milliFlow: Scene Flow Estimation on mmWave Radar Point Cloud for Human Motion Sensing

Fangqiang Ding[1], Zhen Luo[1], Peijun Zhao[2], and Chris Xiaoxuan Lu[3⋆]

[1]University of Edinburgh [2]MIT [3]UCL

Abstract. Human motion sensing plays a crucial role in smart systems for decision-making, user interaction, and personalized services. Extensive research that has been conducted is predominantly based on cameras, whose intrusive nature limits their use in smart home applications. To address this, mmWave radars have gained popularity due to their privacy-friendly features. In this work, we propose milliFlow, a novel deep learning approach to estimate scene flow as complementary motion information for mmWave point cloud, serving as an intermediate level of features and directly benefiting downstream human motion sensing tasks. Experimental results demonstrate the superior performance of our method when compared with the competing approaches. Furthermore, by incorporating scene flow information, we achieve remarkable improvements in human activity recognition and human parsing and support human body part tracking. Code and dataset are available at `https://github.com/Toytiny/milliFlow`.

Keywords: Scene Flow Estimation · Radar Point Cloud · mmWave Human Motion Sensing.

1 Introduction

Perceiving and understanding human behaviours play a pivotal role in human-centred applications such as disaster response [56, 70], surveillance [8, 61] and health monitoring [21, 26, 58]. Conventional methods rely on cameras [32, 68] or wearables [9, 47, 76], which are prone to visual deterioration (such as low lighting conditions, smoke, and fog) and raise privacy concerns, potentially compromising user experience with measurements that are psychologically intrusive. To address these concerns, researchers, on the other side, also propose to use wireless radio frequency (RF) signals bounced off the human body for human sensing [44,73,88–90] which is robust against poor lighting, privacy-preserving and non-intrusive to users. Among many RF techniques, single-chip millimetre wave (mmWave) radar emerges as a low-cost sensor that can provide more trustworthy point clouds of a scene under environment dynamics due to its MIMO transceiver design. For these reasons, there has been a significant increase in the production of single-chip radars [2] and wide deployment in real-world scenarios, ranging from smart buildings [5, 74] to vehicle cabins [16, 67], and to first-responder toolkits [3].

⋆ Corresponding author. Email: xiaoxuan.lu@ucl.ac.uk

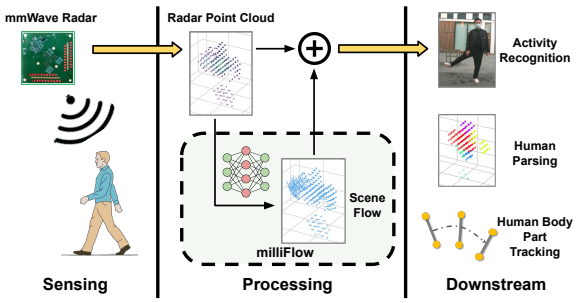

Fig. 1: We propose milliFlow, a scene flow estimation module to provide an additional layer of point-wise motion information on top of the original mmWave radar point cloud in the conventional mmWave-based human motion sensing pipeline.

As motion is naturally continuous and human bodies are non-rigid, point-wise velocity per radar frame is intuitively a strong cue for improving the motion estimation robustness and has been widely used in prior arts [46,69,79,86]. However, fine-grained velocity is difficult to obtain when it comes to mmWave radars. First, while some radars provide the Doppler velocity of a point, the velocity resolution is rather low [38] and thus cannot accurately capture the subtle human body movement, which is usually slower than 0.5m/s in domestic life. Moreover, mmWave radar only senses radial Doppler information but fails to capture the tangential one. For some radars designed for human sensing, e.g., the Vayyar radar [4] used in this work, the Doppler information is even absent in their sensor outputs. Second, extracting motion information from consecutive mmWave radar point clouds is also highly error-prone due to the limitations inherent in low-cost single-chip mmWave radar. These limitations include extremely sparse point clouds due to the target detector, e.g., constant false alarm rate (CFAR) algorithm [66] used on-chip and the presence of ghost points caused by multi-path effect [27]. More recently, it has been found that in a single radar frame, only a subset of body parts reflecting the signal towards the radar can be observed, while other parts deflecting the signal away from the radar are missing from the capture [13]. As a result, some human body parts detected in one frame may disappear in the next frame. For the above reasons, conventional radar point tracking methods [28, 65, 87] struggle to track human motion across frames or give fine-grained velocity in between.

In this work, we propose to estimate and use scene flow as intermediate features to better support radar-based human sensing. Scene flow refers to a set of displacement vectors between two consecutive point-cloud frames describing the motion field of a 3D scene. We hypothesise that scene flow, if estimated accurately, is able to drastically facilitate cross-frame movement analysis by directly exposing per-point motion features, thereby addressing the aforementioned challenges in previous research. Scene flow has long been proven very effective by the computer vision community in image-based human motion sensing applica-

tions [39, 53, 78]. However, estimating scene flow on mmWave radar point cloud is non-trivial because of the inherent sparsity and noise of radar point cloud. Directly applying conventional scene flow estimation methods designed for LiDAR or RGB-D cameras [17, 45, 83, 84] on radar point clouds has been found inadequate and generalises poorly. On the other side, recent mmWave radar scene flow estimation works [23, 24] generally focus on autonomous driving scenarios and rely on deep learning-based pipelines. However, these existing methods cannot be readily applied to our human sensing scenarios because the rigid-body assumption used for autonomous driving scenarios cannot stand in our cases where human subjects have non-rigid motion.

To cope with the above problems, we propose *milliFlow*, as exhibited in Fig. 1, a novel mmWave radar-based scene flow estimation approach for human motion sensing scenarios. Our contributions include:

- To the best of our knowledge, milliFlow is the first-of-its-kind work that aims to estimate mmWave radar-based scene flow for human motion sensing.
- We address the challenges, e.g., sparsity, and lack of temporal cues, for scene flow learning in our cases with a bespoken end-to-end learning network.
- We propose a cross-modal automatic scene flow labelling scheme specific for human motion sensing, avoiding labour-intensive manual labelling.
- We collect a large-scale human motion sensing dataset for evaluation, and perform a comprehensive evaluation of scene flow estimation accuracy as well as the performance on three downstream tasks.

2 Related Work

mmWave Radar-based Human Sensing. The feasibility and versatility of mmWave radar have been extensively demonstrated in various human sensing applications, including vital sign monitoring [7,55,80], signature verification [33,49], fall detection [40,75], human tracking and identification [15,20,31,91,92], gesture recognition [34,51,52,59], activity recognition [6,69,81] and pose estimation, reconstruction [46, 85, 86]. Compared with these applications, our work is unique in that we aim to estimate point-level scene flow vectors instead of providing a holistic output for the whole point cloud. In this way, we can either explicitly augment each radar target with scene flow vectors or implicitly learn robust latent features, which can further benefit many downstream tasks (*cf*. Sec. 4.4).
Scene Flow Estimation on Point Clouds. Recent scene flow works are mostly towards autonomous driving applications and attempt to estimate scene flow on LiDAR point clouds captured from autonomous vehicles. To this goal, different approaches are proposed, including classical methods [19,22,50,62] and deep learning-based ones [10, 54, 83, 84]. Recently, with the advances in point cloud feature learning, deep learning-based methods become more prevalent. According to their learning paradigm, prior works in this thread can be divided into fully-supervised [11, 35, 60, 63, 77, 82, 83], self-supervised [10, 45, 48, 57, 84] and weakly-supervised [25, 29] learning approaches. Apart from the aforementioned works, our work aims to estimate scene flow for human sensing using

mmWave radar. Moreover, our method does not demand any manual annotation efforts, instead, we automatically generate noisy pseudo scene flow labels from corresponding RGB-D images captured by the co-located camera.

Scene Flow Estimation with mmWave Radar. As far as we know, there are limited works that estimate scene flow using mmWave radar. A pioneering work [24] introduced a self-supervised learning method tailored for automotive radar, utilizing unique loss functions and a two-stage network. To improve scene flow performance and enable more downstream applications, a later work [23] exploits cross-modal supervision from co-located sensors (e.g. IMU, LiDAR) on modern autonomous vehicles for radar scene flow learning. However, these methods cannot be transferred to human sensing scenarios due to two reasons. First, the radar used for human-centric applications is different from the automotive radar in many aspects. For example, automotive radar has a lower range but higher angular resolution, making scene flow models bespoken for them hard to generalize to human sensing. More importantly, the human object is a non-rigid body, while in autonomous driving, scene dynamics are usually attributed to rigid body motion (e.g. cars and motorcycles) [22, 25, 29]. Such discrepancy in objective indicates different label and supervision generation schemes.

3 Methodology

3.1 Overview

We formally formulate our scene flow estimation problem in Sec. 3.2. Then we elaborate on the technical challenges of using mmWave radar for scene flow estimation under human motion sensing scenarios in Sec. 3.3. Sec. 3.4 details the design of our neural network, which comprises five sequential modules, tackling the sparsity and noise challenges and compensating for the lack of temporal cues. In Sec. 3.5, we propose a cross-modal automatic scene flow labelling scheme, to efficiently label scene flow for point clouds captured in human sensing scenarios. Sec. 3.6 further introduces our loss function used for training scene flow network.

3.2 Problem Formulation

Here, we consider the problem of scene flow estimation for dynamic 3D point clouds collected by an mmWave radar sensor used in human sensing scenarios. As a general problem setting, the input to point cloud-based scene flow estimation is two consecutive 3D point clouds $\mathcal{P} = \{p_i\}_{i=1}^{N}$ and $\mathcal{Q} = \{q_i\}_{i=1}^{M}$ captured by the same device and the output is a set of 3D vectors $\mathcal{F} = \{f_i\}_{i=1}^{N}$ that align each point p_i in $\mathcal{P}$ to its associated position $p_i^a = p_i + f_i$ in the frame of $\mathcal{Q}$. Different from the task that finds real correspondences between two frames, scene flow estimation only derives per-point 3D displacement for $\mathcal{P}$ and the associate position p_i^a does *not* essentially overlap with any points in $\mathcal{Q}$. Besides 3D coordinates information, each point may also have additional properties given mmWave radar point clouds as input, such as Doppler velocity or intensity value.

Without loss of generality, here we concatenate the per-point incident properties and 3D point coordinates into the 2D matrix and use $\mathcal{X} = \{x_i\}_{i=1}^{N}$, $\mathcal{Y} = \{y_i\}_{i=1}^{M}$ to denote the data from the source and target frame, respectively.

3.3 Technical Challenges

Sparsity and Noise. Due to bandwidth and hardware limitations, the radar raw data has low resolution in both range and angular dimensions from which only outstanding peaks are selected as valid targets. As a result, the point cloud generated by mmWave radar is quite sparse with an average of only around 100 points (in our case) related to the human body, often missing data for specific body parts. Furthermore, the multi-path effect introduces ghost points, adding noise to the already sparse data. This sparsity and noise significantly challenge mmWave radar's ability for scene flow estimation, as the lack of sufficient local geometric cues and the presence of noisy points make it difficult to extract robust local features, essential for accurately tracking movements within the scene.
Lack of Temporal Cues. The radar Doppler velocity measurement, representing the radial velocity of points, could enhance radar scene flow estimation. However, its resolution is limited by hardware and often too low to accurately capture small-scale movements typical in human sensing scenarios. Additionally, some radars, like the Vayyar radar [4] used here, do not measure Doppler velocity due to their specific signal transmitting design, further complicating accurate motion estimation. This limitation, coupled with the absence of consistent radar point data for certain body parts across frames, presents non-trivial challenges for reliable scene flow estimation as they result in the lack of temporal cues.
Scene Flow Annotation. Learning accurate scene flow estimation with deep networks requires point-level scene flow labels for training. However, manual annotation is prohibitively expensive and lacks real-world correspondence. Recent approaches [23, 37, 41] annotate object bounding boxes to generate pseudo scene flow labels, effective in autonomous driving with rigid bodies like cars. Nevertheless, this method is still labour-intensive and less applicable to human sensing due to the non-rigid nature of human movement, posing a challenge in creating accurate training labels for mmWave radar-based scene flow estimation.

3.4 Scene Flow Network

We employ end-to-end trainable deep neural networks for learning point-based scene flow estimation, in line with the state-of-the-art [23, 29, 45, 83]. The network architecture is sketched in Fig. 2. Particularly, we overcome the sparsity and noise challenges by integrating global features with local ones to provide a comprehensive view of each point cloud. Additionally, we address the lack of Doppler velocity and capture inconsistencies by leveraging a GRU network [18] to incorporate temporal information into the scene flow estimation. In subsequent sections, we will describe the details of each module in the network.
Local Feature Abstraction. Given mmWave radar point clouds $\mathcal{X}$ and $\mathcal{Y}$ as inputs, we encode their local features using four parallel SA layers [64] with

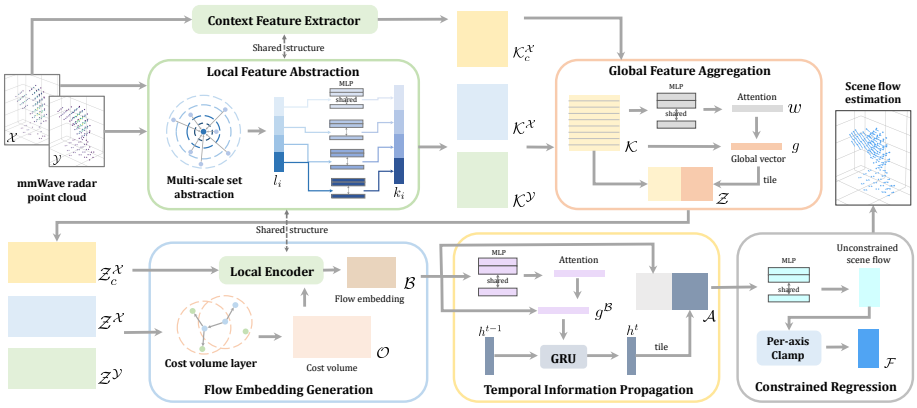

Fig. 2: mmWave-based scene flow network architecture. The network takes consecutive radar point clouds as the input and outputs the scene flow in between.

varying grouping radii, enabling multi-scale local feature extraction to account for radar point clouds' non-uniform density. Each SA layer generates a local feature $l_{i,s}$ for a given scale s, which is then transformed into a higher-level representation $k_{i,s}$ using a shared-weight MLP with scale-specific parameters. For both point clouds, we concatenate these high-level features across scales to form multi-scale local feature sets $\mathcal{K}^{\mathcal{X}}$ and $\mathcal{K}^{\mathcal{Y}}$ (*cf*. Fig. 2). Additionally, a separate context extractor, distinct from but structurally similar to the local encoder, generates context features $\mathcal{K}_c^{\mathcal{X}}$ for $\mathcal{X}$, enhancing the feature representation.

Global Feature Aggregation. In this module, the local feature vector of each radar point is first transformed into a scalar attention weight w_i using an MLP. These weights are normalized to sum to 1 for numerical stability. The global feature vector g is then derived by a weighted sum of all local features, offering an improvement over max-pooling by dynamically adjusting point-wise weights. This global vector is concatenated with each local feature to form local-global representations $\mathcal{Z}^{\mathcal{X}}, \mathcal{Z}^{\mathcal{Y}}$ for $\mathcal{X}$ and $\mathcal{Y}$. Additionally, a global feature for the context $\mathcal{L}_C^{\mathcal{X}}$ is obtained through another MLP, resulting in $\mathcal{Z}_C^{\mathcal{X}}$, as seen in Fig. 2.

Flow Embedding Generation. Given the local-global features of two radar point clouds, i.e., $\mathcal{Z}^{\mathcal{X}}, \mathcal{Z}^{\mathcal{Y}}$, we use the cost volume layer [84] to compute the correlation between them, as seen in Fig. 2. By aggregating the spatial relationship and feature similarities between two frames, the point motions are encoded into the cost volumes, denoted as $\mathcal{O} = \{o_i\}_{i=1}^{N}$. Thanks to the global feature aggregation, the holistic frame information can also be correlated, which yields more robust and stable costs. To further mix it with the context, we then stack the cost volumes $\mathcal{O}$, context features $\mathcal{L}_C^{\mathcal{X}}$ and pass the features into another local encoder to obtain the flow embedding $\mathcal{B} = \{b_i\}_{i=1}^{N}$.

Temporal Information Propagation. Directly propagating 2-dimension flow embedding from previous frames suffers from a) high computation overload, and

b) the change in the number of points. To overcome these issues, we propose to temporally update its global vector $g^{\mathcal{B}}$ instead of the flow embedding itself $\mathcal{B}$ (*cf*. Fig. 2). We apply a GRU network [18] to update it as hidden states temporally and obtain the final global representation $h^t = \text{GRU}(h^{t-1}, g^{\mathcal{B}}; \theta_g)$, where h^{t-1} is the global representation from the last frame and θ_g is the GRU network parameters. Lastly, we concatenate this updated global representation to each point in the flow embedding and denote the final features as $\mathcal{A} = \{a_i\}_{i=1}^{N}$.

Constrained Scene Flow Regression. Given the final features $\mathcal{A}$ produced above, we can use an MLP-based flow regressor to decode per-point scene flow $\mathcal{F} = \{f_i\}_{i=1}^{N}$. However, estimating unconstrained scene flow may lead to non-viable results, *e.g.*, the magnitude of the flow vector exceeds the normal scale of human body movement. To constrain our predictions, we propose to clamp the estimated scene flow before returning it. We set a fixed threshold ϵ and constrain the scene flow on each axis to be within the range of $[-\epsilon, \epsilon]$, as shown in Fig. 2.

3.5 Automatic Scene Flow Labelling

To address the human sensing scenarios where non-rigid motion dominates, we propose a cross-modal automatic scene flow labelling scheme (*cf*. Fig. 3), where pseudo scene flow labels are obtained from the 3D human skeletons. Our motivation is based on the observation that the non-rigid human body can be roughly segmented into multiple skeletons each of which can be seen as a rigid body, such as the neck and thigh bone. To simplify the problem of modelling human body motion, we attribute the human dynamics to the rigid motion of these skeletons and further assume the scene flow of points in the vicinity is induced by them.

Human Skeleton Annotation. In human sensing applications, the human skeleton is usually characterised by the 3D position of its two endpoints, aka. keypoints. To quickly and conveniently annotate such keypoints, we refer to the RGB-D images recorded by the co-located RGB-D camera in this work. Specifically, we first utilize an open-source pose estimation library (e.g. OpenPose [14]) to label 2D keypoints on RGB images and then uplift each 2D keypoint to 3D using its corresponding depth value. Then the human skeleton labels can be obtained using intrinsically connected 3D keypoints, as shown in Fig. 3.

Pseudo Label Generation. Given the 3D skeletons automatically annotated above, we can generate pseudo scene flow labels $\bar{\mathcal{F}} = \{\bar{f}_i \in \mathbb{R}^3\}_{i=1}^{N}$ for radar point clouds. For two consecutive point clouds $\mathcal{P}$ and $\mathcal{Q}$, we first compute the inter-frame transformation matrix for all human skeletons in the source frame. Then we assign each source radar point p_i to its closest skeleton and form the point-skeleton association as exhibited in Fig. 3. In the final, we derive the pseudo scene flow labels for each selected radar point as $\bar{f}_i = (T_j \circ p_i) - p_i$, where $T_j \in$ is the transformation matrix for the skeleton that point p_i is assigned to. $\circ$ is the action that operates homogeneous transformation for 3D points.

Keypoint-Based Label Filtering. To reduce noise from the inaccuracies of 2D keypoint estimation, we filter pseudo scene flow labels using confidence scores and 3D displacement of keypoints. Keypoints with confidence below 0.5 are first removed, and those with displacement over 0.5m between frames are further

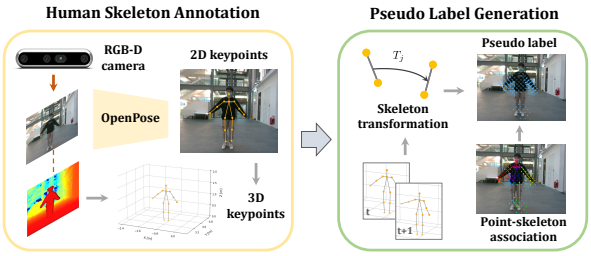

Fig. 3: Automatic scene flow labelling pipeline. With the help of the co-located RGB-D camera, we first label 3D human skeletons and then generate noisy pseudo scene flow labels with respect to the skeleton-based rigid-motion assumption.

filtered out. This process yields a set of valid keypoints per frame, and skeletons with both two end keypoints valid are considered valid. We then generate a mask $\mathcal{M}$, used below to minimise the impact of noise.

3.6 Loss Function

We use the pseudo scene flow labels $\bar{\mathcal{F}}$ to supervise our scene flow network. To mitigate noise, the valid mask $\mathcal{M}$ is used to filter out noisy labels during loss calculation. Initially, our network tended to converge to local minima, producing small and uniform scene flow vectors due to the predominance of small-scale movements in the data. To counteract this, we introduced a weighted loss function that emphasizes points with large-scale movements more significantly. Our adjusted loss is:

$$\mathcal{L} = \alpha_l \mathcal{L}_{large} + \alpha_s \mathcal{L}_{small} \tag{1}$$

where α_l and α_s are hyperparameters that balance large- (*i.e.*, larger than a threshold ζ) and small-scale movement impacts, with the L_2 distance measuring prediction errors. This approach aims to prevent the network from settling into local minima by reducing the influence of static or minor-moving points.

4 Experiment

4.1 Dataset Collection

To facilitate the evaluation of our approach, we collect a large-scale multi-modal human motion sensing dataset with annotation labels for various tasks.

Platform. As exhibited in Fig. 4 (a), we use a commercial Vayyar vTrigB imaging mmWave radar [4] and a RealSense D455 depth camera [1] to capture mmWave radar point clouds and RGB-D images respectively. Both sensors are fixed on a collection board mounted on a tripod to ensure their relative position is unchanged once calibrated. We alternately query the data frame from two sensors and store them in a PC, resulting in synchronized multi-modal data.

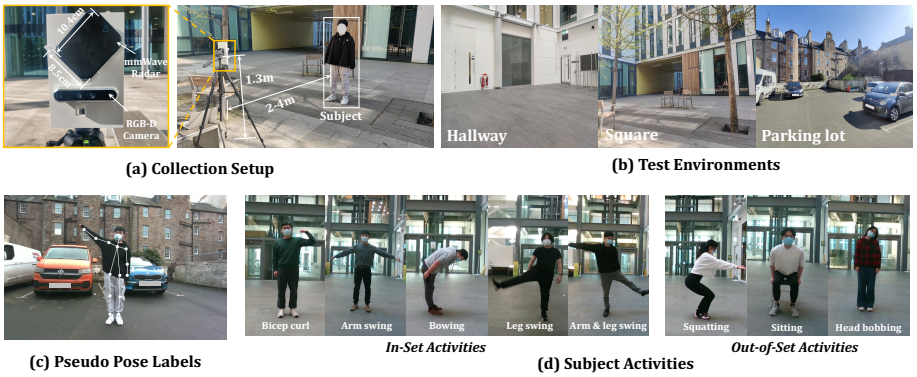

Fig. 4: Collection setup, test environment, subject activities and pseudo pose labels.

Note that the Vayyar radar we used is bespoken designed for fine-grained human sensing. It provides radar data with a higher resolution than most other mmWave radars, *e.g.*, TI radars [72], thus enabling us to estimate fine-grained scene flow for radar points.

Procedure. 12 participants of various genders, ages and heights are recruited for data collection in this work[1]. We choose three sites as our human sensing experiment scenes, as shown in Fig. 4 (b). Each participant is asked to perform 5 'in-set' activities for each scene, and 3 'out-of-set' activities in one scene for the evaluation of the generalization to unseen activities (*cf*. Fig. 4 (d)). Every subject is asked to wear a face mask to protect their identities from being recognised. The distance between subjects and sensors is 2-4m and randomly selected by the participant during collection.

Statistics. After data collection, we crop each sequence to 200 frames for uniform activity distribution. The whole 'in-set' dataset consists of $12 \times 3 \times 5 \times 200 =$ 36k frames in total. To assess the generalization ability to new subjects, we divide the dataset into three parts by subject following the ratio of train:val:test $=$ 3:1:2. The 'out-set' testing set is composed of $12 \times 3 \times 200 =$ 7.2k frames.

4.2 Evaluation Setup

Data Labelling. Our automatic labelling scheme (*cf*. Sec. 3.5) is used to annotate scene flow for the training and validation set by retaining 14 keypoints from OpenPose [14], leading to 13 connections between keypoints (*cf*. Fig. 4 (c)). Invalid skeletons are filtered out, and pseudo scene flow labels are assigned to points on valid skeletons. For the testing set, the same pipeline is used, but all points are annotated, with manual inspection and correction to 2D keypoints. For human activity recognition, activities are manually labelled during recording. Human parsing labels are derived from point-skeleton affiliations in scene

[1] The study has received the ethical approval from the University of Edinburgh, and participant consent forms were signed before the collection.

Table 1: Comparison of scene flow results between ours and state of the arts. ↑ means bigger values are better while ↓ means smaller values are better.

	EPE3D (m) ↓			Acc3D ↑	
Method	All	Moving	Static	Strict	Relax
FlowNet3D [54]	0.293	0.290	0.259	0.016	0.095
PPWC-Net [84]	0.171	0.181	0.128	0.138	0.179
Graph Prior [62]	0.315	0.322	0.283	0.007	0.011
FLOT [63]	0.299	0.307	0.265	0.015	0.094
FlowStep3D [45]	0.243	0.251	0.216	0.062	0.109
NSFP [50]	0.197	0.213	0.167	0.085	0.143
PV-RAFT [83]	0.161	0.170	0.107	0.179	0.292
RaFlow [24]	0.107	0.115	0.094	0.271	0.427
Bi-PFNet [17]	0.159	0.168	0.111	0.153	0.264
milliFlow (ours)	**0.046**	**0.051**	**0.009**	**0.406**	**0.703**

Table 2: Breakdown results of our scene flow network.

		EPE3D (m) ↓			Acc3D ↑	
	Method	All	Moving	Static	Strict	Relax
(a)	Full version	**0.046**	**0.051**	**0.009**	**0.406**	**0.703**
(b)	(a) w/o TP	0.053	0.062	0.018	0.382	0.676
(c)	(b) w/o GA	0.061	0.068	0.025	0.361	0.628
(d)	(c) w/o CF	0.071	0.077	0.028	0.315	0.536
(e)	(d) w/o CR	0.083	0.090	0.034	0.286	0.490

Table 3: Generalization of our model to new activities.

	EPE3D (m) ↓			Acc3D ↑	
Activity	All	Moving	Static	Strict	Relax
Sitting	0.034	0.038	0.004	0.498	0.771
Squatting	0.040	0.047	0.009	0.416	0.688
Head bobbing	0.027	0.031	0.006	0.664	0.859
Average	0.034	0.039	0.006	0.526	0.773

flow labelling (*cf*. Fig. 3), with the valid mask $\mathcal{M}$ applied to filter invalid points during training. The human body part tracking labels are directly obtained from the 3D keypoints identified in the scene flow labelling process.

Evaluation Metric. For scene flow evaluation, we use the EPE3D (m) and Acc3D metrics following [10, 24]. Specifically, we redefined the Acc3D strict and relax requirements by reducing 0.05/0.1m to 0.025/0.05m to adapt to our human-centric scenario. The overall accuracy (OA) (%) is reported for both the human activity recognition and human parsing tasks, while mIoU (%) is used for HP only. To evaluate the human body part tracking, we compute the mean joint localization error (mJE) (m).

4.3 Scene Flow Evaluation

Compared to the State of the Arts. We first compare our scene flow network with state-of-the-art methods of point-based scene flow estimation. Our baselines include 7 deep learning-based methods [17, 24, 45, 54, 63, 83, 84] whose networks are also supervised with pseudo scene flow labels, and two non-learning-based method Graph Prior [62] and NSFP [50] that solves scene flow via online optimization. The evaluation results on the 'in-set' testing set can be seen in Table 1. The scene flow network of milliFlow achieves a cm-level average EPE3D (*i.e*., 4.6cm) and a high relax Acc3D of 70.3%, ranking 1st on all metrics with a large margin compared to the baselines. This satisfactory performance proves the efficacy of our method to address the challenges existing in our scene flow task. Note that our network is trained with pseudo labels automatically generated using RGB-D images (*cf*. Sec. 3.5), which does not demand any manual annotation efforts.

Qualitative Results. Our output examples, illustrated in Fig. 5, effectively demonstrate our method's ability to produce reliable scene flow for diverse activities and subjects across three environments. Despite the inherent sparsity and noise in radar point clouds, our network adeptly learns representative fea-

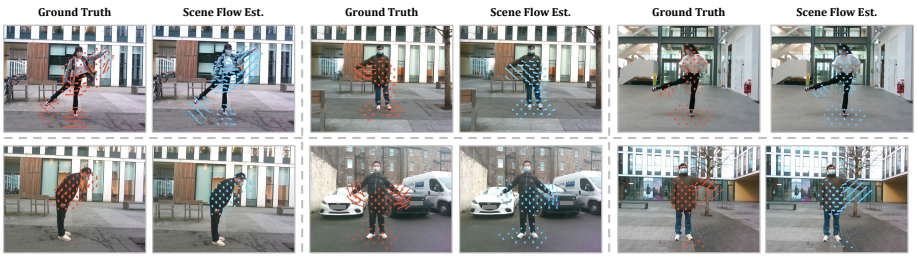

Fig. 5: Qualitative scene flow results. radar points and scene flow vectors are projected onto the image and the colour red is used for the ground truth while blue for ours.

tures through global feature aggregation and robust scene flow vector regression. Notably, even when some body parts are absent in the radar data, our approach compensates by utilizing historical information, thereby benefiting the current scene flow estimation and ensuring robust performance across successive frames.

Runtime Efficiency. We test the runtime efficiency of our scene flow network on a single NVIDIA RTX 3090 GPU. Given sequential testing radar point clouds, we feed them one by one into our trained model for inference. As a result, our model has real-time performance with one inference step in 74ms (∼13.5Hz). Moreover, the maximum allocated GPU memory is only 134 MB during inference. This minimal memory usage enables our network to operate in parallel with downstream networks, underscoring its practicality for real-time applications.

Ablation Study. To validate the effectiveness of key components within our scene flow network, we systematically disabled each, *i.e.*, temporal propagation (TP), global aggregation (GA), context feature (CF), and constrained regression (CR), and observed their effects on performance, as presented in Table 2. Overall, the full version of our network (row (a)) yields the best results and each component helps to elevate the performance on each metric (from row (e) to (a)). By utilizing information from previous frames, temporal propagation significantly improved scene flow estimation, achieving a 13.2% reduction in EPE3D and a 6.3% increase in strict Acc3D accuracy. Global aggregation, employing an attention mechanism, bolstered local features with global context, leading to a 13.1% improvement in EPE3D. The inclusion of context features into our flow embedding, aimed at retaining source frame context, provided a modest but surprising boost, lowering the overall EPE3D by 0.01m. Lastly, the biggest improvement (*i.e.*, a 0.012m decrease on EPE3D) is brought by our constrained regression, in which the scene flow component on each axis is clamped by a fixed threshold (*e.g.* 0.1m). This is reasonable as non-viable results, for example, a scene flow vector with a length of 0.5m, can seriously degrade our results.

Generalization to New Activities. In the above experiments, we evaluate our method on the testing set whose subjects are unseen during training. The results demonstrate the generalization ability of our trained model to new users that perform the same 'in-set' activities. To further test its generalization to new activities, here we evaluate our trained model on the 'out-of-set' testing

Table 4: Evaluation on the benefit of scene flow for the HAR task.

Method	Raw	w. S1	Gain	w. S2	Gain
Ours	47.32	57.88	+10.56	57.78	+10.46
MMPointGNN [30]	52.46	60.16	+7.70	59.94	+7.48
RadHAR [69]	44.65	49.98	+5.33	50.53	+5.88
Average	48.14	56.01	+7.87	56.08	+ 7.94

Table 5: Evaluation on the benefit of scene flow for the HP task.

Method	mIoU (%)	Gain (%)	oA (%)	Gain (%)
Raw	49.09	-	65.75	-
w. S1	52.72	+3.63	69.27	+3.52
w. S2	51.04	+1.95	68.21	+2.46

set in which three performed activities are not included in the training set. We can see from Tab. 3 that, our trained model can still keep an equally good performance when encountering unseen activities. This demonstrates the ability of our model to cope with new activities in human sensing. We also observed that the performance to new activities is better than the overall performance shown in Tab. 1. This is because our subjects either keep most of their bodies static (*i.e.*, heading bobbing) or stay completely static in many frames (*i.e.*, sitting and squatting) when doing unseen activities. As a result, estimating the scene flow of points belonging to them is much easier. We believe that our trained model can also generalize to other daily human activities.

4.4 Downstream Task Evaluation

As a low-level signal in understanding motions, scene flow can directly enhance low-quality radar point clouds with full per-point displacement information between two frames. Moreover, the latent spatial-temporal representations can be implicitly learned by guiding the network to estimate scene flow, which can be used to support other tasks. Therefore, we envision that learning scene flow estimation can benefit a wide range of higher-level downstream tasks in mmWave-based human sensing. To demonstrate the benefit, here we consider three representative downstream tasks, including human activity recognition (HAR), human parsing (HP) and human body part tracking (HBPT) for evaluation.

HAR Evaluation Setup. HAR plays a significant role in a wide range of applications, such as elderly healthcare monitoring [71], smart home [12] and behaviour surveillance [43]. To validate the functionality of scene flow to benefit HAR, we design a HAR base network by combining some key components from Sec. 3.4 and also utilizing the LSTM network [36] to track the temporal relationship before regressing the classification scores. Besides, two state-of-the-art methods [30, 69] bespoken for mmWave-based HAR are selected for a more persuasive comparison. Particularly, we design two strategies to harness the scene flow network as a plug-and-play module to the HAR network. Strategy 1 (S1) directly takes the estimated scene flow as point-level raw features and decorates each radar point with them, while Strategy 2 (S2) leverages the latent representations encoded by scene flow networks to enhance the low-quality radar data.

HAR Evaluation Results. We evaluate our proposed two strategies for scene flow application to HAR task on our 'in-set' testing set. The length of each input sequence T for HAR is set as 20. The evaluation results are shown in Table 4. As we can see, both two proposed strategies can enhance the performance of our base

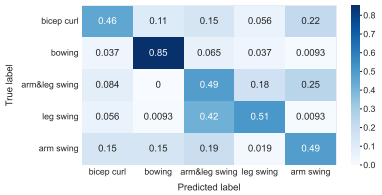

Fig. 6: Confusion matrix of HAR.

Table 6: Scene flow-based HBPT evaluation results.

	Tracking length - mJE (m) ↓			
Activity	1	2	3	4
Arm swing	0.028	0.076	0.097	0.124
Leg swing	0.016	0.071	0.105	0.130
Arm & leg swing	0.030	0.108	0.146	0.178
Average	0.025	0.085	0.116	0.144

network and state-of-the-art methods on HAR, demonstrating their usefulness in helping distinguish between different human activities. It can be also observed that the average accuracy shown in Tab. 4 is relatively low compared to those reported in [30, 69]. We credit this to our subject activities being very similar and inter-included to some extent. For example, the arm & leg swing activity can be separated into arm swing and leg swing, which are also inside our 'in-set' activities. As a result, both ours and the two states of the arts find it more difficult to correctly distinguish between these activities. The breakdown results (ours with S1) on HAR as the confusion matrix is shown in Fig. 6. As we can see, our HAR network achieves a relatively high accuracy value on the bowing activity since it has apparently different motion patterns from others. However, the accuracy values on other activities are not as satisfactory as on bowing. For example, 42% leg swing samples are wrongly predicted as the arm & leg swing activity. This further justifies our explanation that some activities are so similar and thus can easily confuse our HAR networks in classification.

HP Evaluation Human parsing aims to parse human semantic body segments (head, arms, torso, etc.) from sensor data. With radar point clouds as input, our objective of HP is to identify the body parts that correspond to each point. To evaluate the effectiveness of scene flow on this task, we also designed a base network for this downstream task using key components from Sec. 3.4. Both two scene flow application strategies proposed for the HAR above are also used here for the HP task. The number of body segments for parsing is 6, encompassing two arms, two legs, head and torso. As seen in Tab. 5, leveraging scene flow in two strategies also contributes to better performance on the HP task. Directly applying scene flow to points (S1) yields better results than using latent feature recycling (S2), as the former provides explicit per-point motion information. In contrast, latent features are less direct and require additional processing. Some examples of our HP results (with Strategy 1) are exhibited in Fig. 7. Thanks to the enhancement by the scene flow, our method can accurately parse radar point clouds into different body parts, close to the ground truth results.

HBPT Evaluation. Given the initial position of a human body part, our human body part tracking task aims to track its movement in subsequent frames. With point-level scene flow estimation, we first group N_t radar points that can be assigned to the skeleton needed to be tracked (*cf*. Fig. 3). Then we generate natural correspondences $\{p_i, p_i + f_i\}_{i=1}^{N_t}$ using their scene flow and apply the classical Kabsch algorithm [42] to solve the rigid skeleton transformation based

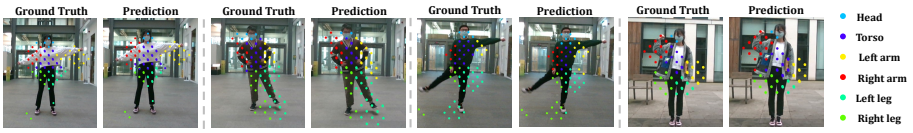

Fig. 7: Human parsing results visualization.

on them. This updates the endpoints' positions, allowing for tracking of the human body part across frames without additional training, relying only on a pre-trained scene flow network. For evaluation, we select the sequences of arm swing, leg swing and arm & leg swing activities in the testing set and take the two arms, legs and both of them respectively as our tracking targets.
For implementation, we divide the long sequences into short clips with a length of 5 frames and track the body parts within each clip. After initializing the ground truth positions in the first frame, we aim to track each skeleton for the next four frames. The HBPT results at different tracking lengths on three activities are reported in Fig. 6. With accurate scene flow estimation, we can effectively track multiple body parts together for different activities. For one-frame tracking, our method achieves an average mJE of $<$3cm on three activities, demonstrating the capability of scene flow to enable the HBPT task. However, the errors become larger as the tracking length increases, which indicates the happening of the tracking drift. This is inevitable for our method as more radar points associated with other skeletons will be wrongly induced for transformation calculation when the tracking continues. We also observe that the performance on the arm & leg swing activity is worse than the other two activities. This is reasonable as we need to track twice the skeletons in this activity as others, where more skeletons may interfere with each other during tracking.

5 Conclusion

In this paper, we introduce *milliFlow*, a deep learning framework designed to estimate scene flow for enhancing 3D mmWave radar point clouds in human motion sensing. Addressing mmWave's inherent instability and sparsity, *milliFlow* integrates multi-scale local and global features with temporal data. We also develop an automated labeling method to reduce the need for costly manual annotation. Extensive testing on our dataset and three downstream tasks shows *milliFlow*'s effectiveness, suggesting its potential as a valuable component in human motion sensing systems, significantly improving their performance.

Acknowledgements

This research is partially supported by the Engineering and Physical Sciences Research Council (EPSRC) under the Centre for Doctoral Training in Robotics and Autonomous Systems at the Edinburgh Centre of Robotics (EP/S023208/1).

References

1. Intel® realsense™ depth camera d455 (2023), https://www.intelrealsense.com/depth-camera-d455/
2. Iwr6843isk evaluation board | ti.com (2023), https://www.ti.com/tool/IWR6843ISK
3. rescueproject (2023), https://rescuerproject.eu/technology-tools/
4. Vayyar imaging - home - vayyar (2023), https://vayyar.com/
5. wholehome-ai-sensor (2023), https://consumer.huawei.com/cn/wholehome/ai-sensor/
6. Ahuja, K., Jiang, Y., Goel, M., Harrison, C.: Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–10 (2021)
7. Alizadeh, M., Shaker, G., De Almeida, J.C.M., Morita, P.P., Safavi-Naeini, S.: Remote monitoring of human vital signs using mm-wave FMCW radar. IEEE Access **7**, 54958–54968 (2019)
8. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. pp. 59–64 (2011)
9. Bannis, A., Pan, S., Ruiz, C., Shen, J., Noh, H.Y., Zhang, P.: IDIoT: Multimodal Framework for Ubiquitous Identification and Assignment of Human-carried Wearable Devices. ACM Transactions on Internet of Things **4**(2), 1–25 (2023)
10. Baur, S.A., Emmerichs, D.J., Moosmann, F., Pinggera, P., Ommer, B., Geiger, A.: SLIM: Self-Supervised LiDAR Scene Flow and Motion Segmentation. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 13126–13136 (2021)
11. Behl, A., Paschalidou, D., Donné, S., Geiger, A.: PointFlowNet: Learning Representations for Rigid Motion Estimation From Point Clouds. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 7954–7963 (2019)
12. Bianchi, V., Bassoli, M., Lombardo, G., Fornacciari, P., Mordonini, M., De Munari, I.: Iot wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment. IEEE Internet of Things Journal **6**(5), 8553–8562 (2019)
13. Cao, D., Liu, R., Li, H., Wang, S., Jiang, W., Lu, C.X.: Cross vision-rf gait re-identification with low-cost rgb-d cameras and mmwave radars. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **6**(3), 1–25 (2022)
14. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
15. Chen, W., Yang, H., Bi, X., Zheng, R., Zhang, F., Bao, P., Chang, Z., Ma, X., Zhang, D.: Environment-aware multi-person tracking in indoor environments with mmwave radars. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **7**(3) (sep 2023)
16. Chen, Y., Luo, Y., Qi, A., Miao, M., Qi, Y.: In-cabin monitoring based on millimeter wave fmcw radar. In: Proceedings of the International Symposium on Antennas, Propagation and EM Theory. pp. 01–03. IEEE (2021)

17. Cheng, W., Ko, J.H.: Bi-PointFlowNet: Bidirectional Learning for Point Cloud Based Scene Flow Estimation. In: Proceedings of the European Conference on Computer Vision. pp. 108–124 (2022)
18. Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In: Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. pp. 103–111 (2014)
19. Chodosh, N., Ramanan, D., Lucey, S.: Re-Evaluating LiDAR Scene Flow. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6005–6015 (2024)
20. Cui, H., Dahnoun, N.: High precision human detection and tracking using millimeter-wave radars. IEEE Aerospace and Electronic Systems Magazine **36**(1), 22–32 (2021)
21. Del Rosario, M.B., Redmond, S.J., Lovell, N.H.: Tracking the evolution of smartphone sensing for monitoring human movement. Sensors **15**(8), 18901–18933 (2015)
22. Dewan, A., Caselitz, T., Tipaldi, G.D., Burgard, W.: Rigid scene flow for 3D LiDAR scans. In: Proceddings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1765–1770 (2016)
23. Ding, F., Palffy, A., Gavrila, D.M., Lu, C.X.: Hidden Gems: 4D Radar Scene Flow Learning Using Cross-Modal Supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–10 (2023)
24. Ding, F., Pan, Z., Deng, Y., Deng, J., Lu, C.X.: Self-Supervised Scene Flow Estimation With 4-D Automotive Radar. IEEE Robotics and Automation Letters pp. 1–8 (2022)
25. Dong, G., Zhang, Y., Li, H., Sun, X., Xiong, Z.: Exploiting Rigidity Constraints for LiDAR Scene Flow Estimation. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 12776–12785 (2022)
26. Fagert, J., Mirshekari, M., Pan, S., Zhang, P., Noh, H.Y.: Gait health monitoring through footstep-induced floor vibrations. In: Proceedings of the 18th International Conference on Information Processing in Sensor Networks. pp. 319–320 (2019)
27. Gennarelli, G., Soldovieri, F.: Multipath ghosts in radar imaging: Physical insight and mitigation strategies. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **8**(3), 1078–1086 (2014)
28. Godrich, H., Chiriac, V.M., Haimovich, A.M., Blum, R.S.: Target tracking in mimo radar systems: Techniques and performance analysis. In: 2010 IEEE Radar Conference. pp. 1111–1116 (2010)
29. Gojcic, Z., Litany, O., Wieser, A., Guibas, L.J., Birdal, T.: Weakly Supervised Learning of Rigid 3D Scene Flow. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 5692–5703 (2021)
30. Gong, P., Wang, C., Zhang, L.: MMPoint-GNN: Graph Neural Network with Dynamic Edges for Human Activity Recognition through a Millimeter-Wave Radar. In: Proceedings of the International Joint Conference on Neural Networks. pp. 1–7 (2021)
31. Gu, T., Fang, Z., Yang, Z., Hu, P., Mohapatra, P.: Mmsense: Multi-person detection and identification via mmwave sensing. In: Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems. pp. 45–50 (2019)
32. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7297–7306 (2018)

33. Han, M., Yang, H., Ni, T., Duan, D., Ruan, M., Chen, Y., Zhang, J., Xu, W.: mmsign: mmwave-based few-shot online handwritten signature verification. ACM Transactions on Sensor Networks (2023)
34. Hazra, S., Santra, A.: Robust gesture recognition using millimetric-wave radar system. IEEE Sensors Letters **2**(4), 1–4 (2018)
35. Hermes, N., Bigalke, A., Heinrich, M.P.: Point cloud-based scene flow estimation on realistically deformable objects: A benchmark of deep learning-based methods. Journal of Visual Communication and Image Representation **95**, 103893 (2023)
36. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
37. Huang, Shengyu and Gojcic, Zan and Huang, Jiahui and Wieser, Andreas and Schindler, Konrad: Dynamic 3D Scene Analysis by Point Cloud Accumulation. In: Proceedings of the European Conference on Computer Vision. pp. 674–690 (2022)
38. Iovescu, C., Rao, S.: The fundamentals of millimeter wave sensors. Texas Instruments pp. 1–8 (2017)
39. Jaimez, M., Souiai, M., Stückler, J., Gonzalez-Jimenez, J., Cremers, D.: Motion cooperation: Smooth piece-wise rigid scene flow from rgb-d images. In: Proceedings of the International Conference on 3D Vision. pp. 64–72. IEEE (2015)
40. Jin, F., Sengupta, A., Cao, S.: mmfall: Fall detection using 4-d mmwave radar and a hybrid variational rnn autoencoder. IEEE Transactions on Automation Science and Engineering **19**(2), 1245–1257 (2020)
41. Jund, P., Sweeney, C., Abdo, N., Chen, Z., Shlens, J.: Scalable scene flow from point clouds in the real world. IEEE Robotics and Automation Letters **7**(2), 1589–1596 (2021)
42. Kabsch, W.: A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A **32**(5), 922–923 (1976)
43. Khurana, R., Kushwaha, A.K.S.: Deep learning approaches for human activity recognition in video surveillance-a survey. In: Proceedings of the International Conference on Secure Cyber Computing and Communication. pp. 542–544. IEEE (2018)
44. Kianoush, S., Savazzi, S., Vicentini, F., Rampa, V., Giussani, M.: Device-free rf human body fall detection and localization in industrial workplaces. IEEE Internet of Things Journal **4**(2), 351–362 (2016)
45. Kittenplon, Y., Eldar, Y.C., Raviv, D.: FlowStep3D: Model Unrolling for Self-Supervised Scene Flow Estimation. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 4114–4123 (2021)
46. Kong, H., Xu, X., Yu, J., Chen, Q., Ma, C., Chen, Y., Chen, Y.C., Kong, L.: m3track: mmwave-based multi-user 3d posture tracking. In: Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services. pp. 491–503 (2022)
47. Lemmens, R.J., Janssen-Potten, Y.J., Timmermans, A.A., Smeets, R.J., Seelen, H.A.: Recognizing complex upper extremity activities using body worn sensors. PloS one **10**(3), e0118642 (2015)
48. Li, R., Zhang, C., Lin, G., Wang, Z., Shen, C.: RigidFlow: Self-Supervised Scene Flow Learning on Point Clouds by Local Rigidity Prior. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 16959–16968 (2022)
49. Li, W., He, T., Jing, N., Wang, L.: mmhsv: In-air handwritten signature verification via millimeter-wave radar. ACM Transactions on Internet of Things (2023)
50. Li, X., Kaesemodel Pontes, J., Lucey, S.: Neural scene flow prior. Advances in Neural Information Processing System **34**, 7838–7851 (2021)

51. Liu, H., Wang, Y., Zhou, A., He, H., Wang, W., Wang, K., Pan, P., Lu, Y., Liu, L., Ma, H.: Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **4**(4) (dec 2020), https://doi.org/10.1145/3432235
52. Liu, H., Zhou, A., Dong, Z., Sun, Y., Zhang, J., Liu, L., Ma, H., Liu, J., Yang, N.: M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar. IEEE Internet of Things Journal **9**(5), 3397–3415 (2021)
53. Liu, P., Reale, M., Yin, L.: 3d head pose estimation based on scene flow and generic head model. In: Proceedings of the IEEE International Conference on Multimedia and Expo. pp. 794–799. IEEE (2012)
54. Liu, X., Qi, C.R., Guibas, L.J.: FlowNet3D: Learning Scene Flow in 3D Point Clouds. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 529–537 (2019)
55. Lv, W., He, W., Lin, X., Miao, J.: Non-contact monitoring of human vital signs using fmcw millimeter wave radar in the 120 ghz band. Sensors **21**(8), 2732 (2021)
56. Mishra, B., Garg, D., Narang, P., Mishra, V.: Drone-surveillance for search and rescue in natural disaster. Computer Communications **156**, 1–10 (2020)
57. Mittal, H., Okorn, B., Held, D.: Just go with the flow: Self-supervised scene flow estimation. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 11177–11185 (2020)
58. Mukhopadhyay, S.C.: Wearable sensors for human activity monitoring: A review. IEEE Sensors Journal **15**(3), 1321–1330 (2014)
59. Palipana, S., Salami, D., Leiva, L.A., Sigg, S.: Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(1) (mar 2021). https://doi.org/10.1145/3448110
60. Pan, Z., Ding, F., Zhong, H., Lu, C.X.: Ratrack: Moving object detection and tracking with 4d radar point cloud. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2024)
61. Paul, M., Haque, S.M., Chakraborty, S.: Human detection in surveillance videos and its applications-a review. EURASIP Journal on Advances in Signal Processing **2013**(1), 1–16 (2013)
62. Pontes, J.K., Hays, J., Lucey, S.: Scene flow from point clouds with or without learning. In: Proceedings of the International Conference on 3D Vision. pp. 261–270 (2020)
63. Puy, G., Boulch, A., Marlet, R.: Flot: Scene flow on point clouds guided by optimal transport. In: Proceedings of the European Conference on Computer Vision. pp. 527–544 (2020)
64. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems **30** (2017)
65. Rohal, P., Ochodnicky, J.: Radar target tracking by kalman and particle filter. In: 2017 Communication and Information Technologies (KIT). pp. 1–4 (2017)
66. Scharf, L.L., Demeure, C.: Statistical signal processing: detection, estimation, and time series analysis. Prentice Hall (1991)
67. Schwarz, C., Zainab, H., Dasgupta, S., Kahl, J.: Heartbeat measurement with millimeter wave radar in the driving environment. In: Proceedings of the IEEE Radar Conference. pp. 1–6. IEEE (2021)
68. Shu, X., Zhang, L., Sun, Y., Tang, J.: Host–parasite: Graph lstm-in-lstm for group activity recognition. IEEE Transactions on Neural Networks and Learning Systems **32**(2), 663–674 (2020)

69. Singh, A.D., Sandha, S.S., Garcia, L., Srivastava, M.: Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In: Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems. pp. 51–56 (2019)
70. Tariq, R., Rahim, M., Aslam, N., Bawany, N., Faseeha, U.: Dronaid: A smart human detection drone for rescue. In: Proceedings of the International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT. pp. 33–37 (2018)
71. Taylor, W., Shah, S.A., Dashtipour, K., Zahid, A., Abbasi, Q.H., Imran, M.A.: An intelligent non-invasive real-time human activity recognition system for next-generation healthcare. Sensors **20**(9), 2653 (2020)
72. Texas Instruments: mmWave Radar Sensors - Overview. `https://www.ti.com/sensors/mmwave-radar/overview.html` (2024), accessed: 2024-02-22
73. Tian, Y., Lee, G.H., He, H., Hsu, C.Y., Katabi, D.: Rf-based fall monitoring using convolutional neural networks. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(3), 1–24 (2018)
74. Tseng, S.P., Li, B.R., Pan, J.L., Lin, C.J.: An application of internet of things with motion sensing on smart house. In: Proceedings of the International Conference on Orange Technologies. pp. 65–68. IEEE (2014)
75. Wang, B., Guo, L., Zhang, H., Guo, Y.X.: A millimetre-wave radar-based fall detection method using line kernel convolutional neural network. IEEE Sensors Journal **20**(22), 13364–13370 (2020)
76. Wang, C., Liu, J., Chen, Y., Xie, L., Liu, H.B., Lu, S.: Rf-kinect: A wearable rfid-based approach towards 3d body movement tracking. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(1), 1–28 (2018)
77. Wang, H., Pang, J., Lodhi, M.A., Tian, Y., Tian, D.: FESTA: Flow Estimation via Spatial-Temporal Attention for Scene Point Clouds. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 14173–14182 (2021)
78. Wang, P., Li, W., Gao, Z., Zhang, Y., Tang, C., Ogunbona, P.: Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 595–604 (2017)
79. Wang, S., Cao, D., Liu, R., Jiang, W., Yao, T., Lu, C.X.: Human parsing with joint learning for dynamic mmwave radar point cloud. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **7**(1), 1–22 (2023)
80. Wang, Y., Wang, W., Zhou, M., Ren, A., Tian, Z.: Remote monitoring of human vital signs based on 77-GHz mm-wave FMCW radar. Sensors **20**(10), 2999 (2020)
81. Wang, Y., Liu, H., Cui, K., Zhou, A., Li, W., Ma, H.: m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8298–8302 (2021)
82. Wang, Z., Li, S., Howard-Jenkins, H., Prisacariu, V., Chen, M.: Flownet3d++: Geometric losses for deep scene flow estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 91–98 (2020)
83. Wei, Y., Wang, Z., Rao, Y., Lu, J., Zhou, J.: PV-RAFT: point-voxel correlation fields for scene flow estimation of point clouds. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 6954–6963 (2021)
84. Wu, W., Wang, Z.Y., Li, Z., Liu, W., Fuxin, L.: PointPWC-Net: Cost Volume on Point Clouds for (Self-) Supervised Scene Flow Estimation. In: Proceedings of the European Conference on Computer Vision. pp. 88–107 (2020)

85. Xue, H., Cao, Q., Ju, Y., Hu, H., Wang, H., Zhang, A., Su, L.: M4esh: mmwave-based 3d human mesh construction for multiple subjects. In: Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems. pp. 391–406 (2022)
86. Xue, H., Ju, Y., Miao, C., Wang, Y., Wang, S., Zhang, A., Su, L.: 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. pp. 269–282 (2021)
87. Yan, J., Jiao, H., Pu, W., Shi, C., Dai, J., Liu, H.: Radar sensor network resource allocation for fused target tracking: a brief review. Information Fusion **86**, 104–115 (2022)
88. Zhang, J., Wei, B., Hu, W., Kanhere, S.S.: Wifi-id: Human identification using wifi signal. In: Proceedings of the International Conference on Distributed Computing in Sensor Systems. pp. 75–82 (2016)
89. Zhang, J., Wu, F., Hu, W., Zhang, Q., Xu, W., Cheng, J.: WiEnhance: Towards data augmentation in human activity recognition using WiFi signal. In: Proceedings of the 15th International Conference on Mobile Ad-Hoc and Sensor Networks. pp. 309–314 (2019)
90. Zhang, J., Wu, F., Wei, B., Zhang, Q., Huang, H., Shah, S.W., Cheng, J.: Data augmentation and dense-LSTM for human activity recognition using WiFi signal. IEEE Internet of Things Journal **8**(6), 4628–4641 (2020)
91. Zhao, P., Lu, C.X., Wang, J., Chen, C., Wang, W., Trigoni, N., Markham, A.: mid: Tracking and identifying people with millimeter wave radar. In: Proceedings of the International Conference on Distributed Computing in Sensor Systems. pp. 33–40 (2019)
92. Zhao, P., Lu, C.X., Wang, J., Chen, C., Wang, W., Trigoni, N., Markham, A.: Human tracking and identification through a millimeter wave radar. Ad Hoc Networks **116**, 102475 (2021)