ft \times 15ft$ and $33ft \times 17ft$ with office (e.g., tables, chairs) and home objects (e.g., sofas, floor lamps). We use the three rooms of different sizes and layouts to demonstrate the effectiveness of the proposed environmental interference removal algorithm. We recruited 7 subjects, including 5 males and 2 females with various heights and weights (e.g., the range of height and weight are $1.63 \sim 1.78m$ and $58 \sim 83kg$ respectively). The experiments involved 17 representative postures: lifting left/right arm to the front for 45/90/180 degree, lifting left/right arm from the side for 45/90/180 degree, lifting left/right leg for 45/90 degree, waving hands, walking, random moving. The volunteers were asked to perform these postures around 1.5m in front of the mmWave device, but we did not limit the specific user position. While we are collecting mmWave data, we use a Microsoft Kinect [33] to record the ground truth 3D coordinates of 17 skeletal joints (i.e., the same set of joints evaluated in WiPose [11]) with a rate of 30 frames per second. The Kinect was placed behind the mmWave device as shown in Figure 1 (a) and extracts 17 skeletal joints with vision-based techniques [39]. Around 640 samples (i.e., mmWave data frames with ground truth skeleton coordinates) were collected per subject per posture. We split the whole dataset (mmWave samples with ground truth joint coordinates) into a training set and a testing set, by using a ratio of 4:1. The deep learning model of *mPose* is trained with the samples and ground truth in the training set.

TABLE I: Overall performance of joint location error.

Location Error	3D Joint	Depth	Azimuth	Elevation
WiPose [11]	36.7mm	-	-	-
mm-Pose [16]	-	32.0mm	75.0mm	27.0mm
RF-Pose3D [40]	76.7mm	42.0mm	49.0mm	40.0mm
mPose	**30.1mm**	**10.7mm**	**16.8mm**	**15.0mm**

Evaluation Metrics. we define the joint localization errors in the 3D space as the projection of the distance between each joint's predicted coordinates and the ground truth on the three axes, depth, azimuth, and elevation. Furthermore, we use the 3D Euclidean distance in millimeters (mm) to quantify the overall joint localization error across the three axes.

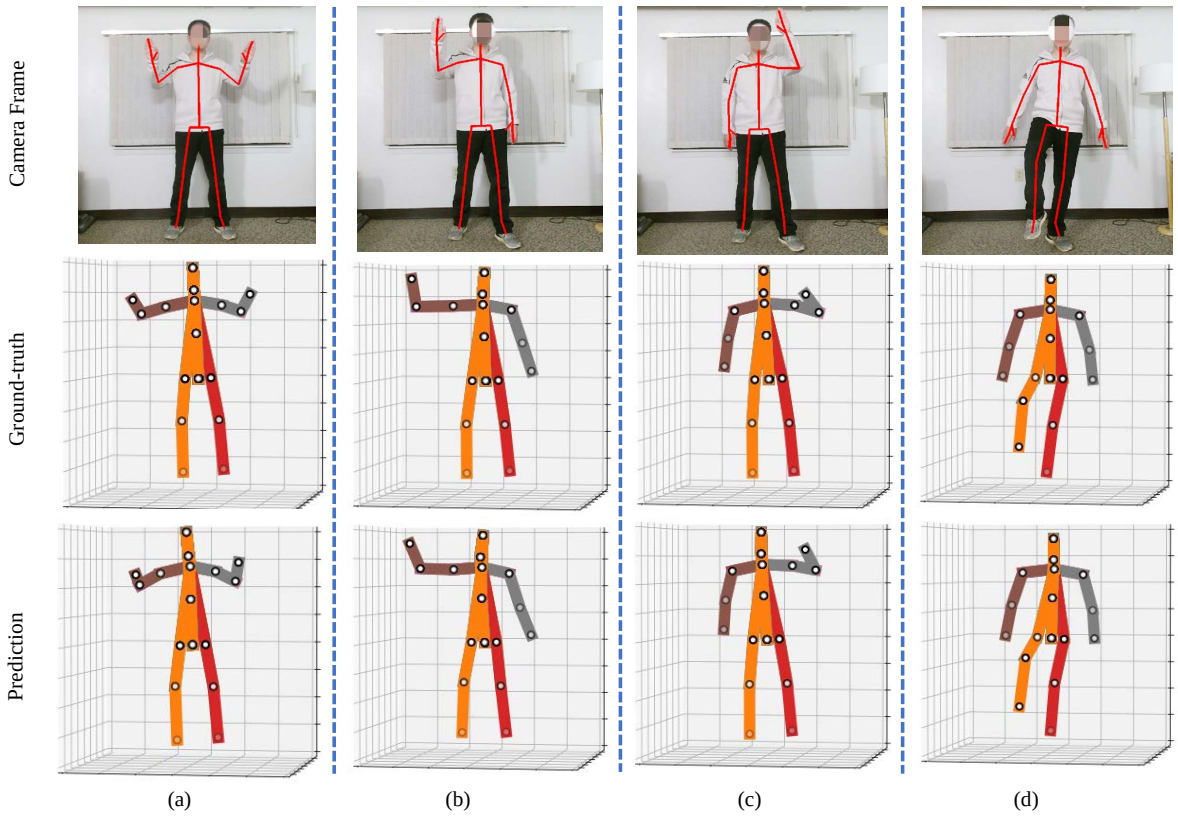


Fig. 8: Reconstructed skeleton postures and ground-truth of four different posture.

B. Performance of 3D Posture Reconstruction

Overall Performance. We first evaluate the overall performance of *mPose* using data collected in the three room environments for 3D skeleton posture reconstruction. Table I shows the average joint localization error across all postures of the 7 subjects. The results show that *mPose* has low errors for all metrics. Specifically, the 3D joint localization error of *mPose* is $30.1mm$, and the errors in depth, azimuth, elevation are $10.7mm$, $16.8mm$, $15.0mm$, respectively. Compared to *mm-Pose* [16], which uses two mmWave devices for posture reconstruction, our system can achieve $21.3mm$, $58.2mm$, and $12.0mm$ improvement in depth, azimuth, elevation with only a single mmWave device. In addition, *mPose* shows $6.6mm$ improvement in joint localization compared to *WiPose* [11], without using spatially distributed antennas/devices. Furthermore, *mPose* has less than half of the 3D joint localization error compared to RF-Pose3D [40], which uses a customized FMCW radio for posture reconstruction. These results demonstrate that *mPose* outperforms the state-of-the-art skeleton reconstruction schemes.

Figure 8 illustrates the differences between the skeleton postures reconstructed by *mPose* and the ground truth. The four postures are: (a) waving hands, (b) lifting the right hand to the front for 180 degrees, (c) lifting the left hand to the front for 180 degrees, (d) lifting the right leg for 90 degrees, respectively. We can find that the reconstructed 3D skeletons are almost the same as the ground truth joint coordinates estimated with Kinect. Figure 9 shows the 3D joint localization error for individual joints. It shows that our system can effectively localize most of the joints with errors lower than $25mm$, except for joint 6 (elbow left), joint 7 (wrist left), joint 9 (elbow right), and joint 10 (wrist right), since these joints are very close to each other in the arm. The results demonstrate the effectiveness of *mPose*.

Impact of Subjects. We further evaluate the generalizability of *mPose* to different subjects. Specifically, we train and test the deep learning model with data collected from each of the 5 subjects. Figure 10(a) shows the joint reconstruction error of individual subjects. We can find that the 3D Euclidean errors are all below $32mm$, which indicate the robustness of *mPose* on different subjects. Furthermore, we observe that *mPose* has low errors in depth, azimuth, elevation, which are below $14mm$, $20mm$, and $19mm$, respectively, indicating that *mPose* is general and can be applied to subjects that have various heights and body sizes.

Impact of Location. Due to the short wavelength, the mmWave signals attenuate rapidly as the propagation distance increases, which introduces uncertainty to our system. Hence, we evaluate the robustness of *mPose* under 3 different distances (i.e., $1.5m$, $2.5m$, and $3.5m$) between the target subject and mmWave device, with the deep learning model trained with the data collected at $1.5m$. Figure 10(b) shows the joint

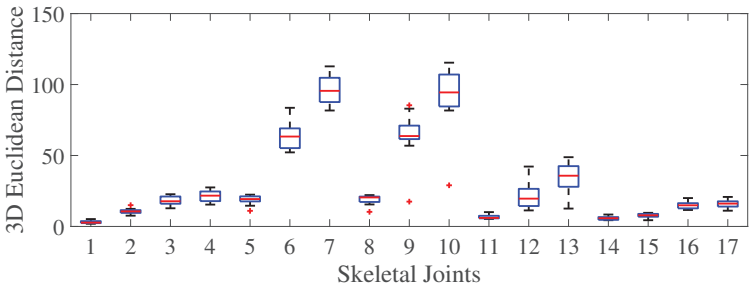
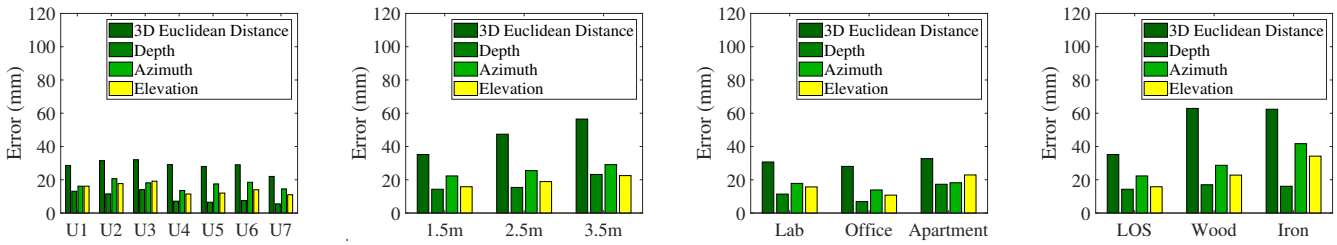


Fig. 9: Performance on reconstructing individual joints.



(a) Different subjects (b) Different distances (c) Different environments (d) NLOS

Fig. 10: Performance under different impacting factors.

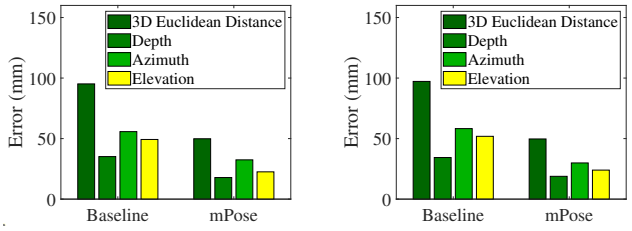
localization errors of $mPose$ under the three distances. We can see that $mPose$ achieves the minimum joint reconstruction error at the distance of $1.5m$, where the training data are collected. We have a similar observation for the errors in depth, azimuth, and elevation. For the other distances, the errors increase a little bit, with a joint localization error of $47.4mm$ at $2.5m$. Even the distance increases to $3.5m$, the joint localization error is still below $60mm$. This is because our environment interference removal algorithm can dynamically detect the subject's location and remove the impact of distance via interpolation, so as to ensure acceptable performance for the skeleton posture reconstruction at different distances.

Impact of Environments. To examine the performance of *mPose* in different environments, we take turns to train and test the deep learning model with data collected in each of the three environments (i.e., lab, office, apartment). As shown in Figure 10(c), we can see that the joint localization errors are below $33mm$ for all environments. The errors in depth, azimuth, and elevation are below $17.3mm$, $18.2mm$, and $22.9mm$, respectively. In addition, we find that *mPose* has relatively higher errors in the apartment. Such an effect is due to the more complex room layout and furniture placement of the apartment, which could introduce more complex multi-path effects of mmWave signals. But even in this case, *mPose* can still achieve a low joint localization error of $32.7mm$. These results demonstrate that *mPose* can achieve satisfactory performance on 3D skeleton reconstruction in different environments.

Impact of Non-Line-of-Sight Conditions. Different from vision-based approaches, mmWave signals can penetrate the obstacle between the mmWave device and the target subject. Hence, we evaluate the performance of *mPose* under the occlusions of metal and wooden objects, simulating object blocking in real-world scenarios (e.g., by home devices or furniture). In the experiment, we place an iron block and a wooden block in front of the mmWave device with a distance of $10mm$, which blocks the LOS between the subject and the mmWave device. Figure 10(d) shows the 3D joint reconstruction error of *mPose* in the LOS and the NLOS scenes. We can see that although the joint localization errors under the NLOS scenarios are larger than that under the LOS scenarios, *mPose* can achieve $62.4mm$ joint localization error when a wooden block placed in front of the mmWave device. In addition, we find that even when the LOS is blocked by iron block, where the mmWave signals are hard to penetrate, *mPose* can still achieve $63.9mm$. These results indicate that our system can work in NLOS scenarios with acceptable performance.

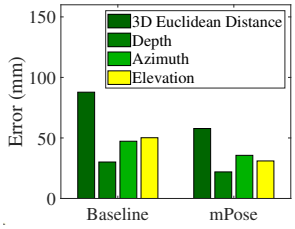
C. Performance of Environment-independent Posture Reconstruction

We next evaluated the transferability of *mPose* across environments. Particularly, we trained the deep learning model with mmWave data from one environment and evaluated it with data collected in another environment, without additional training. Such an evaluation was conducted on every pair of the three environments (i.e., office, lab, and apartment). To show the effectiveness of the proposed environmental artifact removal (Section V-C), we also trained and tested a deep learning model without applying the environmental artifact removal algorithm (i.e., the baseline model). Additionally, we show the overall joint localization error that is calculated by averaging the errors of all joints and all users. For the office and apartment environments, as shown in Figure 11 (a) and (b), we find that *mPose* achieves low joint localization errors of

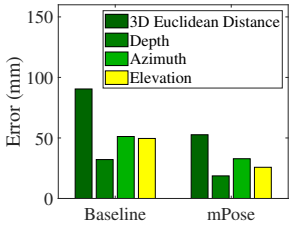


(a) Office to Apartment

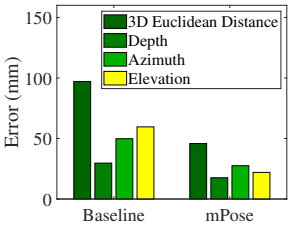
(b) Apartment to Office



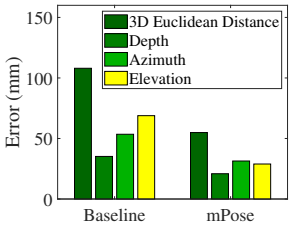
(c) Office to Lab



(d) Lab to Office



(e) Apartment to Lab



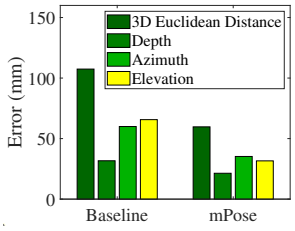
(f) Lab to Apartment

Fig. 11: Average joint localization errors for cross-environment 3D skeleton posture construction.

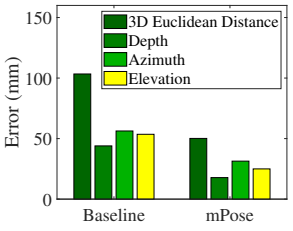
$49.8mm$ and $49.7mm$ under the two training-testing settings. Compared to the baseline model, we find that the proposed environmental artifact removal algorithm greatly reduces the posture reconstruction errors, with over $45.3mm$ and $47.5mm$ improvements for the two settings. The algorithm also reduces the reconstruction errors of depth, azimuth, and elevation, showing its effectiveness to improve the reconstruction in all dimensions. Similarly, as shown in Figure 11 (c) and (d), *mPose* shows lower joint localization errors with the environmental artifact removal algorithm for the office and lab environments, with over $33mm$ and $31mm$ improvements compared to the baseline model for the two training-testing settings. Similar observations can be found in Figure 11 (e) and (f), our system can achieve relatively lower 3D joint localization errors of $68.1mm$ and $69.3mm$ for the domain adaptation between the apartment and the lab. In general, the results demonstrate that the proposed environmental artifact removal algorithm is highly effective in improving the transferability of *mPose* across different environments.

D. Performance of Subject-independent Posture Reconstruction

In real-world scenarios, the posture reconstruction model is usually trained with the dataset from limited users, which makes the model fail to reconstruct posture of users do not involve in the training. To evaluate *mPose* in the subject-independent scenarios, we trained the deep learning model with data collected from one subject and adapt the model to another subject using the domain adaptation technique described in Section V-D. For comparison, we also examined the baseline deep learning model without domain adaptation (i.e., the baseline model). We repeated such an evaluation for all subject pairs and averaged the errors. Figure 12(a) shows the reconstruction errors of *mPose* when training and testing on different subjects in the lab environment. We can observe that *mPose* has $59.6mm$ average joint localization error when using the proposed domain adaptation method, which is improved by $47.7mm$ compared to the baseline model. We also find that the domain adaptation method helps to achieve better performance on individual dimensions in the 3D space, with $10.3mm$, $24.7mm$, and $34.1mm$ improvements in the dimension of depth, azimuth, and elevation, respectively. Similar results can be found in the office environment as shown in Figure 12(b), with $53.2mm$ lower joint localization error after applying the domain adaptation method. The results demonstrate that the proposed domain adaptation method can help to achieve reliable posture reconstruction across subjects with unique body shapes.



(a) Lab



(b) Office

Fig. 12: Average joint localization errors for cross-subject 3D skeleton posture construction.

VII. Conclusion

In this paper, we proposed *mPose* which can continuously track a user's 3D skeleton postures using mmWave radios. The system can be deployed on a single portable COTS mmWave device while achieving a high joint localization accuracy. Through dynamically tracking the user's relative position to the system, *mPose* extracts spatial features associated with the user in a 3D contour, which removes the impacts of environmental factors. A deep regression model based on convolutional neural network is designed to map the extracted spatial features into the skeletal joint coordinates in a 3D space. With the designed domain adaptation method, *mPose* removes subject-specific characteristics entangled in the spatial features and enables robust posture reconstruction across users. Experimental results demonstrated that *mPose* can accurately reconstruct full-body 3D skeleton posture, achieving at least 18% lower prediction errors than existing device-free skeleton posture reconstruction methods with a single mmWave device. Additionally, the experiments confirm the robustness

and generalizability of *mPose* with different environments and different users. We believe that our portable single-mmWave-device solution could enable various emerging applications, including immersive augmented reality or virtual reality, mobile healthcare, and pervasive security monitoring.

VIII. Acknowledgement

We thank Zhenzhe Lin for his help in conducting experiments.

References

[1] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[2] R. Mehrizi, X. Peng, X. Xu, S. Zhang, D. Metaxas, and K. Li, "A computer vision based method for 3d posture estimation of symmetrical lifting," *Journal of Biomechanics*, no. 69, pp. 40 – 46, 2018.

[3] B. Boulay, F. Bremond, and M. Thonnat, "Human posture recognition in video sequence," in *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (IEEE VS-PETS)*, 2003.

[4] VICON, "Motion capture systems — vicon," https://www.vicon.com/, 2019.

[5] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *Proceedings of International Conference on Systems, Man, and Cybernetics (IEEE SMC)*, 2015, pp. 1488–1492.

[6] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the International Conference on Multimedia (ACM MM)*, 2015, pp. 1307–1310.

[7] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures," in *Proceedings of the 20th annual international conference on Mobile computing and networking (ACM MobiCom)*, 2014, pp. 617–628.

[8] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, "Tracking vital signs during sleep leveraging off-the-shelf wifi," in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing (ACM MobiHoc)*, 2015, pp. 267–276.

[9] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the human figure through a wall," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–13, 2015.

[10] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE CVPR)*, 2018, pp. 7356–7365.

[11] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3d human pose construction using wifi," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (ACM MobiCom)*, 2020, pp. 1–14.

[12] "(don't) hold the phone: new features coming to pixel 4," https://www.blog.google/products/pixel/new-features-pixel4/, 2019.

[13] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[14] Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, "Vital sign and sleep monitoring using millimeter wave," *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 2, pp. 1–32, 2017.

[15] X. Yang, J. Liu, Y. Chen, X. Guo, and Y. Xie, "Mu-id: Multi-user identification through gaits using millimeter wave radios," in *Proceedings of International Conference on Computer Communications (IEEE INFOCOM)*, 2020, pp. 2589–2598.

[16] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, 2020.

[17] A. G. Stove, "Linear fmcw radar techniques," in *IEE Proceedings F (Radar and Signal Processing)*, vol. 139, no. 5. IET, 1992, pp. 343–350.

[18] S. Rao, "Introduction to mmwave sensing: Fmcw radars," *Texas Instruments (TI) mmWave Training Series*, 2017.

[19] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research (JMLR)*, vol. 17, no. 1, pp. 2096–2030, 2016.

[20] M. Quwaider and S. Biswas, "Body posture identification using hidden markov model with a wearable sensor network," in *Proceedings of the International Conference on Body Area Networks (ACM ICST)*, 2008, pp. 1–8.

[21] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab, "Human skeleton tracking from depth data using geodesic distances and optical flow," *Elsevier Image and Vision Computing*, vol. 30, no. 3, pp. 217–226, 2012.

[22] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, 2014, pp. 1799–1807.

[23] S. Shen, H. Wang, and R. Roy Choudhury, "I am a smartwatch and i can track my user's arm," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services (ACM MobiSys)*, 2016, pp. 85–96.

[24] E. Valero, A. Sivanathan, F. Bosché, and M. Abdel-Wahab, "Analysis of construction trade worker body motions using a wearable and wireless motion sensor network," *Automation in Construction*, vol. 83, pp. 48–55, 2017.

[25] E. S. Ho, J. C. Chan, D. C. Chan, H. P. Shum, Y.-m. Cheung, and P. C. Yuen, "Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments," *Computer Vision and Image Understanding*, vol. 148, pp. 97–110, 2016.

[26] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

[27] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt, "Real-time pose and shape reconstruction of two interacting hands with a single depth camera," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–13, 2019.

[28] S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, K. Ebihara, and S. Morishima, "Human body postures from trinocular camera images," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, March 2000, pp. 326–331.

[29] O. Postolache, P. S. Girão, R. N. Madeira, and G. Postolache, "Microwave fmcw doppler radar implementation for in-house pervasive health care system," in *2010 IEEE International Workshop on Medical Measurements and Applications*. IEEE, 2010, pp. 47–52.

[30] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," 2017.

[31] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-wifi: Fine-grained person perception using wifi," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5452–5461.

[32] L. Feng, Z. Li, C. Liu, X. Chen, X. Yin, and D. Fang, "Sitr: Sitting posture recognition using rf signals," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 492–11 504, 2020.

[33] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[34] H. Guo, Y. Yu, Q. Ding, and M. Skitmore, "Image-and-skeleton-based parameterized approach to real-time identification of construction workers' unsafe behaviors," *Journal of Construction Engineering and Management*, vol. 144, no. 6, p. 04018042, 2018.

[35] "Vive - tracker," https://www.vive.com/us/vive-tracker/.

[36] P. Caserman, A. Garcia-Agundez, R. Konrad, S. Göbel, and R. Steinmetz, "Real-time body tracking in virtual reality using a vive tracker," *Virtual Reality*, vol. 23, no. 2, pp. 155–168, 2019.

[37] C. Iovescu and S. Rao, "The fundamentals of millimeter wave sensors," *Texas Instruments*, 2017.

[38] Wikipedia, "Huber loss," https://en.wikipedia.org/wiki/Huber_loss, 2019.

[39] "Kinect," https://en.wikipedia.org/wiki/Kinect.

[40] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3d skeletons," in *Proceedings of the Special Interest Group on Data Communication (ACM SIGCOMM)*, 2018, pp. 267–281.