2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

ImmFusion: Robust mmWave-RGB Fusion for 3D Human Body Reconstruction in All Weather Conditions

Anjun Chen[1], Xiangyu Wang[1], Kun Shi[1], Shaohao Zhu[1], Bin Fang[2], Yingfeng Chen[3], Jiming Chen[1], Yuchi Huo[4], and Qi Ye[1]

***Abstract*— 3D human reconstruction from RGB images achieves decent results in good weather conditions but degrades dramatically in rough weather. Complementary, mmWave radars have been employed to reconstruct 3D human joints and meshes in rough weather. However, combining RGB and mmWave signals for robust all-weather 3D human reconstruction is still an open challenge, given the sparse nature of mmWave and the vulnerability of RGB images. In this paper, we present ImmFusion, the first mmWave-RGB fusion solution to reconstruct 3D human bodies in all weather conditions robustly. 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 on a large-scale dataset, mmBody, captured in various environments demonstrate that ImmFusion can efficiently utilize the information of two modalities to achieve a robust 3D human body reconstruction in all weather conditions. In addition, our method's accuracy is significantly superior to that of state-of-the-art Transformer-based LiDAR-camera fusion methods.**

I. Introduction

3D human body reconstruction is widely used in many practical robotic applications [1], [2], [3] like human-robot interaction, pose estimation, and motion capture. Currently, the most popular approach is to reconstruct from RGB images, due to the progress of computer vision technologies. Nevertheless, the performance of reconstruction using RGB images under rough circumstances is still limited, lying that the perception capability of RGB cameras will rapidly deteriorate in poor illumination or inclement weather conditions. On the flip side, as regressing depth from a single image is inherently an ill-posed problem, the 3D reconstruction based on monocular cameras is fairly complicated.

Recently, mmWave radar has gained increasing popularity in wireless sensing areas, like autonomous driving [4], [5], human activity recognition [6], [7], and UAV [8]. Simultaneously, the mmWave radar has demonstrated great potential in human body reconstruction tasks for keeping unaffected by adverse environments. Therefore, its reconstruction results are competitive with or superior to that from RGB images in particular cases, as prior work [9] reveals.

Despite the depth measurement and resistance to extreme weather conditions, reconstruction using mmWave signals suffers from sparsity and multi-path effect, which restricts its high performance in normal scenes. For example, the range and angle resolution of the device used in [9] are only 0.2m and 2 degrees, respectively. Introducing RGB signals into the mmWave system could be a possible solution. Fusing the two modalities to combine their strengths should be the key to realizing robust 3D human body reconstruction in all weather conditions.

However, combining multi-modal information is not trivial. Most existing LiDAR-camera fusion approaches adopt point-to-image projection to fuse point clouds and RGB pixel values by element-wise addition or channel-wise concatenation. Nevertheless, directly attaching the sparse, noisy, randomly missing, and temporally flicking mmWave point cloud to the RGB image would degrade the extracted features [9], especially in harsh circumstances like poor illumination. Therefore, this strategy is not suitable for mmWave-RGB fusion. Alternatively, Transformer [10] opens up new avenues for the research on multi-modal fusion methods [11], [12], [13], [14], [15]. These Transformer fusion frameworks, however, focus on LiDAR-camera fusion-based object detection, which is inapplicable for the mmWave-RGB fusion-based human body reconstruction task.

To address these issues, we present ImmFusion, the first fusion solution to combine the mmWave point clouds and RGB images to robustly reconstruct the 3D human body in all weather conditions. In view of the background noise caused by the multi-path effect, we extract features from point clusters rather than individual points. Besides, in order to address the spatial-temporal misalignment of heterogeneous modalities, we fuse the dense image features with the sparse radar point cluster features through a well-devised Fusion Transformer Module utilizing both local and global features. In addition, we conduct extensive experiments on the large-scale mmWave-human body dataset [9], with 20 subjects captured in 7 scenes, including extreme weather conditions like fog, rain, and night. We evaluate the performance of our fusion model under different scenes, and it outperforms single-modality, point-level, and LiDAR-camera

[1]State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China.
[2]Tsinghua University, Beijing, China
[3]Fuxi AI Lab, NetEase, Hangzhou, China
[4]State Key Lab of CAD&CG, Zhejiang University and Zhejiang Lab, Hangzhou, China
Corresponding author: *Qi Ye* (qi.ye@zju.edu.cn). Qi Ye is with the College of Control Science and Engineering, the State Key Laboratory of Industrial Control Technology, Zhejiang University, and also the Key Laboratory of Collaborative Sensing and Autonomous Unmanned Systems of Zhejiang Province. This work was supported in part by NSFC under Grants 62088101, 62233013, and 62103372. Project page: https://chen3110.github.io/ImmFusion

fusion methods in all weather environments. Our contributions can be summarized as follows:

- We propose ImmFusion, the first state-of-the-art mmWave-RGB fusion method for 3D human body reconstruction in all weather conditions including severe environments like rain, smoke, and poor illumination.
- We employ a well-devised Fusion Transformer Module to effectively fuse the global and local features extracted from mmWave and RGB modalities, and we design an ingenious Modality Masking Module to strengthen the robustness of the model across all scenarios.
- We evaluate the ImmFusion on the large-scale mmWave dataset mmBody [9] and demonstrate that ImmFusion outperforms other non-fusion or LiDAR-camera fusion models in all weather conditions.

II. Related Works

A. Human Mesh Reconstruction from RGB Images

3D human body reconstruction from RGB images can be broadly categorized into parametric [16], [17] and non-parametric approaches [18], [19], [20]. In the former, a body model, such as SMPL [16] or SMPL-X [17], represents the human body and regresses the input images. However, it is still challenging to estimate precise coefficients from a single image [21], [22]. To improve the reconstruction, researchers utilize more visual information or dense relationship maps [23]. Instead, non-parametric approaches directly regress the vertices from an image. Most pioneers choose Graph Convolutional Neural Network [18] to model the local interactions between neighboring vertices with an adjacency matrix. Recently, METRO [19] utilizes a Transformer encoder to jointly model vertex-vertex and vertex-joint interaction globally. Mesh Graphormer [20] puts a step further by proposing a graph-convolution-reinforced transformer encoder and adding image grid features for joints and mesh vertices.

B. mmWave-based Human Sensing

Because mmWave sensors are capable of working in extreme conditions like rain, smoke, and occlusion, they have been widely utilized for sensing applications like human monitoring and tracking [24], detection and identification [7], and behavior recognition [6]. Recently, Xue et al. [25] proposed an accessible real-time human mesh reconstruction solution employing commercial portable mmWave devices. However, this work's datasets are not public, and the capability of the reconstruction from the mmWave signals in bad conditions is not studied. To fill the gap, Chen et al. [9] present a large-scale mmWave human body dataset with paired RGBD images in various environments, which paves the way for further research on combing mmWave radars with RGBD cameras for 3D body reconstruction.

C. Fusion Methods for 3D Object Detection

The bulk of literature is dedicated to mmWave-RGB fusion, which can be broadly divided into data, feature, and decision levels. 1) The data-level fusion can be further categorized into radar-to-camera projection [26], [27] and camera-to-radar mapping [28]. 2) The decision-level fusion methods [29] usually leverage one modality to generate ROI containing valid objects, then make use of the other modality inside the ROI. 3) The feature-level fusion derives from a pioneering framework dubbed AVOD [30], where a series of anchors are predefined in the front-view map and the BEV map, respectively. Then, proposal-wise features are extracted within the region proposals and aggregated by specific fusion operations [31].

Recently, the success of Transformer [10] draws much attention to Transformer-based LiDAR-camera fusion. Specifically, DeepFusion [11] fuses deep camera and LiDAR features instead of decorating raw LiDAR points at the input level. Therein, LearnableAlign is introduced that leverages the cross-attention mechanism to dynamically correlate LiDAR information with the most related camera features. Besides, TokenFusion [12] first prunes single-modal transformers, assuming that they can preserve information better, and further re-utilizes the pruned units for multimodal fusion. Moreover, TransFusion [13] conducts LiDAR-camera fusion with a soft-association approach to cope with inferior image situations. These works, however, differ from ours since we make efforts in constructing a general mmWave-RGB fusion pipeline for 3D human body reconstruction.

III. ImmFusion

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. 1 (a) illustrates the framework of ImmFusion. The structure aims to efficiently diffuse the image and point cloud features at global and local levels to predict the human body mesh. Given a radar point cloud with fixed numbers of points and an image with a size of 224×224, 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 Multi-Layer Perceptrons (MLPs) to upsample the coarse mesh vertices to the full SMPL-X [17] mesh vertices.

A. Preliminary of 3D Human Body Reconstruction

3D human body reconstruction aims to predict the 3D positions of all the joints and vertices. We adopt the non-parametric approach mentioned above for body reconstruction. As our focus is reconstruction, we use 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 $\mathcal{D} = \{P_t, I_t, J_t, V_t\}, t = 0, \ldots, N$, where $P_t \in \mathbb{R}^{1024\times3}, I_t \in \mathbb{R}^{224\times224\times3}$ 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

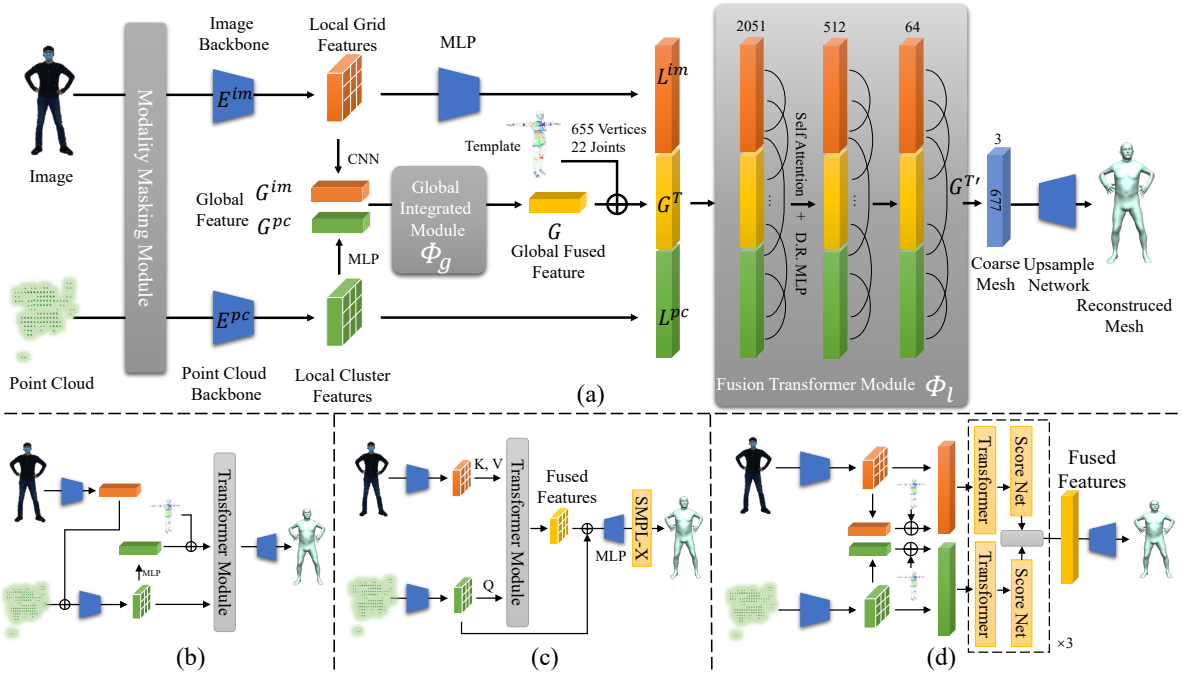

Fig. 1. Comparison of different fusion strategies. (a) Our proposed ImmFusion. D.R. MLP stands for a dimension reduction MLP. (b) Point-level fusion methods [32]. (c) DeepFusion [11]. (d) TokenFusion [12].

$J_t \in \mathbb{R}^{22\times3}, V_t \in \mathbb{R}^{10475\times3}$ are the annotation locations of 22 joints and 10475 vertices at time t, we make efforts to fuse information from two modalities of input to reconstruct 3D human body.

B. Extraction of Global and Local Features

Early point-level fusion works [32], [33] concatenate image features or projected RGB pixels to the point clouds as extended features of the point-based model, as Fig. 1 (b) illustrates. However, this fusion strategy is not suitable for mmWave-RGB fusion due to the sparsity and noise of radar points. As discussed in [9], undesirable issues like randomly missing and temporally flicking would lead to fetching fewer or even wrong image features. Additionally, the low quality of image features in adverse environments like poor lighting would severely degrade the performance of the model. Because of the imbalanced image feature qualities for different body parts, the reconstruction errors of different joints significantly vary. Instead, the point features extracted by networks like PointNet++ [34] are relatively more balanced at different parts. Therefore, we propose fusing the image features and point features to improve the accuracy of the whole body.

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. 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 global feature $G^{im} \in \mathbb{R}^{2048}$ and grid features $L^{im} \in \mathbb{R}^{49\times2051}$ using HRNet [35] E^{im}. (MLPs are used to make features from HRNet the same as that of the point features.)

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,

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

Following [18], we perform positional encoding by attaching 3D coordinates of each joint and vertex in a human template mesh to the global vector $G^T = cat(J^{template}, V^{template}, G)$, where $G^T \in \mathbb{R}^{677\times2051}$. Both local features serve the purpose of providing fine-grained local details for body reconstruction.

C. Transformer Fusion with Global and Local Features

Multi-head attention module [10] is famous for modeling the relationship between information tokens. We adopt this structure to mitigate the feature degradation caused by the sparsity and noise of mmWave signals and the deficiency of RGB information in extreme conditions. We utilize the Fusion Transformer Module Φ_l to combine the strengths of radar points and images, enabling the model to select informative token features from two modalities dynamically,

$$G^{T\prime}, L^{im\prime}, L^{pc\prime} = \Phi_l\left(G^T, L^{im}, L^{pc}\right). \tag{2}$$

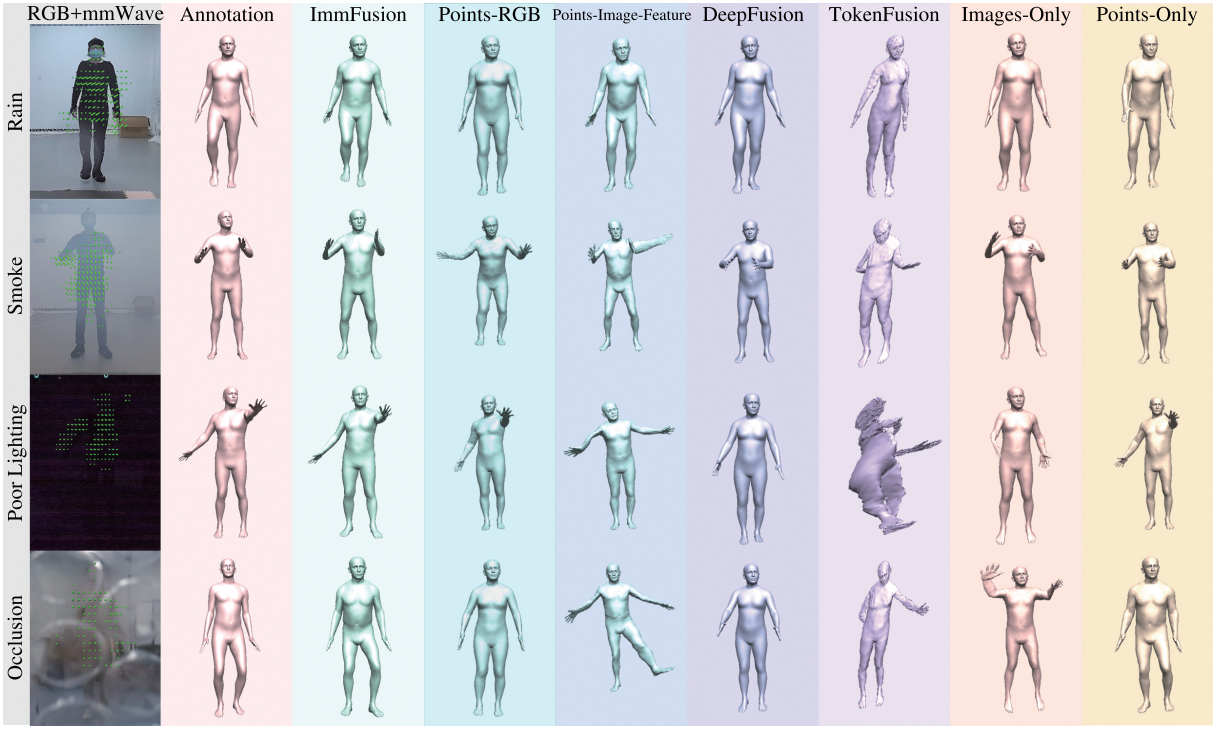

Fig. 2. Qualitative results. Each row represents an adverse weather scene(rain, smoke, poor lighting, and occlusion) and each column shows the reconstructed mesh of the corresponding model, respectively.

where $G^{T'} \in \mathbb{R}^{677\times64}$, $L^{im\prime} \in \mathbb{R}^{49\times64}$ and $L^{pc\prime} \in \mathbb{R}^{32\times64}$. While attending to valid features and restricting undesirable features, the Fusion Transformer Module Φ_l 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 architecture, Graph Convolution [18], to decode the queries $G^{T'}$ containing rich cross-modalities information into 3D coordinates of joints and vertices. Last, a linear projection network implemented using MLPs upsamples the coarse output mesh to the original 10475 vertices.

D. Distortion Solution by Modality Masking

Despite the superiority of the multi-head attention mechanism, the model is prone to struggle with sensor distortions according to [36] 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 model's adaptability across general 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 input token features to simulate self or smoke occlusions and missing parts.

E. Comparison with Relevant Methods

We compare our proposed ImmFusion with other relevant fusion methods. Point-level fusion methods are implemented by decorating point clouds with RGB values fetched by projecting 3D point clouds to the original image plane [35] or image features extracted by the 2D CNN backbone and feeding them directly to the Points-Only pipeline as Fig. 1 (b) shows. We also compare ImmFusion with the LiDAR-camera fusion methods, i.e. DeepFusion [11] and TokenFusion [12], which show the state-of-the-art accuracy in the task of 3D object detection. As DeepFusion is a generic Transformer-based block that is incompatible with the dimension-reduction mechanism, we adopt the parametric reconstruction pipeline by replacing the detection framework of DeepFusion with linear projection to regress SMPL-X parameters. The framework of DeepFusion for 3D body reconstruction is shown in Fig. 1 (c). As for the TokenFusion method, we mainly implement it by referring to the scoring function in [12]. Specifically, we plug the scoring net among the Transformer layers of single-modality models to dynamically predict the importance of joint/vertex tokens and substitute inferior tokens with corresponding ones from the other modality as Fig. 1 (d) suggests.

IV. Experiments

A. Experimental Settings

Dataset We conduct the experiments on the large-scale mmWave 3D human body dataset mmBody [9], which con-

TABLE I

ERRORS (CM) OF DIFFERENT METHODS FOR 3D BODY RECONSTRUCTION IN DIFFERENT SCENES. 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**			
Mean Error	Points-RGB	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
	Points-Image-Feature	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[35]	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[12]	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
	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
	Points-Only	6.3	8.8	6.7	8.9	6.4	8.8	7.8	10.2	8.0	10.6	**6.2**	**8.4**	8.8	12.7	7.2	9.8
	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-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**	**5.9**	**7.4**
Max Error	Points-RGB	24.8	34.3	27.0	38.7	22.6	31.5	29.2	38.2	38.8	54.7	24.4	33.2	36.4	47.9	29.0	39.8
	Points-Image-Feature	13.2	19.2	13.1	19.1	18.2	25.7	22.2	29.1	22.9	32.6	60.2	82.8	72.5	79.6	31.8	41.1
	DeepFusion[35]	10.8	15.6	12.9	18.4	13.9	19.9	22.0	27.9	20.2	28.2	37.8	55.2	41.1	57.1	22.7	31.7
	TokenFusion[12]	12.7	17.8	13.5	19.9	15.0	21.6	21.7	26.9	26.5	36.9	46.3	69.4	38.5	59.4	24.9	36.0
	Images-Only	11.3	15.5	12.1	17.2	14.0	18.9	21.7	27.5	23.2	31.2	43.2	64.5	48.6	73.8	24.9	35.5
	Points-Only	22.8	32.5	28.2	41.8	22.3	31.5	29.4	39.1	25.7	36.6	**21.7**	**30.3**	28.3	38.4	25.5	35.8
	ImmFusion-w/o-LF	14.9	21.5	16.3	23.7	19.1	26.1	23.4	29.7	24.1	31.9	45.3	70.2	37.5	56.0	25.8	37.0
	ImmFusion-w/o-MMM	**10.8**	16.2	10.6	16.0	13.7	20.0	21.4	26.7	21.2	29.7	39.6	56.5	39.3	57.3	22.4	31.8
	ImmFusion-w/o-GIM	10.9	15.3	10.7	14.9	14.1	18.3	21.3	26.2	**20.2**	**27.0**	25.5	33.8	30.5	40.4	19.0	25.1
	ImmFusion	10.9	**14.7**	**10.5**	**14.2**	**13.6**	**18.2**	**20.1**	**24.6**	20.3	27.2	23.1	31.3	**27.0**	**35.8**	**17.9**	**23.7**

sists of a considerable number of synchronized and calibrated mmWave radar point clouds and RGBD images in various conditions and mesh annotations for humans in the scenes. More specifications about the mmWave radar are provided in the product overview [37]. We choose 10 sequences in the lab scenes as the training set while 2 sequences for each scene including labs, furnished, rain, smoke, poor lighting, and occlusion as the test set.

Metrics To evaluate the performance of the reconstruction, we employ the metrics of mean (max) joint error per frame and mean (max) vertex error per frame, which quantify the average (maximum) Euclidean distance between the prediction and the ground truth for joints/vertices in each frame. In comparison to the mean joint/vertex error, the max joint/vertex error is stricter.

Implementation Details Our ImmFusion applies L_1 loss to the reconstructed mesh to constrain the vertices and joints. In addition, the coarse meshes of each layer of the Fusion Transformer Module are also supervised by downsampled ground truth meshes using L_1 loss to accelerate convergence. All the models are implemented using Pytorch and are trained on an Nvidia GeForce RTX 3090. To be fair, we train all the networks for 50 epochs from scratch with an Adam optimizer and an initial learning rate of 0.001.

B. Experimental Results

Fig. 2 shows the reconstructed meshes from ImmFusion for different poses and subjects in the different scenarios. Overall, the reconstructed meshes for most samples are close to the ground truth. Tab. I summarizes the main results of all 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. We provide more qualitative comparisons by visualizing the attention interactions between different token features in our accompanying video.

Comparison with Single-modality Methods To demonstrate the effectiveness of our proposed fusion method, we compare ImmFusion with approaches using single-modality input. We implement single-modality methods by removing one input stream from our proposed ImmFusion pipeline. For the Images-Only input, the feature extractor consists of only the CNN backbone to extract image features. Regarding the Points-Only input, we substitute the CNN backbone with PointNet++. Experimental results demonstrate that ImmFusion is able to integrate the two modalities effectively. As shown in Tab. I, the average of mean joint errors and mean vertex errors for ImmFusion can reach as low as 5.8cm and 7.6cm, decreasing by more than about 1cm and 2cm from that for Images-Only or Points-Only methods. Furthermore, ImmFusion achieves better accuracy of max errors than the other two types of single-modality methods across all scenes, illustrating that ImmFusion can dynamically select preferable information from mmWave point cloud and RGB images. Particularly for poor lighting and occlusion scenes where the RGB camera fails, ImmFusion can work robustly as the mmWave radar uses active lighting of mmWave frequency.

Comparison with Point-level Fusion Methods We compare ImmFusion with point-level fusion methods Points-RGB and Points-Image-Feature implemented by decorating point clouds with RGB values and image features. For the Points-RGB method, this intuitive fusion gains little accuracy improvement in basic scenes and even performs worse than the single-modality methods in adverse scenes, which is mainly due to the under-exploration of the inter-modality interaction. The Points-Image-Feature baseline makes a further step to integrate the multi-modal information, which gains accuracy

in the basic scenes to some extent. However, the inferior image features in the severe scenes cannot be restricted by the attention module in this way, which severely degrades the performance.

Comparison with LiDAR-camera Fusion Methods We also compare ImmFusion with the state-of-the-art fusion methods DeepFusion and TokenFusion. In spite of the simple fusion strategy, ImmFusion achieves more preferable results in all scenarios. DeepFusion tends to lose more global interactions due to the lack of global features. While TokenFusion aims to discard unimportant token features among Transformer layers, it is ineffective to incorporate the single-modality streams at the end of the model, which ultimately leads to unfavorable results as shown in Fig. 2.

C. Ablation Study

We conduct a comprehensive study to validate the effectiveness of local features, Modality Masking Module (MMM), and Global Integrated Module (GIM).

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 backward computation graph. As indicated in Tab. I, the mean and max errors of ImmFusion-w/o-LF are obviously greater than ImmFusion. Despite the assistance of MMM, the errors in extreme conditions like poor lighting or occlusion are even worse than that of the Images-Only model. The max vertex error in poor lighting is up to 70cm, double higher than that of the original ImmFusion. These results strongly support our motivation of utilizing local features to benefit the quality of reconstruction.

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 Tab. I 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. Surprisingly, MMM gains the accuracy of the model across all scenes, which is mainly due to the fact that MMM enforces the Transformer module to lean more attention on the point modality to select helpful features. We train several models with varying maximum masking percentages to choose the best one and the optimal proportion is 30%.

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 Tab. I, ImmFusion-w/o-GIM merely outperforms ImmFusion a little 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.

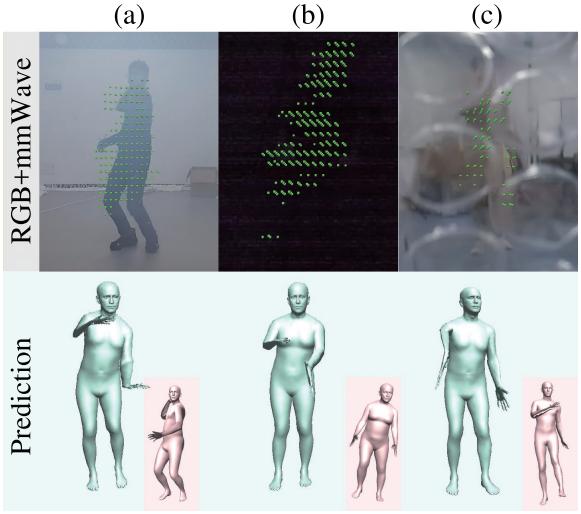

Fig. 3. Failure cases. Columns (a) (b), and (c) show the unfavorable reconstruction results in the smoke, poor lighting, and occlusion scene respectively (Ground truth in pink).

D. Future Work

There are situations when the reconstruction from ImmFusion fails, some of which are exampled in Fig. 3. The reasons for these failures can be mainly attributed to the drawbacks of sensors. The low quality of images in rough conditions and the sparsity of mmWave point clouds severely complicate the regression task since non-parametric methods are prone to give non-smooth meshes when facing these challenges. Adding the third modality like depth images or LiDAR point clouds and exploiting a unified sensor fusion framework may settle this problem. Additionally, constrained by the data collection, we leave the extension of our method to the outdoor scenario as future work.

V. CONCLUSIONS

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. 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.

REFERENCES

[1] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti, "3d reconstruction of freely moving persons for re-identification with a depth sensor," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 4512–4519.

[2] A. Pereira, G. Stillfried, T. Baker, A. Schmidt, A. Maier, B. Pleintinger, Z. Chen, T. Hulin, and N. Y. Lii, "Reconstructing human hand pose and configuration using a fixed-base exoskeleton," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3514–3520.

[3] J. Liu, J. Rojas, Y. Li, Z. Liang, Y. Guan, N. Xi, and H. Zhu, "A graph attention spatio-temporal convolutional network for 3d human pose estimation in video," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 3374–3380.

[4] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, and T.-K. Kim, "Geometry-based distance decomposition for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 172–15 181.

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

[6] Z. Meng, S. Fu, J. Yan, H. Liang, A. Zhou, S. Zhu, H. Ma, J. Liu, and N. Yang, "Gait recognition for co-existing multiple people using millimeter wave sensing," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 01, 2020, pp. 849–856.

[7] 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.

[8] M. Axelsson, M. Holmberg, S. Serra, H. Ovren, and M. Tulldahl, "Semantic labeling of lidar point clouds for uav applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 4314–4321.

[9] A. Chen, X. Wang, S. Zhu, Y. Li, J. Chen, and Q. Ye, "mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar," *arXiv preprint arXiv:2209.05070*, 2022.

[10] 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.

[11] 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 (CVPR)*, 2022, pp. 17 182–17 191.

[12] 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 (CVPR)*, 2022, pp. 12 186–12 195.

[13] 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.

[14] Y. Wang, T. Ye, L. Cao, W. Huang, F. Sun, F. He, and D. Tao, "Bridged transformer for vision and point cloud 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 114–12 123.

[15] 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.

[16] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[17] 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 (CVPR)*, 2019, pp. 10 975–10 985.

[18] 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.

[19] 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 (CVPR)*, 2021, pp. 1954–1963.

[20] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 939–12 948.

[21] 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.

[22] G. Pavlakos, J. Malik, and A. Kanazawa, "Human mesh recovery from multiple shots," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1485–1495.

[23] H. Zhang, J. Cao, G. Lu, W. Ouyang, and Z. Sun, "Learning 3d human shape and pose from dense body parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[24] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mID: Tracking and identifying people with millimeter wave radar," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2019, pp. 33–40.

[25] 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.

[26] Y. Cheng, H. Xu, and Y. Liu, "Robust small object detection on the water surface through fusion of camera and millimeter wave radar," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021, pp. 15 263–15 272.

[27] Y. Long, D. Morris, X. Liu, M. Castro, and P. Narayanan, "Radar-camera pixel depth association for depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 507–12 516.

[28] J. Zhang, M. Zhang, Z. Fang, Y. Wang, X. Zhao, and S. Pu, "Rvdet: Feature-level fusion of radar and camera for object detection," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2822–2828.

[29] 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 (WACV)*, 2021, pp. 1527–1536.

[30] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[31] Y. Kim, J. W. Choi, and D. Kum, "GRIF Net: Gated region of interest fusion network for robust 3d object detection from radar point cloud and monocular image," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 857–10 864.

[32] 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 (CVPR)*, 2021, pp. 11 794–11 803.

[33] 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 (CVPR)*, 2020, pp. 4404–4413.

[34] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[35] 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.

[36] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multi-modal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 682–11 692.

[37] A. Robotics, "4d image radar," 2021. [Online]. Available: https://arberobotics.com/wp-content/uploads/2021/05/4D-Imaging-radar-product-overview.pdf