Human Parsing with Joint Learning for Dynamic mmWave Radar Point Cloud

SHUAI WANG* and DONGJIANG CAO*, Southeast University, China
RUOFENG LIU†, University of Minnesota, United States
WENCHAO JIANG, Singapore University of Technology and Design, Singapore
TIANSHUN YAO, Southeast University, China
CHRIS XIAOXUAN LU, The University of Edinburgh, United Kingdom

Human sensing and understanding is a key requirement for many intelligent systems, such as smart monitoring, human-computer interaction, and activity analysis, etc. In this paper, we present mmParse, the first human parsing design for dynamic point cloud from commercial millimeter-wave radar devices. mmParse proposes an end-to-end neural network design that addresses the inherent challenges in parsing mmWave point cloud (e.g., sparsity and specular reflection). First, we design a novel multi-task learning approach, in which an auxiliary task can guide the network to understand human structural features. Secondly, we introduce a multi-task feature fusion method that incorporates both intra-task and inter-task attention to aggregate spatio-temporal features of the subject from a global view. Through extensive experiments in both indoor and outdoor environments, we demonstrate that our proposed system is able to achieve ~ 92% accuracy and ~ 84% IoU accuracy. We also show that the predicted semantic labels can increase the performance of two downstream tasks (pose estimation and action recognition) by ~ 18% and ~ 6% respectively.

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

Additional Key Words and Phrases: Human Parsing, Joint Learning, Pose Estimation, Millimeter Wave Sensing

ACM Reference Format:
Shuai Wang, Dongjiang Cao, Ruofeng Liu, Wenchao Jiang, Tianshun Yao, and Chris Xiaoxuan Lu. 2023. Human Parsing with Joint Learning for Dynamic mmWave Radar Point Cloud. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 34 (March 2023), 22 pages. https://doi.org/10.1145/3580779

1 INTRODUCTION

Perceiving and understanding human activities plays an increasingly important role in human-centered intelligent applications (surveillance, smart control, AR/VR, fitness tracking, etc). Traditional approaches leverage cameras [41] or body contacted sensors [19], which are susceptible to harsh environment (e.g., poor-illumination, smoke, and fog), incur privacy concern, or introduce intrusive user experience. Recently, researchers propose to use wireless radio frequency (RF) signals reflected off the human body for human sensing [39, 42], which is robust

*Both authors contributed equally to this research.
†Corresponding author. The work was completed when the author was at the University of Minnesota.

Authors' addresses: Shuai Wang, shuaiwang@seu.edu.cn; Dongjiang Cao, djcao@seu.edu.cn, Southeast University, Nanjing, Jiangsu, China; Ruofeng Liu, liux4189@umn.edu, University of Minnesota, Minnesota, Minnesota, United States; Wenchao Jiang, wenchao_jiang@sutd.edu.sg, Singapore University of Technology and Design, Singapore, Singapore; Tianshun Yao, 220215632@seu.edu.cn, Southeast University, Nanjing, Jiangsu, China; Chris Xiaoxuan Lu, xiaoxuan.lu@ed.ac.uk, The University of Edinburgh, Edinburgh, United Kingdom.

2474-9567/2023/3-ART34 $15.00
https://doi.org/10.1145/3580779

against adverse environment, privacy-preserving and non-intrusive to users. Among many RF techniques, single-chip millimeter wave (mmWave) radar emerges as a low-cost sensor that can provide fine-grained 3D point cloud (PC) of the users and thus have been massively produced [2] and increasingly deployed [1, 6] in real-world scenarios such as smart buildings [6], vehicle cabins [33], and first-responder toolkits [3].

Driven by its promise, recent years have witnessed an upsurge of research using mmWave radar that demonstrates its effectiveness in gesture and activity recognition [24], pose estimation [34, 46], identity recognition [16, 48], etc. Despite their success, they commonly fail to explicitly obtain semantic meaning of point clouds, i.e., achieving human parsing (or equivalently human semantic segmentation) that tells *which body part produces each radar point in a point cloud*. The lack of such semantic information significantly limits the mmWave radar becoming an enabling technology for human-centered computing in everyday life. Fine-grained body part information is constantly required in human sensing applications. For example, a radar-equipped Hololens in the first-responder toolkit is supposed to pinpoint the point cloud of the teammate's hands for a rescuer to accurately hand over an item in the smoke, while gait analysis needs to extract point clouds from the specific lower-limbs. Meanwhile, adding the semantic information as an extra channel to inputs can make human sensing system more robust in the wild. In fact, it has been validated by various computer vision tasks that having semantic information in the loop can dramatically improve the accuracy of pose estimation [13, 29, 49], activity recognition [59] and person identification [14]. This benefit is even more prominent for the mmWave radar as the point clouds from such a low-cost RF sensor are intrinsically in lower quality than the images from vision sensors.

To this end, we propose *mmParse*, the first design of human parsing for mmWave point clouds. Specifically, the system takes the radar point cloud captured from a person as the input and feeds it into the deep neural network (DNN) which identifies the body part that corresponds to each point in the cloud. Point cloud annotated with semantic body part labels will be yielded as the final output of our system. In contrast to the literature with a focus on the overall activity or posture of the subjects, our design performs fine-grained "point-wise" analysis of the point cloud to figure out the body part that produces each point. We envision the semantically labeled point cloud can then be utilized by a wide spectrum of human sensing tasks to further unleash the potential of mmWave radars and push the limit of their performance.

Fine-grained point-wised semantic labeling is extremely challenging for radar due to the nature of radar point cloud. First, limited by the antenna size of the single chip design, the point clouds are extremely sparse (only few hundreds of points each frame, among which only dozens of points are correlated to the human body). This sparsity nature makes the detailed human structure difficult to be perceived from the point cloud even with human naked eyes. When training a DNN with such a sparse point cloud, it is also very challenging for the DNN to learn the features containing human structural information (e.g., pose) which are crucial cues for human parsing [52]. Secondly, due to the specular reflection of mmWave waveform on human body, only a subset of body parts that reflect signal towards radar are detected, while other parts that deflect signal off the radar are missing in the capture. The uncertainty in detection further adds to the challenge of the predicting the body part of detected points.

We propose a series of novel designs to address these challenges. First, to tackle the lack of body structure information due to the point cloud sparsity, we exploit a multi-task learning architecture, in which we jointly train our main task (i.e., human parsing) with an auxiliary task (i.e., human pose estimation). Our key insight is that the auxiliary task can efficiently guide the human parsing network to extract high-level structural features representing the subject's pose. Because of the strong correlation between pose and body parsing, these pose-related features help improve the accuracy and robustness of the parsing network in predicting the semantic labels. Secondly, to mitigate the detection uncertainty caused by specular reflection, we draw inspiration from the recent advances in the non-local network (NLN) [43] and design the network module to analyze each point by aggregating features from the global view over the spatio-temporal domains. Our critical observation is the low-cost mmWave radar often captures the point cloud from a different subset of body parts across consecutive

frames. By accumulating the point cloud over time, the network can obtain a holistic picture of the human body as well as the long-term dependencies of each body part in the consecutive frames. This information can help the network predict semantic labels more precisely even when parts of body of are skipped in some local frames. Moreover, We extend the NLN architecture for multi-task feature fusion and present a novel cross-task self-mutual attention mechanism that can further enhance the accuracy of the human parsing.

To summarize, our work makes the following contributions:

- To the best of our knowledge, mmParse is the *first* system that is capable of human parsing for commodity mmWave radar. The semantically annotated point cloud obtained by mmParse can be generically used by a wide spectrum of human sensing tasks to enhance the capability and robustness.
- We address technical challenges of parsing radar point clouds (e.g., sparsity and specularity) with a carefully designed end-to-end deep-learning pipeline containing several novel neural network designs (multi-task learning, non-local network, and cross-task attention).
- We comprehensively evaluate mmParse with 9 hours of radar data over 32 volunteers and 10 types of activities. The result shows that our design can predict the semantic body parts with ~ 92% accuracy and ~ 84% IoU accuracy. Furthermore, to demonstrate the benefits of semantic labels on human sensing tasks, we further utilize the annotated radar point clouds to serve 2 representative downstream tasks (i.e., pose estimation and activity recognition). mmParse improves the accuracy of two tasks by ~ 18% and ~ 6% respectively.

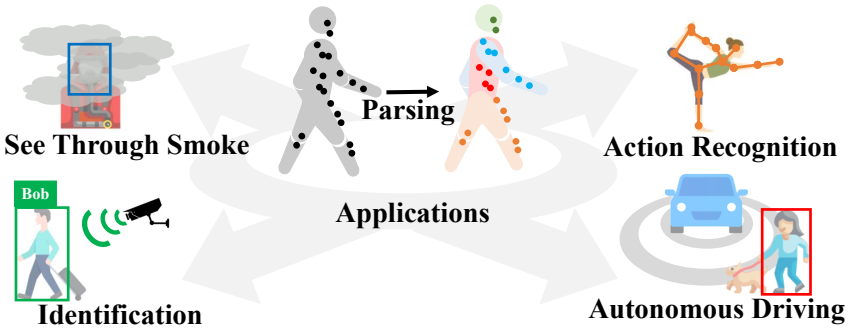


Fig. 1. Human parsing and its applications.

2 MOTIVATION AND CHALLENGE

2.1 The Need of Human Parsing for mmWave Data

Our objective is to identify the semantic body parts (torso, head, arms, legs, etc.) that correspond to each point in the point cloud. This section demonstrates benefits of human parsing in various scenarios. First, point-wised semantic labels are crucial for human users to visually understand radar point clouds. For example, in RESCUER systems of European Union [3], first responders are equipped with both radar and Hololens so that radar points are visualized using AR to augment their vision in adverse conditions such as smoke, darkness, rain, or fog. Semantic annotations for the body parts of interest in the radar point cloud are extremely useful for the rescuers to collaborate with other teammates without draining much cognitive load, e.g., to accurately hand over a tool to the hands of a teammate. Furthermore, raw point clouds of a person from a low-cost mmWave radar suffer from significantly lower resolution and more noise than images. The lack of details renders the reliability and accuracy of human sensing tasks very challenging, largely limiting the wide usage of low-cost radars in ubiquitous human-device interaction, even though a few tasks (e.g., pose estimation) have been be shown feasible. For

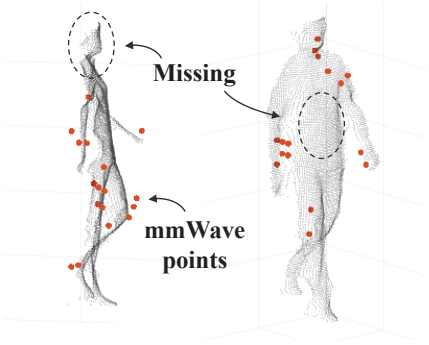


Fig. 2. Sparsity.

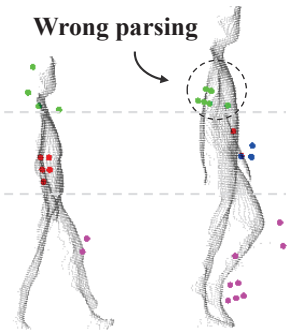


Fig. 3. Wrong parsing result (left).

example, while the Nest Hub [1] has been equipped with the mmWave radar (i.e., the Google Soli), its interaction efficacy with human users is still unsatisfying which lends the Hub itself only a sleep monitor device till now. Human parsing, on the other hand, can produce an extra dimension of point-wised semantics to complement the low-quality radar point cloud. As depicted in Fig.1, we envision that the proposed parsing solution can serve as a generic module to enable a wide range of downstream radar applications for human sensing (e.g., activity recognition, person identification, etc.).

2.2 Challenges of Parsing mmWave Data

2.2.1 Sparsity. As opposed to traditional massive MIMO radar, the commercial mmWave radar that targets human sensing (i.e., TI IWR6843) only has 3×4 antennas in order to be effective in both cost and size, which greatly limits its angular resolution (only 15° in azimuth and elevation). Furthermore, in order to lower the communication overhead, DSP algorithm such as CFAR (Constant False Alarm Rate) is employed in radar to further compress data into discrete points of detection. As a result, the produced radar point clouds (depicted in Fig.2) only contain a few dozens of reflective points of the subject per frame, which is 100 times less than optical sensor (e.g., Lidar). Conventional parsing methods [52] designed for images commonly rely on the shape of body parts and thus are not applicable to the sparse point cloud.

2.2.2 Specular Reflection. The mmWave signal transmitted from radar undergoes a specular reflection (i.e., angle of incidence equals the angle of departure) on the human skin, because the roughness of skin is considerably smaller than the wavelength of mmWave signal and the skin also has high water content [8]. Consequently, for a low-cost mmWave radar with a small antenna aperture, a large portion of reflected signal does not make its way back to the sensor, causing missing body segments in the point cloud. Fig.2 shows two examples where the head and torso are missing in the left and right captures respectively. Determining the body parts solely based on the information in the local frame (e.g., the height of the point) will cause errors such as the one depicted in Fig.3. The torso of the tall subjective is misclassified as the head because the point cloud of the head is missing due to specular reflection.

2.3 Comparison with Pose Estimation

Recent studies have shown the feasibility of pose estimation using mmWave radar [7, 35, 46, 53, 54, 56]. Human parsing in the work is conceptually and technically different from pose estimation. Human parsing is designed to enhance the information in the radar point cloud - it labels each radar point with its semantic body part. Therefore, human parsing can serve a wide range of downstream applications such as action recognition. In contrast, pose estimation data mines the information of point cloud to predict the subject's posture. Technically, human parsing for mmWave is a fine-grained point-wise analysis process, whereas pose estimation is a frame-level prediction

which mainly focus on the overall status of the subject. As a result, the performance of human parsing is not optimal directly with the models of existing design (as we will further illustrate in our evaluation).

In sum, the promising applications, unique technical challenges, and limitations of existing solution motivate us to design a point-wised semantic parsing method for mmWave point cloud.

3 PRELIMINARY

3.1 Principles of mmWave Radar

The single-chip mmWave radar is based on the principles of frequency modulated continuous wave (FMCW) [2] and has the ability to simultaneously measure the range, relative radial speed and angle of the target. Specifically, the FMCW radar repeatedly transmits chirp signals for a short period of time whose frequency increase linearly with time. Then the radar sensor produces Intermediate Frequency (IF) signal by mixing the received signal reflected by objects with the transmitted signals. The IF signal is processed to obtain the three-dimensional information of the object.

3.1.1 Range Measurement. The distance d between the object and the radar is calculated as:

$$d = \frac{f_{IF}\, c\, T_c}{2\, B} \tag{1}$$

where the c is the speed of light, f_{IF} is the frequency of the IF signal, B is the bandwidth swept by chirp, and T_c is the duration of chirp. To measure the range of multiple objects at different ranges, a fast Fourier transform (FFT) is performed on the IF signal (i.e., range-FFT). The result of range-FFT represents the frequency response at different ranges. Thanks to the centimeter level range resolution, it has the ability to detect the position of the human body parts.

3.1.2 Angle Estimation. To depict the exact positions of objects in a spatial Cartesian coordinate system, the angle estimation is indispensable. The mmWave radar uses a linear antenna array to estimate the object angle. After emitting chirps with the same initial phase, RF Front-end simultaneously samples from multiple receiver antennas. Based on the differences in phase of the received signals, the angle of the reflected signal could be estimated. Formally, the angle is calculated as:

$$\theta = \arcsin \frac{\lambda\, \omega}{2\, \pi\, l} \tag{2}$$

where ω denotes the phase difference, l represents the distance between consecutive antennas and λ is the wavelength. Subsequent to range and angle estimation, strong peaks are detected which yield a compact set of 3-D points in a spatial Cartesian coordinate system.

3.2 PointNet.

In our proposed deep learning model, we adopt PointNet [31] as our backbone network for feature extraction. PointNet [31] is a pioneer deep learning method that is originally designed for point cloud mesh. Rather than projecting or quantizing irregular point clouds onto regular grids in 2D or 3D, PointNet can ingest point clouds directly and thus features high computational efficiency. Additionally, it utilizes permutation-invariant operators (e.g., pointwise MLPs and max pooling) to deal with the unordered points, which makes the results invariant to the permutation of the input points. Since radar point cloud is also unordered, we design the feature extraction based on PointNet framework.

3.3 Non-local Network.

Attention mechanism was proposed for language modeling [40]. Given a query and a set of key-value pairs, attention mechanism first computes the attention weight between the query and each key. Then it uses the attention weights to aggregate the values by weighted sum as the output. Similarly, Non-local model [43] was

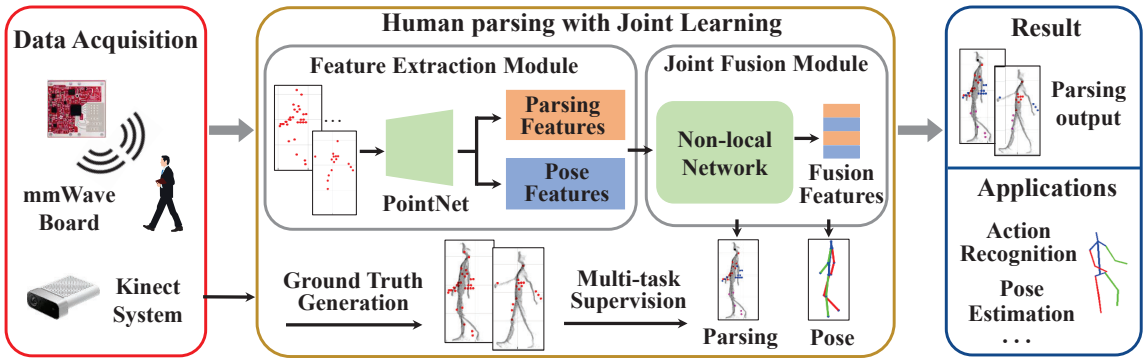


Fig. 4. System architecture. The Kinect system is *only* used in the offline training phase to provide pseudo ground truth but *not* required in the inference phase.

proposed for learning self-attention in 2D or 3D vision modeling to capture long-range dependencies [58]. In this paper, we aggregate the features from a global perspective based on Non-local network.

4 OVERVIEW

4.1 Problem Formulation

In this work, we consider the problem of human parsing for dynamic mmWave radar point cloud. The input is a sequence of mmWave point clouds of the target subject perceived over a period of time denoted as $S = \{C_t\}_{t=1}^{M}$, where M represents the length of the sequence. Each point cloud $C_t = \{p_{i,t}\}_{i=1}^{N}$ contains N radar points. Each point $p_{i,t}$ is a vector of five features, i.e., 3D coordinates $(x_{i,t}, y_{i,t}, z_{i,t})$, intensity $(s_{i,t})$ and velocity $(v_{i,t})$. Our task is to predict the semantic label $L_{i,t} \in \{0, 1, ..., B\}$ for every point, where B is the the number of body parts. For each point cloud C_t, our model outputs the predicted body part label of each point.

4.2 System Overview

From the system perspective, mmParse system consists of three major components depicted in Fig.4:

4.2.1 Data Acquisition. mmWave radar collects raw point clouds (i.e., without semantic labels) for subsequent parsing. Specifically, a radar emits FMCW signals and captures reflections from a person, which generate 3D radar points. Additionally, in the training stage, we use Azure Kinect [4] to produce dense 3D meshes of the subject that are processed to obtain the ground truth of semantic labels as well as pose labels (i.e., joint locations) for the multi-task learning.

4.2.2 Human Parsing. The main contribution of mmParse is the end-to-end trainable deep learning model to estimate the semantic body labels of radar points. In this component, we first conduct feature extraction from point clouds. Then we use the proposed multi-task learning model to jointly perform parsing and pose estimation, which addresses the sparsity of the point cloud. We also exploit Non-local Network for multi-task feature fusion to mitigate the problem of missing body parts due to specular reflection. The details will be described in Section 5.

4.2.3 Parsing Result & Applications. The output of human parsing components is point clouds with annotated semantic labels. The annotated point cloud can be used as the input for various human sensing tasks (e.g., action recognition, pose estimation), which will be evaluated in Section 7.

5 DESIGN

5.1 Design Methodology

As the middle diagram in Fig.4 demonstrates, the critical component of mmParse is a DNN model that can parse 3D mmWave point cloud into body segments. The critical design methodologies of the model are as follows.

The first challenge the network needs to concur is the sparsity of point cloud caused by the limited resolution of the commercial mmWave radar (discussed in the 1^{st} paragraph of Section 2.2). With only a few dozens of points of the subject per frame, it is extremely difficult for a DNN to understand human structure (e.g., the correlation between various body parts). From the model training perspective, the sparse semantic labels alone cannot provide sufficient supervision for the network to learn features that represent human structure. To address the challenge, we propose to jointly train human parsing (i.e., the main task) with other auxiliary tasks. We hypothesize that through the multi-task and multi-label joint learning, the DNN can extract more diverse feature representations from sparse point cloud which are shared among tasks to improve their performance. Among a wide range of human sensing tasks, we choose pose estimation as the auxiliary task because pose provides crucial structural cues for locations of the body parts. In fact, the skeletal key-points (results of pose estimation) and body segments (results of parsing) have strong spatial correlation such that they could be estimated from one another. For example, given the knee's position, the leg's area can be estimated. Conversely, given the area of the leg, you can also predict the position of the knee. Therefore, as the bottom of Fig.4 shows, during the model training, we supervise the network with both parsing labels and additional pose estimation ground truth obtained by Kinect. Note that these pose labels from Kinect is not required in the training and the operation of mmParse only relies on input from mmWave radar.

Another critical challenge lies in the uncertainty of captured body parts due to the specular reflection (discussed in the second paragraph of Section 2.2). We address this issue by designing a network with a global view over a large time window. The benefit is two-fold. First, we observe that due to the specular reflection, consecutive detection of the same body part is often separated by many frames. For example, when a subject walks toward radar, the leg is most likely to be detected each time when the subject makes a specific leg lift pose. A network with a global view can capture such long-distance dependency and aggregate the features of a body part over time. Furthermore, because of human motions, the radar captures different subsets of body parts over time. A global view allows the network to compose these snapshots together and form a holistic picture of the subject, providing a useful context for both parsing and pose estimation. Technically, the global view can be realized by using Non-local network (NLN) technique which uses the self-attention mechanism to capture long-distance spatio-temporal dependencies. We further extend NLN with cross-task attention to aggregate global features across parsing and pose estimation tasks.

The human parsing with joint learning component in Fig.4 gives an overview of our proposed deep learning framework, which is mainly composed of two modules: a Feature Extraction module to extract the high-level representation of parsing and pose estimation from point cloud (details in Section 5.2) and a Joint Fusion module to aggregate features of parsing and pose estimation with self-mutual attention (details in Section 5.3).

5.2 Multi-task Feature Extraction Module

In this section, we extract features for human parsing and pose estimation in parallel. The parsing and pose feature extractor employs a similar network architecture demonstrated in Fig.5, which includes point module, frame module, and frame aggregation. For ease of illustration, we use the human parsing feature extraction as an example while the difference is pointed out in the end.

5.2.1 Parsing Feature Extraction. The point module encodes each point $p_{i,t} = \{x_{i,t}, y_{i,t}, z_{i,t}, s_{i,t}, v_{i,t}\}$ in the point set C_t of t^{th} frame independently using a shared-weighted MLP (Multi-layer Perception). The output (named point feature) is a high-level representation of each point denoted as $e^H_{i,t} = MLP(p_{i,t}; \theta_e)$, where θ_e is the learnable parameters of the MLP and H represents human parsing task.

The point features of all points in a frame are then aggregated into a frame feature to extract global information of the frame. The classic approach (e.g., pointNet [31]) commonly uses the max-pooling operation to obtain frame features due to its permutation invariance. However, max-pooling only reserves the maximum value and discards other information, which causes severe information loss to the sparse radar points.

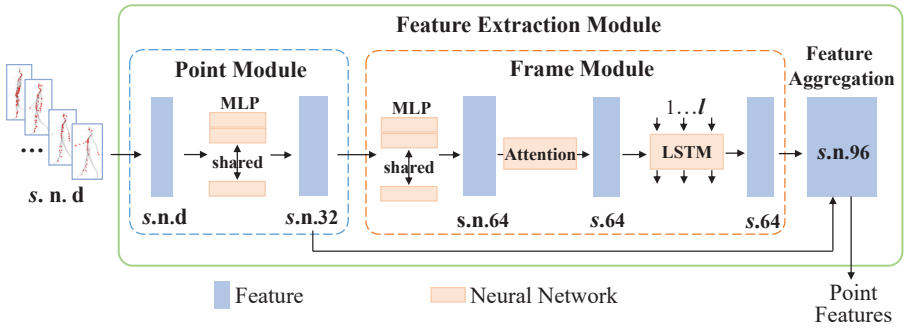


Fig. 5. Feature extraction module. s is the length of frames, n is the number of points in one frame and d is the dimension of the feature.

To minimize the information loss, we replace the max-pooling with attention mechanism [40]. For each frame, attention computes a score for each point and calculates a weighted sum of all scores. The operation is also permutation-invariant and can dynamically adjust the contribution of each point based on the context. Formally, each point feature $e^H_{i,t}$ of point $p_{i,t}$ is first encoded to a higher dimensional representation $h^H_{i,t} = MLP(e^H_{i,t}; \theta_h)$. Then, we use attention function $A()$ to obtain the frame feature of t^{th} frame as follows:

$$g^H_t = \sum_{i=1}^{N} A(h^H_{i,t}; \theta_a) \times h^H_{i,t} \tag{3}$$

where N is the number of points in t^{th} frame and θ_a is the parameter of attention function. Finally, we concatenate the frame feature to point feature and obtain a feature vector for each point denoted as $z^H_{i,t} = [e^H_{i,t}; g^H_t]$. Subsequent operations are conducted on the combined frame vector.

5.2.2 Pose Feature Extraction. We use a slightly different network to extract features for pose estimation. With the frame feature g^P_t of a specific frame (P represents pose estimation), we further exploit the relationships between the adjacent frames to guarantee the smoothness of pose across frames. More specifically, the frame feature g^P_t is processed by a LSTM (Long short-term memory network) to obtain $r^P_t = LSTM(g_t, r_{t-1}; \theta_r)$, where θ_r denotes the parameters of LSTM. r^P_t is concatenated with point feature (i.e., $e^P_{i,t}$) as the pose estimation feature vector $z^P_{i,t}$.

5.3 Multi-task Joint Fusion Module

The feature extraction modules obtain two sets of feature vectors for human parsing (i.e., $z^H_{i,t}$) and pose estimation (i.e., $z^P_{i,t}$) at individual frame C_t. These heterogeneous features need to be elaborately fused such that the structural information in the pose estimation features can benefit the performance of human parsing. Furthermore, we also need to aggregate the features across a large time window for both tasks in order to address the uncertainty of detected body parts. To achieve both goals simultaneously, we design a multi-task feature fusion module using Non-local network with self-mutual attention.

As shown in Fig. 6, we employ two parallel Non-local networks (NLN) for parsing and pose estimation respectively. The parsing NLN takes a sequence of parsing features as the input and performs inta-task self attention (green dot in Fig. 6) to aggregate the features across different frames. This produces a global context for classifying the body parts in each frame, which can address the problem of missing body parts due to the specular reflection in the local frame. More specifically, we first stack the feature map $z^H_{i,t}$ of all points over time into a feature matrix Z^H. As Eq. 4 illustrates, Z^H is linearly transformed into embedding spaces $W^H_\theta Z^H$ and $W^H_\phi Z^H$, where W^H_θ and W^H_ϕ are learnable parameters. The embedding vectors are then dot-product and normalized

by non-linear function σ (e.g., softmax) to produce a intra-task attention matrix a^H. a^H matrix estimates the spatio-temporal correlation among the points in each pair of frames. A similar process is done in the NLN of pose estimation, which produces a self-attention matrix for pose estimation task denoted as a^P.

$$\begin{aligned} a^H &= \sigma[(W_\theta^H Z^H)^T W_\phi^H Z^H] \\ a^P &= \sigma[(W_\theta^P Z^P)^T W_\phi^P Z^P] \end{aligned} \tag{4}$$

To fuse the features across parsing and pose estimation task, we further need to figure out the correlation between parsing and pose features. To achieve this, the pose estimation features are fed into the parsing NLN and their spatio-temporal correlation with parsing features are computed using inter-task cross attention (red dot in Fig.6). As Equation 5 shows, we linearly transform the parsing features Z^H and pose features Z^P into embedding spaces using $W_\theta^{H\to P}$ and $W_\phi^{H\to P}$. The results are then dot-product and normalized to obtain the inter-task attention matrix $a^{H\to P}$, which represents the correlation between parsing and pose estimation features across time and space. The inter-task attention matrix for pose estimation task (denoted as $a^{P\to H}$) is derived in a similar way.

$$\begin{aligned} a^{H\to P} &= \sigma[(W_\theta^{H\to P} Z^H)^T W_\phi^{H\to P} Z^P] \\ a^{P\to H} &= \sigma[(W_\theta^{P\to H} Z^P)^T W_\phi^{P\to H} Z^H] \end{aligned} \tag{5}$$

With the intra-task and inter-task attention matrices, we aggregate the parsing features and pose estimate features in all the frames to predict the body parts of the points in a specific frame. This is done by multiplying Z^H with intra-task as well as inter-task attention matrices, before which Z^H is transformed into embedding spaces $W_g^H Z^H$. This process essentially computes the weighted sum of all features based on their correlation with the current frame, as demonstrated in Equation 6. The aggregated intra-task and inter-task features are concatenated and the result is combined with the origin feature Z^H by element-wise addition and produces the final the aggregated parsing Y^H. The aggregated pose estimation features Y^P is obtained following the same method. Finally, Y^H and Y^P are processed by a MLP and a fully connected neural network (FC) respectively to predict the human body part category $L_{i,t} \in \{0, 1, ..., B\}$, and human skeleton point location $J = \{(x_i, y_i, z_i)|i = 1, ..., D\}$. B

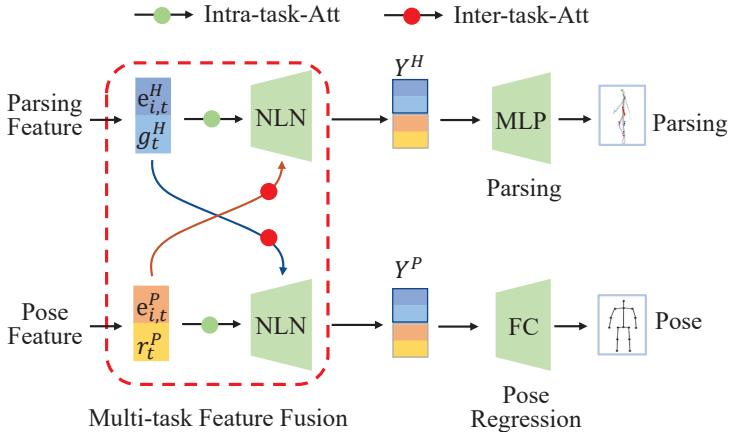


Fig. 6. Joint fusion module.

and D are the number of body part categories and body skeleton points.

$$\begin{aligned} Y^H &= W_y^H[a^H W_g^H Z^H; a^{H\to P} W_g^{H\to P} Z^P] + Z^H \\ Y^P &= W_y^P[a^P W_g^P Z^P; a^{P\to H} W_g^{P\to H} Z^H] + Z^P \end{aligned} \tag{6}$$

5.4 Multi-task Supervision

5.4.1 Loss of Human Parsing. The supervised function of human parsing task is to minimize the error between the predicted class (semantic part) and ground truth class of every point. Given that there are K semantic parts to segment, we minimize the Cross Entropy loss:

$$L_H = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^n \; logp(\hat{y}_k^n) \tag{7}$$

where N is the number of points, K is the number of classes (semantic parts). And y_k^n is the function of 0 and 1, $y_k^n = 1$ if the class of sample n is k, and vice versa. $p(\hat{y}_k^n)$ is the prediction probability that sample n is belong to class k.

5.4.2 Loss of Pose Estimation. The pose estimation task is to minimize the error between the predicted position and ground truth of skeleton joint, given that there are M joints on the skeleton tree, we minimize the Mean Squared Error (MSE) loss:

$$L_P = \frac{1}{M}\sum_{m=1}^{M} \|\hat{p}_m - p_m\| \tag{8}$$

where $\|\cdot\|$ denotes L2-norm, $\hat{p}_m$ and p_m is the predicted position and corresponding ground truth of m^{th} skeleton joint.

The total training objective is:

$$L = \gamma L_H + \beta L_P \tag{9}$$

where γ and β are the hyper parameters. The whole framework is trained end-to-end.

6 IMPLEMENTATION

This section presents the implementation of mmParse, including the experimental setup for data collection, the radar data preprocessing, and details of the neural network.

6.1 Experiment Platform

6.1.1 mmWave Radar Platform. We use a commercial and off-the-shelf millimeter-wave radar IWR6843-BOOST [2] for the radar data acquisition. The radar operates in a frequency band from 60 GHz to 64 GHz with a wavelength of $\sim$ 4mm. The device has three tx antennas and four rx antennas that form a 60° FoV (Field of view) in both azimuth and elevation with a $\sim 15^\circ$ angular resolution. We follow the standard FMCW processing chain provided by TI to produce 3D point cloud. For reproducibility, the detailed configuration parameters of the device are provided as follows. The radar transmits 10 frames per second with 32 chirps per frame. The start frequency of the chirp is set to 60.065 GHz and the bandwidth is set to 3194.88 MHz. The frequency slope is set to be 12.5 MHz/us.

6.1.2 Kinect Platform. To collect the ground truth of human parsing and pose estimation, we use Azure Kinect which is equipped a RGB-D camera that can collect fine-grained 3D mesh of the subjects. Because of its much better resolution (typical systematic error $< 11mm + 0.1\%$ of distance) than mmWave radar ($\sim 4cm$), Kinect has been used as a sensor to collect ground-truth in multiple sensing tasks [42, 47]. In order to further improve the trustworthiness of the ground-truth, We calibrate the coordinates and timestamp of radar with Kinect. To further

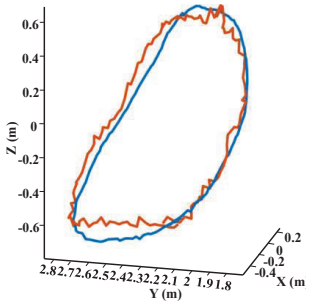


Fig. 7. Validation of the ground truth data.

validate the credibility of ground-truth data, we quantitatively analyze the calibration errors of mmWave radar and Kinect. Fig.7 shows the output of mmWave Radar (in orange) and Kinect (in blue) for the experiment that the subject writes the letter "O" in mid-air using corner reflector. The two curves have a large overlap and the qualitatively average error is 4.3 cm, which shows that using Kinect as the ground-truth device is credible. Then we label mmWave point cloud using the body part identified from the dense 3D mesh. More specifically, For each mmWave point, we find the nearest Kinect point and label mmWave point with the body part of the Kinect point. If the distance between the mmWave point and kinect point is larger than 30cm, the mmWave point is labelled as a noisy point. In addition, we use Body Tracking SDK [9] of Kinect to obtain the positions of 17 skeleton points for pose estimation.

6.2 Data Acquisition and Preprocessing

Data Acquisition. We recruited a total of 32 participants for data collection. The participants consist of 17 males and 15 females with ages varying from 17 to 29, heights from 158 cm to 186 cm, and weights from 45 kg to 80 kg. During the experiment, each participant is asked to perform 10 different activities, including: (1) walking on the spot; (2) rotating body; (3) clapping; (4) swinging arms upward and downward; (5) swinging arms horizontally; (6) swing left or right arm; (7) walking back and forth; (8) walking back and forth with arms swing; (9) lunging with left leg; (10) lunging with right leg. To evaluate the effectiveness across different environments, we collect data in three different sites including both indoor and outdoor scenarios as shown in Fig.8. Fig.8(a) is a big hallway in an office building with dim light, there are some tables and chairs in this scene. Fig.8(b) is an ordinary classroom with light from the lamp tube, where desks and chairs are around. And Fig.8(c) is a road with trees on both sides. The device is placed at one end of the areas and participants are asked to perform activities within the activity area with a $1.5m \sim 4.5m$ distance from the device. In each experiment, the participant keeps performing for 5 minutes, which produces more than 3000 frames of radar data for each activity.

(a) Hallway (indoor) (b) Room (indoor) (c) Road (outdoor)

Fig. 8. Experimental scenarios.

Preprocessing. Due to the beam spreading, diffraction and reflection from ambient objects such as walls, floors, and ceilings, the propagation of mmWave signals between objects and transceivers tends to travel through multiple paths [28]. Consequently, unwanted points often appear in the radar point cloud which are widely known as the 'ghost points' [27]. In order to mitigate the impact of these noisy points and separate human subjects out, we preprocess raw radar point cloud with the following steps.

First, we remove unlikely points based on their 3D locations and our prior knowledge about the areas where people are commonly located. For instance, when the device is installed 1 meter above the ground, any point with a height above 2.5 meters or below -1 meter are unlikely to come from human bodies and thus can be safely removed. Next, we apply the DBScan algorithm [10] to acquire the cluster of points of a person such that the near-person noise can be suppressed. DBScan is a density-aware clustering algorithm that can divide a point cloud based on the distance and the density described based on a set of neighborhoods in the 3D space. Our implementation carefully follows [55] to cluster radar points belonging to subject.

6.3 Network Setting and Training

In this section, we describe the details of the mmParse model, training, and testing procedure. In feature extraction module, we implement two multi-layer perceptions (MLP) and the size of each layer are (6,12,24) and (32,48,64) respectively. We adopt batch normalization followed by ReLU activation functions after all layers. The attention operation in feature extraction is implemented by fully connected layer of size (64,1). The LSTM has 3 layers and the size of each layer is 64. In joint fusion module, we utilize 1D NONLocalBlock [43] with dot-product. We implement MLP $(64, partnum)$ and FC $(64, 17 * 3)$ for parsing and pose estimation respectively.

For every activity, 75% of the data records of collected from each subject are used for model training, and the remaining 25% serve as the testing set. The training records are segmented to 5-second fragments, containing a sequence of 50 point cloud frames. During the training, we set the batch size to 32 and the learning rate to 0.0002. The training procedure takes 5000 epochs. The hyper-parameters γ and β in the loss functions (Equation 9) are set as follows: $\gamma = 1$ and $\beta = 5$. Our model is implemented with Python 3.7 and PyTorch 1.8.0, and trained with NVIDIA RTX 3090.

7 SYSTEM EVALUATION

This section presents the performance evaluation of mmParse. We start with the evaluation methodology including evaluation metrics (Section 7.1) and competing approaches (Section 7.2). The overall performance is reported in Section 7.3, followed by separate evaluations of each critical design component in Section 7.4. We then extensively study the impact of various factors on the performance in Section 7.5.

7.1 Evaluation Metrics

To quantify the performance of our proposed approach, we adopt the following metrics that are commonly used to evaluate the accuracy of body parsing in the literature [13, 22].

Overall Accuracy (oA). Overall accuracy (oA) [31] formulates human parsing problem as a per-point classification problem, and oA is defined as the proportion of correctly labeled points among all the points in the point cloud. This metric mainly measures the overall performance over the entire point cloud.

Mean Intersection over Union (mIoU). mIoU [30] is a widely adopted metric for human parsing. Given test samples with annotations, the IoU for a given body part is defined as the percentage $\frac{|A \cap B|}{|A \cup B|}$, where A is the set of points predicted as stemming from this body part and B is the ground-truth set of points for this part. From the view of a single body part, IoU measures how well and how completely it is segmented. Averaging this metric over all semantic bodys yields the mean IoU (mIoU) score.

Mean Dice score (mDice). With same notations as above, Dice [30] for a given body part is defined as $\frac{2|A \cap B|}{|A|+|B|}$. Averaging over all semantic parts yields the global performance mean Dice (mDice). Semantic segmentation can

be seen as a 1-vs.-all classification problem for each class, and the Dice amounts to the mean value of the task's precision and recall (e.g., F1 score). Since IoU and Dice metrics are complementary [30], we report both of them in our evaluation.

Average Joint Localization Error (AJE). We jointly train the parsing network with a pose estimation network to obtain body structural features. To evaluate the effectiveness of the feature extraction, we examine the accuracy of pose estimation using these features. The AJE metric is defined as the average Euclidean distance between the predicted skelection key points (i.e., joint locations) and their ground truths [35, 46].

7.2 Baseline

We compare our approach with the following baselines to demonstrate the effectiveness of mmParse. Since mmParse is the first design of human parsing for mmWave point cloud, we develop three baseline approaches based on the popular network architectures for deep learning designs on mmWave point cloud. And in order to prove the superiority of our designed model over existing pose estimation works, we compare our proposed method with three representative pose estimation models.

PointNet++ + RNN (P+R). The enhanced version of PointNet (i.e., PointNet++ [32]) is used for feature extraction of individual frames. PointNet++ can capture local structure in the point cloud with a sampling layer, a grouping layer and a PointNet layer that abstracts fine geometric structures from the neighborhood points in a hierarchical way. Further, by stacking several set abstraction levels, PointNet++ can obtain multi-scale features. Secondly, the feature vectors are fed into RNN to exploit their temporal and spatial correlation. Lastly, a multi layer perception (MLP) aggregates the feature to predict the semantic labels.

DGCNN + RNN (DGR). DGCNN [44] is a graph CNN network designed for learning tasks on point clouds including classification and segmentation. It addresses the lack of topological information in the point cloud by designing a model to recover the inherent topology. DGCNN is followed by RNN and MLP layers.

Voxelization + 3DCNN + RNN (VCR). Voxelization [55] is another popular method for feature extraction from mmWave point cloud. It first maps the point cloud to 3D voxel grid, and then the 3D voxel grid is converted into feature vector by 3DCNN.

mm-Pose [35]. mm-Pose is a novel approach to estimate human skeletons using an mmWave radar. A novel low-size high-resolution radar-to-image representation is presented and a forked CNN architecture was used to predict the real-world position of the skeletal joints in 3-D space. To show the superiority of the proposed point-wised mmParse model over the single-task frame-wised pose estimation model, we implement mm-Pose and port it to human parsing task to support the point-wised semantic parts as the outputs.

mmMesh [46]. To demonstrate the superiority and difference of the proposed mmParse over the existing work of pose estimation, we implement the latest pose estimation model using mmWave point cloud (i.e., mmMesh) and extend it to support the point-wised semantic parts as the outputs. mmMesh utilizes the PointNet as the backbone network and designs an anchor point module to address the misalignment of the sparse point cloud with the human body parts. Specifically, We develop the same deep learning model as mmMesh while the output is point-wised semantic parts rather than frame-wised skeleton.

Point-convolution-based (PCB) [56]. PCB is also a pose estimation model and extracts point cloud features based on point-by-point convolution. we also implement the model and extend it to support the point-wised semantic parts as the outputs.

7.3 Overall Performance

This section provides the overall performance of mmParse and compares them with baseline approaches. Depending on the need of applications, our design can segment the point cloud into 3 parts (torso, arms, legs), 6 parts (head, torso, left arm, left leg, etc.), or 10 parts (head, torso, upper left arm, lower left arm, lower left leg, upper right leg, lower right leg, etc.). Table.1 reports the results for these options. mmParse achieves 92.06% overall

Table 1. Overall performance of mmParse.

Models	10 body parts			6 body parts			3 body parts		
	oA(%)	mIoU(%)	mDice(%)	oA(%)	mIoU(%)	mDice(%)	oA(%)	mIoU(%)	mDice(%)
P+R[32]	80.68	76.82	80.23	83.16	76.59	79.24	87.31	76.03	79.06
DGR[44]	80.48	76.70	79.46	84.92	77.04	80.03	87.98	76.32	79.64
VCR[55]	81.03	77.03	80.98	84.26	76.58	79.53	89.06	77.48	81.23
mm-Pose[35]	77.83	74.91	77.71	80.64	75.67	77.42	84.86	73.81	78.64
mmMesh[46]	80.92	76.94	80.68	84.44	76.84	79.83	88.02	76.49	81.06
PCB[56]	74.45	72.35	73.63	79.19	73.68	75.65	82.48	70.53	73.05
mmParse	**87.23**	**84.30**	**86.54**	**89.35**	**83.26**	**86.31**	**92.06**	**83.78**	**87.94**

accuracy for 3 body parts parsing and 89.35% for 6 parts option. In the most challenging setting (10 body parts), it can still maintain 87.23% accuracy. Compared with baselines of P+R, DGR and VCR, the performance of our method outperforms the baseline approaches by at least 6.27%, mmParse also improves mIoU and mDice by up to 7.63% and 8.88%, the results show that mmParse is effective in predicting semantic body label and various strategies we proposed can enhance the performance of human parsing. Compared with baselines of mm-Pose, mmMesh and PCB, the performance of our method outperforms the baseline approaches by at least 6.31% for oA metric, showing that the multi-task learning architecture and the joint fusion module based on point-wised self-mutual attention mechanism in our model has the advantage than the existing single-task frame-wised pose estimation model.

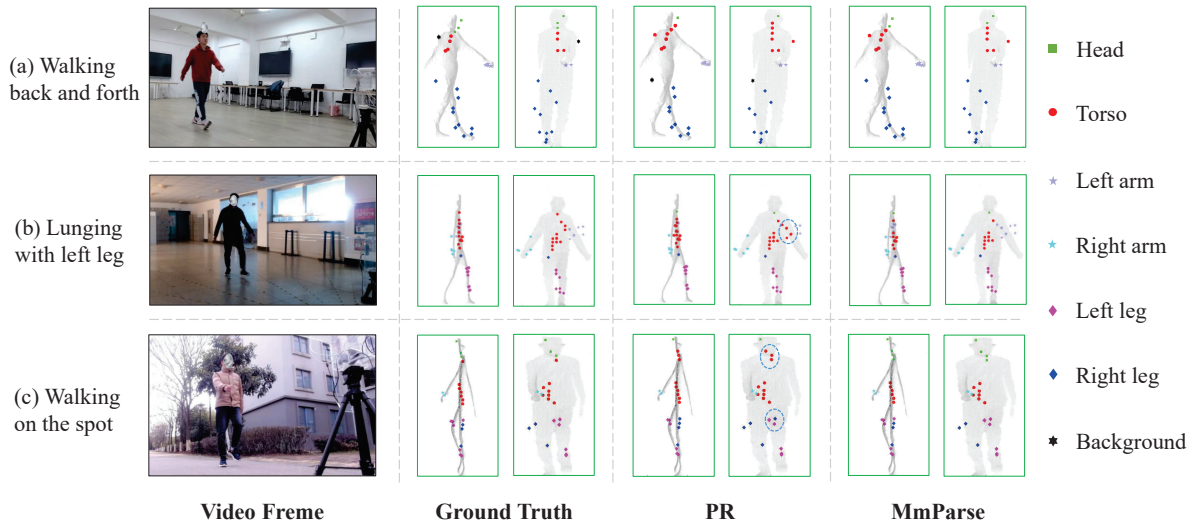


Fig. 9. Qualitative results (parsing into six body parts).

Fig.9 demonstrates a few examples of the parsing outputs. We depict points with different colors and shapes to indicate semantic labels. The rows (a)-(c) correspond to 3 relatively complex activities including walking back and forth, lunging with left leg, and walking on the spot. The first column in each row shows the video frame when the subject is conducting the activity. The second and the fourth column compare the corresponding ground truth human parsing generated by the Kinect system and the predicted human parsing based on our proposed mmParse model. Despite the sparsity of points and missing body parts (e.g., the left leg in Fig.9(a), the right leg in Fig.9(b) and the left arm in Fig.9(c)), our system is still able to predict the different body parts correctly. In contrast, the baseline (**PR**) fails to label points on the upper left arm in the lunging and points on the head for walking on spot. These improvements are achieved by two critical designs, (i) *multi-task* learning architecture in our model encodes the body structure information in the network, (ii) the joint fusion module based on self-mutual attention mechanism captures the global view over the spatio-temporal domains.

To further illustrate the effectiveness of multi-task learning, we also analyse the result of the auxiliary pose estimation task. As Table.2 depicts, the mean joint localization error is 2.320*cm*. It outperforms the baseline approaches which reduces the joint localization error by 0.95cm at their best, showing that the sharing the features between human parsing and pose estimation is beneficial for both tasks.

Table 2. Overall performance of pose estimation.

Models	P+R[32]	DGR[44]	VCR[55]	mm-Pose[35]	mmMesh[46]	PCB[56]	**mmParse**
AJE(cm)	3.276	2.991	2.874	3.148	2.983	2.455	**2.320**

7.4 Ablation Study

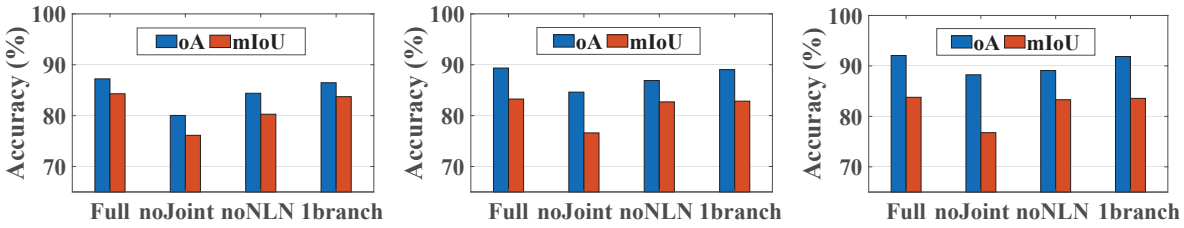
In Section 5, several designs of the deep learning model were introduced. To evaluate their effectiveness and impact on the results, we measure the performance of the system when specific components are disabled. Specifically, we conduct experiments with the following 4 different settings.

Full version (Full): All the components in Section 5 are used.

w/o joint learning (noJoint): The branch of pose estimation task in the network is disabled. The network is trained only with the parsing task. This setting examines the performance gain from the auxiliary task.

w/o self-mutual attention with non-local network (noNLN): The non-local network (NLN) in the multi-task joint fusion module is replaced by RNN. This setting examines the benefits of NLN (e.g., learning long-distance spatio-temporal correlation among radar frames).

1 branch joint learning (1 branch): Our network consists of two separate network branches for parsing and pose estimation tasks. Another popular multi-task learning strategy is to make two tasks share the neural network layers for both feature extraction and fusion, which we denote as 1 branch joint learning. The setting compares our two branch architecture with the one with a single branch.



(a) Parsing into 10 parts. (b) Parsing into 6 parts. (c) Parsing into 3 parts.

Fig. 10. Ablation study of mmParse.

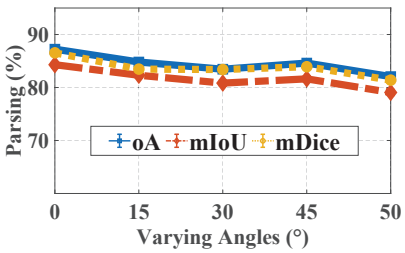


Fig. 11. Evaluation of different angles.

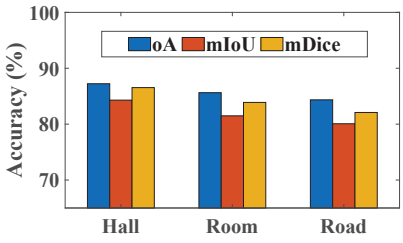


Fig. 12. Evaluation of different scenes.

Models	10 body parts oA(%)	mIoU(%)	mDice(%)
P+R[30]	74.83	73.07	74.62
DGR[41]	74.80	73.92	74.70
VCR[51]	75.84	74.69	75.17
mm-Pose[34]	72.28	67.93	71.49
mmMesh[45]	74.39	72.35	73.01
PCB[55]	69.36	66.01	68.57
mmParse	**84.65**	**82.40**	**84.34**

Fig. 13. Performance of cross-subject.

The results are shown in Fig.10. The full version setting achieves the best results, confirming that every design components are useful for the overall human parsing task. Among all components, joint learning with pose estimation plays the most important role . For example, there is 7.2% overall accuracy loss in **noJoint** setting when parsing into 10 parts. We therefore conclude that training human parsing with a pose estimation task is an effective strategy for the low-quality radar point cloud. Furthermore, the result of **noNLN** demonstrates that NLN can also improve the performance, especially for 10 body parts setting. Hence, the long-distance correlation between parsing and pose estimation features across time and space provide useful contexts for predicting body part labels. Finally, the performance drops shown in **1 branch** proves that using two parallel networks for parsing and pose estimation is necessary.

7.5 Sensitivity Analysis

7.5.1 Impact of View Angles. We evaluate the performance against various view angles. To simulate different view angles of the mmWave radar and the subjects, candidates are asked to perform activities with a varying offset (from 0 to 1.2 meters) from the radar's mid-line. By doing this, the angle between the target and the device varies from 0° to 60° with the interval of 15°, and ~ 60° is the maximum angle of the device's FoV. We use the 75% of data collected for training and the rest for testing. The result of parsing into 10 body parts is shown in Fig.11, the three curves of different metrics demonstrate that our system shows good robustness when the angle changes, and mmParse achieves stable parsing performance accuracy (79.2%) at a very large angle 50° (1.2 meters offset).

7.5.2 Impact of Different Scenes. As aforementioned, mmWave signal travel through multiple paths and the signals arriving at the radar usually carry information that is specific to the environment where the activities are recorded. We investigate the the robustness of our design in the unseen environment. Specifically, we use the data collected in the hallway (depicted in Fig.8(a)) for training. In test phase, we collect data in two new scenes: room (Fig.8(b)) and outdoor (Fig.8(c)), and evaluate the performance on these unseen environments. We evaluate the performance when parsing into 10 boay parts. As shown in Fig.12, the model shows good robustness for the new environment in general. The performance drops slightly for outdoor because of two interfering factors. First, electromagnetic environments indoor and outdoor might be very different. Second, the ground truth from Kinect might be influenced by outdoor light condition. To further improve the robustness, we might collect more data from various environments and apply recent domain adaptive wireless sensing techniques [18], which is left to future work.

7.5.3 Performance Evaluation for Cross-subject Human Parsing. To test the generalization ability of the trained model, we evaluate the performance on new subjects that do not appear in the training stage. Specifically, we train the model with the data of 22 subjects and use the remaining data of other 10 subjects for testing. The performance of parsing into 10 body parts is provided and compared with baselines in Fig.13. Our methods show

Table 3. Comparison of per-bodypart IoU of human parsing.

Models	head	torso	ul arm	ll arm	ur arm	lr arm	ul leg	ll leg	ur leg	lr leg	bkg	avg
P+R[32]	83.21	77.25	73.16	70.34	69.82	66.12	71.37	87.99	70.91	86.52	88.29	76.82
DGR[44]	84.73	78.53	72.67	70.65	70.09	64.98	72.71	85.55	71.26	85.56	89.19	76.90
VCR[55]	83.09	79.32	73.75	69.62	69.86	65.95	70.93	87.89	70.73	86.57	89.59	77.03
mm-Pose[35]	83.52	80.74	70.32	66.01	65.77	61.19	67.21	86.37	69.28	86.39	87.23	74.91
mmMesh[46]	85.68	78.69	72.82	69.12	69.06	65.28	72.75	86.07	71.72	85.36	89.79	76.94
PCB[56]	78.55	77.18	62.48	65.01	67.22	69.00	65.37	76.26	69.96	76.42	88.43	72.35
mmParse	**89.28**	**85.22**	**81.12**	**79.14**	**79.61**	**76.21**	**81.75**	**91.26**	**81.24**	**91.59**	**90.92**	**84.30**

stable performance in all the three metrics when meets 'new' users. It is noteworthy that mmParse significantly outperforms the baselines by 10.33% on these unseen subjects. This demonstrates an important advantage of our multi-task learning design - the features learned under the supervision of multiple tasks suffer less from outfitting and thus generalize better to the unseen data.

7.5.4 Per-bodypart analysis. The previous discussion mainly focuses on the average accuracy of parsing. This section breaks down the overall results into each individual body parts to examine the consistency of the accuracy over various parts. Table.3 shows the IoU accuracy of each body parts, where mmParse is consistently more accurate than baselines across various body parts. It achieves the best accuracy on head, low left leg, and low left right leg. Low left and right arms are most challenging due to their small sizes, while mmParse improves the accuracy by more than 10% in these error-prone body parts.

7.6 Case Study

mmParse annotates the raw point cloud with semantic labels, which can be leveraged by a wide spectrum of downstream huamn sensing tasks. To demonstrate the benefits, we further utilize the annotated radar point clouds to support two representative downstream tasks, i.e., pose estimation and action recognition.

7.6.1 Case 1: Pose estimation. **Pose estimation.** To demonstrate the effectiveness of the proposed human parsing for downstream task of pose estimation, we implement the latest pose estimation work using mmWave point cloud (i.e., mmMesh [46]) and extend it to support the extra semantic labels of body parts as part of the inputs. mmMesh utilizes the PointNet as the backbone network and designs an anchor point module to address the misalignment of the sparse point cloud with the human body parts. To make it a fair comparison, we develop the same deep learning model as mmMesh [46] while train and test it with our own mmWave dataset as well as our own ground truth labels. Specifically, mmMesh obtains the ground-truth joint position using the VICON system whereas we use Kinect system and its Body Tracking SDK [9] to label 3D positions of 17 skeleton points. The data of 22 subjects in Section.6.2 are used for training, and the remain data of 10 subjects are used for testing.
Performance. We evaluate the accuracy of pose estimation under three settings. First, we test the baseline approach, i.e., pose estimation without body part labels. Then, we conduct the pose estimation with the labels predicted by mmParse. Finally, we also repeat the experiment with the ground truth semantic labels obtained from Kinect. This represents the ideal situation where human parsing is error-free. The mean Euclidean distance

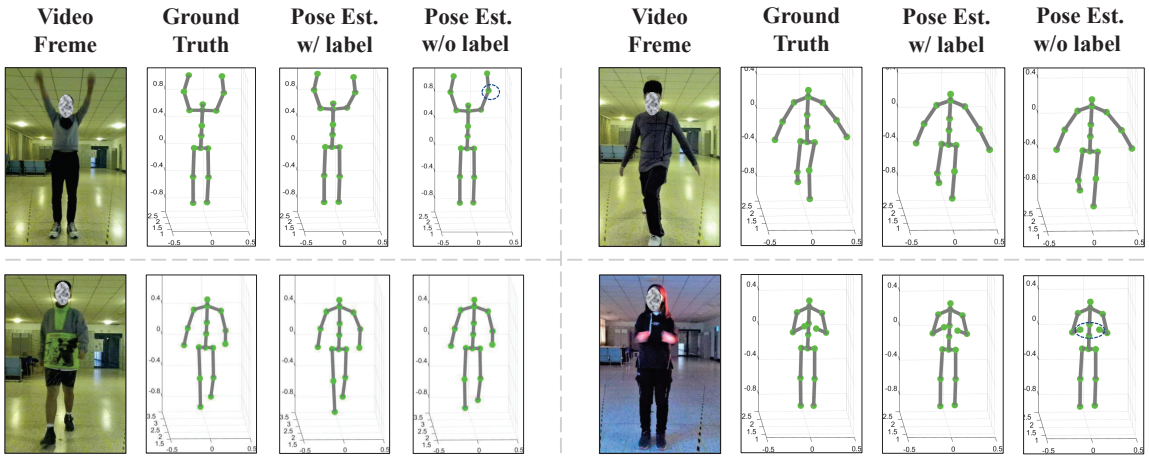


Fig. 14. 3D pose estimation results. Circled points are the main performance gain over the baseline.

between the predicted skeleton key point and the ground truths is used as the evaluation metric. The quantitative result of pose estimation is shown in Fig.15. mmParse manages to reduce the error by $\sim 5mm$, leading to a 18% accuracy improvement of the pose estimation. In addition, the performance with predicted label is very close to the one with the ground truth label. The important observation indicates that the semantic label with minor errors can still benefit pose estimation task. Four examples of pose estimation are shown in Fig.14. The skeleton structures predicted using our method proximate the growth truth well and are more accurate than the baseline in a few joint locations.

7.6.2 Case 2: Action Recognition. **Action Recognition.** We implement an action recognition network with a deep neural network based backbone (e.g., PointNet [31] and LSTM [15]) which takes the mmWave point cloud as input. The training and testing are performed with the dataset introduced in Section 6.2, which consists of 10 different activities denoted as (1) $\sim$ (10). The ground truth of actions is recorded during data collection.

Performance. Fig.16 compares recognition accuracy under three settings (i.e., recognition without labels, with predicted body part labels, and with ground truth labels). When the extra body part labels from mmParse are provided to the recognition model, the average accuracy and the F1 score are both improved by more than 5%. It is because body parts labels can provision extra semantic information, which complements the limited geometry

Models	AJE(cm)
w/ GT label	**2.286**
w/ predicted label	**2.332**
w/o label	2.868

Fig. 15. Performance of pose estimation.

Models	F1 Score (%)	Accuracy (%)
w/ GT label	**95.69**	**95.62**
w/ predicted label	**94.87**	**94.73**
w/o label	89.32	89.25

Fig. 16. Performance of action recognition.

Fig. 17. Confusion matrix of action recognition.

information conveyed in the sparse point cloud. We further break down the accuracy results to each activity. Fig.17 depicts the confusion matrix of action recognition, which shows that the accuracy values of actions are above 91%. Furthermore, by adding body part label, we find that the model converges faster during training with the same parameters.

To sum up, the performance of two representative human sensing tasks are both enhanced by body part labels predicted by mmParse, demonstrating that the fine-grained information of human parsing is useful to downstream tasks. Therefore, we envision that mmParse has the potential to serve as a generic enabler to a wide range of applications using mmWave radars.

8 DISCUSSION AND FUTURE WORK

This work presents the first proof-of-principle human parsing for mmWave point clouds. We notice that there are limitations that need to be further investigated in future extensions.

Number of people in the FoV. Constrained by the low-level HW/SW configurations of the mmWave radar used in this work (e.g., the maximum number of points per frame and limited communication bandwidth), our experiments are conducted with a single person in the sensors' FoV. In our future work, we plan to implement mmParse with more powerful commodity mmWave radars and evaluate our design on a greater number of people in the FoV.

Parsing for static human. In this paper, we prototype mmParse with the low cost commercial mmWave radars, which are limited in the resolution and thus provide very few points when the subject is stationary. Therefore, our evaluation of the system assume that the subject is performing activities. In the future, we plan to extend mmParse to emerging mmWave imaging radar [5] with much higher resolutions and evaluate the effectiveness of the proposed mmParse of static subjects.

More accurate ground-truth. In this paper, we utilize the low-cost RGB-D cameras as the source of ground-truth data, and the imperfect output from the Azure Kinect may affect the real label, which would limit the performance of the mmParse model. In the future, we plan to utilize more than one Kinect to collect the ground-truth.

9 RELATED WORK

9.1 mmWave-based Human Sensing

Recent advances have demonstrated that mmWave radar is feasible in various human sensing tasks, such as human monitoring and tracking [45, 50], pose estimation [7, 35, 46, 53, 54, 56], skeleton reconstruction [34, 35, 46], behavior recognition [24, 36, 51], human detection and identification [11, 16, 20, 48], and human acoustic sensing [21, 26] . Instead of proposing not yet another human sensing application, our mmParse provides a generic framework to enhance the amount of information in the sparse point cloud data (semantic information). Therefore, mmParse is a complementary to these wide range of applications and has the potential to benefit many sensing tasks as we demonstrate in the case study.

9.2 Human Pose Estimation Using Radar

In recent years, many FMCW radar sensing systems have been developed to estimate human pose [7, 35, 46, 53, 54, 56]. Among them, [53] addresses 2D pose estimation while RF-Pose [54] can estimate 3D pose of multi-person simultaneously. Both works are prototyped using specially designed device with a large antenna array. Recently, [35, 56] detects and tracks frame-wised human skeletons and [46] constructs human 3D mesh with the COTS mmWave radar. In contrast to these works, our design focus on predicting the point-wised body part labels of each point, which is a method to enhance the information in the radar point cloud. Technologically, We focus on fine-grained point-wised features rather than the overall activity or posture feature of the subjects. As a result, the performance of human parsing is not optimal directly with the models of existing pose estimation design.

9.3 Human Parsing

Recently, several works explore various deep learning models to directly parse human semantic body parts from images [13, 22, 49, 52], videos [25, 38, 57], and point cloud [12, 17, 23, 37]. Despite the great success achieved by vision based approaches, the performance of camera can be severely impaired by bad illumination, occlusion and blurry. In contrast, mmWave radar is robust to these interference and thus are used for scenarios with hash environment (e.g., fire fighter). Moreover, vision based devices are unacceptable in camera-restricted scenario (e.g., living, shower, and restroom). In contrast, our mmWave based approach can not only avoid the camera-restriction issue but also be immune to the poor lighting conditions.

Technically, these existing solutions of human parsing [12, 17, 23, 37] mainly apply to the image data with dense colorful pixels that can capture appearance and shape of the whole human body. In contrast, the mmWave radar data is sparse and noisy. Therefore, mmParse presents several new designs to address these unique challenges in the mmWave radar scenario.

10 CONCLUSION

This paper presents a novel system design mmParse that predicts the semantic label of a person's dynamic mmWave point clouds. To tackle the lack of body structure information due to the point cloud sparsity, we exploit a multi-task learning architecture. mmParse also features an effective heterogeneous feature fusion method based on intra-task and inter-task attention with a global view over the spatio-temporal domains. In addition, we utilize the annotated radar point clouds to serve 2 representative downstream tasks. We envision that parsing for mmWave point cloud could be utilized by a wide spectrum of human sensing tasks to further unleash the potential of mmWave radars and push the limit of their performance.

ACKNOWLEDGMENTS

This work was supported in part by 2030 major projects of scientific and technological innovation under Grant No. 2021ZD0114200, National Natural Science Foundation of China under Grant No. 62272098.

REFERENCES

[1] 2022. Google's Soli radar returns to track sleep on the new Nest Hub. https://techcrunch.com/2021/03/16/googles-soli-returns-to-track-sleep-on-the-new-nest-hub/.
[2] 2022. IWR6843. https://www.ti.com/tool/IWR6843ISK.
[3] 2022. rescuerproject. https://rescuerproject.eu/technology-tools/.
[4] 2022. Sensor Capture + Azure Kinect + Refinement Workflow. https://www.depthkit.tv/tutorials/azure-kinect-microsoft-volumetric-capture-depth-workflow-depthkit.
[5] 2022. Vayyar Imaging Radar. https://vayyar.com/care/b2c/.
[6] 2022. wholehome-ai-sensor. https://consumer.huawei.com/cn/wholehome/ai-sensor/.
[7] 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.
[8] Sherif Sayed Ahmed and Lorenz-Peter Schmidt. 2012. Illumination of humans in active millimeter-wave multistatic imaging. In *2012 6th European Conference on Antennas and Propagation (EUCAP)*. IEEE, 1755–1757.
[9] Justin Amadeus Albert, Victor Owolabi, Arnd Gebel, Clemens Markus Brahms, Urs Granacher, and Bert Arnrich. 2020. Evaluation of the pose tracking performance of the azure kinect and kinect v2 for gait analysis in comparison with a gold standard: A pilot study. *Sensors* 20, 18 (2020), 5104.
[10] Derya Birant and Alp Kut. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering* 60, 1 (2007), 208–221.
[11] Dongjiang Cao, Ruofeng Liu, Hao Li, Shuai Wang, Wenchao Jiang, and Chris Xiaoxuan Lu. 2022. Cross Vision-RF Gait Re-identification with Low-cost RGB-D Cameras and mmWave Radars. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2022).
[12] Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. 2022. Pq-transformer: Jointly parsing 3d objects and layouts from point clouds. *IEEE Robotics and Automation Letters* (2022).

[13] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. 2014. Towards unified human parsing and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 843–850.

[14] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. 2020. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14225–14233.

[15] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610.

[16] Tianbo Gu, Zheng Fang, Zhicheng Yang, Pengfei Hu, and Prasant Mohapatra. 2019. Mmsense: Multi-person detection and identification via mmwave sensing. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*. 45–50.

[17] Jing Huang and Suya You. 2016. Point cloud labeling using 3d convolutional neural network. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2670–2675.

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

[19] Ryanne JM Lemmens, Yvonne JM Janssen-Potten, Annick AA Timmermans, Rob JEM Smeets, and Henk AM Seelen. 2015. Recognizing complex upper extremity activities using body worn sensors. *PloS one* 10, 3 (2015), e0118642.

[20] Hao Li, Ruofeng Liu, Shuai Wang, Wenchao Jiang, and Chris Xiaoxuan Lu. 2022. Pedestrian Liveness Detection Based on mmWave Radar and Camera Fusion. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 262–270.

[21] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. VocalPrint: exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 312–325.

[22] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[23] Fangyu Liu, Shuaipeng Li, Liqiang Zhang, Chenghu Zhou, Rongtian Ye, Yuebin Wang, and Jiwen Lu. 2017. 3DCNN-DQN-RNN: A deep reinforcement learning framework for semantic parsing of large-scale 3D point clouds. In *Proceedings of the IEEE international conference on computer vision*. 5678–5687.

[24] 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. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4, 4 (2020), 1–28.

[25] Si Liu, Xiaodan Liang, Luoqi Liu, Ke Lu, Liang Lin, Xiaochun Cao, and Shuicheng Yan. 2015. Fashion parsing with video context. *IEEE Transactions on Multimedia* 17, 8 (2015), 1347–1358.

[26] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 97–110.

[27] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. 2020. See through smoke: robust indoor mapping with low-cost mmwave radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 14–27.

[28] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB de Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. 2020. milliEgo: single-chip mmWave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 109–122.

[29] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. 2018. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 502–517.

[30] Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. 2021. Multi-View Radar Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15671–15680.

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.

[32] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* 30 (2017).

[33] Kun Qian, Zhaoyuan He, and Xinyu Zhang. 2020. 3D point cloud generation with millimeter-wave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[34] Arindam Sengupta, Feng Jin, and Siyang Cao. 2020. NLP based skeletal pose estimation using mmWave radar point-cloud: A simulation approach. In *2020 IEEE Radar Conference (RadarConf20)*. IEEE, 1–6.

[35] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. 2020. mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sensors Journal* 20, 17 (2020), 10032–10044.

[36] Karly A Smith, Clément Csech, David Murdoch, and George Shaker. 2018. Gesture recognition using mm-wave sensor for human-car interface. *IEEE sensors letters* 2, 2 (2018), 1–4.
[37] Matteo Terreran, Leonardo Barcellona, Daniele Evangelista, and Stefano Ghidoni. 2021. Multi-view Human Parsing for Human-Robot Collaboration. In *2021 20th International Conference on Advanced Robotics (ICAR)*. IEEE, 905–912.
[38] Yan Tian, Guohua Cheng, Judith Gelernter, Shihao Yu, Chao Song, and Bailin Yang. 2020. Joint temporal context exploitation and active learning for video segmentation. *Pattern Recognition* 100 (2020), 107158.
[39] Yonglong Tian, Guang-He Lee, Hao He, Chen-Yu Hsu, and Dina Katabi. 2018. RF-based fall monitoring using convolutional neural networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.
[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[41] Chaohui Wang, Nikos Komodakis, and Nikos Paragios. 2013. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding* 117, 11 (2013), 1610–1627.
[42] Chuyu Wang, Jian Liu, Yingying Chen, Lei Xie, Hong Bo Liu, and Sanclu Lu. 2018. RF-kinect: A wearable RFID-based approach towards 3D body movement tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–28.
[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.
[44] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)* 38, 5 (2019), 1–12.
[45] Yong Wang, Wen Wang, Mu Zhou, Aihu Ren, and Zengshan Tian. 2020. Remote monitoring of human vital signs based on 77-GHz mm-wave FMCW radar. *Sensors* 20, 10 (2020), 2999.
[46] 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.
[47] Chao Yang, Xuyu Wang, and Shiwen Mao. 2020. Rfid-pose: Vision-aided three-dimensional human pose estimation with radio-frequency identification. *IEEE Transactions on Reliability* 70, 3 (2020), 1218–1231.
[48] Xin Yang, Jian Liu, Yingying Chen, Xiaonan Guo, and Yucheng Xie. 2020. MU-ID: Multi-user identification through gaits using millimeter wave radios. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2589–2598.
[49] Dan Zeng, Yuhang Huang, Qian Bao, Junjie Zhang, Chi Su, and Wu Liu. 2021. Neural Architecture Search for Joint Human Parsing and Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11385–11394.
[50] Yunze Zeng, Parth H Pathak, Zhicheng Yang, and Prasant Mohapatra. 2016. Human tracking and activity monitoring using 60 GHz mmWave. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 1–2.
[51] 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.
[52] Ziwei Zhang, Chi Su, Liang Zheng, and Xiaodong Xie. 2020. Correlating edge, pose with parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8900–8909.
[53] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.
[54] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 267–281.
[55] 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 *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 33–40.
[56] Jinxiao Zhong, Liangnian Jin, and Ran Wang. 2022. Point-convolution-based human skeletal pose estimation on millimetre wave frequency modulated continuous wave multiple-input multiple-output radar. *IET Biometrics* 11, 4 (2022), 333–342.
[57] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. 2018. Adaptive temporal encoding network for video instance-level human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*. 1527–1535.
[58] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. 2019. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 593–602.
[59] Maryam Ziaeefard and Robert Bergevin. 2015. Semantic human activity recognition: A literature review. *Pattern Recognition* 48, 8 (2015), 2329–2345.