Multi-Subject 3D Human Mesh Construction Using Commodity WiFi

YICHAO WANG, Florida State University, USA
YILI REN, University of South Florida, USA
JIE YANG*, University of Electronic Science and Technology of China, China

This paper introduces MultiMesh, a multi-subject 3D human mesh construction system based on commodity WiFi. Our system can reuse commodity WiFi devices in the environment and is capable of working in non-line-of-sight (NLoS) conditions compared with the traditional computer vision-based approach. Specifically, we leverage an L-shaped antenna array to generate the two-dimensional angle of arrival (2D AoA) of reflected signals for subject separation in the physical space. We further leverage the angle of departure and time of flight of the signal to enhance the resolvability for precise separation of close subjects. Then we exploit information from various signal dimensions to mitigate the interference of indirect reflections according to different signal propagation paths. Moreover, we employ the continuity of human movement in the spatial-temporal domain to track weak reflected signals of faraway subjects. Finally, we utilize a deep learning model to digitize 2D AoA images of each subject into the 3D human mesh. We conducted extensive experiments in real-world multi-subject scenarios under various environments to evaluate the performance of our system. For example, we conduct experiments with occlusion and perform human mesh construction for different distances between two subjects and different distances between subjects and WiFi devices. The results show that MultiMesh can accurately construct 3D human meshes for multiple users with an average vertex error of $4cm$. The evaluations also demonstrate that our system could achieve comparable performance for unseen environments and people. Moreover, we also evaluate the accuracy of spatial information extraction and the performance of subject detection. These evaluations demonstrate the robustness and effectiveness of our system.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: WiFi Sensing, 3D Human Mesh, Multi-subject Scenarios, Channel State Information (CSI), Deep Learning

ACM Reference Format:
Yichao Wang, Yili Ren, and Jie Yang. 2024. Multi-Subject 3D Human Mesh Construction Using Commodity WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 23 (March 2024), 25 pages. https://doi.org/10.1145/3643504

1 INTRODUCTION

In recent years, tremendous efforts have been put toward 3D human mesh construction, including 3D poses and body shapes, which aim to adapt 3D human mesh construction to a broader range of emerging commercial applications. For example, 3D human meshes can be used in animation and film to create lifelike characters and special effects [3]. 3D human mesh construction is also a key component of augmented reality (AR)/ virtual reality (VR) technology [23], allowing users to interact with digital environments in a more natural and intuitive

[*] Corresponding author: Jie Yang

Authors' addresses: Yichao Wang, Florida State University, 1017 Academic Way, Tallahassee, Florida, USA, 32306, ywang@cs.fsu.edu; Yili Ren, University of South Florida, 4202 E. Fowler Avenue, Tampa, Florida, USA, 33620, yiliren@usf.edu; Jie Yang, University of Electronic Science and Technology of China, Chengdu, Sichuan, China, jie.yang@uestc.edu.cn.

2474-9567/2024/3-ART23
https://doi.org/10.1145/3643504

way. Moreover, 3D human mesh construction can be a preliminary task that leads to or enables other tasks, such as human-robot interaction and neurodegenerative condition diagnosis [13].

To realize 3D human mesh construction, traditional approaches mainly rely on computer vision-based techniques, which construct 3D human mesh by utilizing RGB images collected by single or multiple cameras. However, the computer vision-based approach encounters various challenges in non-line of sight (NLoS) or poor lighting conditions and also incurs large errors when subjects wear baggy clothes [4, 17]. Recent works seek to utilize different sensing modalities to address the challenges faced by computer vision-based approaches. Particularly, owing to the ability that could traverse occlusions and illuminate human subjects and physical objects in dark environments, Radio Frequency (RF) sensing-based methods have attracted more attention, which utilizes the reflections of RF signals for sensing various human activities and objects [6, 8]. Some RF-based systems have been proposed to demonstrate the capability of constructing 3D human mesh by using RF signals [41, 43]. However, these systems are implemented based on dedicated hardware (e.g., USRP) and specialized RF signals (e.g., FMCW), which limits the potential of their consumer-oriented use or wide deployment due to their high cost. Most recently, Wi-Mesh [34] offers an appealing alternative, which reuses commodity WiFi, originally designed for communication, to construct 3D human mesh for potential mass adoption in intelligent wireless environments. Although Wi-Mesh has shown promising performance, its application scope is limited as it only focuses on the mesh construction of a single subject. This makes it unsuitable for more challenging and prevalent multi-subject scenarios, such as multi-subject motion monitoring, multi-subject gaming, *etc.*

In this paper, we propose MultiMesh, a multi-subject 3D human mesh construction system based on commodity WiFi devices. Specifically, MultiMesh can simultaneously construct accurate 3D human meshes for multiple subjects by reusing commodity WiFi devices and signals. Thus, it could be a cost-effective option for widespread adoption in the smart home and Internet of Things (IoT) environments. In addition, compared with computer vision-based systems, our system is not limited by lighting conditions and occlusion scenarios. However, realizing such a 3D human mesh estimation system is non-trivial and there are several challenges that need to be addressed.

The first challenge is how to separate multiple subjects who are very close to each other. WiFi signals reflected off all human subjects in the environment are combined and then received by the WiFi device, making it hard to separate the signals of each person. Inspired by existing approaches such as Wi-Mesh [34], we can leverage multiple antennas on the receiver to form a two-dimensional (i.e., L-shaped) antenna array to estimate the two-dimensional angle of arrival (2D AoA) (i.e., azimuth and elevation) of signal reflections. Such spatial information provides the commodity WiFi with the opportunity to separate multiple users in the azimuth-elevation space. Nevertheless, due to the limited number of antennas on commodity WiFi devices (e.g., several antennas on the WiFi 6 device), the resolvability for people who are very close to each other is insufficient. Therefore, we further propose to exploit additional information of signals from more dimensions [39] which can dramatically improve the resolvability. Thus, we can enhance the separation accuracy to separate the subjects who are close to each other and also identify the body shape and deformations in multiple dimensions. Specifically, we further incorporate the angle of departure (AoD) of the transmitting signal and the length/delay of the single propagation path (i.e., time of flight (ToF)) into the 2D AoA.

The second challenge is the dynamic interference caused by indirect reflections. In particular, we define indirect reflection as the signals that are bounced off one person, then reflected by other people or the environment, and finally received by the receiver. We refer to direct reflections as the signals that are directly reflected by the target human body from the transmitter, and then received by the receiver. Such direct reflections contain meaningful information about human subjects. However, indirect reflections will disturb direct reflections of the human body and thus hinder the 3D human mesh construction. In addition, indirect reflections are hard to be removed as they are associated with the combination of dynamic human movements and the static environment. Our insight to tackle this challenge is that indirect reflection has a longer propagation path and different angles compared with direct reflections of the human body. Thus, we can leverage multi-dimensional information of signal propagation (e.g., ToF and AoD) to distinguish indirect reflections from signals reflected off human subjects and mitigate indirect reflections.

The third challenge is the signal reflections from faraway subjects are very weak. Consequently, it is hard to detect and distinguish the signals of the faraway subject from the noise at any given point in time, especially when there are other subjects that are near the WiFi device presenting strong signals. We refer to this challenge as the near–far problem. To address this challenge, we leverage the fact that the noise-like signal reflections of human subjects are continuous in time and spatial domains, whereas the noises tend to have a random distribution. Hence, we can differentiate the weak signal reflections of faraway subjects from the noises in the continuous spatial-temporal domain. In particular, we utilize a dynamic tracking algorithm to track and predict the trajectory of each user, which could discriminate between noise signals and human subjects.

At last, after we separate multiple subjects and obtain clean 2D AoA images for each subject in the physical space, we feed the images to a deep learning model for the 3D human mesh construction. In the deep learning model, convolutional neural network (CNN) is utilized to extract the high-level spatial features, and the Gated Recurrent Unit (GRU) is leveraged to learn the temporal features from a series of consecutive frames. We also utilize the self-attention mechanism to enhance the contribution of the important frames in the final representation. Then, we divide the human body into five regions, including the torso, left arm, right arm, left leg, and right leg, to learn the deformation of these regions separately. Finally, we adopt the Skinned Multi-Person Linear model (SMPL) model [22] for the final human mesh construction.

We evaluate the system with multiple people in various indoor environments including the conference room, classroom, and laboratory. We perform multi-subject 3D mesh construction across different and unseen environments, as well as for unseen people or the same people with different activities. Moreover, we conduct experiments with occlusion and also perform human mesh construction for different distances between subjects and different distances between the subjects and the WiFi devices. We also evaluate the accuracy of the AoA and ToF extraction, as well as the performance of subject detection. We highlight our main contributions as follows:

- We design a multi-subject 3D human mesh construction system using commodity WiFi devices, expanding the applications of WiFi-based human mesh construction from single-subject to multi-subject scenarios.
- We propose a multi-subject separation method that leverages the 2D AoA, AoD, and ToF information of reflected signals to jointly separate subjects who are close to each other and significantly improve the resolvability of WiFi sensing.
- We mitigate dynamic interferences by removing indirect reflections. We also distinguish the weak signals of faraway subjects with the noises to better detect the signal reflections of subjects.
- We conduct extensive experiments in various environments. We conduct experiments with occlusion and implement human mesh construction for different distances between subjects and different distances between the subjects and the WiFi devices. These evaluations demonstrate that our system can accurately construct 3D human meshes for multiple subjects with unseen people and environments.

2 RELATED WORK

WiFi Sensing. As WiFi techniques and devices become pervasive in daily life and industrial manufacturing, researchers tend to make sense of WiFi and extend their capability from communication to ubiquitous sensing for the physical world. For instance, tremendous efforts have been made to reuse commodity WiFi devices that already exist in the environment for building smart home applications, such as large-scale human activities recognition [33, 35], small-scale human motion detection [9, 30], vital sign monitoring [20, 21, 42], indoor localization [44], person identification [31], and object sensing [26]. Recent researchers focus on estimating 2D or 3D human pose by utilizing commodity WiFi [15, 32]. Different from the applications mentioned above, our work explores achieving 3D human mesh construction using commodity WiFi.

Computer Vision-based 3D Human Mesh Construction. With the proliferation of deep learning algorithms and statistical body models, 3D human mesh construction has drawn considerable attention and has been widely explored. Some pioneering work [2, 29] extract the silhouette and joint information from an image to optimize a statistical body model for the 3D human mesh construction. Recently, there are also some existing works focusing on the task for 3D mesh construction based on images [4, 16, 17]. Moreover, substantial methods are proposed to recover human mesh using RGB video [24, 25, 27]. However, these vision-based approaches all rely on RGB cameras, which have the inherent limitation that their accuracy decreases significantly in NLoS or poor lighting conditions.

RF-based 3D Human Mesh Construction. In recent years, some wireless sensing systems have been developed to infer 3D human mesh. The pioneering work, RF-Avatar presented by Zhao et al. [43] utilizes FMCW RADAR [1] to estimate the 3D human mesh based on the RF reflection from the human body. Then Xue et al. proposed mmMesh [41], which utilizes point cloud data directly exported from the mmWave RADAR to construct the human mesh. However, these systems are all based on dedicated and specialized hardware, which are not scalable for mass deployment due to their high cost. Moreover, Wi-Mesh [34] is developed to estimate 3D human mesh by using the two-dimensional angle of arrival (2D AoA) of the WiFi signal reflections based on commodity WiFi. While Wi-Mesh is only suitable for single-subject scenarios. In contrast, our work explores how to construct human mesh for multiple subjects.

3 SYSTEM DESIGN

3.1 Design Challenges

Our proposed MultiMesh system aims to address the following challenges.

Separation of multiple subjects. Previous WiFi-based 3D human mesh construction system only focuses on a single subject. In contrast to dedicated hardware and signals (e.g., FMCW Radar), commodity WiFi has a limited number of antennas and bandwidth for the separation of multiple subjects who are close to each other in the physical space.

Indirect reflection interference. Indirect reflection represents the signal first reflecting off one human body and then reflecting off other human bodies and the environment before received by the receiver. It interferes with direct reflections of the human body and is received by the receiver. Movements of the human body make indirect reflections highly dynamic and affect the performance of subject separation, especially in multi-subject scenarios.

Dealing with the near-far problem. Signal reflections of faraway subjects have very weak signal strength which could be similar to the noise. It is difficult to detect and distinguish signal reflections of faraway subjects from noises. Thus, it could lead to missing detection of the subjects.

3.2 System Overview

The goal of our work is to construct the 3D human mesh effectively for each subject in multi-subject scenarios. As shown in Figure 1, the commodity WiFi transmitter sends out WiFi signals and we collect the Channel State Information (CSI) measurements of WiFi signals by using the WiFi receiver when multiple subjects are conducting various activities in the sensing area.

We first preprocess the CSI measurements of received signals to remove the random phase offsets. Then, we utilize the L-shaped antenna array on the receiver which is similar to existing work, to estimate the azimuth and elevation (2D AoA) information for separation of multiple subjects. In order to enhance the resolvability of the WiFi signals which have the same 2D AoA, we further incorporate the AoD of transmitting signals and ToF which can represent the length of the propagation path to separate very close subjects.

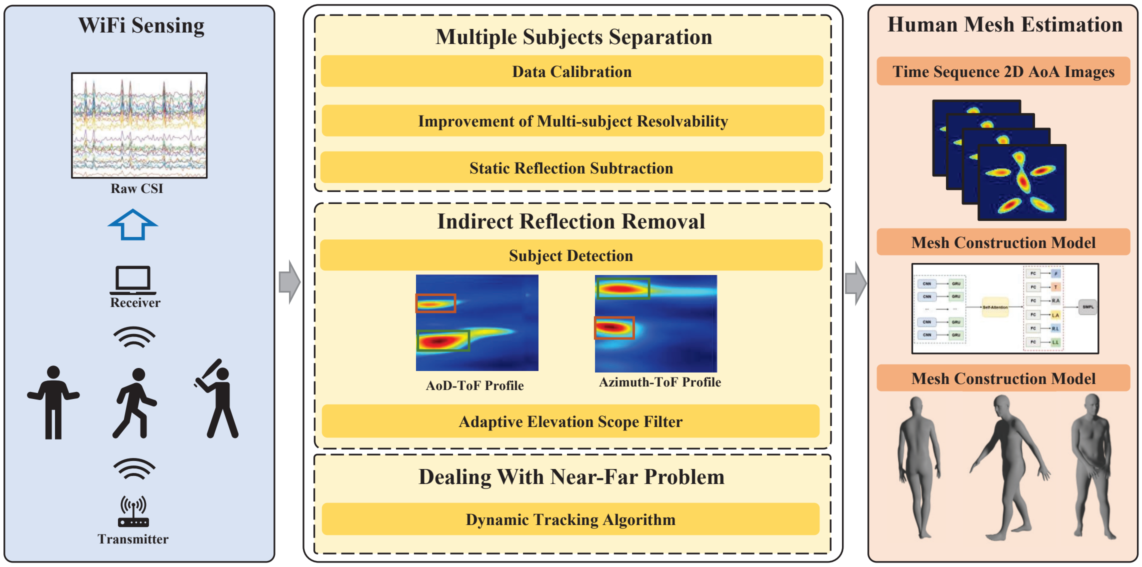

WiFi Sensing
Raw CSI
Receiver
Transmitter
Multiple Subjects Separation
Data Calibration
Improvement of Multi-subject Resolvability
Static Reflection Subtraction
Indirect Reflection Removal
Subject Detection
AoD-ToF Profile
Azimuth-ToF Profile
Adaptive Elevation Scope Filter
Dealing With Near-Far Problem
Dynamic Tracking Algorithm
Human Mesh Estimation
Time Sequence 2D AoA Images
Mesh Construction Model
CNN
GRU
Self-Attention
FC
β
T
R.A
L.A
R.L
L.L
SMPL
Mesh Construction Model

Fig. 1. System Overview.

After that, we mitigate the impact of indirect reflection as it has different propagation paths (i.e., longer propagation distance and different angle of departure) compared with direct reflections from the human body. Thus, we leverage a deep learning detector and propagation path information to detect each subject based on the Azimuth-ToF profile and AoD-ToF profile. The detector could eliminate the interference of indirect reflection which has a larger ToF value than the signals reflected straightly from the human body.

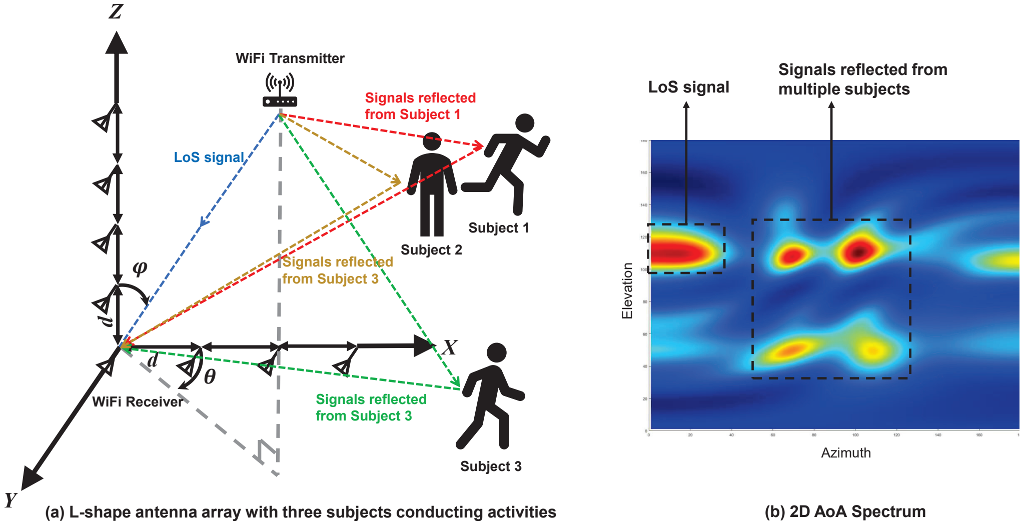
Next, a dynamic tracking algorithm is leveraged to handle the near-far problem. As the dynamic human subject has a predictable trajectory whereas the noise appears randomly, we utilize both appearance and motion correlation between consecutive frames to track each subject and remove the random noise. Furthermore, an adaptive filter is designed to prune the elevation scope of each subject. Finally, based on the filtered azimuth, elevation, AoD, and ToF information, we can generate clean 2D AoA images for each subject.

Then, a deep learning model is utilized to reconstruct the 3D human mesh from the input 2D AoA images. Among the framework of the deep learning model, the CNN is utilized to extract the spatial features, and the GRU is adopted to derive the temporal features, while the self-attention is utilized to highlight the contribution of the important frames. To effectively learn the deformation of different parts of the human body, we divide the whole body into five regions. At last, the SMPL model is adopted to generate the realistic 3D human mesh representation.

In our system, the commodity WiFi devices in the environment could be reused for potential mass adoption. Moreover, our system could also estimate human mesh under NLoS and poor lighting scenarios, where the camera-based systems do not work well.

3.3 Data Calibration

The raw CSI measurements are affected by random phase offsets caused by the sampling time offset (STO) and packet detection delay (PDD) discrepancies across packets, which are a consequence of imperfections in commodity WiFi hardware. To accurately estimate AoA and ToF information, which is derived from the phase shifts across multiple subcarries as well as multiple antennas of the receiver, it is necessary to preprocess the CSI


Z
WiFi Transmitter
Signals reflected from Subject 1
LoS signal
Subject 1
Subject 2
φ
Signals reflected from Subject 3
d
X
d
θ
WiFi Receiver
Signals reflected from Subject 3
Subject 3
Y
LoS signal
Signals reflected from multiple subjects
Elevation
Azimuth

(a) L-shape antenna array with three subjects conducting activities

(b) 2D AoA Spectrum

Fig. 2. L-shaped antenna array and the 2D AoA (azimuth θ and elevation ϕ) estimation [34].

measurements and eliminate these random phase offsets. In specific, we adopt a linear fit method proposed in [18] to sanitize the random phase shift. The optimal linear fit method is denoted as follows:

$$\sigma = \underset{\alpha}{argmin} \sum_{x,y=1}^{X,Y} \sum_{z=1}^{Z} (\Psi(x,y,z) + 2\pi f_{\delta}(z-1)\alpha + \beta)^2, \tag{1}$$

where f_{δ} represents the frequency difference of the adjacent OFDM subcarriers, $\Psi(x,y,z)$ is the unwrapped phase of the CSI at the x^{th} subcarrier of a packet, which is transmitted from the y^{th} transmitting antenna and received at the z^{th} receiving antenna, α is the common slope of the received phase responses for all antennas, and β is the offset. The σ includes the time delay of each WiFi packet. Finally, the calibrated CSI phase $\bar{\Psi}(x,y,z)$ can be generated by removing the time delay as $\bar{\Psi}(x,y,z) = \Psi(x,y,z) - 2\pi f_{\delta}(z-1)\sigma$.

3.4 Improvement of Multi-subject Resolvability

To achieve the separation of multiple subjects, inspired by Wi-Mesh [34], we can leverage an L-shaped antenna array to distinguish different users in the 2D AoA (i.e., azimuth-elevation) space. However, it is hard to separate multiple people who are very close. Hence, we propose to improve the resolvability of commodity WiFi sensing by combining AoD and ToF information with 2D AoA as the signals that have the same 2D AoA could be separated in the ToF or AoD dimension.

Leveraging spatially distributed multiple receiving antennas. As current WiFi technology could support multiple antennas at each WiFi device, we explore utilizing such spatially distributed antennas to separate the signal reflections from different directions. Thus, such spatial information makes it possible for the separation of multiple users in the environment.

In particular, inspired by previous work [34] that explores the two-dimensional angle of arrival (2D AoA) (i.e., azimuth and elevation) of the WiFi signals, we also form multiple antennas on the receiver into an L-shaped antenna array to extract spatial information which can enable WiFi devices to roughly separate the silhouettes of multiple human subjects. The L-shaped antenna array consists of two uniform linear subarrays with an equal distance d aligned with the X-Z axis to receive the WiFi signals as shown in Figure 2(a). Each uniform linear subarray has N omnidirectional antennas. The human body as well as other static objects (e.g., walls and furniture) in the environment will reflect the WiFi signals when they travel through space. We can calculate the 2D AoA of the signal reflections based on the phase shift of the received signals at multiple antennas. The phase shift of the received signal between the adjacent antennas with the corresponding subarray can be denoted as:

$$\Phi_x(\varphi_l, \theta_l) = e^{\frac{-j2\pi d}{\lambda}\sin(\varphi_l)\cos(\theta_l)}, \tag{2}$$

$$\Phi_z(\varphi_l) = e^{\frac{-j2\pi d}{\lambda}\cos(\varphi_l)}, \tag{3}$$

where Φ_x is the phase difference between subarray across the X axis, and Φ_z is the phase difference between subarray across the Z axis. λ is the wavelength of transmitted WiFi signal, φ_l and θ_l denote the elevation and azimuth angle of the l^{th} signal, respectively. Based on these phase differences, both the azimuth and elevation of the signal reflections can be derived by using the MUSIC algorithm [28]. Such spatial information can roughly separate multiple subjects in the azimuth-elevation space.

However, due to the limited number of antennas on commodity WiFi devices, the corresponding 2D AoA spatial resolution is still insufficient to separate multiple human subjects who are close to each other. For example, as shown in Figure 2(b), we can only observe two subjects in the 2D AoA image as subject 1 and subject 2 are very close to each other in Figure 2(a). Besides, an L-shaped antenna array with nine antennas and the MUSIC algorithm can provide an angular resolution of about 0.3 radians [14, 40]. It could separate two subjects with a distance of $50cm$ between them with a probability of 50% in a typical larger room environment according to our simulation, shown in Figure 3.

Leveraging spatially distributed multiple transmitting antennas. For those signals that have similar 2D AoA and cannot be separated, we leverage multiple antennas on the transmitter (i.e., transmitting antennas) to generate the angle of departure (AoD) of signals. In particular, we incorporate the phase shifts introduced by the spatially distributed transmitting antennas. In this work, The transmitting antenna forms a linear antenna array. We describe the phase shift $\Psi(\omega)$ across transmitting antennas as follows:

$$\Psi(\omega) = e^{-j2\pi f d\sin(\omega)/c}, \tag{4}$$

where ω is the AoD of the signal, d is the distance between two adjacent transmitting antennas, and f is the frequency of the signal. Also, we can use the MUSIC algorithm for joint estimation of azimuth, elevation, and AoD. Assuming that the transmitter is equipped with three antennas, it can separate two subjects with a distance of $30cm$ between them with a probability of 50% in a typical larger room environment, when we add AoD into our simulation.

Leveraging frequency-distributed subcarriers. Besides 2D AoA and AoD of signals, our work further incorporates the time of flight (ToF) of signals derived by multiple OFDM subcarriers to increase the resolvability of WiFi sensing and thus to better separate multiple human subjects. It is because different OFMD subcarriers will cause phase shifts for the same reflection as well due to frequency differences. Specifically, we integrate the phase shifts associated with the different frequencies of the OFDM subcarriers. For evenly distributed subcarriers, the phase shift across two adjacent subcarriers can be described as follow:

$$\Omega(\tau) = e^{-j2\pi f_\delta \tau_l/c}, \tag{5}$$

where τ_l is the ToF of the l^{th} propagation path.

Therefore, by combining multiple antennas distributed in the spatial domain and multiple OFDM subcarriers distributed in the frequency domain, we can use the MUSIC algorithm for joint estimation of four-dimensional (4D) information (i.e., azimuth, elevation, AoD, and ToF) to significantly improve the resolvability of commodity WiFi sensing. Thus, we can separate multiple human subjects in the same environment at the same time, and also

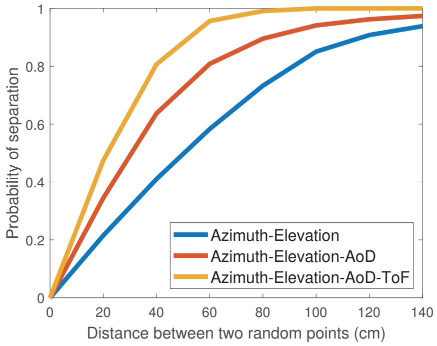


Fig. 3. Improvement of resolvability by incorporating AoD and ToF.

Table 1. Probability of inseparability for two random points between different distances.

	$20cm$	$40cm$	$60cm$	$80cm$	$100cm$
Azimuth-Elevation	0.79	0.59	0.42	0.27	0.15
Azimuth-Elevation-AoD	0.66	0.36	0.19	0.10	0.058
Azimuth-Elevation-AoD-ToF	0.53	0.19	0.044	0.0091	0.00005

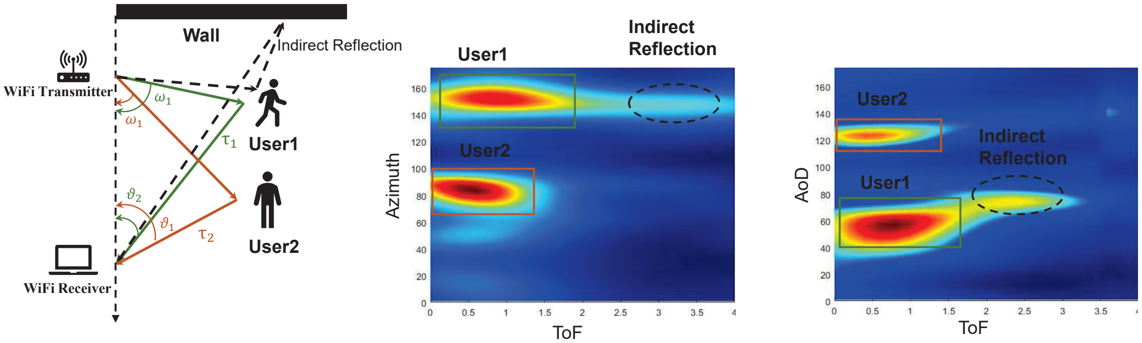
obtain a better illustration of the shape and deformation of the human body for 3D human mesh construction. The estimation is presented as follows by maximizing the spatial spectrum function:

$$P(\theta, \varphi, \omega, \tau) = \frac{1}{A^H(\theta, \varphi, \omega, \tau)E_N E_N^H A(\theta, \varphi, \omega, \tau)}, \tag{6}$$

where A is the steering vector for four-dimensional information and E_N is the noise vector subspace [28]. After we further incorporate ToF into our simulation, we can separate two subjects with a distance of just $20cm$ between them with a probability of 50%, shown in Figure 3.

Illustration of resolvability improvement. To illustrate the more detailed effectiveness of incorporating AoD and ToF to improve resolvability, we conducted a simulation in a practical setting. Our simulated environment is a room with dimensions of $9m \times 9m \times 3.6m$, with a transmitter and a receiver positioned at opposite ends. The transmitter has a linear antenna array with three antennas, while the receiver is equipped with an L-shaped antenna array with nine antennas. The bandwidth of WiFi signals is 40MHz. These can provide a raw resolution of azimuth or elevation of approximately 0.6 radians [14], and a raw resolution of AoD of 1.2 radians. Additionally, the raw resolution of ToF is approximately 7.5m [38]. However, the MUSIC algorithm used in this work is a super-resolution algorithm that can enhance these raw resolutions by a factor of approximately 2× [40].

To evaluate our proposed solution, we randomly sampled two points in the simulated space and determined if the two points could be separated based on the resolution of the MUSIC algorithm. As shown in Figure 3, we can observe that the curve dramatically moves towards the upper left corner of the figure when we incorporate more information from other dimensions. This shows that the resolvability has been significantly improved. According to Table 1, we can also observe that AoD and ToF can remarkably reduce the inseparability of signals. For example, when the distance between two points is $60cm$, the probability of inseparability is reduced by a factor of 2.2 after we add ToF. Moreover, if we combine all multi-dimensional information, the probability of inseparability is reduced by almost a factor of 10.



(a) Sensing Scenario (b) Azimuth-ToF Profile (c) AoD-ToF Profile

Fig. 4. Illustration of subject detection.

3.5 Static Reflection Subtraction

Based on the derived four-dimensional information, we can capture the spatial information of all the moving subjects, static objects, and line-of-sight signals in the sensing environment. However, both the signals reflected off the static objects and the line-of-sight signals can interfere with the signals of human subjects. To eliminate the impact of the static environment, our system should remove the LoS signals and reflections from static objects such as walls or furniture.

Due to the fact that static environments remain relatively stationary to WiFi devices, their distance and orientation with respect to the devices stay consistent over time. Consequently, the four-dimensional information of these static reflections remains constant over time. To eliminate the interference from static environments, we subtract the four-dimensional information from previous frames. In contrast to Wi-Mesh[34] that deducts the previous frame, our work subtracts a series of previous sequence frames to eliminate static reflections. This is because the time span occupied by two consecutive frames is very short, when the movement of the subject is not so dramatic, subtracting two consecutive frames will also remove most of the motion information. Therefore, we seek to extend the time span and utilize more previous frames. And since each previous frame has a varying degree of influence on the current result over time, we assign a weight to each previous frame. The subtraction method is denoted as follows:

$$F_r = F_c - a_1 F_1 - a_2 F_2 - ... - a_n F_n, \tag{7}$$

where F_c is the four-dimensional information of the current frame, F_n is the n^{th} previous frame, and F_r is the four-dimensional information after the background removal process. a_n is the weight attributes to each n^{th} previous frame, where we assign more weight to the more recent frame and $a_1 + a_2 + ... + a_n = 1$. In our work, we subtract four consecutive frames and $a_1 = 0.4$, $a_2 = 0.3$, $a_3 = 0.2$, $a_4 = 0.1$.

3.6 Indirect Reflection Removal

Subject Detection We aim to obtain the clean version of signal reflections (i.e., direct reflection) of each subject as the representation of their bodies which can be fed into the later human mesh construction model. Although we eliminate the reflections from the static objects in the environment, the remaining reflections still contain indirect reflections which are the signals that bounce off the human body and are further reflected from the static objects in the environment or other moving subjects.

To address the indirect reflection problem, we leverage the insight that indirect reflections typically follow propagation paths with different lengths and have distinct AoD compared to signals directly reflected from the

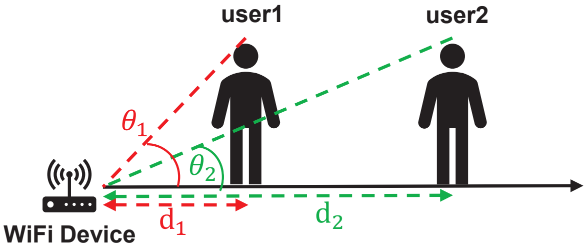


Fig. 5. Illustration of elevation filter.

human body. And the ToF information directly correlates with the distance of the signal propagation path. As a result, we delve into the ToF and AoD information to effectively distinguish between reflections originating from human subjects and indirect reflections. This distinction is made possible by generating two 2D angle-range profiles using the derived four-dimensional spatial information. The two profiles are the Azimuth-ToF profile and the AoD-ToF profile, which help us differentiate and address the source of signal reflections. Figure 4(a) shows that there are two subjects conducting activities in front of the WiFi devices while they also lead to indirect reflections. As shown in Figure 4(b) and (c), the clean signal reflections of these two subjects are illustrated in both the Azimuth-ToF profile and the AoD-ToF profile with the bounding boxes, and the indirect reflection are also distinguishable with respect to the reflections direct from the human body on the Azimuth-ToF profile and the AoD-ToF profile.

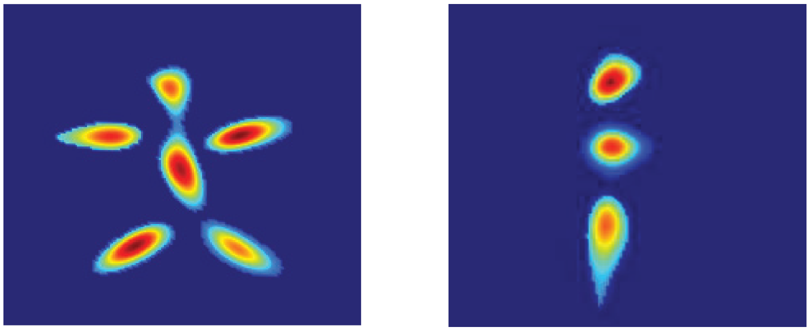
Based on the two aforementioned angle-range profiles, our system leverages the YOLACT [5] as the backbone framework for detecting potential users and mitigating interference arising from indirect reflections. In specific, YOLACT is a real-time instance segmentation model which first generates a set of prototype masks and then combines the mask coefficients generated for each instance with the prototype masks to predict the segmentation result. After the detection process, it will output the bounding box of each frame for all the potential users, which represents the more precise AoD, azimuth, and ToF information for each user. Moreover, the irrelevant information caused by indirect reflections will also be ignored as the subject detection process only focuses on the information of each target subject.

Adaptive Elevation Scope Filter We have obtained clean azimuth, AoD, and ToF information from previous steps. In this step, we adopt an adaptive filter to extract the accurate elevation scope of each subject. As illustrated in Figure 5, the human subject has larger elevation scope when he or she is closer to the WiFi devices and vice versa. Therefore, once the range information is obtained, we can obtain the corresponding reasonable elevation scope and eliminate interferential elevations. By using the derived range (i.e., ToF) information and the constraint of human height (i,e, usually from 1.5m to 2.0m), we design an adaptive elevation scope filter that derives the elevation scope negatively related to the ToF information of each subject.

3.7 Dealing With Near-far Problem

Due to the near-far problem, the strength of the signal reflected from the subject away from the WiFi devices is relatively weaker than the closer ones, which may be confused with random noise. Therefore, it is possible that the subject away from the WiFi devices is not detected in certain frames during the subject detection process. Moreover, there also exists the possibility that some random noises are detected as potential subjects by the detector. Towards separating multiple subjects effectively and constructing corresponding mesh for each subject accurately, we need to explore a solution to resolve the near-far problem.

Based on this insight that human movement is characterized by continuous, predictable patterns over time and space, as opposed to the random and sporadic nature of disturbed noises, we observe that there exists high



(a) 2D AoA Image for User1 who is jumping.

(b) 2D AoA Image for User2 who is standing still.

Fig. 6. Extracted 2D AoA images for two subjects.

coherence of the motion sequence when the subject engages in activities. In essence, a moving subject normally follows a discernible and predictable trajectory, with each subsequent motion highly related to the preceding one. Conversely, noise exhibits random and unrelated behavior.

Therefore, we adopt the output format of the DeepSORT [36], which could predict and track each user dynamically. In specific, DeepSORT utilizes two branches to track the subjects: the appearance branch and the motion branch. In the appearance branch, a deep appearance descriptor is utilized to extract the appearance feature of each input bounding box, and the cosine distance matches the extracted feature with the new detections. In our work, the extracted Azimuth-ToF and AoD-ToF profile represents the spatial information about the top view of human motion, which also has high dependencies between the consecutive frames. Therefore, we could employ the deep appearance descriptor to derive the features from the Azimuth-ToF and AoD-ToF profiles and match them with the relevant detected subject. In the motion branch, the Kalman filter algorithm is utilized to predict the position of the current bounding box and update its state based on the Mahalanobis distance. This branch helps to filter out unlikely associations, which are random noises. The Hungarian algorithm is then used to solve the association task, allowing us to identify each subject from all the potential users and eliminate random noise.

Finally, after we derive the precise azimuth, elevation, AoD, and ToF related to each subject, we accumulate the 2D AoA (azimuth and elevation) values in the ToF and AoD dimensions to generate clean 2D AoA images of each subject. Then, we utilize a threshold to highlight the reflections from the human body which have relatively high signal strength. Figure 6 shows the illustration of 2D AoA images for two subjects in which the silhouettes of the human bodies and the deformation of poses can be observed.

3.8 Human Mesh Construction Model

After obtaining the 2D AoA images for each subject, we further feed the 2D AoA images to a deep learning model to learn and predict the 3D human mesh. As the granularity of the human body surface is much smaller than the wavelength of the WiFi signal, the human body can be regarded as specular about WiFi signals. In consequence, when the WiFi signals reach the human body, some parts of the signals may be directly reflected to the receiver, while other parts may be scattered away from the receiver. Thus, we may only capture the spatial information about a subset of the human body in a single 2D AoA image but miss other parts that disperse the signals away. To overcome this problem, we aggregate multiple 2D AoA images in time sequence to recover the spatial information of the whole human body. Note that we utilize the deep learning model to learn the relation about such a combination. Accordingly, the input of our model is a series of 2D AoA images which are

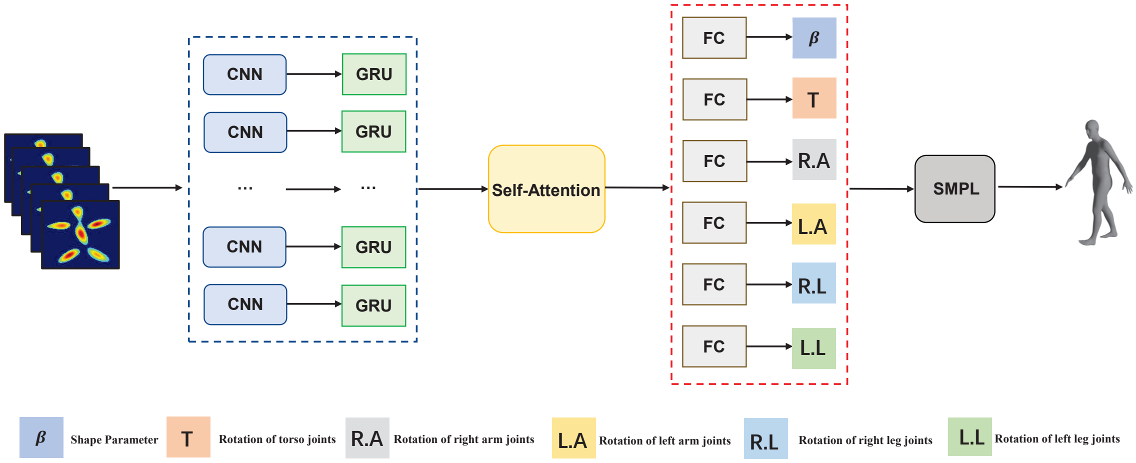

CNN
GRU
CNN
GRU
...
...
CNN
GRU
CNN
GRU
Self-Attention
FC
β
FC
T
FC
R.A
FC
L.A
FC
R.L
FC
L.L
SMPL
β Shape Parameter
T Rotation of torso joints
R.A Rotation of right arm joints
L.A Rotation of left arm joints
R.L Rotation of right leg joints
L.L Rotation of left leg joints

Fig. 7. Mesh construction model.

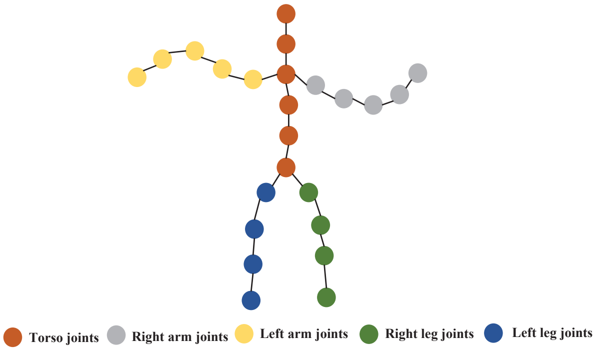


Fig. 8. Illustration of body regions.

transformed into a 3D matrix. In specific, the three dimensions represent the number of frames, the ranges of the elevation, and the azimuth angles of the 2D AoA image, respectively.

The overview framework of our deep learning model is shown in Figure 7. First, we feed a sequence of frames into a convolutional neural network (CNN)-based feature extractor, which extracts the higher-level spatial features per frame. As the dynamic human activities are continuous and the next move is highly relevant to the previous motion, the high-level vector of human body representation has high temporal correlations among the consecutive frames. Therefore, we send these above-extracted representation vectors to a multi-layer GRU to yield a temporal feature based on the previous frame. Whereas, the current result is affected to varying degrees by a series of previous frames over time about dynamic activities. Thus, the self-attention mechanism is adopted to dynamically learn the importance of each previous frame and assign corresponding weights. After that, we can highlight the contribution of the important frames and generate the global representation of all the input frames.

To learn the deformation of 3D dynamic motion effectively, we split the entire body into five local body regions: torso, left arm, right arm, left leg, and right leg, as shown in Figure 8. This is because joint positions within each group are highly correlated, while the joint positions between the groups are relatively less related. We feed the global representation to multiple fully connected (FC) blocks to predict the joint rotation of each body region as well as the body shape parameter. Finally, the shape and pose parameters are sent to the gender-neutral SMPL model to represent the 3D human mesh.

Human Mesh Model To construct realistic 3D human mesh, we adopt the SMPL model to represent human mesh as one of our system components. SMPL is a vertex-based linear 3D human body model that leverages two types of low-dimensional parameters, which are pose and shape parameters, to represent the human body. The pose parameter θ is represented by one global 3D rotation vector and relative 3D rotation of 23 joints in axis-angle representation, which controls the deformations in the joints of the human body. The shape parameter $\beta \in \mathbb{R}^m$ controls the statistical shape features of the subject and is irrelevant to various poses, which depicts the top m coefficients of a Principal Component Analysis (PCA) shape space. Given the aforementioned parameters as input, the SMPL model provides an efficient mapping function $M(\beta, \theta)$ which digitizes the shape and pose parameters into a triangulated mesh with 6890 vertices that fully characterize the 3D surface and poses of the human body.

Loss Function. We propose a loss function consisting of the shape losses $\boldsymbol{L_s}$ and pose losses $\boldsymbol{L_p}$. And the pose losses are the losses of summation of five body components. We use the L_1 norm to evaluate differences between the predicted parameter and the ground truth as follows:

$$\boldsymbol{Loss} = \frac{1}{F}\sum_{f=1}^{F} \|K - GT(K)\|_{L_1}, \tag{8}$$

Where F is the number of frames about the input 2D AoA image sequence. And $GT(K)$ is the corresponding ground truth of the predicted parameters K.

Our network is trained with the loss function as follows:

$$\boldsymbol{L}_{SMPL} = \lambda_J \boldsymbol{L}_p + \lambda_V \boldsymbol{L}_s, \tag{9}$$

Where we weight the shape losses $\boldsymbol{L_s}$ and pose losses $\boldsymbol{L_p}$ with parameter λ_V and λ_J, respectively.

4 PERFORMANCE

In this section, various experiments are conducted to test the performance of MultiMesh in different real-world environments.

4.1 Experimental Setup

4.1.1 Devices. In the experiments, MultiMesh is implemented with the commodity WiFi devices, (i.e., Dell LATITUDE laptops) as the WiFi transmitter and WiFi receiver. Figure 9 shows the experimental scenario, where the WiFi receiver is equipped with an L-shaped antenna array of nine antennas, which could achieve both azimuth and elevation AoA estimations. The L-shaped antenna array consists of two uniform linear subarrays in the orthogonal direction which both have two Intel 5300 Network Interface Cards (NICs). Moreover, a signal splitter is utilized to stitch NICs with shared antennas to simulate possible antenna configuration of the new generation WiFi devices. And the WiFi transmitter consists of three linearly-spaced antennas. The antennas on both the receiver and transmitter are all equally spaced with half a wavelength apart ($2.8cm$). The WiFi transmitter is configured to generate WiFi signals with a bandwidth of 40 MHz. The default transmitting packet rate is set to 1000 packets per second. Linux 802.11 CSI tool [11] is utilized to collect CSI measurements of 30 OFDM subcarriers. We leverage the network time protocol (NTP) to enable synchronization for all devices.

4.1.2 Ground Truth Mesh Construction. In our work, the SMPL model is utilized to represent the ground truth 3D human mesh. The SMPL model takes the pose information and body shape information as input and outputs the corresponding 3D human mesh. We utilize the camera to record the ground truth of the human body and activities. Specifically, we adopt the vision-based approach in [18] to obtain high-resolution ground truth of the

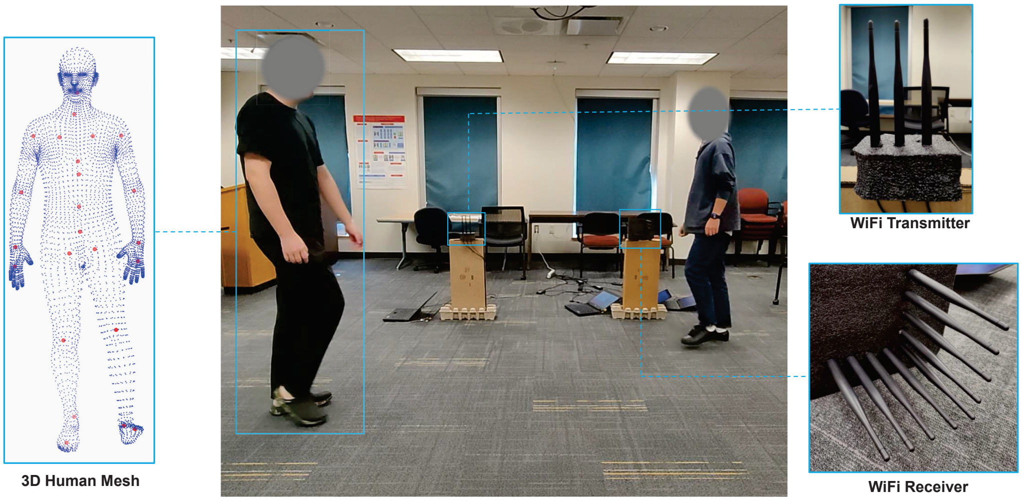


Fig. 9. The illustration of experimental scenario.

pose information and utilize the VideoAvatar [4] to capture the body shape information as the ground truth. In addition, We calculate the 3D position of human joints by averaging the ground truth pose information recorded by two cameras which are placed in opposite positions.

4.1.3 Data Collection. In our experiments, we invite 14 volunteers of different genders, weights, and heights to perform diverse daily activities in the sensing area, including walking back and forth in straight, walking in a circle, walking with random arm motions, sitting down and standing up, rotating torso and random arm motions in the place. There are either two or three volunteers performing these aforementioned activities simultaneously in three real-world environments, including a classroom, a laboratory, and a conference room. We perform experiments in both occluded and unoccluded scenes. We also evaluate our system for different distances between two subjects and different distances between subjects and WiFi devices. We collect around ninety million WiFi CSI packets to train and test our system for 3D human mesh construction totally. The data collection was approved by the IRB of the authors' institution.

4.1.4 Model Settings and Training. In this section, we present the detail of our deep learning model settings and how we train our model. The ResNet[12] is adopted as our feature extractor. The GRU has two layers with the number of hidden states as 2048. And two fully-connected layers of size 2048, as well as $tanh(\cdot)$ activation, are utilized in the self-attention module. During the model training, we train our model with 80% of the data with 14 subjects and leverage the rest 20% data for evaluation. The initial learning rate is set to 0.0001 with periodical decay and the batch size is set to 16. And the hyperparameter λ_V and λ_J are set to 1 and 0.01, respectively. We utilize PyTorch to implement our deep learning model and train the model on NVIDIA RTX 3090 GPU.

4.1.5 Baselines. Since no existing work utilizes commodity WiFi to construct multi-subject human mesh, we compare our model with the baselines by removing the spatial information in the framework of the proposed model as follows:

Baseline A. we only utilize the azimuth and ToF information of WiFi signals to construct human mesh in this baseline. In specific, we only extract the azimuth and tof information of each subject as one 2D energy map from

Table 2. Overall system performance of two subjects.

	PVE (*cm*)	MPJPE (*cm*)	PA-MPJPE (*cm*)
Baseline A	9.93	8.91	4.45
Baseline B	6.29	5.62	2.76
Baseline C	4.93	4.05	2.37
MultiMesh	4.01	3.51	1.90

Table 3. Overall system performance of three subjects.

	PVE (*cm*)	MPJPE (*cm*)	PA-MPJPE (*cm*)
Baseline A	11.26	10.25	5.15
Baseline B	8.01	7.23	3.56
Baseline C	6.54	5.18	2.81
MultiMesh	5.39	4.65	2.43

the Azimuth-ToF profiles and then feed the extracted information to our proposed deep-learning network to predict the human mesh for each subject.

Baseline B. In this baseline, we derive the azimuth, AoD, and ToF information of each subject from the Azimuth-ToF and AoD-ToF profiles, and also send this information to our proposed deep-learning network for human mesh construction.

Baseline C. We leverage the azimuth, elevation, and ToF information of the reflected signals in this baseline. We utilize the 2D AoA images as input for our proposed deep-learning network for human mesh construction.

4.1.6 Metrics. To evaluate our system performance, we use the following metrics: per vertex error (PVE), mean per joint position error (MPJPE), and Procrustes aligned mean per-joint position error (PA-MPJPE). Among them, PVE is the average vertex error by averaging Euclidean distance between the predicted human mesh vertices and the corresponding vertices on the ground truth mesh. MPJPE is the average Euclidean distance between the predicted joint locations of the human mesh and the ground truth after root matching. PA-MPJPE is a variant of the MPJPE metric that is defined as the average Euclidean distance between the joint locations of the predicted human mesh and the ground truth after aligning the predicted 3D poses to the ground truth poses using Procrustes analysis method [10]. Procrustes Analysis removes the effects of translation, rotation, and scale. Therefore, PA-MPJPE focuses on the reconstructed 3D mesh/pose itself.

4.2 Overall Performance

We first quantitatively evaluate the overall performance of our proposed system compared with the aforementioned baseline. The results are illustrated in Table 2 and Table 3. As we can see, the average PVE, MPJPE, and PA-MPJPE of our system is $4.01cm$, $3.51cm$, $1.90cm$ for two subjects scenarios, and $5.39cm$, $4.65cm$, $2.43cm$ for three subjects scenarios. And our system achieves the best results compared with the other three baselines. From the results we can observe that Baseline A achieves the worst performance, this is because it only utilizes 2D information, azimuth, and ToF, which is less accurate for separating multiple subjects and reconstructing an accurate human mesh. And Baseline B performs better than Baseline A, the reason is that Baseline B utilizes more dimensional spatial information than Baseline A. While Baline B and Baseline C utilize the same number of dimensional spatial information, Baline C is better than Baseline B, this is because Baseline C extends 1D AoA estimation to 2D AoA estimation (azimuth and elevation), which could uniquely identify the body shape and

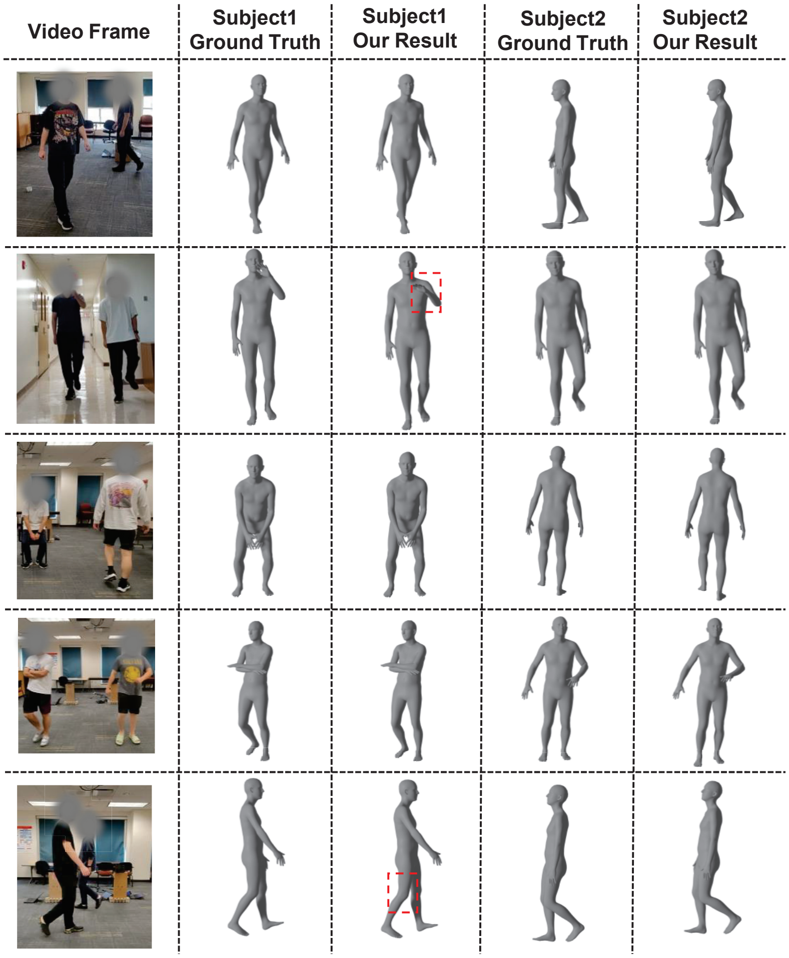


Fig. 10. The examples of constructed mesh in two subjects scenarios.

deformations in 2D space. Our system performs the best since our system jointly estimates 4D spatial information (azimuth, elevation, ToF, and AoD) and thus can significantly improve signal resolvability. Moreover, we can utilize more dimensional information than these three Baselines to separate signals from different subjects in multi-subject scenarios. Therefore, our system could better separate different subjects and estimate the human mesh for each subject benefiting from the high signal resolvability. When we compare the results from Table 2 and Table 3, we can observe that the performance of the system decreases as the number of people increases. This is because when there are more moving subjects in the sensing area, the multi-path effect among the subjects will become more severe, which makes it more difficult for our system to generate high-quality 4D spatial metric and further affect the accuracy of generated human mesh. Nevertheless, our system could achieve accurate 3D human mash constructions under two or even more subject scenarios, demonstrating the robustness of our system regarding the number of subject changes.

Then, we qualitatively evaluate the overall performance of MultiMesh when multiple subjects are conducting various activities in diverse environments. As illustrated in Figure 10, the first column shows the reference video

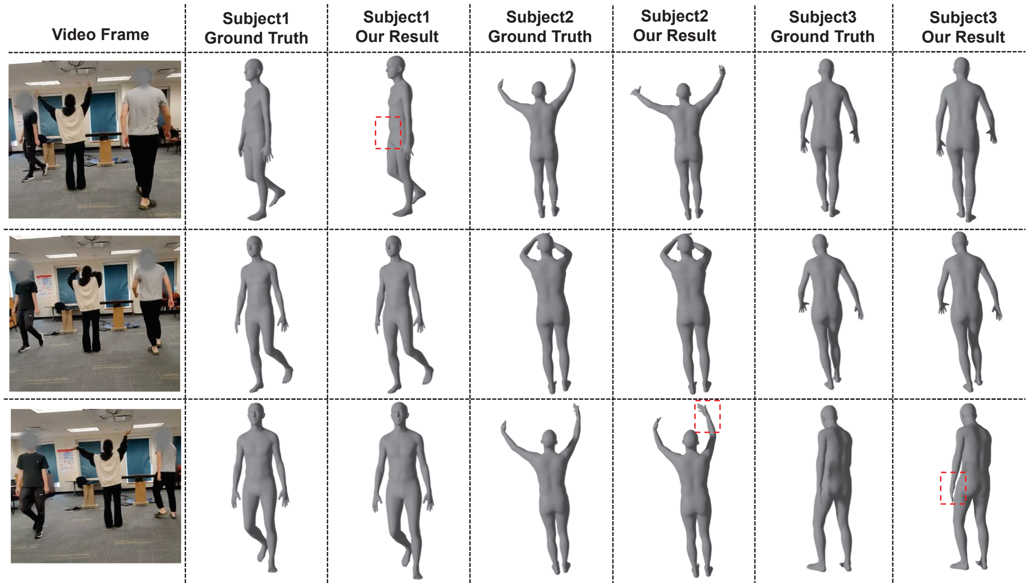


Fig. 11. The examples of constructed dynamic mesh in three subjects scenarios.

frames where the subjects are conducting activities. And the second and fourth columns present the corresponding ground truth meshes generated by the vision-based method. While the third and fifth columns show the human meshes constructed by our system. We can see that the meshes generated by our model match the ground truth very well. Moreover, we also illustrate the constructed human mesh and corresponding ground truth in three subject scenarios in Figure 11. We utilize red dotted boxes to emphasize mispredicted and distorted body parts, with the majority of these inaccuracies occurring at the extremities of the arms or legs. This is primarily due to relatively weaker signal reflections from these body areas. Nonetheless, it is worth noting that the overall constructed 3D meshes align closely with the ground truth data. Nevertheless, the results still demonstrate that our system is capable of constructing 3D human mesh accurately and effectively under diverse environments, activities, and subjects.

4.3 System Performance for Unseen Subjects

To study the robustness of our proposed system for unseen subjects, we conduct a cross-subjects experiment where the subject in the training set and evaluation set is totally non-overlapped. As shown in Table 4, the average PVE, MPJPE, and PA-MPJPE of our system for unseen subjects is 5.16cm, 4.61cm, 2.26cm for two subjects scenarios, and 6.90cm, 6.01cm, 2.73cm for three subjects scenarios. Compared to overall performance, dealing with subjects that have never been seen in the training period only introduces a small increase in vertex errors (1.05cm and 1.51cm). The results demonstrate that our proposed system could effectively handle unseen subjects and activities.

4.4 System Performance for Unseen Environments

In practical scenes, our system should be able to adapt to diverse and unseen indoor environments. We also conduct experiments to evaluate the robustness of our system in the unseen environment. Specifically, we utilize the data collected in the lab and classroom to train our model and evaluate our system in the unseen environment

Table 4. System performance for unseen subjects.

	PVE (*cm*)	MPJPE (*cm*)	PA-MPJPE (*cm*)
Two-subjects	5.16	4.61	2.26
Three-subjects	6.90	6.01	2.73

Table 5. System performance for unseen environments.

	PVE (*cm*)	MPJPE (*cm*)	PA-MPJPE (*cm*)
Two-subjects	4.51	3.98	2.04
Three-subjects	6.30	5.61	2.46

Table 6. System performance in occluded scenarios.

	PVE (*cm*)	MPJPE (*cm*)	PA-MPJPE (*cm*)
Two-subjects	6.49	5.84	2.49
Three-subjects	8.24	7.03	3.12

(conference room). The results are illustrated in Table 5. In specific, we can observe that our system could deal with unseen environments with the average PVE, MPJPE, and PA-MPJPE of 4.51*cm*, 3.98*cm*, 2.04*cm*, and 6.30*cm*, 5.61*cm*, 2.56*cm* for two and three subjects scenarios, respectively. The results demonstrate the robustness of our system in a new environment.

4.5 Performance of Experiments in Occluded Scenarios

To investigate the system performance when the subjects are occluded by obstacles, we conduct experiments where the WiFi receiver and transmitter are placed in one room and the subjects conduct activities in another adjacent room to evaluate our system under the occluded scenarios. As shown in Table 6, the average PVE, MPJPE, and PA-MPJPE of our system in occluded scenarios is 6.49*cm*, 5.84*cm*, 2.49*cm*, and 8.24*cm*, 7.03*cm*, 3.12*cm* for two and three subjects scenarios, respectively. The performance degrades a little compared to non-occluded scenarios. The reason is that the strength of the WiFi signal will be impaired when it passes through the obstacles, leading to low SNR of the received signals. Figure 12 illustrates the constructed human mesh in the occluded scenario. We can observe that the results are aligned with the subjects in the video frame. The performance verify that our system could work in NLoS conditions.

In real-life implementation, the target subjects can be blocked by the furniture relative to the WiFi devices. Therefore, we also evaluate the extraction of AoA and ToF information from WiFi signals and the overall performance in the complex scenes with the furniture (e.g., desk, chair) occluded and compare the performance with common scenes without occlusion. As shown in Figure 13 and Figure 14, the AoA estimation errors are below 11.8° for the 80% of the cases in complex scenes, and there is less than 8.2° of angle estimation error for 80% of the cases for common scenes. The ToF estimation errors are below 4.8*ns* for 80% of the cases in complex scenes with the furniture occluded, and they are 3.2*ns* for common scenes without occlusion. And as illustrated in the Table 7, the PVE, MJPJE, and PA-MPJPE under complex scenes are 5.54*cm*, 4.82*cm*, and 2.52*cm*. While, the PVE, MJPJE, and PA-MPJPE in common scenes are 4.01*cm*, 3.51*cm*, and 1.90*cm*. Compared to the common scenes, the accuracy of AoA and ToF extraction slightly decreases in complex scenes due to lower SNR. However, the

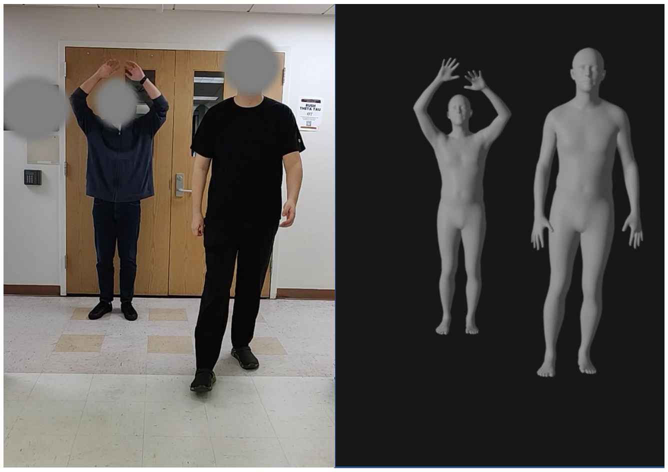


Fig. 12. The examples of constructed mesh in occluded scenarios.

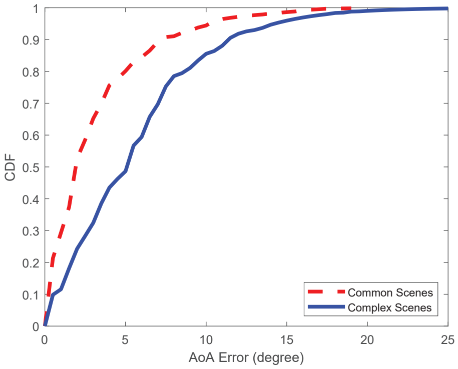


Fig. 13. AoA estimation error for complex scenes and common scenes.

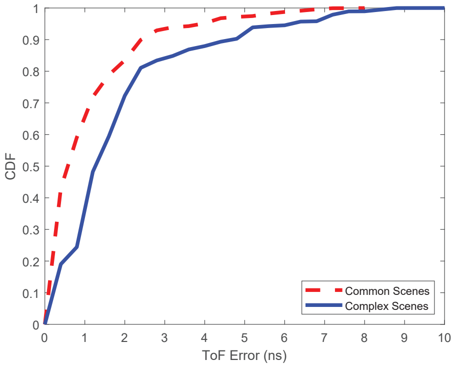


Fig. 14. ToF estimation error for complex scenes and common scenes.

Table 7. System performance for complex scenes and common scenes.

	PVE (cm)	MPJPE (cm)	PA-MPJPE (cm)
Comlex Scenes	5.54	4.82	2.52
Common Scenes	4.01	3.51	1.90

results still demonstrate that our system could extract relatively accurate AoA and ToF information and construct human mesh for multiple subjects in complex scenes.

4.6 Performance of AoA and ToF Extraction

We also conduct experiments to evaluate the extraction of AoA and ToF information from WiFi signals. The CDFs of AoA and ToF estimation errors are shown in Figure 15 and Figure 16, we can observe that our system achieves

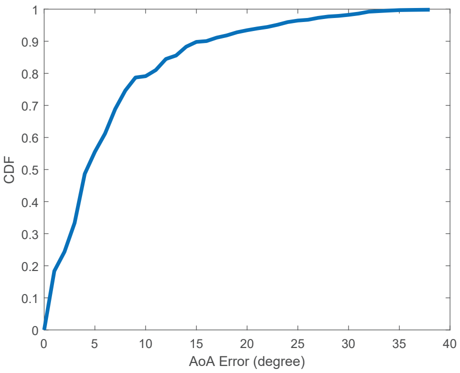


Fig. 15. AoA estimation error.

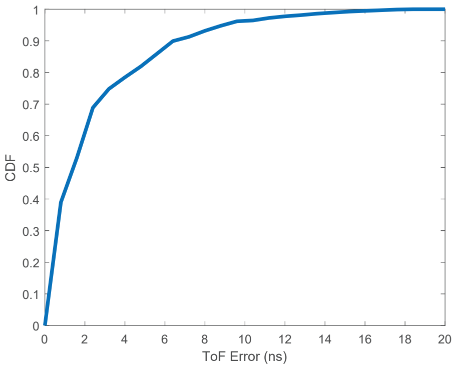


Fig. 16. ToF estimation error.

Table 8. System performance for different sensing distances.

	PVE (cm)	MPJPE (cm)	PA-MPJPE (cm)
$2m$	3.86	3.23	1.75
$4m$	4.41	3.79	2.10
$6m$	4.96	3.95	2.23

10.2° AoA estimation error at 80^{th} percentile, and 4.1 ns ToF estimation error at 80^{th} percentile when the signals of different individuals can be separated. The results demonstrate our system could extract accurate AoA and ToF information about each individual. This is because we jointly estimate four-dimensional information including azimuth, elevation, ToF, and AoD in our work, which significantly improves the signal resolvability.

4.7 Impact of Sensing Distance

As our system should adapt to different sizes of sensing environments, we further evaluate the impact of the sensing distance. Specifically, we evaluate the distance between the WiFi devices and two subjects performing activities including $2m$, $4m$, and $6m$. As shown in Table 8, the PVE is $3.86cm$, $4.41cm$, $4.96cm$, the MPJPE is $3.23cm$, $3.79cm$, $3.95cm$, and the PA-MPJPE is $1.75cm$, $2.10cm$, $2.23cm$ for these three sensing distances, respectively. We can observe that the error increases as the sensing distance becomes larger. The reason is that the WiFi signals will suffer attenuation in long propagation, and the signals are even reflected by the human body. However, our system could still achieve good performance with a sensing distance of $6m$, which could roughly produce the reflection path length corresponding to $13m$. The results demonstrate that our system is robust to sensing distance variation and could achieve a considerable performance in a typical size of smart and IoT environments.

4.8 Impact of Distance between Different Subjects

To evaluate our system performance under different distances between two subjects, we conducted the experiment with two people $100cm$, $50cm$, and $10cm$ apart from each other. The results are shown in the Table 9, we can observe that the PVE is $4.12cm$, $4.68cm$, and $5.68cm$, the MPJPE is $3.57cm$, $3.92cm$, and $4.72cm$, the PA-MPJPE is $2.02cm$, $2.21cm$, and $2.41cm$ when the distance between the two people is $100cm$, $50cm$, and $10cm$, correspondingly. We can see that the system performance becomes worse as the distance between two people increases. This is because, when two people perform activities very closely, the WiFi signals reflected from them will interfere with

Table 9. System performance for two people with different distances.

	PVE (cm)	MPJPE (cm)	PA-MPJPE (cm)
$10cm$	5.68	4.72	2.41
$50cm$	4.68	3.92	2.21
$100cm$	4.12	3.57	2.02

Table 10. System performance for different distances between subjects and WiFi device.

	PVE (cm)	MPJPE (cm)	PA-MPJPE (cm)
$50cm$	4.25	3.81	2.12
$100cm$	4.12	3.57	2.02
$150cm$	4.45	3.76	2.19
$200cm$	4.51	3.81	2.21
$300cm$	5.13	4.26	2.43
$500cm$	6.58	5.29	2.97

each other, leading to low accuracy of spatial information extraction of each individual. However, our system could still achieve considerable performance when two people are 10 cm apart.

4.9 Impact of Distance between Subject and WiFi Devices

We also evaluate the impact of different distances between the subject and WiFi devices. In specific, we conduct experiments with different distances between the subjects and the WiFi devices as $50cm$, $100cm$, $150cm$, $200cm$, $300cm$, and $500cm$, and the distance between two subjects is fixed at 100cm. As illustrated in Table 10, the PVE is $4.25cm$, $4.12cm$, $4.45cm$, $4.51cm$, $5.13cm$, and $6.58cm$, the MPJPE is $3.81cm$, $3.57cm$, $3.76cm$, $3.81cm$, $4.26cm$, and $5.29cm$, the PA-MPJPE is $2.12cm$, $2.02cm$, $2.19cm$, $2.21cm$, $2.43cm$, and $2.97cm$ for the distance between subject and the WiFi device as $50cm$, $100cm$, $150cm$, $200cm$, $300cm$, and $500cm$, correspondingly. We can observe that the overall system performance deteriorates as the subjects move farther away from the devices. This degradation occurs due to the attenuation of the WiFi signal's strength over longer distances, resulting in a lower SNR. Moreover, it will be more difficult for our system to distinguish the reflected signals from two subjects when they are getting closer due to the fixed signal resolvability. Nevertheless, our system maintains a relatively high level of performance when the distance between the subjects and the device remains below 300 cm, roughly equivalent to the dimensions of a typical indoor room.

4.10 Results of Subject Detection

To evaluate the performance of the subject detection process, we utilize both the Average Precision (AP) [19] and AP@70 to measure the results of bounding box prediction. The AP metric assesses subject detection performance by considering both precision and recall rates concerning the predicted bounding boxes and the ground truth bounding boxes, using specific Intersection over Union (IoU) thresholds. More precisely, it computes the average area under the Precision-Recall Curve within a range of IoU thresholds, from 0.5 (AP@50) to 0.95 (AP@95), with an increment of 0.05. Notably, AP@70 is calculated using an IoU threshold of 0.7, where a predicted bounding box must exhibit a substantial overlap with the ground truth bounding box to be deemed a correct prediction. Specifically, we calculate the bounding box prediction results for different distances between two subjects as well as different distances between the subjects and the WiFi devices.

Table 11. Subject detection results for two people with different distances.

	AP	AP@70
$10cm$	0.572	0.736
$50cm$	0.642	0.824
$100cm$	0.710	0.868

Table 12. Subject detection results for different distances between subjects and WiFi device.

	AP	AP@70
$50cm$	0.669	0.849
$100cm$	0.710	0.868
$150cm$	0.652	0.843
$200cm$	0.631	0.814
$300cm$	0.597	0.765
$500cm$	0.534	0.702

Table 13. Overall system performance comparison for data calibration.

	PVE (cm)	MPJPE (cm)	PA-MPJPE (cm)
System with data calibration	4.01	3.51	1.90
System without data calibration	12.21	10.54	9.03

The results of different distances between the two subjects are shown in Table 11, and the results of different distances between the subjects and WiFi devices are presented in Table 12. From the results, we can observe that our system encounters challenges in detecting multiple subjects when they are in close proximity (e.g., 10 cm) to each other or positioned at a considerable distance from the WiFi devices (e.g., 500 cm). Nonetheless, it is worth emphasizing that the accuracy of subject detection remains notably high in the majority of scenarios. This demonstrates the effectiveness and robustness of our system's subject detection process.

4.11 Ablation Study

To evaluate the effectiveness of our system flow, we utilize the ablation study to evaluate two stages of our system, including data calibration, and static reflection subtraction.

First, to evaluate the effectiveness of the data calibration process, we compare the overall system performance with and without the data calibration process. Table 13 shows the overall system performance comparison referring to the data calibration process. The results demonstrate that our data calibration process that utilizes an optimal linear fit method could remove the random phase offsets effectively, which is beneficial to further AoA and ToF estimation, and improve the overall performance.

To evaluate how the static reflection subtraction process facilitates our system, we compare the overall system performance with and without the static reflection subtraction method. The comparison results are shown in Table 14. The results demonstrate that the static reflection subtraction method could mitigate the influence of the static environments and further enhance the final overall performance.

Table 14. Overall system performance comparison with and without static reflection subtraction.

	PVE (cm)	MPJPE (cm)	PA-MPJPE (cm)
System with static reflection subtraction	4.01	3.51	1.90
System without static reflection subtraction	4.76	3.92	2.33

5 DISCUSSION

Constructing multi-subject 3D human mesh based on commodity WiFi is a challenging task. The performance of various experiments demonstrates that MultiMesh is capable of achieving promising results. However, the current proposed system still has some limitations for practical implementation.

Crowded scenario. Our current design can construct accurate human mesh for multiple subjects. However, this does not mean that the number of subjects in the sensing area is unlimited. For the crowded indoor environment, our system still has large errors when estimating the human mesh for each subject when they are very close and fully overlap relative to the WiFi devices. The reason is that the multi-path effect will become more severe when multiple subjects are overlapped, which will decrease the SNR of the received WiFi signals. Therefore, it is hard for our system to detect and track each subject when they are very close and fully overlap, and further not able to estimate accurate human mesh. However, when multiple people are performing activities in the indoor environment, their positions change dynamically over time. Therefore, the time when two people are fully overlapped and close to each other only occupies a short period of time, the impact of full overlap on the overall performance of our system will be kept limited. And as the applications of WiFi are rapidly evolving, the new generation of WiFi will support more number of antennas [7]. In future work, we could utilize more antennas as well as more WiFi devices to significantly improve the WiFi signal resolvability and tackle the problem of crowded scenarios.

Impact of other living things. In real-life implementation, the influence of pets is inevitable in indoor environments. For pets with small bodies, as the shape of the pet is much smaller than the human subject, the reflection signals from the human are much stronger than the signals from the small pet. Therefore, we can easily distinguish them from the Azimuth-ToF and Azimuth-AoD profiles. Therefore, they have a very limited impact on our system performance. However, for large pets, the reflection signals from them may be the same or even stronger than the signals reflected from the human body. It is hard to distinguish them based on the Azimuth-ToF and Azimuth-AoD profiles. They may be mispredicted as humans, which will decrease the performance of our system. In our future work, one possible way to solve this problem is that we could add a network block to distinguish if the signals are from the human body or other living things by the gait pattern, as demonstrated in the related work [37].

6 CONCLUSION

In this paper, we propose MultiMesh, which leverages the commodity WiFi to achieve 3D human mesh construction for multiple subjects. Specifically, we investigate the possibility of using the 2D AoA generated by an L-shaped antenna array, and further utilizing the AoD and ToF information to improve the resolvability of WiFi signal for precise multiple subjects separation. Our system proposes to mitigate the interference of indirect reflections by utilizing information from various signal dimensions. And we leverage the coherence of human motion to tackle the near-far problem. Based on the extracted clean information, we generate the 2D AoA image for each subject. A deep learning model is utilized to learn the correlation between the 2D AoA image and the corresponding human mesh. We conduct extensive experiments in real-world multi-subject scenarios. For instance, we conduct experiments with occlusion and perform human mesh construction for different distances between two subjects and different distances between subjects and WiFi devices to evaluate our system. Moreover, we also evaluate the accuracy of spatial information extraction and the subject detection performance of our system. The results demonstrate that MultiMesh could utilize commodity WiFi to construct accurate 3D human mesh effectively and robustly.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful feedback. This work was partially supported by the NSF Grants CNS-2131143 and CNS-1910519.

REFERENCES

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.

[2] Ankur Agarwal and Bill Triggs. 2005. Recovering 3D human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence* 28, 1 (2005), 44–58.

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*. 408–416.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European conference on computer vision*. Springer, 561–578.

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. 2019. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9157–9166.

[6] Zhe Chen, Chao Cai, Tianyue Zheng, Jun Luo, Jie Xiong, and Xin Wang. 2021. Rf-based human activity recognition using signal adapted convolutional neural network. *IEEE Transactions on Mobile Computing* 22, 1 (2021), 487–499.

[7] Cailian Deng, Xuming Fang, Xiao Han, Xianbin Wang, Li Yan, Rong He, Yan Long, and Yuchen Guo. 2020. IEEE 802.11 be Wi-Fi 7: New challenges and opportunities. *IEEE Communications Surveys & Tutorials* 22, 4 (2020), 2136–2166.

[8] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-net: A unified meta-learning framework for RF-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 517–530.

[9] Ruiyang Gao, Mi Zhang, Jie Zhang, Yang Li, Enze Yi, Dan Wu, Leye Wang, and Daqing Zhang. 2021. Towards position-independent sensing for gesture recognition with Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–28.

[10] John C Gower. 1975. Generalized procrustes analysis. *Psychometrika* 40 (1975), 33–51.

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

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 630–645.

[13] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[14] Donny Huang, Rajalakshmi Nandakumar, and Shyamnath Gollakota. 2014. Feasibility and limits of wi-fi imaging. In *Proceedings of the 12th ACM conference on embedded network sensor systems*. 266–279.

[15] 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 *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[16] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7122–7131.

[17] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2252–2261.

[18] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. Spotfi: Decimeter level localization using wifi. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 269–282.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[20] Jian Liu, Yingying Chen, Yan Wang, Xu Chen, Jerry Cheng, and Jie Yang. 2018. Monitoring vital signs and postures during sleep using WiFi signals. *IEEE Internet of Things Journal* 5, 3 (2018), 2071–2084.

[21] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. 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*. 267–276.

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.

[23] Virtual Medicine. 2017. Human Anatomy VR. https://www.oculus.com/experiences/gear-vr/1658650407494367/.

[24] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2020. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *Acm Transactions On Graphics (TOG)* 39, 4 (2020), 82–1.

[25] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. 2022. Human mesh recovery from multiple shots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1485–1495.

[26] Yili Ren, Sheng Tan, Linghan Zhang, Zi Wang, Zhi Wang, and Jie Yang. 2020. Liquid Level Sensing Using Commodity WiFi in a Smart Home Environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.

[27] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. 2016. General automatic human shape and motion capture using volumetric contour cues. In *European conference on computer vision*. Springer, 509–526.

[28] Ralph Schmidt. 1986. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation* 34, 3 (1986), 276–280.

[29] Leonid Sigal, Alexandru Balan, and Michael Black. 2007. Combined discriminative and generative articulated pose and non-rigid shape estimation. *Advances in neural information processing systems* 20 (2007).

[30] Sheng Tan and Jie Yang. 2016. WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition. In *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*. 201–210.

[31] Fei Wang, Jinsong Han, Feng Lin, and Kui Ren. 2019. Wipin: Operation-free passive person identification using wi-fi signals. In *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.

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

[33] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. 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*. 617–628.

[34] Yichao Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Wi-Mesh: A WiFi Vision-based Approach for 3D Human Mesh Construction. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 362–376.

[35] Yan Wang, Jie Yang, Hongbo Liu, Yingying Chen, Marco Gruteser, and Richard P Martin. 2013. Measuring human queues using WiFi signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. 235–238.

[36] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3645–3649.

[37] Chenshu Wu, Beibei Wang, Oscar C Au, and KJ Ray Liu. 2022. Wi-fi can do more: toward ubiquitous wireless sensing. *IEEE Communications Standards Magazine* 6, 2 (2022), 42–49.

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

[39] Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. 2019. mD-Track: Leveraging multi-dimensionality for passive indoor Wi-Fi tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[40] Jie Xiong, Karthikeyan Sundaresan, and Kyle Jamieson. 2015. Tonetrack: Leveraging frequency-agile radios for time-based indoor wireless localization. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 537–549.

[41] 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 *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 269–282.

[42] Youwei Zeng, Dan Wu, Jie Xiong, Jinyi Liu, Zhaopeng Liu, and Daqing Zhang. 2020. MultiSense: Enabling multi-person respiration sensing with commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[43] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10113–10122.

[44] Xiuyuan Zheng, Hongbo Liu, Jie Yang, Yingying Chen, Richard P Martin, and Xiaoyan Li. 2013. A study of localization accuracy using multiple frequencies and powers. *IEEE Transactions on Parallel and Distributed Systems* 25, 8 (2013), 1955–1965.