m^3Track: mmWave-based Multi-User 3D Posture Tracking

Hao Kong
Shanghai Jiao Tong University
Shanghai, China
hao.kong@sjtu.edu.cn

Xiangyu Xu
Southeast University
Nanjing, China
xy-xu@seu.edu.cn

Jiadi Yu*
Shanghai Jiao Tong University
Shanghai, China
jiadiyu@sjtu.edu.cn

Qilin Chen
Shanghai Jiao Tong University
Shanghai, China
1017856853@sjtu.edu.cn

Chenguang Ma
Ant Financial Services Group
Hangzhou, China
chenguang.mcg@antfin.com

Yingying Chen
Rutgers University
New Brunswick, NJ, USA
yingche@scarletmail.rutgers.edu

Yi-Chao Chen
Shanghai Jiao Tong University
Shanghai, China
yichao@utexas.edu

Linghe Kong
Shanghai Jiao Tong University
Shanghai, China
linghe.kong@sjtu.edu.cn

ABSTRACT

Nowadays, the market of 3D human posture tracking has extended to a broad range of application scenarios. As current mainstream solutions, vision-based posture tracking systems suffer from privacy leakage concerns and depend on lighting conditions. Towards more privacy-preserving and robust tracking manner, recent works have exploited commodity radio frequency signals to realize 3D human posture tracking. However, these studies cannot handle the case where multiple users are in the same space. In this paper, we present a mmWave-based multi-user 3D posture tracking system, m^3Track, which leverages a single commercial off-the-shelf (COTS) mmWave radar to track multiple users' postures simultaneously as they move, walk, or sit. Based on the sensing signals from a mmWave radar in multi-user scenarios, m^3Track first separates all the users on mmWave signals. Then, m^3Track extracts shape and motion features of each user, and reconstructs 3D human posture for each user through a designed deep learning model. Furthermore, m^3Track maps the reconstructed 3D postures of all users into 3D space, and tracks users' positions through a coordinate-corrected tracking method, realizing practical multi-user 3D posture tracking with a COTS mmWave radar. Experiments conducted in real-world multi-user scenarios validate the accuracy and robustness of m^3Track on multi-user 3D posture tracking.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing*.

KEYWORDS

mmWave Radars, 3D Posture Tracking, Multi-User Scenarios

*Jiadi Yu is the corresponding author

MobiSys '22, June 25-July 1, 2022, Portland, OR, USA

ACM ISBN 978-1-4503-9185-6/22/06…$15.00
https://doi.org/10.1145/3498361.3538926

ACM Reference Format:
Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m^3Track: mmWave-based Multi-User 3D Posture Tracking. In *The 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22), June 25-July 1, 2022, Portland, OR, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3498361.3538926

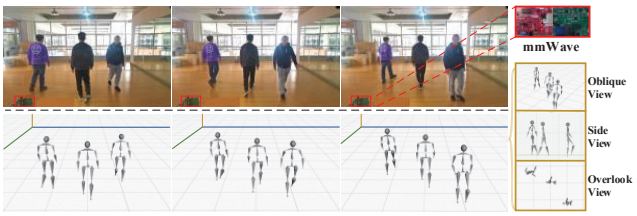


Figure 1: Illustration of m^3Track system.

1 INTRODUCTION

Recent years have witnessed a rapid development of 3D human posture tracking technology, which generates dynamic skeletons that follow a person as she/he moves, walks, or sits. Together with the increasing popularity of IoT devices, the market of 3D human posture tracking nowadays has been largely extended from a few specialized scenarios (e.g., filmmaking, athletic training, and military applications) to a broader range of commercial applications, including virtual reality (VR)/augmented reality (AR), motion-sensing gaming, smart-home control, etc. Along with this trend, the deployment of 3D human posture tracking shifts from expensive or intrusive wearable sensors [11, 26] to non-intrusive vision-based techniques [1, 20], which dominates current 3D human posture tracking markets. However, vision-based approaches depend on the lighting conditions of environments. Moreover, they suffer from privacy leakage concerns, which are increasingly being taken seriously nowadays.

With the great efforts towards more robust and privacy-preserving 3D human posture tracking approaches, researchers have exploited radio frequency (RF) signals, such as Wi-Fi [10] and mmWave [23, 24, 32], to track 3D human postures. But all these approaches only focus on tracking a single user's posture, not covering more challenging yet useful multi-user scenarios. Although some studies [27, 39] enable human tracking in multi-user scenarios, they can only track the locations of users and cannot generate dynamic skeletons

that follow their movements. Some pioneer studies [36–38] realize 3D posture tracking for multiple users through specially-designed RF facilities, which demonstrates the possibility of capturing the 3D posture information of multiple users with RF signals. However, although with some successful experiences, they strictly rely on the specialized hardware (requiring to equip at least 20 antennas), making it difficult to be widely deployed.

Towards this end, a robust and practical multi-user 3D human posture tracking system based on the commercial off-the-shelf (COTS) RF signals is highly desirable. Such a system can be easily deployed in real-world scenarios, enabling a broad array of real-world applications including multi-user gaming, multi-object motion tracking, etc. To achieve multi-user 3D posture tracking with a single COTS mmWave radar, we face several challenges in practice. First, we need to accurately separate multiple users and capture their posture information individually. Second, the system should reconstruct fine-grained 3D posture of each user through the implicit mmWave signals. Third, the system needs to track the positions of multiple users' 3D postures simultaneously.

In this paper, we focus on tracking the 3D postures of multiple users, and propose a mmWave-based multi-user 3D posture tracking system, m^3Track, which leverages a single COTS mmWave radar to track 3D postures of multiple users simultaneously. First, m^3Track utilizes designed chirp signals to sense all users through spectrum convolution and then separates the users on mmWave signals through a minimum variance distortionless response-based approach. Based on the mmWave signal of each separated user, m^3Track extracts spatial and temporal features that describe the shape and motion of a user, and further reconstructs the 3D human posture of each user through a deep learning model, i.e., a forked-ConvLSTM. Next, m^3Track maps the reconstructed 3D postures of all users into real-world 3D space by finding the minimal mapping errors between the reconstructed postures and the generated point clouds of users, and continuously tracks the positions of multiple users through a proposed coordinate-corrected extended Kalman filter. Finally, we validate the accuracy and robustness of m^3Track by conducting experiments with a COTS mmWave radar in 6 real-world environments. The results show that m^3Track can effectively track 4 users at the same time with the average joint tracking error of 42.4mm and the localization error of 21.5mm. An example of m^3Track is illustrated in Figure 1.

We highlight our main contributions as follows:

- We design a multi-user 3D posture tracking system using a single COTS mmWave radar, m^3Track, facilitating a broad range of practical posture tracking applications for multi-user scenarios.
- We propose a multi-user separation method to effectively separate multiple users on mmWave signals, and build a ConvLSTM-based deep learning model to realize 3D human posture reconstruction for each separated user.
- We present a point cloud-based mapping method to map all 3D human postures into real 3D space, and propose a coordinate-corrected tracking method to realize practical multi-user 3D posture tracking.
- We conduct extensive experiments in real-world multi-user scenarios, and the results validate the accuracy and robustness of m^3Track in tracking 4 users simultaneously.

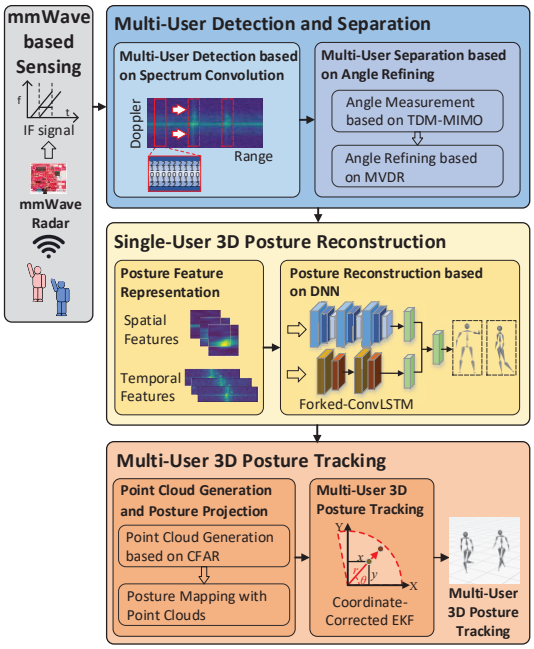


Figure 2: System framework of m^3Track.

2 SYSTEM OVERVIEW

To realize multi-user 3D posture tracking in real-world scenarios, we propose m^3Track, which leverages a single commercial off-the-shelf (COTS) mmWave radar to reconstruct and track 3D human postures of multiple users simultaneously. Figure 2 shows the system framework of m^3Track, which consists of four modules:

mmWave-based Sensing. In the module, we design chirp mmWave signals to sense objects in the sensing area. A mmWave radar propagates the chirp mmWave signals, and the signals are reflected by the objects in the sensing area. Then, the reflected signals are collected by the receive antennas of the same mmWave radar to sense all objects in the sensing area.

Multi-User Detection and Separation. This module is designed to detect and separate multiple users on mmWave signals. m^3Track first detects all users through a designed spectrum convolution method, and further separates the users on mmWave signals through minimum variance distortionless response approach.

Single-User 3D Posture Reconstruction. This module aims to reconstruct the 3D human posture of each user in multi-user scenarios. From the separated mmWave signal profiles of each user, m^3Track extracts both spatial and temporal features that describe a user's shape and motion, respectively. Then, m^3Track reconstructs 3D human posture for each user through a designed deep neural network, i.e., a forked-ConvLSTM.

Multi-User 3D Posture Tracking. In the module, m^3Track realizes 3D posture tracking for multiple users. First, in multi-user scenarios, m^3Track generates the point clouds of all users, and maps the reconstructed 3D postures of all users with the generated point clouds in 3D space by finding the minimal mapping errors between postures and point clouds. Then, m^3Track continuously tracks the positions of multiple users through a proposed coordinate-corrected extended Kalman filter, realizing practical 3D posture tracking for all users.

3 MMWAVE-BASED SENSING

Before multi-user 3D posture tracking using mmWave signals, m^3Track first needs to transmit and collect mmWave signals from a COTS mmWave radar to sense objects in the sensing area.

For a typical COTS mmWave radar, it leverages frequency-modulated continues wave (FMCW) technique to transmit multiple chirps with linearly increased frequency from the transmit antennas. The signals propagate in front of the radar and are reflected by the objects in the sensing area. Then, the reflected signals are captured by the receive antennas of the same radar. To reveal the sensing information, the transmitted and received signals are mixed to generate the intermediate frequency (IF) signal:

$$x_{IF}(t) = A_r \cdot e^{j2\pi \cdot \tau(r,c)\left[f_0 + \frac{B}{T_c}t - \frac{B}{2T_c}\tau(r,c)\right]}, \quad (1)$$

where f_0 is the start frequency, B is the signal bandwidth, T_c is the chirp duration, A_r is the amplitude coefficient that represents the attenuation of mmWave signal, and $\tau(r, c)$ is the delay of the received signal with respect to the transmit signal which is determined by the distance r of the targets and the speed c of mmWave signal. Through analyzing the IF signal, the mmWave radar senses all objects in the sensing area.

To achieve effective sensing for multi-user 3D posture tracking, we design the chirp signals for m^3Track based on the fundamental of mmWave signals. According to Eq. (1), the resolution to detect different objects, i.e., range resolution D_{res}, can be calculated as $D_{res} = \frac{c}{2B}$, where c is the speed of light and B is the bandwidth of signals. Similarly, the resolution to measure the movement of objects, i.e., the speed resolution V_{res}, is denoted as $V_{res} = \frac{\lambda}{2T_f}$, where λ is the wavelength associated with the average frequency of the transmitted mmWave, and T_f is the duration of an FMCW signal frame. According to the resolution analysis, a higher average frequency and bandwidth of the mmWave signal lead to a better resolution for detecting multiple objects and capturing their movements. Thus, we leverage the highest average frequency and bandwidth in the mmWave radar, i.e., a $77 \sim 81GHz$ frequency range. With the frequency range, m^3Track achieves a range resolution of $3.75cm$ and speed resolution of $0.02m/s$, which enables to detect multiple objects and capture their movements at cm-level.

Therefore, based on mmWave, m^3Track further builds a practical multi-user 3D posture tracking system using mmWave signals.

4 MULTI-USER DETECTION AND SEPARATION

Based on the collected mmWave signals, all the dynamic objects and static objects in a sensing environment are captured. To realize multi-user 3D posture tracking, m^3Track first removes the static objects and detects all potential users, and further separates multiple users from each other on mmWave signals.

4.1 User Detection on mmWave Signals

Since the IF signal of mmWave reveals the range of sensed objects, m^3Track can detect all the objects (including dynamic and static objects) through range analysis. Specifically, m^3Track extracts the range of sensed objects through the frequency f of the IF signal. The range of an object is represented as $r = \frac{cfT_c}{2B}$, which denotes a linear relationship between the object's range r and the frequency f of the IF signal. Through performing FFT (Range-FFT) to the IF signal, m^3Track obtains a *Range-Profile* for the objects sensed by mmWave radar. Hence, all the objects in the sensing environment are first detected by m^3Track, which contains both dynamic objects (i.e., users) and static objects.

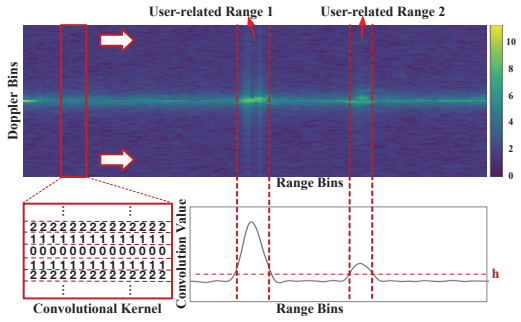
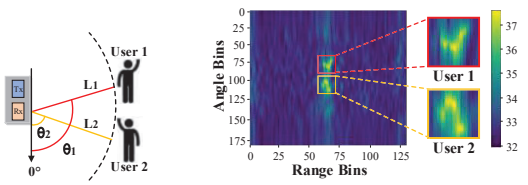


Figure 3: Illustration of user detection.

Because the range analysis in m^3Track detects not only dynamic objects but also static objects, it is necessary to remove the static objects and detect users on mmWave signals for realizing human posture tracking. Usually, most background objects are static, while users are not completely static. Even if a user does not move, the tiny breathing and heartbeat movements can still be captured as subtle changes on mmWave signals. Based on the intuition, we remove static objects and detect target users by measuring the movement of sensed objects. To detect users, we leverage Doppler responses, which vary with the speed of movement, to filter out static objects while retaining the target dynamic users. For extracting Doppler responses, m^3Track measures the speed of sensed objects by phase changes ω of IF signal. The speed of an object towards the mmWave radar can be denoted as $v = \frac{\lambda}{4\pi T_c}\omega$, where λ is the wavelength of the mmWave signal. Similar to Range-FFT, another FFT operation (Doppler-FFT) is conducted on the *Range-Profile* to generate a *Range-Doppler-Profile*, which measures the speed of all objects in different ranges.

With the range and speed, m^3Track further distinguishes users and background objects. To achieve it, a convolution-based approach, i.e., spectrum convolution, is proposed to detect all users on the *Range-Doppler-Profile*, as shown in Figure 3. Specifically, a specialized convolutional kernel is designed to slide along the range bins of the *Range-Doppler-Profile*, and a convolution operation is performed for each sliding step to detect all users. As Figure 3 shows, the convolutional kernel has a specific width and length, in which the width is set as the same width of the *Range-Doppler-Profile*, and the length is set as 14. Setting the length to 14 leads to a $0.5m$ range resolution, which is slightly larger than the typical range of a human body, so the convolutional kernel can cover a person on mmWave signals. For the parameters of the kernel, we set them in the center row of the kernel as 0, and other rows increase linearly from the center row to the edge row. The intuition is that the ranges with higher Doppler responses correspond to dynamic objects, while the ranges with little Doppler responses reflect the static environmental objects. Thus, the convolutional kernel design assigns more weights to the high Doppler response ranges, which focuses on dynamic objects such as users. Meanwhile, through weighting less to the range of little Doppler responses, m^3Track can ignore the static object captured by mmWave signals.



(a) Users of the same range. (b) MVDR-based *Range-Angle-Profile* viewd from elevation.

Figure 4: Illustration of user separation.

After the spectrum convolution, the convolution value with respect to different ranges can be calculated. Then, m^3Track utilizes an empirical threshold h to detect the ranges that contain users. Figure 3 illustrates a case where there are two users in range 1 and range 2. It can be observed that the two users are effectively detected through the convolution-based method. In addition, we can see the convolution value of user 2 is far less than user 1, because user 1 is moving while user 2 is holding still. But for user 2, m^3Track can still detect the user through the continual breathing and heartbeat movements, which has a higher convolution value than all the background static objects.

User Joining and Leaving. User joining and leaving in multi-user scenarios is a common and recurring scene. In m^3Track, users who enter or leave the sensing area engender convolution value changes in the *Range-Doppler-Profile*. With the threshold h, m^3Track is able to judge the number of users in the sensing area in real-time. Hence, based on the real-time judgment of user number, m^3Track can detect all users in the sensing area.

4.2 User Separation on mmWave Signals

Although users are detected on mmWave signals based on the user-related ranges, there could be more than one user in the same range, causing overlapping on the range dimension of mmWave signals, as shown in Figure 4(a). However, it is intuitive that the users with the same range are different in angles toward the radar, making them separable on mmWave signals. Hence, to extract the mmWave profiles of each individual user in multi-user scenarios, we further exploit the angle relative to the radar for separating users.

With the assistant of multiple transmit and receive antennas on mmWave radar, the angle of sensed objects are extracted by angle of arrival (AoA) estimation. The angle of an object with respect to the mmWave radar (with azimuth angle θ and elevation angle ϕ) is represented as $\theta = \sin^{-1}\left(\frac{\lambda\omega_a}{2\pi d_1}\right)$ and $\phi = \sin^{-1}\left(\frac{\lambda\omega_e}{2\pi d_2}\right)$, where d_1 and d_2 are physical distances of receive antennas on azimuth and elevation directions, respectively, ω_a and ω_e are the phase difference of MIMO channels on azimuth and elevation directions, respectively. After AoA estimation, an *Angle-Profile* is generated, which reveals the angle of all sensed objects. With the *Range-Profile* and *Angle-Profile*, a *Range-Angle-Profile* is then obtained for separating users.

m^3Track uses the TDM-MIMO scheme of the mmWave radar to calculate the angle of multiple users. The COTS mmWave radar (i.e., TI IWR1443BOOST) is equipped with 3 transmit antennas (Tx) and 4 receive antennas (Rx). The antennas can be expanded to a 2D multiple-input and multiple-output (MIMO) array with 3×4 Tx-Rx pairs, as shown in Figure 5. With the antenna array design, m^3Track is able to estimate the angle of targets in both azimuth and elevation directions. In m^3Track, the 2D MIMO array is activated with time-division multiplexing (TDM), which means that, in the transmit end, alternate time slots are dedicated to Tx1, Tx2, and Tx3, and all 4 Rx are activated to receive the mmWave signals. With the TDM-MIMO scheme, m^3Track overcomes the hardware drawback of insufficient antenna array by generating virtual antenna arrays, providing the basis for extracting the angles of multiple users on mmWave signals.

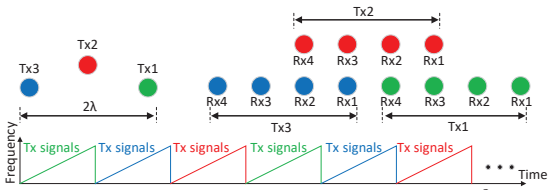


Figure 5: Illustration of antenna array for m^3Track.

Furthermore, we utilize minimum variance distortionless response (MVDR) [6] to refine the angles of different users. The basic idea of MVDR is to mitigate the interference and noise from other angles while obtaining distortionless responses to the angle of view. Combined with the angle of mmWave signals obtained by TDM-MIMO, m^3Track is able to achieve a finer angle extraction. In MVDR, given a_θ as the signal steer vector corresponding to the angle of arrival θ, the MVDR weight is calculated as:

$$w = \frac{R^{-1}a_{\theta,\phi}}{a^H_{\theta,\phi}R^{-1}a_{\theta,\phi}}, \tag{2}$$

where R is the correlation matrix of the antenna array. Then, the output signal power of the antenna array using the optimum weight vector from the MVDR method is:

$$P_{MVDR}(\theta,\phi) = \frac{1}{a^H_{\theta,\phi}R^{-1}a_{\theta,\phi}}, \tag{3}$$

which is calculated by detecting the peaks in the angular spectrum. By replacing the original *Angle-Profile* with the signal power calculated as Eq.(3) for each range bin, m^3Track is able to achieve fine-grained angle extraction. Figure 4(b) shows the MVDR-based *Range-Angle-Profile* for two users of the same range viewed from elevation. Although the users are in the same range, they are separated on angles in the MVDR-based *Range-Angle-Profile*. This demonstrates that the MVDR-based angle solution is effective in separating multiple users on mmWave signals.

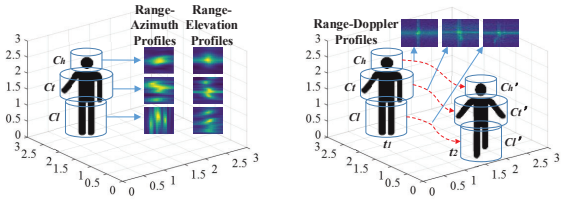
Therefore, based on range and angle, m^3Track separates and extracts mmWave profiles for each user in multi-user scenarios, which are further leveraged for multi-user 3D posture tracking.

5 SINGLE-USER POSTURE RECONSTRUCTION

After separating each user on mmWave signals in multi-user scenarios, m^3Track needs to reconstruct the 3D posture of each user for realizing the 3D posture tracking of multiple users.

5.1 Posture Feature Representation

For 3D human posture reconstruction, m^3Track first obtains spatial features that depict the shape of a user, and temporal features that describe the motion of a user, from mmWave signals.



(a) Spatial features. (b) Temporal features.

Figure 6: Illustration of posture feature representation for spatial and temporal features.

Spatial Features. In multi-user detection and separation, m^3Track calculates the range and angle (including azimuth angle and elevation angle) of each user with respect to the radar. With the range and angle, m^3Track obtains feature patterns of each user in the *Range-Angle-Profiles*. Given the range as r_i, the azimuth angle as θ_i, and the elevation angle as ϕ_i, the coordinate of the central point for a user i can be denoted as (r_i, θ_i, ϕ_i). To obtain the shape patterns of a user, we utilize the central point as an anchor point, and create several cylinders to cover the human body, as shown in Figure 6(a). Specifically, a 3-cylinder model $M(C_h, C_t, C_l)$ covers the human body: a head cylinder $C_h(r_i, \theta_i, \phi_i)$ that relates to the head-neck, a torso cylinder $C_t(r_i, \theta_i, \phi_i)$ relates to chest-arm, and a leg cylinder $C_l(r_i, \theta_i, \phi_i)$ relates to the lap-leg of the user. The radius and height of the three cylinders change adaptively to cover the corresponding body part of a user. To locate the three cylinders in the human body for different postures and human subjects, we fix the height of the torso cylinder around the anchor point and expand the height of the other two cylinders proportionally, because the vertical length of the body torso varies less for different postures and human subjects. In specific, with the height of the torso cylinder set to 60cm for the body, the other two cylinders expand vertically to the margin of the pattern. For each cylinder, m^3Track locates the region, and calculates 2 *Range-Angle-Profiles* from mmWave signals corresponding to azimuth and elevation angles, respectively, as shown in Figure 6(a). Since the profiles from different cylinders are eventually stitched together in the following neural network model, it does not matter if the proportions of the three cylinders have some differences from the actual human body regions. In total, 6 *Range-Angle-Profiles* are obtained for each user as the spatial features. Compared to directly calculating the *Range-Angle-Profiles* of the entire human body, calculating the partial *Range-Angle-Profiles* amplifies the feature representation of specific body parts, which better describes the detailed shape of a user in the 3D space. m^3Track uses the spatial features as the basis for 3D posture reconstruction.

Temporal Features. Besides spatial features, temporal features are also critical for 3D posture reconstruction, which describes the motion of different body parts for each user. To obtain the temporal features, m^3Track also leverages the 3-cylinder model. For each cylinder, a *Range-Doppler-Profile* is calculated to describe the motions for the corresponding body parts of a user. Specifically, for 2 consecutive time slots t_1 and t_2, the *Range-Doppler-Profiles* describe the motions of different body parts for a user across time are calculated, as shown in Figure 6(b). In total, 3 *Range-Doppler-Profiles* are obtained from mmWave signals for each user as the temporal features, which describe the motions of different body parts for each user. m^3Track also utilizes the temporal features as the basis for 3D posture reconstruction.

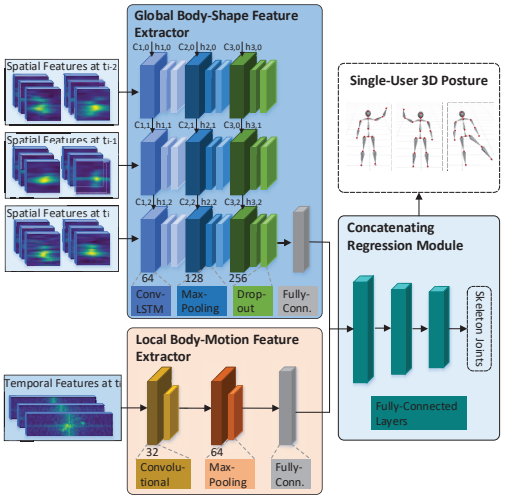


Figure 7: Neural network structure of m^3Track.

Therefore, the shape and motion patterns of each user in multi-user scenarios are obtained in the spatial features and temporal features respectively, which are further leveraged for 3D posture reconstruction.

5.2 3D Posture Reconstruction

We design a deep learning model, i.e., a forked-ConvLSTM, to map the spatial and temporal features to a user's skeleton joint coordinates for 3D posture reconstruction. Figure 7 illustrates the architecture of the proposed deep learning model. The model takes the sequence of spatial features and the corresponding temporal features as inputs of two different branches. A *Global Body-Shape Feature Extractor* and a *Local Body-Motion Feature Extractor* are designed to extract the feature embeddings underlying spatial features and temporal features respectively. A *Concatenating Regression Module* is designed to concatenate the feature embeddings from the outputs of the two feature extractors, and predict the 3D coordinates of skeleton joints for a user to reconstruct 3D posture. Since 3D posture reconstruction is treated as a regression problem, the deep learning model can reconstruct different postures, enabling the universality of 3D posture reconstruction.

Global Body-Shape Feature Extractor. The Global Body-Shape Feature Extractor is designed to learn body-shape feature embeddings from a user's spatial features. The spatial features (i.e., *Range-Angle-Profiles*) at a specific moment represent the global spatial reflection of a target user's shape in space, i.e., they correspond to the instantaneous state of the user's body shape. For the next moment, the body shape in *Range-Angle-Profiles* changes from the state of the previous moment, which means that each part of the *Range-Angle-Profiles* is related to the previous state. Hence, the spatial features are correlated in both time and space, which inspires us to leverage convolutional operation to learn regional characteristics in space, and utilize memory cells to describe the sequential characteristics in time. Based on the above analysis, we leverage ConvLSTM [30], which is the integration of convolutional neural network (CNN) and long short-term memory (LSTM), as the basis to extract global body-shape features.

The architecture of the Global Body-Shape Feature Extractor is shown in the top part of Figure 7, which consists of three layers of ConvLSTM. The input of multiple profiles derived from the posture feature representation is stitched to the input layer. To describe the sequential characteristics in time, the spatial features of t_{i-2}, t_{i-1}, and t_i are simultaneously fed into the network, and the output of the last node in the third layer is taken as the feature extraction result of t_i. In addition, a convolutional operation is utilized to yield regional attention for each spatial feature while maintaining the sequential relationship. The output can be denoted as:

$$Z' = \mathrm{FC}\left(\mathrm{CL}\left(F_{\mathrm{m}}\left(t\right), F_{\mathrm{m}}\left(t-1\right), \cdots, \Theta'\right)\right), \tag{4}$$

where Z' is the output feature embeddings, $\mathrm{FC}(\cdot)$ is the fully-connected layer, $\mathrm{CL}(\cdot)$ is the convolutional long short-term memory network operation, and Θ' is the trainable parameters. At the end of the feature extractor, the feature embedding is stretched using a fully connected network to facilitate the subsequent feature splicing operation. Finally, the feature embeddings in spatial features describing the global body shape of a user are extracted, which are further fed to the following Concatenating Regression Module.

Local Body-Motion Feature Extractor. The Local Body-Motion Feature Extractor is designed to learn the feature embeddings of local body motions of a user from the temporal features. The temporal features, i.e., *Range-Doppler-Profiles*, are correlated with a series of local motions, which describe the motion of corresponding body parts in terms of velocity. Hence, the Local Body-Motion Feature Extractor is designed to extract feature embeddings associated with Doppler responses, which describes the motions of different body parts of a user. As shown in the bottom part of Figure 7, the module is composed of a 2-layer convolutional neural network to extract feature embeddings of local body-motions from the *Range-Doppler-Profiles* $F_d(t)$ of a user. The profiles derived in the posture feature representation are stitched to the input layer. The output of this module is a feature embedding of a user's local body motions during the corresponding time period, which is formulated as:

$$Z = \mathrm{Conv}(F_d(t), \Theta), \tag{5}$$

where Z is the output feature embedding, $\mathrm{Conv}(\cdot)$ is the convolution operation for temporal features, and Θ is the trainable parameters. Based on the feature extraction in this module, m^3Track extracts the feature embeddings in temporal features describing the local body motions of a user, which are further fed to the following Concatenating Regression Module.

Concatenating Regression Module. In the Concatenating Regression Module, the feature embeddings extracted from the previous two modules are combined to encode a stitching vector, which aims to predict the skeleton joint coordinates of a user. The Concatenating Regression Module consists of three fully-connected layers, as shown in the right part of Figure 7. The first layer encodes a stitching vector that embeds the feature embeddings obtained from the previous two modules. Based on the stitching vector, the second and the third layers predict the coordinates of skeleton joints in the 3D Cartesian coordinate system. The operation of the Concatenating Regression Module can be expressed as:

$$\hat{P} = \mathrm{G}\left(\mathrm{Concat}\left(Z, Z'\right), \hat{\Theta}\right), \tag{6}$$

where $\hat{P}$ is the output of the predicted 3D skeleton joints, $\mathrm{G}(\cdot)$ is the coding regression network operation, $\mathrm{Concat}(\cdot)$ is the feature connection operation, and $\hat{\Theta}$ is the trainable parameters. Finally, the regression module outputs the predicted 3D coordinates of skeleton joints of a user, once the whole deep learning model is trained.

Loss Function Design. To train the deep learning model for obtaining the 3D skeleton joints of a user, we design a loss function according to the distance between the predicted skeleton joints and the ground truth label obtained by Kinect. The loss function is designed based on Smooth L_1 loss [3], i.e.,

$$L_p = \begin{cases} \frac{1}{2}(P-\hat{P})^2 & |P-\hat{P}| \le \delta \\ \\ \delta(|P-\hat{P}| - \frac{1}{2}\delta) & |P-\hat{P}| > \delta \end{cases}, \tag{7}$$

where P is the ground truth of 3D skeleton joint coordinates, and $\hat{P}$ is the output of predicted 3D skeleton joint coordinates. To prevent non-convergence of the model caused by the high penalty coefficient of outliers, an outlier threshold δ is introduced into the loss function for outlier joint detection. The ground truth of 3D skeleton joint coordinates is calibrated by subtracting the coordinates center of the user's joints, which obtains the relative coordinates of the user's joints. This helps to train a robust and accurate skeleton joint estimation model, because the influence of a user's absolute position is eliminated by the operation of ground truth data. By training the parameters Θ, Θ' and $\hat{\Theta}$ with the loss function, the deep learning model finally outputs the 3D coordinates of each user's skeleton joints in multi-user scenarios.

Different from classification models, the proposed deep learning model establishes the space mapping between mmWave profiles and skeleton joint coordinates in regression. Hence, for any mmWave profiles induced by human postures, m^3Track is able to predict the 3D coordinates of corresponding skeleton joints, which achieves the universality towards different postures. With the 3D skeleton joint coordinates of each user, the 3D posture of each user in multi-user scenarios is reconstructed.

6 MULTI-USER 3D POSTURE TRACKING

Although m^3Track reconstructs the 3D postures for each user in multi-user scenarios, the reconstructed 3D postures of the users are not accurately mapped to real-world 3D space. To realize 3D posture tracking, m^3Track needs to map the reconstructed postures of all users into real-world 3D space, and track the position of each user in the 3D space for multi-user 3D posture tracking.

6.1 Posture Mapping with Point Cloud

To map the postures of all users into 3D space, m^3Track first generates point clouds to acquire specific position information of all users in the 3D space, and further maps the reconstructed 3D postures of users with the generated point clouds in the 3D space.

For mmWave-based point cloud generation, m^3Track generates the point clouds of all users based on the constant false alarm rate algorithm (CFAR) [22], which is a classical target detection method to generate point clouds for radar systems. After the point clouds of all users are generated, m^3Track further maps the reconstructed posture of each user with the generated point clouds in 3D space. However, we are unaware of the mapping relations between the postures and the point clouds in 3D space. To map the postures

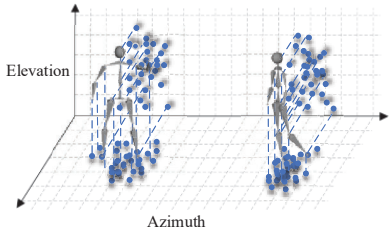


Figure 8: Multi-user 3D posture mapping.

with the point clouds for posture tracking, we establish the mapping relation by finding the minimal mapping errors between postures and point clouds, as shown in Figure 8. Specifically, for every point cloud, m^3Track clusters the points to multiple clusters through K-Means algorithm [14]. Each cluster has a cluster center, which corresponds to a posture joint. If there is a minimal distance error between all the joints and cluster centers, m^3Track considers that the posture is accurately mapped with the point cloud. Hence, m^3Track maps the 3D postures of all users with the point clouds by finding the minimal mapping error between postures and point clouds, which is denoted as:

$$m(i,k) = \arg\min \sum_{i=1}^{P} \sum_{j=1}^{J} (Joint_{k,j} + \hat{s} - X_{i,j})^2, \tag{8}$$

where $m(i,k)$ is the optimal mapping relation between point clouds and reconstructed postures, P is the number of observable point clouds, J is the number of observable joints, $Joint_{k,j}$ is the j-th joint for k-th posture, $\hat{s}$ is the correction displacement, and $X_{i,j}$ is the j-th cluster center for the i-th point cloud. By solving the optimization problem, we can obtain an optimal mapping relation between all the postures and point clouds, so the reconstructed 3D postures of all users can be mapped into real-world 3D space.

6.2 3D Posture Tracking

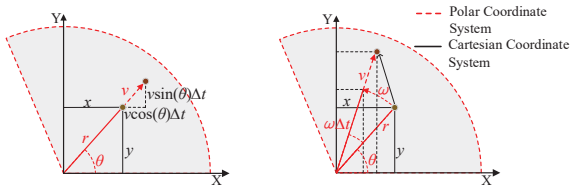
Based on the mapped 3D postures of all users, m^3Track further tracks the positions of the users in 3D space to realize multi-user 3D posture tracking through a proposed coordinate-corrected extended Kalman filter (coordinate-corrected EKF).

EKF [21] is a classical state estimation method, which can be utilized to track targets' positions. The basic idea of KEF is to combine the estimated position and measured position for deriving an accurate position of a target, which mitigates measurement errors and interferences during tracking. However, the mmWave radar works in the polar coordinate system while the 3D human postures are reconstructed in the Cartesian coordinate system. Hence, we present a coordinate-corrected extended Kalman filter, which applies EKF on the two coordinate systems, for accurately tracking the position of users in the Cartesian coordinate system.

Specifically, in the coordinate-corrected EKF, we first define a state function to describe the current state of the tracking target, which can be expressed as:

$$s_t = (x \quad y \quad v \quad \theta \quad \omega), \tag{9}$$

where (x, y) is the position of the target in the Cartesian coordinate system, v is the Doppler (i.e., radial speed) in the polar coordinate system, θ is the deflection angle in the polar coordinate system, and ω is the deflection angle speed in the polar coordinate



(a) Deflection angle speed $\omega = 0$. (b) Deflection angle speed $\omega \neq 0$.

Figure 9: Relations in two coordinate systems.

system. Figure 9 shows geometrical relations of the target's movement in the two coordinate systems. With the initial definition and the geometrical relations, we estimate the state function in every step Δt as:

$$s_{t+\Delta t} = \begin{cases} \begin{pmatrix} v\cos(\theta)\Delta t + x \\ v\sin(\theta)\Delta t + y \\ v \\ \theta \\ \omega \end{pmatrix} & \text{if } \omega = 0, \\ \begin{pmatrix} (\frac{x}{\cos(\theta)} + v\Delta t) \cdot \cos(\omega\Delta t + \theta) \\ (\frac{y}{\sin(\theta)} + v\Delta t) \cdot \sin(\omega\Delta t + \theta) \\ v \\ \omega\Delta t + \theta \\ \omega \end{pmatrix} & \text{if } \omega \neq 0. \end{cases} \tag{10}$$

After the estimation of the state function at $t+\Delta t$, m^3Track further calculates the covariance of the state function. Then, m^3Track measures the ground-truth value of state function at $t + \Delta t$, i.e., $z_{t+\Delta t}$, and also calculates its covariance correspondingly. Based on the above variables, m^3Track combines the estimation and measurement to derive an accurate state of the target, i.e.,

$$\hat{s}_{t+\Delta t} = K_{t+\Delta t} \cdot z_{t+\Delta t} + (I - K_{t+\Delta t}H)s_{t+\Delta t}, \tag{11}$$

where $K_{t+\Delta t}$ is the Kalman Gain calculated by the two covariances, H is a transformation matrix, and I is a unit matrix. Hence, with the combination of the estimation and measurement, m^3Track obtains a more accurate state (i.e., position) of the target in $\hat{s}_{t+\Delta t}$. Through the iteration of the above process, the position trajectories of user objects are accurately and continuously tracked.

With the position trajectories, m^3Track continuously tracks the positions of the mapped postures in real-world 3D space. However, in practical scenarios, people may block each other sometimes, so that the mmWave signals cannot sense the blocked user due to signal propagation blockage. To obtain the walking trajectory of a transiently blocked user, m^3Track estimates the position of the blocked user based on the estimation step of the coordinate-corrected EKF, i.e., (x, y) in $s_{t+\Delta t}$. Hence, m^3Track still captures the position trajectory of each user when they are transiently blocked. Given that the crossover scenes are brief during walking, even if the posture of the blocked user is missing, m^3Track can still continuously track the user's position, which enables robust and practical 3D posture tracking in multi-user scenarios.

7 EVALUATION

In the section, we conduct experiments to evaluate m^3Track's performance in real environments.

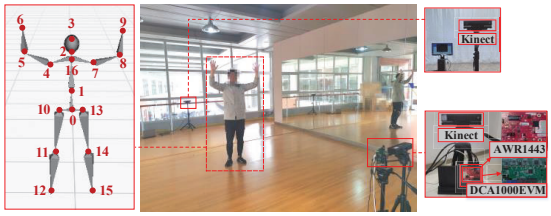


Figure 10: Experimental settings in the lab.

7.1 Evaluation Setup

Testbed. m^3Track is implemented with a single commercial off-the-shelf (COTS) mmWave radar, i.e., Texas Instruments(TI) AWR1443 mmWave radar [7], as the sensing front. The mmWave radar is equipped with three onboard transmit antennas and four receive antennas. It is configured to generate mmWave chirp signals with a bandwidth of $77 \sim 81GHz$ and a signal frame of 128 pulses within $50ms$. The sampling rate is 512 sampling points for each pulse. The mmWave radar is connected with a TI DCA1000EVM data capture card [8] to achieve high-speed data transmission between the mmWave radar and the back end. The back end is a DELL G15 laptop for reading and processing mmWave data. We utilize two Kinect V2 [16] to capture the ground truth of human postures. With the RGB and infra-red cameras, the Kinect devices can capture the depth images of users and generate human skeleton joints.

Environmental Settings. The experiments are conducted in different environments, including indoor areas (e.g., lab, corridor, meeting room), and outdoor spaces. Figure 10 shows the environmental settings in the lab. In specific, a mmWave radar is placed in the environment, continuously emitting FMCW signals towards the orientation of the radar's antenna panel, and continuously receiving the FMCW signals reflected by users. The two Kinect devices are placed in opposite positions to capture the ground truth postures, where one Kinect is placed closely around the mmWave radar and the other is placed on the opposite side of the mmWave radar with a distance of $8m$. The mmWave radar and Kinect devices are placed in the same horizontal plane of $1.3m$. The users perform activities in front of the mmWave radar. The distance of users toward the mmWave radar is from a minimum of $1.2m$ to a maximum of $7m$. The settings of device placement in other environments are the same as that in the lab.

Data Collection. In terms of posture tracking, we select 17 skeleton joints from the human body, as shown in Figure 10. With the key joint nodes supporting the human body, m^3Track is able to reconstruct the human postures as users move, walk, or sit. We recruit 15 volunteers to participate in the experiments. The volunteers naturally perform various daily activities in the sensing area, including in-place activities (such as lifting arms, lifting legs, squatting, etc.), and walking activities (such as walking forward, walking back, walking across, etc.). The ground truth posture joints captured by Kinect are calibrated by subtracting the coordinates center of a user's joints to obtain relative coordinates.

Model Training and Testing. We leverage 2 users' data as the training data for the deep learning model. The 2 users perform in-place activities and walking activities in the lab to provide training data. As mentioned in Section 5.2, since the deep learning-based human posture reconstruction is implemented for an individual user, the 2 users' mmWave data are collected individually, each of which contains 7200 frames collected within the training data collection process. Together with the skeleton joint coordinates from Kinect as ground truth, the 2 users' data are fed to the deep learning model for training. During the training of the system, the learning rate of the deep learning model is set to 0.001. The batch size is 32. The number of epochs is 200. We use Keras to implement the deep learning model in the backend laptop with Intel i7-11800H and NVIDIA RTX3060. The other 13 users' data are leveraged for evaluating the system's performance. The evaluation process lasts about 27 hours within 19 days in different environments, including the lab, corridor, meeting room, and outdoor space. The 13 users perform arbitrary daily activities including in-place activities and walking activities, which are not necessarily the same as those performed during training. The users are grouped with different user numbers (i.e., 1 to 4) in each evaluation scenario for evaluating multi-user 3D posture tracking capability.

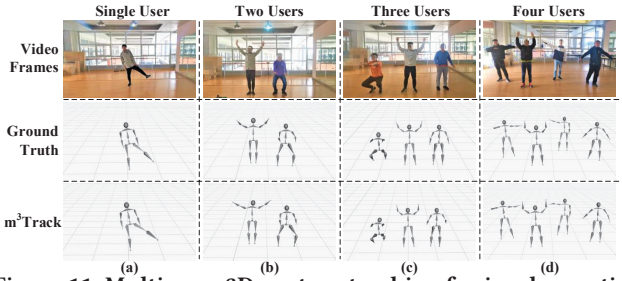


Figure 11: Multi-user 3D posture tracking for in-place activities.

7.2 Overall Performance

We first evaluate the overall performance of m^3Track on multi-user 3D posture tracking by intuitively exhibiting the reconstructed postures in 3D space. Figure 11 shows 3D postures tracking results for in-place activities with corresponding ground truth and video frames in the lab. It can be observed from Figure 11 that m^3Track accurately reconstructs and tracks the 3D posture of each user. Specifically, we can see from Figure 11(a) that the skeleton of the single user is effectively reconstructed by the proposed m^3Track compared with the ground truth. Then, when multiple users simultaneously perform activities in the sensing environment, as shown in Figure 11(b), 11(c), and 11(d), m^3Track is also capable of constructing the skeletons of each user and accurately tracking the posture dynamics with little interference. The examples demonstrate that m^3Track can effectively reconstruct and track the 3D human postures of multiple users.

Besides posture tracking for in-place activities, we also evaluate the performance of m^3Track on tracking the walking users. Figure 12 shows the consecutive 3D human postures reconstructed by m^3Track when tracking three walking users in the lab. It can be first seen that m^3Track effectively reconstructs the 3D postures for the walking users, including the details of arm swing and leg stepping, compared to the real scenes and ground truth. In addition, the reconstructed postures follow the real scenes of users' positions continuously. The results demonstrate that the proposed tracking solution is effective in mapping the reconstructed postures into real-world 3D space and continuously tracking the walking trajectories of each user. Therefore, m^3Track is able to realize practical multi-user 3D posture tracking.

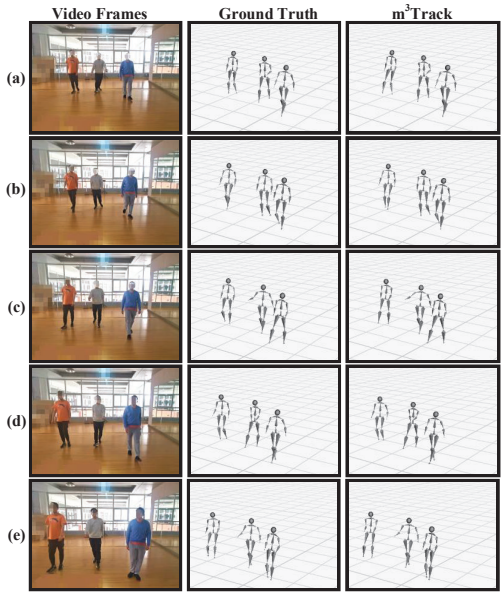


Figure 12: Multi-user 3D posture tracking for walking activities.

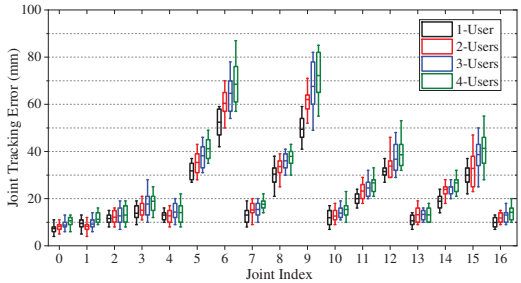


Figure 13: Joint tracking errors.

7.3 Quantitative Results

We also quantitatively evaluate m^3Track by measuring the joint tracking errors, which is the Euclidean distance of skeleton joint coordinates between tracked postures and ground truth postures, in depth (X), azimuth (Y), and elevation (Z). Table 1 shows the joint tracking errors for different user scenarios. Specifically, the overall joint tracking error is 32.4mm, 34.9mm, 38.6mm, and 42.4mm for 1-user, 2-user, 3-user, and 4-user scenarios, respectively, demonstrating an effective multi-user 3D posture tracking of m^3Track. Compared to single-user scenarios, tracking the postures of multiple users simultaneously only introduces a small increase in tracking errors. Besides the overall joint tracking errors, the errors in depth, azimuth, and elevation have a similar tendency under different user numbers. The results demonstrate that m^3Track effectively extends 3D posture tracking from single-user scenarios to multiple-user scenarios.

Table 1: Tracking errors under different user numbers.

Users	1	2	3	4
Overall	32.4mm	34.9mm	38.6mm	42.4mm
Depth	14.4mm	15.5mm	16.9mm	17.4mm
Azimuth	22.2mm	24.6mm	28.4mm	32.5mm
Elevation	18.7mm	19.3mm	20.0mm	20.9mm

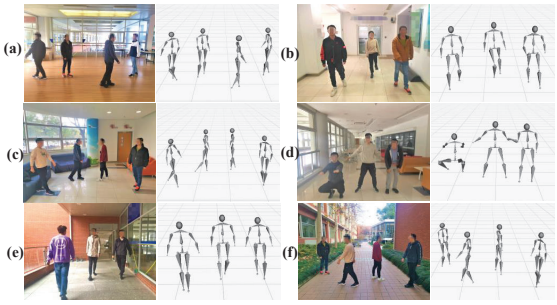


Figure 14: Multi-user 3D posture tracking in different environments.

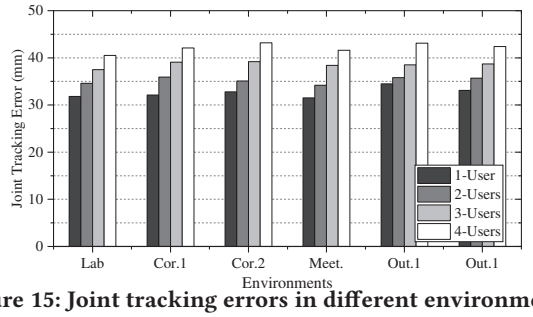


Figure 15: Joint tracking errors in different environments.

Furthermore, we study the tracking errors of each joint in detail. Figure 13 shows the detailed joint tracking errors under different user numbers. The corresponding relationship between the 17 joints and the human body is shown in Figure 10. It can be observed that there is a variation in tracking errors among different joints. Specifically, the joints corresponding to the arms (No. 5, 6, 8, 9) and the legs (No. 11, 12, 14, 15) have relatively higher tracking errors, and also relatively large increases from single-user scenarios to multi-user scenarios, compared to other joints. This is because the movements of the limbs are more complex and variable, and sometimes they are obscured by other body parts in multi-user scenarios. However, although tracking limbs have relatively higher errors, the maximum error on tracking hands is still less than 90mm, which demonstrates the effectiveness of m^3Track on tracking different parts of the human body.

7.4 Performance in Different Environments

To demonstrate the feasibility and robustness of m^3Track in broader scenarios, we further evaluate the 3D posture tracking performance of m^3Track in different environments. The experimental environments include a lab, two corridors, a meeting room, and two outdoor spaces, with different space sizes and environmental layouts.

Figure 14 shows the examples of multi-user 3D posture tracking in different environments. For each user in the environment, m^3Track effectively tracks the 3D posture of the user, which is less affected by the environmental layouts. For example, in Figure 14(b), although there are walls right next to the users, m^3Track is still able to track the postures of the users without being affected by the multipath reflection of the walls. Moreover, in the two outdoor spaces, i.e., Figure 14(e) and 14(f), the postures of all the users are tracked accurately, indicating the effectiveness of multi-user 3D posture tracking in outdoor spaces.

Then, we quantitatively evaluate the performance of multi-user 3D posture tracking in different environments. Figure 15 shows

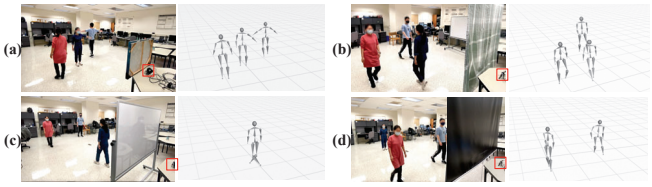


Figure 16: Multi-user 3D posture tracking in occluded scenarios.

the joint tracking errors under different user scenarios in the 6 environments, respectively. It can be observed that the joint tracking errors among different environments have few differences. Specifically, the standard deviation of joint tracking errors in different environments is only 3.2*mm*. Hence, m^3Track achieves effective and robust multi-user 3D posture tracking in different environments, which can work on a broad range of scenarios.

7.5 Performance in Occluded Scenarios

To investigate the capability of m^3Track working in non-line-of-sight (NLOS) conditions, we experiment on 3D posture tracking in occluded scenarios. Specifically, we place 4 different barriers between mmWave radar and users as occlusions, i.e., a mural painting, a cloth screen, a whiteboard, and a projector screen, respectively. With the mmWave radar placed right behind a barrier, m^3Track senses and tracks users behind the barrier.

Figure 16 shows the scenarios of 4 different barriers and corresponding results of m^3Track. For the mural painting and cloth screen, m^3Track is capable of tracking the 3D postures of all users, as shown in Figure 16(a) and 16(b). The average joint tracking errors under the two barriers are 45.7*mm* and 44.2*mm* respectively, which is close to that in non-occluded scenarios. However, for the barriers of whiteboard and projector screen, m^3Track loses the details of the closer users and the main information of further users, leading to an incomplete multi-user 3D posture tracking results, as shown in Figure 16(c) and 16(d). The reason is that the barriers of complex structures and hardly-penetrated materials cause significant amplitude attenuation and phase change of signals, so the range, angle, and Doppler of users could not be accurately measured. The results demonstrate that m^3track only works under barriers of simple structures and easily-penetrated materials.

7.6 Performance of User Joining and Leaving

We evaluate the 3D posture tracking performance of m^3Track in user joining and leaving scenarios. In the experiment, with 2 users already existing in the sensing area, another user walks into the area and then leaves. Figure 17 shows the reconstructed 3D posture and corresponding video frames when the user joins and leaves in a multi-user scenario. It can be observed that m^3Track is able to accurately reconstruct and track all the users in the sensing area when a user enters and leaves. Specifically, when the user walks into the sensing area, as shown from Figure 17(a) to 17(b), m^3Track detects the presence of the user at a specific frame and starts tracking the user's posture. Later, as the user leaves in the sensing area shown from Figure 17(c) to 17(d), m^3Track loses the information about the user at a certain frame. The above results demonstrate that the proposed m^3Track is able to accurately track multiple users in user joining and leaving scenarios.

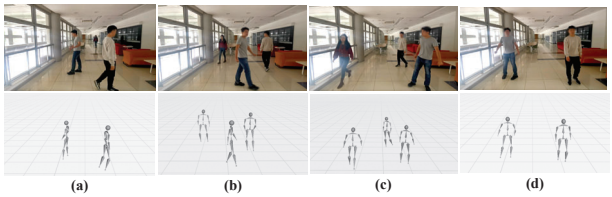


Figure 17: Multi-user 3D posture tracking for user joining and leaving.

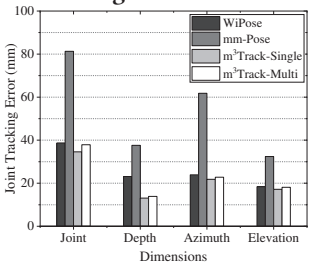


Figure 18: Joint tracking errors for m^3Track, *WiPose*, and *mm-Pose*.

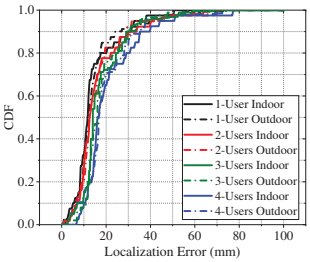


Figure 19: Localization errors in different environments.

7.7 Comparison with SOTA Systems

We compare the performance of posture tracking of the proposed system with two state-of-the-art (SOTA) systems, i.e., *WiPose* [10], and *mm-Pose* [24]. *WiPose* uses commodity WiFi devices, extracting CSI data and 3D velocity profiles, to construct 3D skeletons for a single user through an RNN-based deep learning model. And *mm-Pose* uses COTS mmWave radar to extract mmWave range-angle patterns for estimating and constructing a single user's skeleton through a CNN-based learning model. Besides the two systems, there are other RF-based skeleton construction and tracking systems (e.g., *RF-Pose3D* [38]), but it is difficult to implement these systems for comparison due to the dedicated hardware design. Hence, we leverage the two SOTA systems that work on COTS devices for comparison. Specifically, we implement *WiPose* and *mm-Pose* using WiFi testbed and mmWave radar respectively. Since the two systems are designed for single-user scenarios, we compare the joint tracking errors by collecting the same single-user experimental data. In addition, we also exhibit the average joint tracking errors of m^3Track in multi-user scenarios for comparison.

Figure 18 shows the joint tracking errors of *WiPose*, *mm-Pose*, m^3Track in single-user scenarios, and m^3Track in multiple-user scenarios, respectively. It can be observed that m^3Track has the minimum errors on posture reconstruction compared to other systems. Specifically, compared to *WiPose* that has an average 38.7*mm* joint tracking error, m^3Track achieves similarly effective performance for both single-user scenarios and multi-user scenarios. However, *WiPose* cannot construct the postures of multiple users simultaneously, while m^3Track expands human posture tracking to multi-user scenarios. For another COTS mmWave-based method *mm-Pose*, it has relative high joint tracking errors in joints (81.3*mm*) and other dimensions (37.6*mm*, 61.8*mm*, and 32.4*mm*). This is because it only leverages a CNN to extract spatial features but discards the temporal relations between different moments. In contrast, m^3Track incorporates convolutional operation and temporal relations together, which achieves more effective multi-user 3D posture tracking based on the same COTS mmWave radar.

Figure 20: Posture tracking in adjacent distances.

7.8 Localization Performance

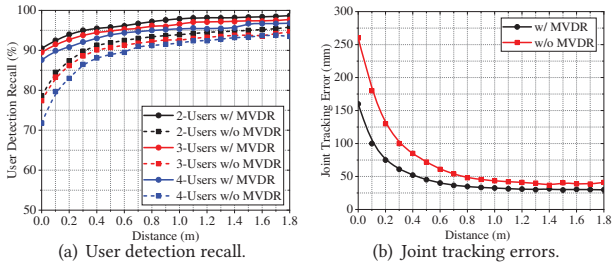
Since m^3Track constantly localizes all the users and tracks their walking trajectories in the 3D space, it is necessary to evaluate the localization performance for multiple users in m^3Track. To cover a wide range of scenes, the experiment is conducted in an indoor environment (the lab) and outdoor space (as shown in Figure 14(f)). In the experiment, we calculate the geometric center of the mapped posture in 3D space as the predicted location of each user. Since it is difficult to get the accurate user locations through video frames, the ground truth of user location is marked on the floor, and the users follow the marked trajectories for the evaluation.

Figure 19 shows the cumulative distribution function (CDF) of localization errors under different user numbers in two environments respectively. It can be observed that the user numbers and environments have little impact on the localization performance of m^3Track. Specifically, m^3Track achieves a median localization error of 18.9mm with a standard deviation of 11.4mm for single-user scenarios, and 21.5mm with a standard deviation of 13.5mm for 4-user scenarios, which have little differences. Compared with *RF-Pose3D* [38] that achieves 17mm, 28mm, and 23mm localization error on X, Y, and Z for multiple users through specially-designed RF radar, our system achieves a comparable localization performance with only a COTS mmWave radar. The results demonstrate that the proposed m^3Track can accurately localize each user and continuously track the user's trajectory in multi-user scenarios.

7.9 Impact of Distance between Users

We further evaluate the impact of distance between users on separating multiple users and tracking users' postures. Since we utilize a MVDR approach to improve the signal-to-noise rate for refining the angle of different users and thus separating users, we evaluate the impact of user distance with and without the MVDR approach, respectively. Figure 20 shows the examples of posture reconstruction for two users with and without the MVDR approach respectively, where the two users stand closely and their arms are adjacent. It can be observed that the m^3Track with the MVDR-based approach can effectively separate the two users and accurately reconstruct the posture of each user. However, if the MVDR approach is not employed, the two users are not well separated in the adjacent body part. This indicates that multiple users can be separated and tracked at close distances with the MVDR approach.

Furthermore, we evaluate the user detection recall, which is the fraction of users that are detected and separated over the total amount of users, to measure m^3Track's ability on separating all the users without misses in different distances. The distance of users is measured by the linear distance of users' nearest body parts. Figure 21(a) shows the user detection recall in different distances for different user numbers. As the distance between users decreases, the user detection recall also decreases, which indicates that close



(a) User detection recall. (b) Joint tracking errors.

Figure 21: Performance under different user distances.

distance affects multi-user separation. However, with MVDR applied, even if the users are adjacent, m^3Track can still separate each user with an average of 90.5%, 89.5%, and 87.6% accuracy for 2-user, 3-user, and 4-user scenarios respectively, which demonstrates the robustness of m^3Track on detecting and separating multiple users.

Moreover, we evaluate joint tracking errors under different distances with and without the MVDR-based approach respectively, as shown in Figure 21(b). We can see that the tracking errors are low and stable beyond a distance of 0.6m, and increase rapidly as the distance gets smaller. However, with the MVDR-based approach, m^3Track can realize multi-user 3D human posture tracking with only 51.7mm errors on average even if the distance between users is only 0.4m. According to [25], a suitable face-to-face communication distances of people is between 0.46m and 1.22m. Hence, m^3Track can effectively track multiple users' postures under the majority of suitable communication distances.

7.10 Impact of Distance to Radar

We further quantitatively evaluate joint tracking errors of posture tracking when users are of different distances to radar. In the experiment, we leverage the data in 4-user scenarios under the 1.2m to 7m experimental space, record the straight-line distance between each user and the radar, and measure joint tracking errors for each user in different distance ranges. Table 2 shows the average joint tracking errors for each user at different distance ranges to radar. We can see that with the increase of distance, the joint tracking errors increase slightly due to the attenuation of mmWave signals in long propagation, but the difference in joint tracking errors between the optimal and worst distance range is only around 20mm on average. The result demonstrates that m^3Track is less sensitive to different distances to radar.

7.11 Impact of Body Shape and Clothes

We further evaluate the performance of m^3Track when users are of different body shapes and wear different clothes. In the experiment, we recruit 2 users, in which one female and one male. The female user is of 1.53m height and 50Kg weight while the male user is of 1.81m height and 82Kg weight. The two users wear heavy coats and light clothes in the experiment, respectively. Figure 22 shows the examples of 3D posture reconstruction for the 2 users. We can see that m^3Track effectively reconstructs the postures of users who are of different body shapes. This is because m^3Track estimates the skeleton joint coordinates of 3D postures, and the 3D

Table 2: Joint tracking errors of different distances to radar.

Distance(m)	1.2-2.0	2.0-3.0	3.0-4.0	4.0-5.0	5.0-6.0	6.0-7.0
Error(mm)	39.3	38.4	40.1	42.0	48.6	58.7

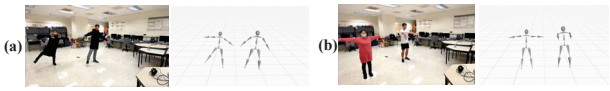


Figure 22: Multi-user 3D posture tracking with different body shapes and clothes.

modeling of such coordinates exhibits a user's body shape. Also, it can be observed that users with different clothes are effectively reconstructed. The reason is that the mmWave signals can effectively propagate cloth materials and thus 3D human posture reconstruction is less affected by clothes. Therefore, m^3Track is capable of reconstructing users with different body shapes and clothes.

7.12 Time Consumption

We evaluate the time consumption of m^3Track to explore whether it is real-time for practical use. Using the laptop with Intel i7-11800H and NVIDIA RTX3060 as the back end, the time consumption of each module for tracking 4 users is measured, as shown in Table 3. Among them, the mmWave-based sensing module is the most time-consuming process, which costs $251ms$ to read IF signals using UDP protocol. In addition, the prediction of skeleton joints using the forked-ConvLSTM model also spends $214ms$. In total, m^3Track costs $745ms$ to output the tracking results for each frame. Such a processing time is close to real-time for general applications. Furthermore, if a shorter response time is needed to facilitate more delicate and real-time applications, m^3Track can be implemented on devices with more computing power.

8 RELATED WORK

In this section, we review works related to m^3Track.

mmWave Sensing. With the development of $60GHz$ WiFi and 5G cellular communication, millimeter wave technology (mmWave) becomes ubiquitous in daily life and industrial manufacturing. Due to its high popularity, researchers are attracted to design mmWave-based sensing solutions to extend their capabilities from cellular communication to ubiquitous sensing for the physical world. For example, many efforts have been made to utilize mmWave signals for promoting the realization of smart homes through mmWave-based behavior recognition [15, 35], acoustic sensing [12, 31], vital sign monitoring [33, 34], human identification [4], etc. Other researches focus on facilitating the deployment of mmWave sensing in industrial manufacturing, such as vibration measurement [9], and car imaging [5]. All these studies have proved the capability of millimeter waves in sensing areas.

Human Posture Reconstruction. Reconstructing human postures has drawn considerable concern and has been widely explored. A popular approach to realizing human posture reconstruction is based on dedicated vision cameras [1, 20], but they depend on lighting conditions and suffer from privacy concerns. Towards privacy-preserving manner, others leverage RF signals to realize single-user posture reconstruction in contact-free manner, such as Wi-Fi [10] and mmWave [23, 24, 32]. However, all the works only enable single-user posture or mesh reconstruction, while the more complex and common multi-user scenarios cannot be achieved. Although some recent pioneer studies [36–38] realize multi-user posture reconstruction, they strictly rely on specialized hardware design (e.g., requiring to equip at least 20 antennas), making it difficult to be practically deployed. Along the direction, a human posture reconstruction method for multi-user scenarios using a COTS RF device is highly desirable.

Indoor Localization and Tracking. Indoor localization and tracking in a device-free manner have been widely studied in recent years, supporting a wide range of applications that requires knowing the location of users. Among existing methods, wireless signals have been extensively utilized, emerging a variety of indoor localization works. For example, some studies leverage Wi-Fi signals for achieving localization and tracking in indoor environment [13, 17–19, 28, 29]. However, WiFi-based approaches are usually susceptible to environmental interferences. Besides Wi-Fi signals, the emerging mmWave signals are also widely exploited for indoor localization and tracking in device-free manner [2, 27, 27, 39, 39]. Although multi-user scenarios are considered, these researches are mostly limited to localizing users, which cannot track the dynamic postures of users as they move, walk, or sit.

Table 3: Time consumption of different modules.

Module	Sense	Separate	Reconstruct	Track	Total
Time(ms)	251	153	214	127	745

9 DISCUSSION

In this section, we discuss several practical issues.

Overlap Between Users. The overlap between users is a practical situation in multi-user scenarios. In m^3Track, the overlap between users causes the loss of mmWave signals for sensing blocked users. For example, if user A is behind user B, user B blocks the propagation of mmWave signals, so that m^3Track cannot transmit the mmWave signals to user A for sensing the user. Hence, m^3Track is unable to reconstruct the 3D human postures of blocked users in overlap scenarios. However, if users are only transiently blocked by others, m^3Track can potentially utilize the coordinate-corrected EKF to track the position trajectories of blocked users.

Complex Indoor Structure. m^3Track works in different environments and even occluded scenarios. However, these environments are of wide area and with less furniture. Like other wireless signals, the propagation of mmWave signals is also affected by complex environmental structures. An indoor environment with more complex structures, such as a classroom or a canteen, may cause significant signal amplitude attenuation and phase change, leading to inaccurate sensing and tracking for the users.

10 CONCLUSION

In this paper, we propose m^3Track, which utilizes a single COTS mmWave radar to realize 3D postures tracking for multi-user scenarios. m^3Track first detects and separates all users on mmWave signals, and designs a novel deep learning model to reconstruct 3D posture for each user. After that, m^3Track maps the reconstructed 3D postures of all users into real-world 3D space, and continuously tracks the positions of the users, realizing practical multi-user 3D posture tracking. Extensive experiments in real-world multi-user scenarios validate the accuracy and robustness of m^3Track.

ACKNOWLEDGMENTS

This research was sponsored by NSFC (No. 62172277, 62141220, 61972253, U1908212, 72061127001).

REFERENCES

[1] Bernard Boulay, François Bremond, and Monique Thonnat. 2003. Human posture recognition in video sequence. In *Proc. IEEE VS-PETS '03.*

[2] Han Cui and Naim Dahnoun. 2021. High Precision Human Detection and Tracking Using Millimeter-Wave Radars. *IEEE Aerospace and Electronic Systems Magazine* 36, 1 (2021), 22–32.

[3] Ross Girshick. 2015. Fast r-cnn. In *Proc. IEEE ICCV '15.* 1440–1448.

[4] Tianbo Gu, Zheng Fang, Zhicheng Yang, Pengfei Hu, and Prasant Mohapatra. 2019. mmSense: Multi-Person Detection and Identification via mmWave Sensing. In *Proc. ACM mmNets@MobiCom '19.* Los Cabos, Mexico, 45–50.

[5] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. 2020. Through Fog High-Resolution Imaging Using Millimeter Wave Radar. In *Proc. IEEE/CVF CVPR '20.* WA, USA, 11461–11470.

[6] Emanuël Anco Peter Habets, Jacob Benesty, Israel Cohen, Sharon Gannot, and Jacek Dmochowski. 2009. New insights into the MVDR beamformer in room acoustics. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 1 (2009), 158–170.

[7] Texas Instruments. 2021. AWR1443 Single-chip 76-GHz to 81-GHz automotive radar sensor integrating MCU and hardware accelerator. [Online]. Available: https://www.ti.com/product/AWR1443.

[8] Texas Instruments. 2021. DCA1000EVM Real-time data-capture adapter for radar sensing evaluation module. [Online]. Available: https://www.ti.com/tool/DCA1000EVM.

[9] Chengkun Jiang, Junchen Guo, Yuan He, Meng Jin, Shuai Li, and Yunhao Liu. 2020. mmVib: micrometer-level vibration measurement with mmwave radar. In *Proc. ACM MobiCom '20.* London, United Kingdom, 45:1–45:13.

[10] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using wifi. In *Proc. ACM MobiCom '20.* London, United Kingdom, 23:1–23:14.

[11] Wenchao Jiang and Zhaozheng Yin. 2015. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proc. ACM MM '15.* Brisbane Australia, 1307–1310.

[12] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, and Wenyao Xu. 2020. VocalPrint: exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proc. ACM SenSys '20.* Japan, 312–325.

[13] Xiang Li, Shengjie Li, Daqing Zhang, Jie Xiong, Yasha Wang, and Hong Mei. 2016. Dynamic-music: accurate device-free indoor localization. In *Proc. ACM UbiComp '16.* Heidelberg, Germany, 196–207.

[14] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition* 36, 2 (2003), 451–461.

[15] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. 2020. Real-time Arm Gesture Recognition in Smart Home Scenarios via Millimeter Wave Sensing. *Proc. ACM IMWUT* 4, 4 (2020), 140:1–140:28.

[16] Microsoft. 2021. Kinect for Windows. [Online]. Available: https://developer.microsoft.com/en-us/windows/kinect/.

[17] 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 *Proc. ACM MobiHoc '17.* Chennai, India, 6.

[18] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2.0: Passive Human Tracking with a Single Wi-Fi Link. In *Proc. ACM Mobisys '18.* Munich, Germany, 350–361.

[19] Muhannad Quwaider and Subir Biswas. 2008. Body posture identification using hidden Markov model with a wearable sensor network. *Bodynets* 8 (2008), 1–8.

[20] Zhou Ren, Junsong Yuan, Jingjing Meng, and Zhengyou Zhang. 2013. Robust part-based hand gesture recognition using kinect sensor. *IEEE transactions on multimedia* 15, 5 (2013), 1110–1120.

[21] Maria Isabel Ribeiro. 2004. Kalman and extended kalman filters: Concept, derivation and properties. *Institute for Systems and Robotics* 43 (2004), 46.

[22] Hermann Rohling. 1983. Radar CFAR thresholding in clutter and multiple target situations. *IEEE transactions on aerospace and electronic systems* 4 (1983), 608–621.

[23] Arindam Sengupta, Feng Jin, and Siyang Cao. 2020. NLP based Skeletal Pose Estimation using mmWave Radar Point-Cloud: A Simulation Approach. In *Proc. IEEE RadarConf '20.* Florence, USA, 1–6.

[24] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. 2019. mm-Pose: Real-Time Human Skeletal Posture Estimation using mmWave Radars and CNNs. *CoRR* abs/1911.09592 (2019).

[25] Agnieszka Sorokowska, Piotr Sorokowski, Peter Hilpert, Katarzyna Cantarero, Tomasz Frackowiak, Khodabakhsh Ahmadi, Ahmad M Alghraibeh, Richmond Aryeetey, Anna Bertoni, Karim Bettache, et al. 2017. Preferred interpersonal distances: a global comparison. *Journal of Cross-Cultural Psychology* 48, 4 (2017), 577–592.

[26] Jianwu Wang, Zhichuan Huang, Wenbin Zhang, Ankita Patil, Ketan Patil, Ting Zhu, Eric J Shiroma, Mitchell A Schepps, and Tamara B Harris. 2016. Wearable sensor based human posture recognition. In *Proc. IEEE Big Data '16.* IEEE, Washington DC, USA, 3432–3438.

[27] Chenshu Wu, Feng Zhang, Beibei Wang, and K. J. Ray Liu. 2020. mmTrack: Passive Multi-Person Localization Using Commodity Millimeter Wave Radio. In *Proc. IEEE INFOCOM '20.* IEEE, Toronto, ON, Canada, 2400–2409.

[28] Jiang Xiao, Kaishun Wu, Youwen Yi, Lu Wang, and Lionel M Ni. 2013. Pilot: Passive device-free indoor localization using channel state information. In *Proc. IEEE ICDCS '13.* IEEE, Philadelphia, Pennsylvania, USA, 236–245.

[29] Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. 2019. mD-Track: Leveraging multi-dimensionality for passive indoor Wi-Fi tracking. In *Proc. ACM MOBICOM '19.* Los Cabos, Mexico, 1–16.

[30] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Proc. NeurIPS '15.* Montreal, Quebec, Canada, 802–810.

[31] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. WaveEar: Exploring a mmWave-based Noise-resistant Speech Sensing for Voice-User Interface. In *Proc. ACM MobiSys '19.* Seoul, Republic of Korea, 14–26.

[32] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proc. ACM MobiSys '21.* Wisconsin, USA, 269–282.

[33] Zhicheng Yang, Parth H Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2016. Monitoring vital signs using millimeter wave. In *Proc. ACM MobiHoc '16.* Paderborn, Germany, 211–220.

[34] Zhicheng Yang, Parth H Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2017. Vital sign and sleep monitoring using millimeter wave. *ACM Transactions on Sensor Networks* 13, 2 (2017), 1–32.

[35] Renyuan Zhang and Siyang Cao. 2018. Real-time human motion behavior detection via CNN using mmWave radar. *IEEE Sensors Letters* 3, 2 (2018), 1–4.

[36] 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 *Proc. IEEE CVPR '18.* Salt Lake City, UT, USA, 7356–7365.

[37] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. 2019. Through-wall human mesh recovery using radio signals. In *Proc. IEEE/CVF ICCV '19.* Seoul, Korea (South), 10113–10122.

[38] 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 *Proc.ACM SIGCOMM '18.* Budapest, Hungary, 267–281.

[39] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. 2019. mID: Tracking and Identifying People with Millimeter Wave Radar. In *Proc. IEEE DCOSS '19.* Santorini, Greece, 33–40.