Towards Weather-Robust 3D Human Body Reconstruction: Millimeter-Wave Radar-Based Dataset, Benchmark, and Multi-Modal Fusion

Anjun Chen, Xiangyu Wang, Kun Shi, Yuchi Huo, Jiming Chen, *Fellow, IEEE,* and Qi Ye

***Abstract*—3D human reconstruction from RGB images achieves decent results in good weather conditions but degrades dramatically in rough weather. Complementarily, mmWave radars have been employed to reconstruct 3D human joints and meshes in rough weather. However, combining RGB and mmWave signals for weather-robust 3D human reconstruction is still an open challenge, given the sparse nature of mmWave and the vulnerability of RGB images. The limited research about the impact of missing points and sparsity features of mmWave data on reconstruction performance, as well as the lack of available datasets for paired mmWave-RGB data, further complicates the process of fusing the two modalities. To fill these gaps, we build up an automatic 3D body annotation system with multiple sensors to collect a large-scale mmWave dataset. The dataset consists of synchronized and calibrated mmWave radar point clouds and RGB(D) images under different weather conditions and skeleton/mesh annotations for humans in these scenes. With this dataset, we conduct a comprehensive analysis about the limitations of single-modality reconstruction and the impact of missing points and sparsity on the reconstruction performance. Based on the guidance of this analysis, we design ImmFusion, the first mmWave-RGB fusion solution to robustly reconstruct 3D human bodies in various weather conditions. Specifically, our ImmFusion consists of image and point backbones for token feature extraction and a Transformer module for token fusion. The image and point backbones refine global and local features from original data, and the Fusion Transformer Module aims for effective information fusion of two modalities by dynamically selecting informative tokens. Extensive experiments demonstrate that ImmFusion can efficiently utilize the information of two modalities to achieve robust 3D human body reconstruction in various weather environments. In addition, our method achieves superior accuracy compared to that of the state-of-the-art Transformer-based LiDAR-camera fusion methods.**

***Index Terms*—3D human body reconstruction, mmWave-RGB fusion, human body dataset.**

I. Introduction

3D human body reconstruction has been studied extensively and has wide applications, such as XR technologies, autonomous driving, outdoor robotics, search and rescue, etc. With easy access and low cost, RGB cameras are one of the most common sensor modalities for the reconstruction [1]. Nevertheless, the performance of reconstruction using RGB images under adverse circumstances is still limited, as the perception capability of RGB cameras rapidly deteriorates in poor illumination or inclement weather conditions [2]. In addition, recovering 3D information from a single 2D RGB image is inherently an ill-posed inverse problem [3] due to depth ambiguity.

Millimeter wave (mmWave) imaging radar is a newly emerging sensing technology to capture 3D or more high-dimensional scene information with relative lower cost than LiDAR. It has gained increasing popularity in wireless sensing areas in recent years, such as autonomous driving [11]–[13], human activity recognition [14], [15], and SLAM [16], [17]. Moreover, mmWave radar can sense low-visibility environments such as dense fog, smoke, snowstorm, rain, etc. [18], which makes it appealing for many applications that require working in various weather conditions. Despite these capabilities, point clouds generated from mmWave radar suffer from sparsity due to low spatial resolution, large missing parts due to specular reflection, and high-level noise due to multi-path effects. Despite existing applications in large scenes and motion classification, these defects can hinder its application in fine-grained 3D human body reconstruction. A promising solution is to combine RGB images with mmWave signals, as the sparse and noisy mmWave radar point cloud could be complemented by the high-resolution and high-quality RGB images. Therefore, fusing the two modalities to combine their strengths is essential to realize robust 3D human body reconstruction in various weather conditions.

However, mmWave-RGB fusion faces many challenges: 1) research works and public data are limited for studying the fine-grained 3D human reconstruction from point clouds with characteristics of sparsity and missing points like mmWave signals; 2) the quality of mmWave signals for 3D body reconstruction compared with RGB images or point clouds from depth cameras are not well studied, which makes it difficult to design a fusion strategy for the modality; 3) despite some works on LiDAR-RGB fusion, the performance of these fusion strategies for mmWave and RGB is questionable due to the significant defects of sparsity and missing mmWave points. In this work, we aim to address these challenges.

Currently, research on the human body reconstruction from mmWave radar is limited. Some works pioneer in the exploration of human body reconstruction from wireless signals

Anjun Chen, Xiangyu Wang, Kun Shi, Jiming Chen, and Qi Ye are with the State Key Laboratory of Industrial Control Technology, Zhejiang University. Email: {anjunchen, xy_wong, kuns, cjm, qi.ye}@zju.edu.cn. Yuchi Huo is with the State Key Lab of CAD&CG, Zhejiang University and Zhejiang Lab. Email: eehyc0@zju.edu.cn. Corresponding author: *Qi Ye*.

This work was supported in part by NSFC under Grants 62088101, 62233013, 61790571, 62103372, and the Fundamental Research Funds for the Central Universities.

TABLE I
COMPARISON OF HUMAN BODY DATASETS WITH WIRELESS SIGNALS. NO. MOTIONS INDICATE THE NUMBER OF MOTIONS.

Datasets	Signals	Labels	No. Motions	Public	Scenes				
					Furnished	Rain	Smoke	Poor Lighting	Occlusion
Person-in-WiFi [4]	Wi-Fi	2D Skeletons	/	×	×	×	×	×	×
RF-Pose [5]	RF Signal	2D Skeletons	/	×	×	×	×	✓	✓
RF-Pose3D [6]	RF Signal	3D Skeletons	/	×	×	×	×	×	✓
RF-MMD [7]	RF Signal	3D Skeletons	35	×	×	×	×	✓	✓
RF-Avatar [8]	RF Signal	Mesh	/	×	×	×	×	×	✓
mmMesh [9]	mmWave	Mesh	8	×	✓	×	×	✓	✓
mRI [10]	RGB(D), mmWave	3D Skeletons	12	✓	×	×	×	×	×
Our mmBody	RGB(D), mmWave	Mesh, Skeletons	100	✓	✓	✓	✓	✓	✓

[4]–[9]. Despite the inspiring exploration, there is no public mmWave radar dataset available for the community to study the problem. Additionally, these works have not quantitatively evaluated the accuracy of reconstructing 3D human mesh from mmWave signals in different scenarios and how they perform compared with RGB and depth cameras. Furthermore, among the defects of sparsity, missing parts, and high noise, key factors influencing the reconstruction quality from mmWave signals and the fused mmWave-RGB signals, are not identified, which are important for the design of fusion algorithms.

Though there is little work on mmWave-RGB fusion, LiDAR-camera fusion has been studied in some computer vision problems. Early LiDAR-camera fusion approaches [19], [20] adopt point-to-image projection to combine point clouds and image pixel values/features through element-wise addition or channel-wise concatenation. These approaches heavily rely on the local projection relationship between the point clouds and images, which can break down if one of the modalities is compromised or fails. Undesirable issues like low density, random incompleteness, and temporal fluctuation of mmWave point clouds can result in the retrieval of inadequate or incorrect features from corresponding images. More recently, several customized Transformer-based structures [21]–[23] have been proposed for multi-modal fusion. These fusion frameworks, however, focus on LiDAR-camera fusion-based object detection, which is inapplicable for the mmWave-RGB fusion-based human body reconstruction task. Furthermore, the degradation of modality features in challenging environments, such as low lighting and smoke conditions, can extremely impair performance.

In this paper, we make efforts in filling these gaps and addressing the challenges in three aspects: 1) proposing a dataset for the study of human body reconstruction from mmWave signal, 2) comparing the reconstruction quality of different sensor inputs in different environments and analyzing the impact of characteristics of noisy points (sparsity and missing points) on the reconstruction performance, and 3) proposing a novel fusion strategy tailored for RGB fusion with sparse and noisy signals like mmWave point clouds.

mmWave Dataset. We first design a data collection system with automatic 3D body mesh annotation, which is realized by fitting the SMPL-X body model [24] to markers attached to subjects using MoSh++ [25]. Using this system, we collect a large-scale mmWave 3D human body dataset (denoted as mmBody) with 100 motions captured from 20 volunteers in 7 different scenes. The statistics and visualization of the dataset in Table I and Fig. 2 reveal that our dataset makes a significant advancement in terms of completeness and diversity of scenarios, shapes, and poses. In addition to the mmWave signals, we also collect synchronized and calibrated RGB(D) images for mmWave-RGB fusion for the 3D body reconstruction in different weather conditions.

mmWave Quality Evaluation. With this dataset, we conduct extensive experiments to evaluate 3D body reconstruction performance using different single-sensor inputs (mmWave signals, depth from TOF sensors [26] and RGB images) in different scenarios including extreme weather conditions like smoke, rain, and night. To further analyze the characteristics of mmWave data, we investigate the impact of missing points and sparsity on the reconstruction accuracy by comparing the reconstruction from mmWave signals with that from depth point clouds.

mmWave-RGB Fusion. With the guidance of these analyses, we present ImmFusion, the first fusion solution to combine the mmWave point clouds and RGB images to robustly reconstruct the 3D human body in various conditions. Due to the noisy mmWave point clouds, in our framework, different from fusion via projection, we do not establish the connection explicitly via the spatial relation of two modalities. Instead, we resort to well-devised Transformer-based fusion modules to dynamically fuse the information from different modalities based on their feature strengths. Additionally, in contrast to previous fusion methods that regard point clouds as the main modality, our framework does not assume a main modality and treats features from different modalities as equal tokens (like words in NLP). The corrupted tokens from one modality could possibly be remedied by others or disregarded to accommodate the sparsity and missing parts. Experimental results demonstrate that ImmFusion can effectively mitigate sensor defects and fuse information from the two modalities to achieve robust 3D human body reconstruction in various environments.

The rest of this paper is organized as follows: Section II gives a brief overview of related works on 3D human reconstruction, sensor fusion, and human body datasets. Section III introduces our data collection system and our large-scale mmWave-RGB human body dataset, mmBody. Section IV

presents our proposed mmWave-RGB fusion method, ImmFusion. Section VI elaborates experimental results and analysis. Section VII finally concludes the paper.

II. Related Works

A. Human Body Reconstruction

3D human body reconstruction has been researched for many years. Most prevailing reconstruction approaches leverage RGB images. Learning-based methods to solve this problem can be broadly divided into two categories: parametric and non-parametric approaches. For parametric methods, a mapping function from the input to the output representation of the body, e.g. 2D/3D skeletons [27], [28], and the parameters of SMPL or SMPL-X [29], [30] is learned. Despite greatly reducing regressing parameters, it is still challenging to estimate precise coefficients from a single image [31], [32]. To improve the reconstruction, researchers make efforts by utilizing multi-view information [33]–[35], dense depth maps [36], [37] or sequential videos [38]–[40]. On the other hand, non-parametric approaches directly regress the vertices of the 3D mesh from the input image. Most pioneers choose Graph Convolutional Neural Network [41] to model the local interactions between neighboring vertices with an adjacency matrix. More recent approaches, such as METRO [42] and Mesh Graphormer [43], utilize transformer encoders to jointly model the relationships between vertices and joints.

Recently, millimeter wave (mmWave) sensors have gained popularity for their ability to work in challenging conditions such as rain, smoke, and occlusion. Several wireless systems have been developed to reconstruct the human body and the mmWave-based system is one of them. The mmWave sensing has been widely adopted to enable various human sensing works, such as human monitoring and tracking [44], [45], human detection and identification [46], [47], and gesture recognition [48]. For human pose estimation, several pioneering works [4], [5], [49] have been proposed to recover human skeletons from RF and Wi-Fi signals. Works on the full-body reconstruction including shape estimation from wireless signals are limited. Zhao *et al.* [8] reconstruct the 3D human mesh by utilizing RF signals, which demonstrates that wireless signals contain sufficient information for the estimation of the pose and shape of the human body. To make the reconstruction more accessible, Xue *et al.* [9] present a real-time human mesh estimation system using commercial portable mmWave devices. However, the datasets are not public in these works and the capability of the reconstruction from the combination of multi-sensor signals is not studied.

B. Multi-Modal Fusion

Existing fusion methods can be broadly classified into three categories: decision-level, data-level, and feature-level fusion. Generally, how to overcome disparateness and exploit the synergy of heterogeneous modalities are the foremost considerations. To this end, most of the methods resort to investigating elaborately-designed modal alignment schemes. Decision-level fusion [50], [51] usually utilizes information from one modality to generate regions of interest containing valid objects. However, such coarse-grained fusion strategies may not fully release the potential of multiple modalities. Data-level fusion [19], [20] commonly entails the coordinate projection technique, which is easily affected by sensor misalignment and defective image features. Feature-level fusion [52], [53] typically involves the fusion of proposal-wise features in multi-modal feature maps, while determining the optimal weighting for features of each modality is challenging. All these conventional approaches make efforts in modal alignment schemes, while most of them are short of efficiency, adaptability, and compatibility.

Recently, promising performance has been achieved by Transformer-based fusion, which sheds light on the possibility of leveraging the Transformer structure as a substitute for manually designed alignment operations. Specifically, DeepFusion [21] uses a learnable alignment mechanism to dynamically correlate LiDAR information with the most relevant camera features. TokenFusion [22] prunes feature tokens among single-modal Transformer layers to preserve better information and then re-utilizes the pruned tokens for multi-modal fusion. CAT-Det [54] jointly encodes intra-modal and inter-modal long-range contexts to explore multi-modal information for detection. TransFusion [55] employs a soft-association approach to process inferior image situations. Some recent works [56]–[58] propose to formulate unified end-to-end multi-sensor fusion frameworks for 3D detection. These works, which focus on fusion-based object detection, however, differ from ours since we make efforts in constructing the mmWave-RGB fusion pipeline for 3D human body reconstruction. Additionally, in contrast to previous fusion methods that deteriorate severely when they are applied to noisy point clouds, our framework can effectively accommodate the sparsity and missing parts.

C. Human Body Datasets

In recent years, various human body datasets have been introduced to advance the research on human body reconstruction. These datasets provide annotated RGB(D) images with ground truth in the form of skeletons or mesh. However, most of the frequently used datasets, such as CMU MoCap [59], MPI-INF-3DHP [60], NTU RGB+D [61], 3DPW [62], and Human3.6M [63], do not include scenes captured in adverse environments due to the degradation of RGB(D) cameras in such conditions.

With the demand for autonomous driving, some mmWave-based datasets [64], [65] have been proposed recently for object detection and semantic understanding. However, as exhibited in the Table I, no public mmWave-based datasets for 3D body reconstruction are available, which limits the development of this field to some extent. Our proposed annotation system and large-scale multi-modal human body dataset can pave the way for further research on combing mmWave radars with RGBD cameras for 3D body reconstruction in various weather conditions.

III. mmBody Dataset

In this section, we first introduce our method of building an automatic annotation system and then present our mmWave

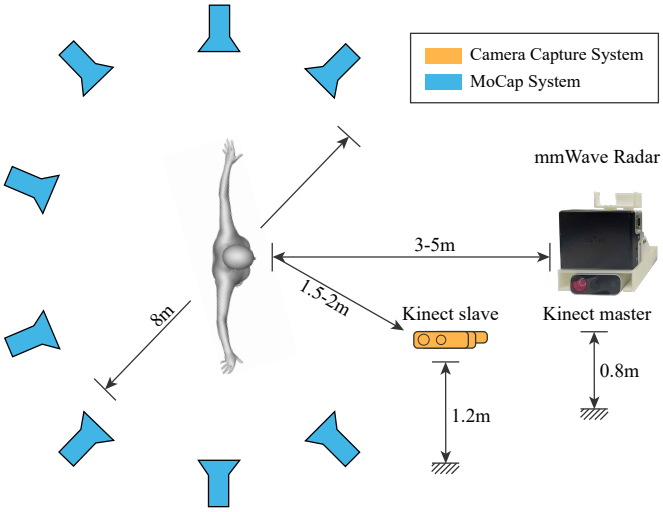

Fig. 1. Hardware system. It consists of three parts: a mmWave capture system to record the body motions, a MoCap system to label both human joint locations and full-body meshes, and a camera capture system to obtain RGB(D) images.

human body dataset. The hardware system consists of three parts: a mmWave capture system to record the body motions, a MoCap system to label both human joint locations and full-body meshes, and a camera capture system to obtain RGB(D) images. The spatial arrangement of each component of the hardware system is roughly set as shown in Fig. 1. The hardware system is explained in Section III-A. Then the synchronization and calibration among the systems follow in Section III-B. In Section III-C, the acquisition of the full-body mesh annotation is discussed. Section III-D shows statistics of our benchmark. Section III-E gives a comparison of the pose and shape space of mmBody with popular human body datasets captured using MoCap or RGB(D) images.

A. Hardware System

mmWave System. The spatial sensing via mmWave is achieved by transmitting wireless signals and receiving their reflections from environments via antenna arrays. The frequency shift between transmitting and receiving signals and the difference in arriving time between antennas determine the range measurement and the angle measurement, respectively. For more details on the mmWave spatial sensing mechanism, we refer readers to the technical report [66].

In our work, we choose the Phoenix type mmWave radar produced by Arbe Robotics[1] for its high resolution, which works at 10 to 30 FPS. An antenna array of 48 transmitting channels by 48 receiving channels enables it to reach 0.4 meters for the range resolution and about 2.0 degrees for the angle resolution. It has an onboard processor to convert the original signals into point clouds which we use as input for 3D body reconstruction. More specifications of the radar are provided in the product overview [67].

The mmWave radar is placed on a 3D printing holder with a depth camera (Azure Kinect[2]) beneath it, shown in Fig. 1. The holder is fixed on a tripod about 0.8 meters above the ground. The 3D point clouds for the motion of subjects are captured at a distance of 3 to 5 meters away from the radar as the mmWave radar works well at a distance of 3 meters away. The mmWave radar captures the scene at about 14 FPS.

Camera Capture System. The camera capture system aims to get the RGB(D) images. The system consists of 2 Azure Kinects: a master Kinect is placed right under the mmWave radar, and the other slave one is located on one side of the radar-body line, 1.5-2 m from the body. As the quality of the depth images degrades with distance, we place the slave Kinect closer to the subject. The Kinects are connected using synclines. Azure Kinects provide color images and depth images at a speed of 30 FPS.

MoCap System. The MoCap system aims to provide the 3D body skeletons and full-body meshes. It is the main annotation system for our dataset collection. Our OptiTrack[3] MoCap system consists of 8 cameras and markers attached to the human body. Cameras are evenly fixed on 8 tripods around a circular field with a radius of 8 meters at the height of 2.5 m, all looking at the center of the field. The system provides high-quality maker locations (accuracy of 0.8 mm) at a speed of 300 FPS at most. The number of markers is 37 and most markers are attached near human joints.

B. Calibration and Synchronization

Calibration. We set the mmWave radar coordinate frame as the target coordinate frame and transform the labels obtained from the MoCap system and the camera capture system to it.

The calibration between the mmWave radar and the camera capture system is achieved in two steps. The first step is the calibration of the Azure Kinect sensors. It is calibrated using a 1m $\times$ 1m Aruco tag, and the transformation matrix is obtained via the Colored ICP algorithm [68]. The second step is the calibration between the mmWave radar and one of the Azure Kinect sensors. Following [17], we place the mmWave radar and the sensor on a 3D printing holder. The transformation matrix between the two sensors is set beforehand. The calibration between the mmWave radar and the MoCap system is achieved by placing markers on the radar and using the position of markers located by the OptiTrack system to calculate the transformation matrix.

Synchronization. As the three systems work at different operation systems, synchronization is needed. The camera capture system and the mmWave radar are connected to the same laptop and therefore can be synchronized by the system time. For the MoCap system (running on another PC), we synchronize it to the camera capture system via the local network connection. For the mmWave radar, we can only get the timestamp of receiving the point clouds and therefore, are not able to get the exact capture time between the timestamps for two frames. The miss-alignment between the mmWave radar and other data is manually checked and adjusted slightly for each sequence.

[1]https://arberobotics.com
[2]https://azure.microsoft.com/en-us/products/kinect-dk
[3]https://optitrack.com

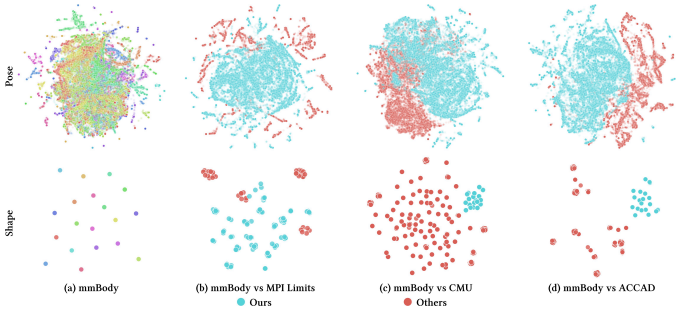

Fig. 2. 2D-TSNE embedding of poses and shapes of mmBody and other datasets. The color of the dots in (a) represents different subjects.

C. Full Body Annotation

To obtain full-body mesh annotations, we use MoSh++ [25] to fit the parameterized body representation, i.e. SMPL-X [24] to marker locations from the MoCap system. The SMPL-X model is defined as a function $\mathcal{M}(\beta, \alpha, \gamma)$, where β represents shape parameters, α body pose, hand pose, and facial expression parameters, and γ translation. For the body pose parameters, the first 3 dimensions represent the global rotation of joints, and the rest represent the rotations of 21 body joints. For the hand pose and expression parameters, we use their template values and keep them fixed. In the following paper, we use α for body pose only. The output of the function $\mathcal{M}(\beta, \alpha, \gamma)$ is a triangulated mesh.

D. Build a Complete and Concise Benchmark

We collected more than 100k frames covering 100 motions of 20 volunteers in 6 different environments. Among the 100 motions, there are 16 static poses, 9 torso motions, 20 leg motions, 25 arm motions, 3 neck motions, 14 sports motions, 7 daily indoor motions, and 6 kitchen motions. Among the 20 volunteers, there are 10 females and 10 males (physical gender), with weights ranging from 42kg to 75kg and heights ranging from 159cm to 183cm. The 7 different scenes include 2 different labs, a furnished lab, poor lighting, rain, smoke, and occlusion environments. In the furnished scene, the furniture is randomly placed behind the human activity area in the lab. For the smoke and rain scene, we simulate smoke/fog weather using smoke cakes and rain weather using a shower head. In the occlusion scene, the mmWave radar and Kinect master are covered with different materials (plastic wrapping paper and foam board). Only Kinect master and radar are influenced in the rain, smoke, and occlusion scene while Kinect slave is not interfered. The collection and use of our data adhere to the ethical guidelines strictly. All human subject data collections are under full acknowledgment and agreement.

E. Comparison with Other Datasets

To show the coverage of our dataset better, we compare the pose space and the shape space of mmBody with three popular datasets for human body reconstruction using MoCap or RGB(D) images, i.e. the CMU dataset [59], the MPI Limits, [69] and the ACCAD dataset [70]. The comparison of the 2D TSNE of SMPL-X poses and shapes of these datasets is shown in Fig. 2. The SMPL-X parameters of the other three datasets are from the AMASS [71], a large and varied database of human motion. Fig. 2 (a) shows the TSNE visualization of our SMPL-X body shape space and pose space which demonstrates the completeness and evenness of the pose and shape in our dataset. Fig. 2 (b) compares the pose and shape space with the MPI Limits dataset [69] (referred to as PosePrior in AMASS). The AMASS provides 35 motions of the MPI Limits, at a total length of 20.82 minutes. This dataset aims to model the pose priors over 3D human pose and the subjects are instructed to perform extreme poses. The TSNE embedding reflects the extent of these limits that mmBody fails to reach but mmBody covers a very even space within these limits. The shape space of mmBody has a border coverage. Fig. 2 (c) compares the pose and shape space with the CMU dataset [59]. The CMU MoCap dataset contains 2605 trials in 6 categories and 23 subcategories. The AMASS provides SMPL-X parameters containing 2083 motions of 106 subjects, at a total length of 551.56 minutes. Though our dataset only consists of 100 motions, about 5% of the CMU motions, the pose space covers a similar large space. Fig. 2 (d) compares the pose and shape space with the ACCAD dataset [70]. The AMASS provides 252 motions of 20 subjects of the ACCAD, at a total length of 26.74 minutes. The ACCAD contains daily motions which mmBody covers, and stage actions like dance and performance which mmBody does not cover. The shape space is larger than mmBody.

IV. mmWave-RGB Fusion

In this section, we present our proposed method ImmFusion for 3D human body reconstruction with both RGB images and mmWave point clouds as input. Fig. 3 (a) illustrates the framework of ImmFusion.

A. Problem Formulation

ImmFusion aims to predict the 3D positions of the joints and vertices of human meshes from mmWave point clouds and RGB images. We adopt the non-parametric approach mentioned above for body reconstruction. As our focus is reconstruction, we utilize the bounding boxes automatically annotated from the ground truth mesh joints to crop the region of interest containing only the body part. Given a dataset $\mathcal{D} = \{P, I, J, V\}, t = 0, \ldots, N$, where $P \in \mathbb{R}^{1024\times3}, I \in \mathbb{R}^{224\times224\times3}$ are the cropped body region of the mmWave radar point cloud with 1024 points and the RGB image with a size of 224×224, and $J \in \mathbb{R}^{22\times3}, V \in \mathbb{R}^{10475\times3}$ are the XYZ-coordinate annotations of 22 joints and 10475 vertices, the global/local point and image features are firstly extracted by the image and point backbone, respectively. Next, the two global features are incorporated as one global feature vector and embedded with SMLP-X template positions. Then, all global/local features are tokenized as input of a multi-layer Fusion Transformer Module to dynamically fuse the information of two modalities and directly regress the coordinates of 3D human joints and coarse mesh vertices. Last, we employ a two-stage coarse-to-fine mechanism to upsample the coarse mesh vertices to the full SMPL-X [24] mesh vertices.

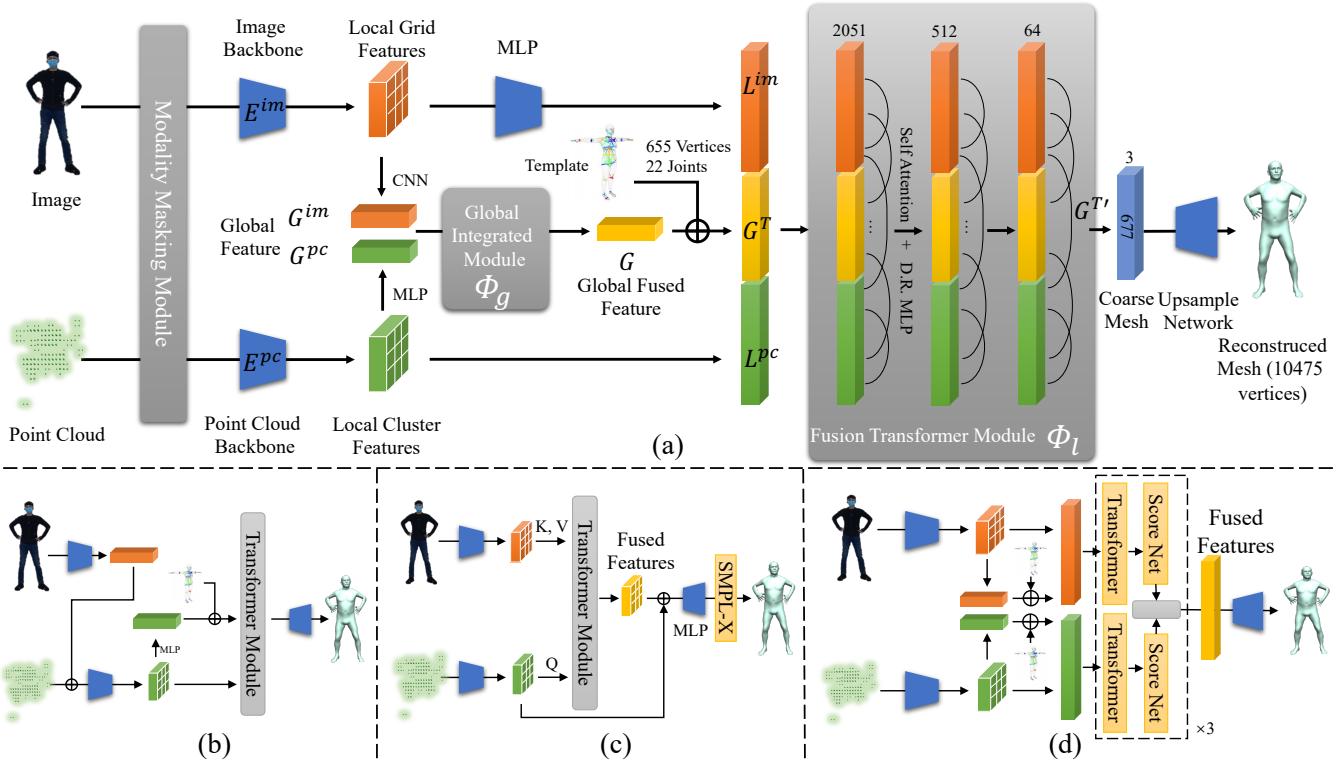

Fig. 3. Comparison of different fusion strategies. (a) Our proposed ImmFusion. We first extract global and local features from each of the masked modalities using corresponding backbones. Next, we utilize the Global Integrated Module to incorporate global features. Then, we employ the Fusion Transformer Module to fuse global and local features and to regress locations of joints and vertices. D.R. MLP stands for a Dimension Reduction MLP. (b) Points-Image-Feature method [20]. (c) DeepFusion [21]. (d) TokenFusion [22].

B. Extraction of Global and Local Features

We extract global and local features for the image and point cloud inputs to help extract global contextual dependencies and model local interactions. Specifically, we directly feed point clouds and images to the commonly used point and image backbones to extract global and local features. Either backbone can be substituted with alternative options as necessary. With this feature extraction mechanism, the deficiencies of RGB images in adverse weather conditions can be effectively constrained at the global level, while the defects of radar point clouds in normal environments can be well compensated by local image features as demonstrated in our ablation study. For brevity, we leave out the subscript t in the following parts.

For the point cloud data, we obtain cluster features $L^{pc} \in \mathbb{R}^{32\times(3+2048)}$, from a radar point cloud P using PointNet++ E^{pc}, where 32 denotes the number of seed points sampled by Farthest Point Sample (FPS), 3 denotes the spatial coordinate, and 2048 denotes the dimension of features extracted from the grouping local points. A global feature vector $G^{pc} \in \mathbb{R}^{2048}$ is further extracted from cluster features L^{pc} using an MLP. For image data, we acquire the grid features $L^{im} \in \mathbb{R}^{49\times2051}$ using HRNet [72] E^{im}, where 49 denotes the number of grid features from the last convolution block of HRNet. The global feature $G^{im} \in \mathbb{R}^{2048}$ is extracted from the grid features using a CNN layer. MLPs are used to make the dimensions of local features the same as those of the point features.

C. Accommodation Corrupted Features with Self-Attention Fusion

Traditional point-based fusion works [20], [73] concatenate image features or projected RGB pixels to the point clouds as extended features of points, as Fig. 3 (b) illustrates. However, this early fusion strategy is not suitable for mmWave-RGB fusion due to the sparsity and noise of radar points. As our analysis in Section V reveals, point-based reconstruction methods are inevitably affected by the issues of missing points and sparsity. On the other hand, undesirable issues like low density, randomly missing, and temporally flicking of radar point cloud would lead to fetching fewer or even wrong image features. The low quality of image features in adverse environments like poor lighting would further degrade the performance of the model severely. Multi-head attention module [74] is famous for modeling the relationship between information tokens. Cross-attention fusion methods [21], [55] employ a Transformer-based module to fuse image and point features as illustrated in Fig. 3 (c). Specifically, it converts the point features into the queries and image features into the keys/values and then aggregates image features to the point features. However, this mechanism requires to treat point clouds as the main modality, which cannot handle the corrupted features of the two modalities either. For TokenFusion [22] shown in Fig. 3 (d), it aims to substitute unimportant modality tokens detected by Score Nets with projected features from the other modality. The Score Net is implemented using a four-layer MLP to dynamically

score the feature tokens. Similar to point-based fusion, this projection-based design is also ineffective in incorporating corrupted modalities in adverse environments.

To mitigate the feature degradation caused by the sparsity and noise of mmWave signals and the deficiency of RGB information in extreme conditions, we formulate our fusion problem into the self-attention framework by exacting *words* (local features) and *sentences* (global features) from different inputs and designing the interaction modules between these *words* and *sentences*. This structure allows our fusion framework to effectively select informative token features from the two input modalities based on their feature strengths instead of the spatial affinity and to dynamically fuse these features. Even if some *words* (local features) is masked out (missing points), the model can utilize global information and other features to complete them.

Specifically, the two global features are fused into a global feature $G \in \mathbb{R}^{2048}$ by Global Integrated Module (GIM) Φ_g implemented using a tiny Transformer module,

$$G = \Phi_g(G^{im}, G^{pc}), \tag{1}$$

where Φ_g is a three-layer attention module ending with a sum operation to integrate the global features.

After Φ_g, similar to [41], we perform positional embedding by attaching the 3D coordinates of 22 joints and 655 vertices in a coarse mesh downsampled from a SMPL-X template mesh to the global vector $G^T = cat(J^{template}, V^{template}, G)$ to simplify the training, where $G^T \in \mathbb{R}^{677\times2051}$. Both local features serve the purpose of providing fine-grained local details for body reconstruction. In addition, the adoption of this non-parametric mechanism enables interactions between vertices, joints, local features, and global features, which can further enhance the reconstruction performance of ImmFusion.

Subsequently, we utilize the Fusion Transformer Module Φ_f to combine the strengths of radar points and images, enabling the model to select informative token features from two modalities dynamically:

$$G^{T'}, L^{im\prime}, L^{pc\prime} = \Phi_f\left(G^T, L^{im}, L^{pc}\right), \tag{2}$$

where $G^{T'} \in \mathbb{R}^{677\times64}$, $L^{im\prime} \in \mathbb{R}^{49\times64}$ and $L^{pc\prime} \in \mathbb{R}^{32\times64}$. Φ_f is implemented with a three-layer Transformer module that uses several attention heads in parallel to fuse global and local features. While attending to valid features and restricting undesirable features, the Fusion Transformer Module Φ_f adaptively adopts cross attention between joint/vertex queries G^T generated from global features G and point/image token features from local features L^{im} L^{pc} to aggregate relevant contextual information. Simultaneously, the self-attention mechanism reasons interrelations between each pair of candidate queries. Then, we adopt a dimension-reduction graph convolution [41] architecture to decode the queries $G^{T\prime}$ containing rich cross-modalities information into 3D coordinates of joints and vertices following [43]. The Dimension Reduction (DR) MLP substantially reduces the training parameters while improving performance. The Graph Convolution (GC) can effectively model interactions between joints and vertices. Lastly, a linear projection network implemented using MLPs upsamples the coarse output mesh to the original 10475 vertices.

D. Data Imbalance Solution by Modality Masking

Despite the superiority of the multi-head attention mechanism, the model is prone to struggle with data imbalance for multi-modal input according to [75] due to the bias of training data (without data under adverse conditions), which makes Transformer focus all attention on the single modality that performs better under normal circumstances as demonstrated in our experiments. To effectively activate the potential of the model across all scenarios, we design a Modality Masking Module (MMM) to mask one of the input modalities randomly and thus enforce the model to learn from the other modality in various situations. As a result, MMM enables the Fusion Transformer Module to overcome the training data bias problem and consider both modalities, which further facilitates the model to perform better across all scenarios in our experiments. In addition to the modality masking, we also randomly mask some percentages of joint/vertex token features G^T to simulate self or smoke occlusions and missing parts. For the mask proportion, we set it to 30% in the following experiments as it achieves the best accuracy.

E. Training Loss

Our ImmFusion applies L_1 loss to the reconstructed mesh to constrain the 3D vertices V and joints J. In addition, the coarse meshes V_{d1}, V_{d2} are also supervised by downsampled ground truth meshes using L_1 loss to accelerate convergence. The total loss of ImmFusion is calculated by:

$$\begin{aligned}\mathcal{L} = \lambda\|J - \bar{J}\|_1 + \mu(\|V - \bar{V}\|_1 + \|V_{d1} - \bar{V}_{d1}\|_1 \\ + \|V_{d2} - \bar{V}_{d2}\|_1),\end{aligned} \tag{3}$$

where λ and μ denote the weight of each component, and variables with overline represent the ground truth.

V. Evaluation on mmWave Point Clouds

Despite some inspiring exploration of human body reconstruction from the wireless signals [4], [5], these works have not evaluated the accuracy quantitatively of reconstructing 3D human mesh from commercial mmWave radar devices in different scenarios and how they perform compared with RGB and depth cameras. Therefore, in this section, we make efforts in answering the following questions. 1) Can mmWave radars work robustly in different environments as claimed for 3D body reconstruction? 2) Can they achieve comparable accuracy with RGB cameras or depth cameras? 3) If not, what are the key factors leading to inferior performance?

Dataset. With the dataset collected above, we can then evaluate 3D body reconstruction performance in different scenarios using different sensor inputs. The dataset is split into training and testing sets as Table II shows. We choose 20 sequences from 10 subjects recorded in the lab scenes as the training set, and 2 sequences for each scene including labs, furnished, rain, smoke, poor lighting, and occlusion as the test set. Each sequence contains about 2000 frames of data.

Methods. To evaluate the performance of 3D body reconstruction with different single modalities, we implement single-modality methods by removing one input stream from our proposed ImmFusion pipeline (see Section IV). For the Images-Only input, the feature extractor consists of only the CNN

TABLE II
TRAINING SET AND TESTING SET. */* DENOTES THE NUMBER OF SEQUENCES/NUMBER OF SUBJECTS.

Scenes	Lab1	Lab2	Furnished	Poor Lighting	Rain	Smoke	Occlusion
train	10/4	10/6	/	/	/	/	/
test	2/2	2/2	2/2	2/2	2/2	2/2	2/2

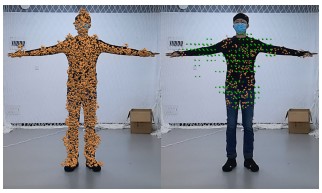

Fig. 4. Comparison of original depth point cloud (left) and noisy depth point cloud after processing (right). Radar points are in green and depth points are in orange.

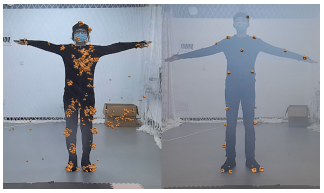

Fig. 5. Depth point clouds are significantly affected in the rain (left) and smoke (right) scenes.

backbone to extract image features. Regarding the Radar-Only method, we substitute the CNN backbone with PointNet++ and the image input with the radar point cloud. In addition to the RGB image and radar input, we also evaluate the reconstruction from the depth point cloud, i.e. Depth-Only method, to make a comprehensive analysis.

Metrics. To evaluate the performance of the reconstruction, we employ commonly used metrics, Mean Per Joint Position Error (MPJPE) and Mean Per Vertex Error (MPVE), which quantify the average Euclidean distance between the prediction and the ground truth for joints/vertices in each frame. For the BEHAVE dataset, we additionally employ Procrustes Analysis [76] MPJPE (PA-MPJPE) and MPVE (PA-MPVE) to evaluate the alignment accuracy.

A. Reconstruction from mmWave

The point clouds generated by the mmWave radar are usually very sparse, and contain many missing parts and noise resulting from the multi-path effect. Particularly, with such low-resolution point clouds, its ability to reconstruct the full 3D body is questioned. Our experiment results show that the 3D body can be reconstructed from the mmWave radar signals in spite of the sparsity. The mean joint error and the mean vertex error can reach 7cm and 9cm, which is comparable with that from RGB images.

B. Robustness in Different Environments

As discussed in the Section I, each modality is constrained by the limitations of its respective sensor. With the inherent defects of each modality, its ability to individually reconstruct the full 3D body in different environments is questioned.

Table III presents the quantitative results for methods with different single-modality inputs. As can be observed from the table, each modality possesses unique strengths and limitations. For instance, the RGB image modality (Images-Only in Table III) performs well in basic scenes, while it naturally degrades significantly in poor lighting and occlusion scenarios. The mean vertices error reaches 14.1cm and 16.6cm. On the other hand, with the susceptibility to the point cloud sparsity and random missing points, the mmWave modality (Radar-Only) performs poorly in basic scenes, but is relatively robust to be resistant to the effects of adverse environments. Its mean joint error achieves the best performance (8.3cm) in the smoke scene. The depth modality (Depth-Only) provides dense depth information and demonstrates a relatively high precision in basic scenes, but is severely impacted by rain and smoke. The mean vertices error in the smoke scene reaches 15.1cm. Therefore, fusing multiple single-modality inputs to combine their complementary strengths is beneficial to improve the performance of 3D body reconstruction.

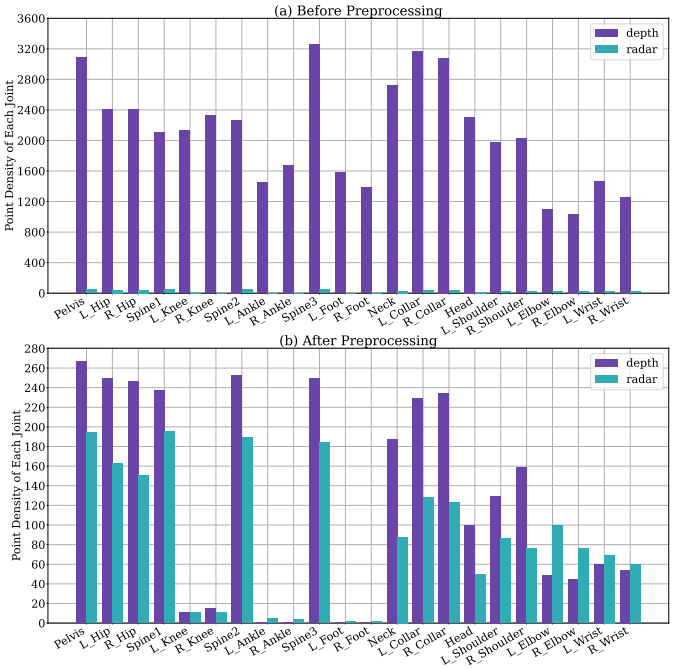

Fig. 6. Average number of points of depth and radar data within a radius of 0.15m around different joint locations.

C. Challenges of Reconstruction from mmWave

Though mmWave exhibits robustness in different adverse weather conditions, there still exists a gap (about 3 cm for mean joint error) from the depth point clouds in normal scenarios. By examining the failure cases of the reconstruction from the mmWave radar, the reasons for these failures may be attributed to sparse point clouds and large missing parts. (1) **Sparsity of radar points clouds** (see the input radar point clouds in Fig. 4): each frame of the mmWave radar only contains about 1k human points (at a distance of 3-5 m) due to the bandwidth and antennas of the Arbe Phoenix radar while a depth image contains up to 200k. Fig. 6 (a) presents the average number of points of depth and radar data within a radius of 0.15m around different joint locations. It can be observed that the number of points of the depth point cloud is much greater than that of the radar point cloud for all joints. (2) **Large missing parts**: some parts of the human body, such as the head and limbs, may not have radar points due to the specularity of mmWave signals, as shown in Fig. 4. These cases pose particular challenges different

TABLE III
ERRORS (CM) OF DIFFERENT METHODS FOR 3D BODY RECONSTRUCTION IN DIFFERENT SCENES ON THE MMBODY DATASET. FOR THE TWO COLUMNS OF EACH SCENE, THE FIRST COLUMN IS FOR JOINT ERROR AND THE SECOND VERTEX ERROR.

Scenes		Basic Scenes						Adverse Environments								Average	
		Lab1		**Lab2**		**Furnished**		**Rain**		**Smoke**		**Poor lighting**		**Occlusion**			
Single Modality	Images-Only	4.1	5.5	4.0	5.3	5.4	6.8	5.9	7.4	8.5	11.2	9.9	14.1	11.3	16.6	7.0	9.6
	Radar-Only	6.1	8.2	6.6	9.3	6.8	9.1	6.8	8.9	**8.3**	**10.5**	6.4	8.4	**7.5**	**9.6**	6.9	9.1
	Depth-Only	**3.1**	**4.0**	**3.3**	**4.3**	**3.7**	**4.4**	6.1	7.8	10.9	15.1	**4.1**	**4.7**	9.5	14.2	**5.8**	**7.8**
	Depth-Only-128	3.8	4.8	3.7	4.7	4.2	5.0	**5.7**	**6.9**	10.2	14.2	4.8	5.5	9.8	14.7	6.0	8.0
	Preprocessed-Depth-1024	5.6	6.8	4.4	5.7	4.7	5.7	7.4	10.1	11.1	14.6	5.5	6.6	9.5	14.2	6.9	9.1
	Preprocessed-Depth-512	5.8	7.1	4.7	5.7	5.1	6.2	6.7	9.1	10.7	15.0	5.8	6.8	9.5	14.1	6.9	9.1
	Preprocessed-Depth-128	6.0	7.4	5.2	6.4	5.3	6.7	6.5	8.7	11.5	15.7	6.1	7.5	9.6	14.1	7.2	9.5
	METRO [42]	4.9	7.0	4.4	6.4	6.5	8.8	7.2	9.4	8.4	11.6	13.6	17.9	19.1	26.0	9.2	12.5
	CLIFF [77]	4.0	5.4	3.9	5.1	5.3	6.7	6.0	7.6	8.7	11.6	11.9	17.1	11.4	16.7	7.3	10.0
	Zuo *et al.* [78]	8.0	10.6	8.8	11.5	8.5	12.1	9.2	11.2	8.7	12.2	8.4	11.8	9.5	12.3	8.7	11.7
	P4Transformer [79]	7.8	9.5	8.0	9.9	8.2	10.4	8.8	10.2	8.7	10.0	7.5	9.5	10.7	14.1	8.5	10.5
mmWave-RGB Fusion	Points-RGB [19]	6.7	9.3	6.7	8.7	6.6	8.9	7.7	10.1	11.3	14.8	7.0	9.4	12.0	17.2	8.3	11.2
	DenseFusion [52]	5.8	8.5	5.7	8.2	6.1	7.9	7.4	9.1	9.5	10.9	10.9	14.5	10.2	14.4	7.9	10.5
	Points-Image-Feature [20]	4.4	6.1	4.2	5.4	6.0	8.0	6.4	8.5	8.0	10.9	13.0	19.6	18.4	20.7	8.6	11.3
	DeepFusion [21]	5.1	6.5	5.7	6.8	6.7	8.2	7.0	8.2	9.6	12.1	13.4	16.9	13.3	17.8	8.7	10.9
	TokenFusion [22]	4.3	6.0	4.0	5.3	5.6	7.0	6.0	7.4	9.4	12.9	11.3	15.7	10.8	14.9	7.4	9.9
	ImmFusion (Ours)	**4.1**	**5.4**	**3.7**	**4.7**	**5.2**	**6.4**	**5.6**	**6.8**	**7.6**	**9.8**	**6.8**	**9.0**	**7.8**	**11.0**	**5.9**	**7.4**

from point clouds from the Kinect, thus more sophisticated algorithms are required to deal with these challenges. To verify the hypotheses, we conduct further analysis to investigate the impact of missing points and sparsity on reconstruction performance.

Sparsity. The number of points of the depth point cloud is much greater than that of the radar point cloud for all joints as shown in Fig. 6 (a), particularly in the lower body region where radar points are mostly absent. To verify our hypotheses, we downsample the depth point clouds to 128 points to validate the impact of sparsity on the reconstruction performance. As indicated in Table III, the errors of Depth-Only-128 are higher than that of Depth-Only in the basic scenes. However, in scenes with rain or smoke, the impact of the environment on the Depth-Only-128 method is relatively small. As shown in Fig. 5, the depth point clouds are significantly affected by rain and smoke, resulting in a noisy subset of points remaining. Therefore, the number of corrupted input points decreases after downsampling, leading to an improvement of performance.

Missing Points. We conduct further analysis to investigate the impact of missing points on reconstruction performance. We preprocess the depth and radar point clouds to reduce their gap on missing points and sparsity. Specifically, we utilize ground truth joint locations to remove most of the depth point clouds in the lower body region of the human body, and then randomly remove points near 1-5 limb joints to simulate the random missing characteristics of the radar point clouds. Subsequently, the radar and remaining depth point clouds are padded or downsampled to 1024 points as input. The average numbers of points of depth and radar point clouds for every joint are approximately close as shown in Fig. 6 (b).

The experimental results confirm our hypothesis. As reported in Table III, compared to the Depth-Only method, Preprocessed-Depth-1024 exhibits a significantly higher error in all scenes. In addition to padding the depth point cloud to 1024 points, we also downsample it to 512 and 128 points to investigate the impact of sparsity on performance. With increasing sparsity of the point cloud, the reconstruction error also increases. For instance, the error of the Preprocessed-Depth-128 is as high as that of the Radar-Only method in the lab1 scene.

In conclusion, our experiments demonstrate that: 1) it is feasible to reconstruct detailed 3D human bodies from mmWave point clouds; 2) compared with RGB and depth data, mmWave radar demonstrates higher errors in the basic scenes but exhibits stable performance in adverse weather conditions; 3) as revealed in our analysis, the sparsity and missing points of point clouds can severely impact the reconstruction performance. Meanwhile, these issues can also impair the effectiveness of traditional point-based fusion methods as demonstrated in Table V. Therefore, to accommodate these challenges and push the performance border of 3D body reconstruction further, we propose ImmFusion to combine the mmWave point clouds and RGB images to robustly reconstruct the 3D human body in various weather conditions. In contrast to projection-based methods, our well-devised self-attention Transformer modules can effectively fuse image and point cloud features and predict precise human body mesh. The self-attention mechanism allows the model to effectively select informative token features from arbitrary input modalities, and to dynamically fuse these features. The corrupted tokens from one modality could possibly be remedied by others or disregarded to accommodate the sparsity and missing parts of mmWave point clouds.

VI. EXPERIMENTS AND ANALYSIS

To evaluate our proposed fusion method, we conduct experiments on our collected mmBody dataset and the other public multi-modal human dataset, BEHAVE [80], to demonstrate its adaptability. To be fair, all the models are implemented using Pytorch and are trained on an Nvidia GeForce RTX 3090. We train all the networks for 50 epochs from scratch with an Adam optimizer and an initial learning rate of 0.001. The loss weights of λ and μ in our experiments are 1000 and 100.

A. Experimental Results for ImmFusion

Fig. 7 shows the reconstructed meshes from ImmFusion for different poses and subjects in the different scenarios. Overall,

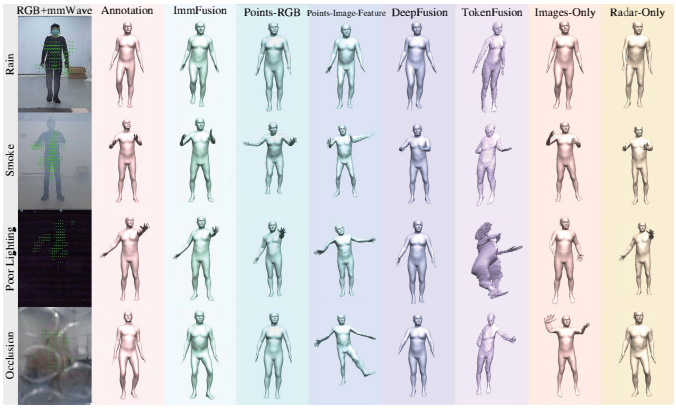

Fig. 7. Qualitative results. Each row represents an adverse weather scene (rain, smoke, poor lighting, and occlusion) and each column shows the reconstructed mesh, respectively.

the reconstructed meshes for most samples are close to the ground truth. Table III summarizes the main results of different fusion models tested on the mmBody dataset. Compared with existing fusion solutions and baselines, our approach can better exploit the complementary nature of two modalities: in addition to eliminating the negative effects of one modality on the other one, it also enhances the performance of one single modality by utilizing the complementary feature of the other.

Comparison with Single-Modality Methods. To demonstrate the effectiveness of our proposed fusion method, we compare ImmFusion with approaches using single-modality input. In addition to the methods implemented using the ImmFusion pipeline, we also evaluate SOTA single-modality methods on our dataset, including image-based methods METRO [42] and CLIFF [77], as well as point-based methods by Zuo *et al.* [78] and P4Transformer [79]. Experimental results demonstrate that ImmFusion is able to integrate strengths and mitigate defects of two modalities effectively. As shown in Table III, the average of mean joint errors and mean vertex errors of ImmFusion can reach as low as 5.8cm and 7.4cm, decreasing by more than about 1cm and 2cm from that of Images-Only or Radar-Only methods in Table III. Furthermore, ImmFusion achieves better accuracy than the other SOTA single-modality methods across all scenes, illustrating that ImmFusion can dynamically select preferable information from mmWave point cloud and RGB images. Particularly in poor lighting and occlusion scenes where the RGB camera fails, ImmFusion can work robustly as the mmWave radar emits active signals in the mmWave frequency range, which are independent of external light conditions and can penetrate through occlusions. Meanwhile, ImmFusion can effectively overcome the sparsity and missing points issues of the radar data in the basic scenes.

Comparison with Point-Level Fusion Methods. We conducted a comparative study between ImmFusion and point-level fusion methods, i.e. Points-RGB [19], DenseFusion [52], and Points-Image-Feature [20], which are implemented by augmenting point clouds with RGB values and image features. Results show that the intuitive fusion approach of Points-RGB yields minimal accuracy improvement in basic scenes and even performs worse than single-modal methods in adverse scenes. This can be mainly attributed to the limited exploration of inter-modality interactions. Issues like severe sparsity and missing points of radar point cloud and deficiency of RGB images in adverse environments can not be well settled in this fusion way. On the other hand, DenseFusion and Points-Image-Feature methods take a step further in integrating multi-modal features, resulting in some accuracy improvement in basic scenes. However, the inferior image features in severe scenes cannot be effectively constrained, leading to a significant degradation in performance.

TABLE IV
RESULTS (MM) ON THE BEHAVE [80] DATASET.

Methods	MPJPE ↓	MPVE ↓	PA-MPJPE ↓	PA-MPVE ↓
Mesh Graphormer [43]	65.35	83.81	39.23	64.42
VoteHMR [36]	63.34	72.28	40.33	52.25
CHORE [81]	-	-	-	55.80
CONTHO [82]	-	-	-	**49.90**
ImmFusion (Ours)	**54.56**	**72.11**	**38.68**	51.53

TABLE V
PERFORMANCE OF DIFFERENT FUSION METHODS FOR RGB IMAGES AND DEPTH (AND NOISY) POINT CLOUDS ON THE BEHAVE DATASET [80]. NOISY KINECT DEPTH IS DOWNSAMPLED KINECT DEPTH POINT CLOUDS WITH MISSING POINTS (SEE SECTION V FOR THE PROCESSING).

Methods	Image w/ Kinect Depth		Image w/ Noisy Kinect Depth	
	MPJPE ↓	MPVE ↓	MPJPE ↓	MPVE ↓
Points-Image-Feature [20]	62.34	82.12	75.01 (21% ↑)	96.52 (17% ↑)
DeepFusion [21]	59.64	76.14	70.73 (20% ↑)	88.28 (16% ↑)
TokenFusion [22]	56.24	74.23	79.83 (41% ↑)	95.14 (28% ↑)
ImmFusion (Ours)	**54.56**	**72.11**	**58.72** (**7%** ↑)	**76.12** (**5%** ↑)

Comparison with LiDAR-Camera Fusion Methods. We also compare ImmFusion with the state-of-the-art fusion methods DeepFusion [21] and TokenFusion [22]. DeepFusion exhibits inferior performance compared to ImmFusion in all scenarios. This can be mainly attributed to the lack of global features, which leads to reduced global interactions during the fusion stage. Regarding TokenFusion, it has been observed to exhibit suboptimal performance in challenging environments. We suspect the reason that TokenFusion aims to discard unimportant token features among Transformer layers, which is ineffective in incorporating the single-modality streams at the end of the model, which ultimately leads to unfavorable results as shown in Fig. 7. Furthermore, compared to the other two fusion methods, ImmFusion demonstrates superior robustness in poor lighting and occlusion scenes. Specifically, the mean joint errors of ImmFusion in these scenes only increase by about 3cm compared to the same scenes without poor lighting and occlusion. For reference, the errors of DeepFusion increase by about 8cm, and TokenFusion increases by about 7cm.

Performance on the Other Dataset. We further validate ImmFusion on the public BEHAVE [80] dataset, and the results are summarized in Table IV and Table V. The BEHAVE dataset is a challenging large-scale human-object interactions dataset that presents difficulties such as object occlusions and variations in background environments. As this dataset does not provide the mmWave data, we utilize the depth point cloud as the input of the point stream. We compare ImmFusion with the state-of-the-art single-modality reconstruction methods,

TABLE VI
ABLATION STUDY ON THE MMBODY DATASET.

Methods	Basic Scenes						Adverse Environments								Average	
	Lab1		**Lab2**		**Furnished**		**Rain**		**Smoke**		**Poor Lighting**		**Occlusion**			
ImmFusion-ResNet [83]	4.6	6.1	4.2	5.6	5.6	7.6	6.1	7.5	8.8	11.7	7.0	9.5	8.6	12.1	6.4	8.6
ImmFusion-CLIPResNet [84]	4.1	5.4	3.8	5.0	5.3	7.0	5.2	6.2	8.4	10.3	6.7	8.8	8.1	11.3	6.0	7.8
ImmFusion-PointNeXt [85]	**3.9**	**5.3**	**3.7**	4.8	**4.8**	**6.0**	**5.0**	**6.0**	**7.1**	**9.4**	**6.2**	**8.8**	**7.7**	**10.5**	**5.5**	**7.3**
ImmFusion-w/o-DR	4.1	5.4	3.7	4.7	5.1	6.0	5.7	6.9	7.5	9.6	7.2	9.3	8.0	11.2	5.9	7.7
ImmFusion-w/o-GC	4.2	5.5	3.8	4.9	5.3	6.5	5.8	6.9	7.7	10.2	7.0	9.5	8.8	12.7	6.1	8.0
ImmFusion-w/o-LF	4.9	6.5	4.7	6.0	6.0	7.8	6.7	8.1	8.5	10.9	10.9	15.5	10.4	14.4	7.4	9.9
ImmFusion-w/o-MMM	4.1	5.7	3.8	5.0	5.3	7.0	6.0	7.2	7.9	10.1	9.7	13.6	10.7	14.1	6.8	9.0
ImmFusion-FD	4.1	5.4	3.8	4.9	5.2	6.4	5.7	6.9	7.7	9.9	7.2	9.7	10.3	12.8	6.3	8.0
ImmFusion-w/o-GIM	4.1	5.5	3.7	4.8	5.3	6.6	6.1	7.3	7.7	9.7	7.6	9.5	9.6	14.9	6.3	8.3
ImmFusion	4.1	5.4	3.7	**4.7**	5.2	6.4	5.6	6.8	7.6	9.8	6.8	9.0	7.8	11.0	5.8	7.6

Mesh Graphormer [43], CHORE [81], CONTHO [82], and VoteHMR [36]. We also evaluate other fusion methods, i.e. Points-Image-Feature [20], DeepFusion [21], and TokenFusion [22] on this dataset. As we can see in Table IV, ImmFusion achieves lower errors compared to the single-modality and other fusion methods, which demonstrates its effectiveness in combining information from image and depth modalities. Additionally, ImmFusion can handle the fusion with noisy point clouds much more effectively compared with other traditional LiDAR-Camera fusion methods. As demonstrated in Table V, when other fusion methods are applied to sparse point clouds with large missing points, their performances deteriorate rapidly. In contrast, ImmFusion decreases slightly.

Computational Overhead. Our model consumes affordable computational resources and can achieve real-time performance. The total parameter count of ImmFusion is 228.3M and the running speed can achieve 14.6 fps.

B. Ablation Study

We conduct comprehensive ablation studies on different backbones, DR, GC, local features (LF), MMM, feature dropout (FD), and GIM on the mmBody dataset.

Ablation on Different Backbones. We study the behavior of extracting global and local features by using different modality backbones. We use the original ResNet-50 [83] and ResNet-50 pre-trained using CLIP [84] (CLIPResNet) for the image backbone and PointNeXt [85] for the point backbone. In Table VI, we observe that ImmFusion achieves inferior performance when using the original ResNet-50. However, after pre-training with numerous vision-language data, ImmFusion-CLIPResNet achieves competitive results. Additionally, utilizing the superior point backbone PointNeXt can also bring improvements.

Effectiveness of Dimension Reduction MLP. Due to the non-parametric approach our ImmFusion employs, FTM needs to process a large number (775) of input tokens. DR MLP reduces the training parameters while improving performance in adverse environments as demonstrated in Table VI.

Effectiveness of Graph Convolution. Despite the proficiency of Self-Attention in extracting long-range dependencies, it demonstrates less efficiency in capturing fine-grained information within intricate data structures like 3D meshes [43]. GC in our proposed network can effectively model interactions between joints and vertices. ImmFusion outperforms ImmFusion-w/o-GC in most scenes.

TABLE VII
ABLATION STUDY ON THE MASK RATIO.

Mask Ratio	MPJPE ↓	MPVE ↓
10%	6.1	7.9
30%	**5.8**	**7.6**
50%	6.0	7.8

Effectiveness of Local Features. The local features, which directly affect the quality and details, play a very important role in reconstruction tasks. To analyze the effectiveness of the local features, we compared the results of the original ImmFusion with its variation ImmFusion-w/o-LF, in which the cluster features and grid features are removed from the input of FTM. As indicated in Table VI, the mean errors of ImmFusion-w/o-LF are obviously greater than ImmFusion.

Effectiveness of Modality Masking Module. An important question is whether MMM is valid. The results of single-modality methods, i.e. Images-Only and Points-Only in Table III report that RGB images have better accuracy than mmWave point clouds in the basic scenes due to the high resolution. Therefore, the training set only consisting of basic data would force the Transformer module to pay more attention to the image modality, which leads to a rapid decline of the performance in the poor lighting and occlusion scenes. Clearly, MMM eliminates the bias of training data and significantly improves the performance in extreme scenes as the result of ImmFusion-w/o-MMM demonstrates. We further compare our masking strategy with dropout at the feature level. We can see that ImmFusion achieves better results since MMM enforces the Transformer module to lean more attention on the effective modality to select helpful features. We train several models with varying maximum masking ratios to choose the best one and the optimal proportion is 30% as shown in Table VII.

Effectiveness of Global Integrated Module. In ImmFusion, GIM serves as a mixer to integrate global features of mmWave and RGB input. Instead of naive element-wise addition or channel-wise concatenation, GIM contains learnable parameters to control the weights of global features from different modalities. Among all types of scenes in Table VI, ImmFusion-w/o-GIM merely outperforms ImmFusion a lit-

tle in the smoke environment, where the valid information proportion of RGB v.s mmWave is balanced, misleading the model to select useless features from the global feature. In other situations, especially in the poor lighting and occlusion scene, ImmFusion-w/o-GIM clearly underperforms compared with ImmFusion, proving the importance of GIM.

VII. Conclusions and Limitations

In this paper, we introduce ImmFusion, a multi-modal fusion model which combines mmWave and RGB signals for robust all-weather 3D human body reconstruction. In addition to the good results in basic scenes, ImmFusion shows great robustness in severe environments like rain, smoke, poor lighting, and occlusion due to the effectiveness of the attention mechanism and the Modality Masking Module. To evaluate our method, an automatic capture and annotation system is built up with multiple sensors. We collect a large-scale multi-modal 3D body reconstruction dataset and close the gap of no available public datasets to study the problem of reconstructing the 3D human body from multi-view multi-modal inputs in different scenes. Experimental results suggest that ImmFusion can efficiently fuse the information of mmWave and RGB signals. In addition, we investigate various fusion approaches and demonstrate that ImmFusion outperforms single-modality, point-level, and LiDAR-camera fusion methods in all basic scenes and the majority of adverse environments.

Though with the masking module, our model has gained a certain level of generalization ability: it exhibits relatively high error in furnished, smoke, and occlusion conditions on the mmBody dataset due to sensor defects and data imbalance. Contrastive and predictive multi-modal pre-training are promising solutions. Additionally, constrained by the data collection, we could only capture data in the indoor scenes with fixed sensors. With the advancement of autonomous driving and mobile robotics, information fusion from various modalities with dynamic multiple viewpoints in outdoor environments is an important problem. We leave the extension of our method to such scenarios as future work.

References

[1] D. S. Alexiadis, A. Chatzitofis, N. Zioulis, O. Zoidi, G. Louizis, D. Zarpalas, and P. Daras, "An integrated platform for live 3d human reconstruction and motion capturing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 798–813, 2016.

[2] S. He, K. Shi, C. Liu, B. Guo, J. Chen, and Z. Shi, "Collaborative sensing in internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, 2022.

[3] A. Chen, X. Wang, K. Shi, S. Zhu, B. Fang, Y. Chen, J. Chen, Y. Huo, and Q. Ye, "Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2752–2758.

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

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

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

[7] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, "Making the invisible visible: Action recognition through walls and occlusions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 872–881.

[8] M. Zhao, Y. Liu, A. Raghu, T. Li, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human mesh recovery using radio signals," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 113–10 122.

[9] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "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*, 2021, pp. 269–282.

[10] S. An, Y. Li, and U. Ogras, "mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 414–27 426, 2022.

[11] K. Shi, Z. Shi, C. Yang, S. He, J. Chen, and A. Chen, "Road-map aided gm-phd filter for multivehicle tracking with automotive radar," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 1, pp. 97–108, 2021.

[12] S. Yao, R. Guan, X. Huang, Z. Li, X. Sha, Y. Yue, E. G. Lim, H. Seo, K. L. Man, X. Zhu *et al.*, "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Vehicles*, 2023.

[13] S. Yao, R. Guan, Z. Peng, C. Xu, Y. Shi, Y. Yue, E. G. Lim, H. Seo, K. L. Man, X. Zhu *et al.*, "Radar perception in autonomous driving: Exploring different data representations," *arXiv preprint arXiv:2312.04861*, 2023.

[14] Y. Cheng and Y. Liu, "Person reidentification based on automotive radar point clouds," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[15] S. Zou, Y. Xu, C. Li, L. Ma, L. Cheng, and M. Vo, "Snipper: A spatiotemporal transformer for simultaneous multi-person 3d pose estimation tracking and forecasting on a video snippet," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[16] M. Aladsani, A. Alkhateeb, and G. C. Trichopoulos, "Leveraging mmwave imaging and communications for simultaneous localization and mapping," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4539–4543.

[17] C. X. Lu, M. R. U. Saputra, P. Zhao, Y. Almalioglu, P. P. de Gusmao, C. Chen, K. Sun, N. Trigoni, and A. Markham, "milliego: single-chip mmwave radar aided egomotion estimation via deep sensor fusion," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 109–122.

[18] K. Garcia, M. Yan, and A. Purkovic, "Robust traffic and intersection monitoring using millimeter wave sensors," *Texas Instruments*, 2018.

[19] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[20] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 794–11 803.

[21] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 182–17 191.

[22] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 186–12 195.

[23] J. Wang and X. Tan, "Mutually beneficial transformer for multimodal data fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[24] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 975–10 985.

[25] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[26] S. B. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor-system description, issues and solutions," in *2004 conference on*

computer vision and pattern recognition workshop. IEEE, 2004, pp. 35–35.

[27] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7025–7034.

[28] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma, "Dr-pose3d: Depth ranking in 3d human pose estimation," *arXiv preprint arXiv:1805.08973*, 2018.

[29] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7122–7131.

[30] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "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*, 2019, pp. 2252–2261.

[31] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5253–5263.

[32] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3d people in depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 243–13 252.

[33] A. Sengupta, I. Budvytis, and R. Cipolla, "Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 094–16 104.

[34] Z. Li, M. Oskarsson, and A. Heyden, "3d human pose and shape estimation through collaborative learning and multi-view model-fitting," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1888–1897.

[35] B. Huang, J. Ju, Y. Shu, and Y. Wang, "Simultaneously recovering multi-person meshes and multi-view cameras with human semantics," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[36] G. Liu, Y. Rong, and L. Sheng, "Votehmr: Occlusion-aware voting network for robust 3d human mesh recovery from partial point clouds," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 955–964.

[37] P. Hu, E. S.-L. Ho, and A. Munteanu, "3dbodynet: fast reconstruction of 3d animatable human body shape from a single commodity depth camera," *IEEE Transactions on Multimedia*, vol. 24, pp. 2139–2149, 2021.

[38] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao, "Video-based outdoor human reconstruction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 760–770, 2016.

[39] J. Yang, X. Guo, K. Li, M. Wang, Y.-K. Lai, and F. Wu, "Spatio-temporal reconstruction for 3d motion recovery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1583–1596, 2019.

[40] Y. Sun, Q. Bao, W. Liu, T. Mei, and M. J. Black, "Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8856–8866.

[41] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4501–4510.

[42] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1954–1963.

[43] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 939–12 948.

[44] Z. Liu, J. Huang, J. Han, S. Bu, and J. Lv, "Human motion tracking by multiple rgbd cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 2014–2027, 2016.

[45] Y. Wang, W. Wang, M. Zhou, A. Ren, and Z. Tian, "Remote monitoring of human vital signs based on 77-ghz mm-wave fmcw radar," *Sensors*, vol. 20, no. 10, p. 2999, 2020.

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

[47] H. Nodehi and A. Shahbahrami, "Multi-metric re-identification for online multi-person tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 147–159, 2021.

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

[49] C. Xie, D. Zhang, Z. Wu, C. Yu, Y. Hu, and Y. Chen, "Rpm 2.0: Rf-based pose machines for multi-person 3d pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[50] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[51] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1527–1536.

[52] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7345–7353.

[53] J. Liu, W. Sun, C. Liu, X. Zhang, S. Fan, and W. Wu, "Hff6d: Hierarchical feature fusion network for robust 6d object pose tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7719–7731, 2022.

[54] Y. Zhang, J. Chen, and D. Huang, "Cat-det: Contrastively augmented transformer for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 908–917.

[55] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.

[56] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 172–181.

[57] J. Yan, Y. Liu, J. Sun, F. Jia, S. Li, T. Wang, and X. Zhang, "Cross modal transformer: Towards fast and robust 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 268–18 278.

[58] H. Wang, H. Tang, S. Shi, A. Li, Z. Li, B. Schiele, and L. Wang, "Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6792–6802.

[59] C. G. Lab, "Cmu graphics lab motion capture database," http://mocap.cs.cmu.edu/, 2000.

[60] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. [Online]. Available: http://gvv.mpi-inf.mpg.de/3dhp_dataset

[61] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[62] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 601–617.

[63] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[64] Y. Wang, G. Wang, H.-M. Hsu, H. Liu, and J.-N. Hwang, "Rethinking of radar's role: A camera-radar dataset and systematic annotator via coordinate alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2815–2824.

[65] S. Yao, R. Guan, Z. Wu, Y. Ni, Z. Huang, R. W. Liu, Y. Yue, W. Ding, E. G. Lim, H. Seo *et al.*, "Waterscenes: A multi-task 4d radar-camera fusion dataset and benchmarks for autonomous driving on water surfaces," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[66] M. A. Richards *et al.*, *Fundamentals of radar signal processing*. Mcgraw-hill New York, 2005, vol. 1.

[67] A. Robotics, "4d image radar," 2021. [Online]. Available: https://arberobotics.com/wp-content/uploads/2021/05/4D-Imaging-radar-product-overview.pdf
[68] J. Park, Q.-Y. Zhou, and V. Koltun, "Colored point cloud registration revisited," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 143–152.
[69] I. Akhter and M. J. Black, "Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1446–1455.
[70] O. ACCAD, "Accad," https://accad.osu.edu/research/motion-lab/system-data, 2022.
[71] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of Motion Capture As Surface Shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5442–5451.
[72] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
[73] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas, "Imvotenet: Boosting 3d object detection in point clouds with image votes," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2020, pp. 4404–4413.
[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[75] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 682–11 692.
[76] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, pp. 33–51, 1975.
[77] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "Cliff: Carrying location information in full frames into human pose and shape estimation," in *ECCV*, 2022.
[78] X. Zuo, S. Wang, Q. Sun, M. Gong, and L. Cheng, "Self-supervised 3d human mesh recovery from noisy point clouds," *arXiv preprint arXiv:2107.07539*, 2021.
[79] H. Fan, Y. Yang, and M. Kankanhalli, "Point 4d transformer networks for spatio-temporal modeling in point cloud videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 204–14 213.
[80] B. L. Bhatnagar, X. Xie, I. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Behave: Dataset and method for tracking human object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.
[81] X. Xie, B. L. Bhatnagar, and G. Pons-Moll, "Chore: Contact, human and object reconstruction from a single rgb image," in *European Conference on Computer Vision*. Springer, 2022, pp. 125–145.
[82] H. Nam, D. S. Jung, G. Moon, and K. M. Lee, "Joint reconstruction of 3d human and object via contact-based refinement transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 218–10 227.
[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[84] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
[85] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem, "Pointnext: Revisiting pointnet++ with improved training and scaling strategies," *Advances in neural information processing systems*, vol. 35, pp. 23 192–23 204, 2022.