DAP: Doppler-aware Point Network for Heterogeneous mmWave Action Recognition

Jiaying Lin[1], Shiman Wu[2], Jinfu Liu[3], Can Wang[4,5], and Mengyuan Liu[1†]

[1]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, China
[2]Huazhong University of Science and Technology, Wuhan, China
[3]DJI Technology Co., Ltd., Shenzhen, China
[4]China Energy Engineering Group Co., Ltd., Beijing, China
[5]Kiel University, Kiel, Germany

Abstract. Millimeter-wave (mmWave) radar provides privacy-preserving sensing and is valuable for human action recognition (HAR). Existing mmWave point cloud datasets are limited in scale and mostly collected under homogeneous single-source settings, preventing current methods from handling real-world distribution shifts caused by heterogeneous radar sources, such as different devices and frequency bands. To address this, we introduce UniMM-HAR, the largest and first mmWave point cloud HAR dataset for heterogeneous multi-source scenarios, standardizing three distinct radar configurations to realistically evaluate cross-source generalization. We further propose the Doppler-aware Point Cloud Network (DAP-Net) to tackle heterogeneity challenges. DAP-Net enhances intra-modal representations and performs cross-modal alignment to learn source-invariant action semantics. Leveraging action-consistent spatio-temporal Doppler patterns as anchors, the Dual-space Doppler Reparameterization (D^{2}R) module performs sample-adaptive geometric densification and Doppler-guided feature recalibration, while the Text Alignment Module (TAM) provides stable semantic anchors via a pretrained textual space. Experiments show that DAP-Net significantly outperforms existing methods under heterogeneous radar settings, achieving state-of-the-art accuracy and strong cross-source robustness. Code and dataset are publicly available at https://github.com/jolin830/DAP-Net.

Keywords: Radar Point Cloud · Heterogeneous mmWave Action Recognition · mmWave Doppler

1 Introduction

Human action recognition (HAR) is fundamental to video understanding and intelligent homes [19, 25, 35, 36]. Most existing approaches rely on camera-based

† Corresponding author: liumengyuan@pku.edu.cn.

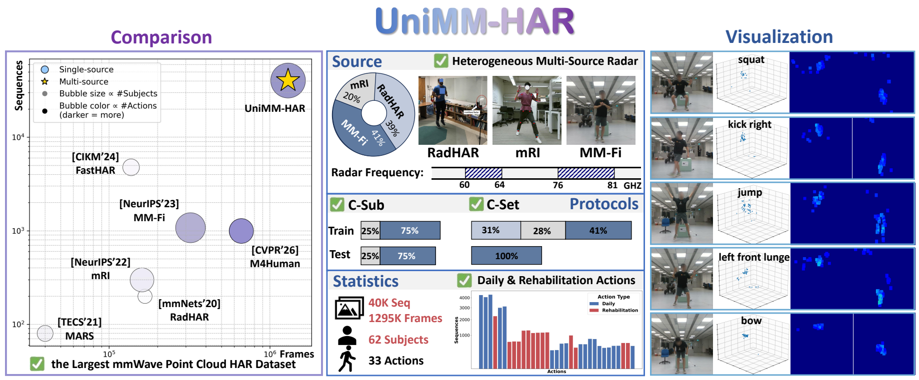

Fig. 1: UniMM-HAR is currently the largest mmWave point cloud human action recognition dataset and the first unified benchmark for heterogeneous multi-source distributions, providing Cross-Subject and Cross-Set evaluation protocols and covering both daily and rehabilitation actions. The samples in UniMM-HAR are collected from radar devices with different models and operating frequencies.

modalities, including RGB [6, 16], and skeleton [9, 33, 71]. These methods inevitably expose identifiable visual information and remain highly sensitive to illumination variations and occlusion, which constrains their deployment in privacy-sensitive environments [66, 67]. In contrast, millimeter-wave (mmWave) radar captures human motion through electromagnetic reflections rather than appearance cues, thereby inherently preserving privacy while offering superior robustness to low-light conditions and visual obstructions, making it more suitable for continuous and unobtrusive monitoring [32, 43, 51, 60].

In recent years, mmWave-based Human action recognition has advanced rapidly, largely driven by public radar datasets, with point cloud datasets being particularly valuable due to their explicit 3D structure and rich geometric semantics [1, 54–56, 63, 72]. However, existing mmWave point cloud HAR datasets are mostly limited in scale and collected under homogeneous single-source settings, where source refers to radar configurations such as device model and operating frequency band. Accordingly, current methods are primarily designed and validated under this simplified assumption. In contrast, real-world mmWave sensing scenarios are typically heterogeneous and multi-source, as different applications may deploy diverse radar devices, and practical settings often involve device upgrades and multi-device collaboration. Therefore, a practically deployable HAR system must be capable of adapting to distribution shifts arising from different radar sources. The core challenge of heterogeneous multi-source data lies in the fact that the same human action, when captured by different radar sources, can exhibit significant distribution shifts in terms of point density, spatial distribution, velocity statistics, and noise characteristics. These shifts are not merely additive noise but structural variations tightly coupled with device-specific properties. Models trained on single-source data tend to overfit source-specific statistical patterns rather than learning stable, source-agnostic action semantics, leading to substantial performance degradation under distribution changes and hindering deployment in multi-source environments [37,58,80]. Therefore, it is necessary to construct a large-scale mmWave point cloud dataset that covers heterogeneous multi-source distributions and to explore effective modeling strategies tailored to such heterogeneous settings, in order to enhance cross-source generalization capability.

Motivated by this, we introduce UniMM-HAR, a large-scale heterogeneous radar point cloud HAR dataset, along with the Doppler-aware Point Cloud Network (DAP-Net). To our knowledge, UniMM-HAR is the largest mmWave point cloud HAR dataset and the first unified benchmark from heterogeneous multi-source data as Fig.1. It standardizes data from three heterogeneous radar configurations and provides cross-subject (C-Sub) and cross-set (C-Set) protocols to evaluate generalization under inherent inter-source distribution shifts. Compared with single-source benchmarks, UniMM-HAR more rigorously tests whether models learn source-agnostic and semantically meaningful representations. To address heterogeneous distribution shifts, DAP-Net performs intra-modal enhancement and cross-modal alignment. It leverages the action-consistent spatio-temporal Doppler patterns and introduces the Dual-space Doppler Reparameterization (D^2R) module to decouple source-specific interference in geometric and feature spaces. Furthermore, a lightweight Text Alignment Module (TAM) aligns mmWave features with a pretrained textual semantic space, serving as a stable semantic anchor to improve cross-source robustness.

DAP-Net is built upon the D^2R module, which consists of Doppler-guided Geometry Reparameterization (DGR) and Motion-aware Feature Recalibration (MFR) for geometric and feature-level modeling, respectively. The reparameterized point clouds are processed by a point cloud backbone, and the extracted features are further aligned via TAM. In the geometric space, given the sparsity and noise of mmWave point clouds, DGR performs Doppler-guided, sample-adaptive densification over motion-relevant regions while suppressing device-induced noise and outliers, instead of applying a fixed processing strategy. In the feature space, MFR dynamically recalibrates feature channels conditioned on Doppler cues, enhancing motion-related responses and mitigating source-specific bias. This sample-wise reparameterization enables more consistent cross-source representations and is implemented as a parameter-efficient, plug-and-play module compatible with various backbones. To further enhance semantic consistency, TAM projects mmWave features into a pretrained textual semantic space constructed from action prompts and aligns them via similarity-based learning, serving as a stable semantic anchor for cross-source robustness.

In summary, this work has the following contributions:

- We construct and release UniMM-HAR, the largest mmWave point cloud HAR dataset to date and the first unified benchmark designed for heterogeneous multi-source distributions. It more realistically simulates multi-source deployment scenarios in real-world applications and provides a foundation for cross-source generalization research.

Table 1: Comparison of UniMM-HAR with existing mmWave point cloud HAR datasets. For fair comparison, only the single-person subset of mmMulti [81] is evaluated. RF-BW: radar frequency band and bandwidth (GHz). Multi-Source: whether the dataset is collected from heterogeneous radar source. †: non-public data. *: originally designed for other tasks. Source datasets of UniMM-HAR are in light gray.

Dataset	Statistics				Radar			Actions		Protocols	
	# Act.	# Sub.	# Seq.	# Frame	Device	RF-BW	Multi-Source	Daily	Rehab	C-Sub	C-Set
RadHAR [56]	5	2	200	167k	TI IWR1443	76–81	✗	✓	✗	✗	✗
*MARS [2]	10	4	80	40k	TI IWR1443	76–81	✗	✓	✗	✗	✗
m-Activity† [63]	5	9	N/A	15k	TI IWR1443	76–81	✗	✓	✗	✗	✗
mRI [1]	12	20	300	160k	TI IWR1443	76–81	✗	✓	✓	✓	✗
miliPoint [12]	49	11	N/A	545k	TI IWR1843	76–81	✗	✓	✗	✗	✗
MM-Fi [72]	27	40	1080	320k	TI IWR6843	60–64	✗	✓	✓	✓	✓
*mmParse† [62]	10	32	N/A	N/A	TI IWR6843	60–64	✗	✓	✗	✗	✗
FastHAR [55]	7	5	4776	137k	TI IWR6843	60–64	✗	✓	✗	✗	✗
*milliflow [15]	7	12	N/A	43k	Vayyar vTrigB	62–69	✗	✓	✗	✓	✗
mmMulti† [81]	6	4	2520	N/A	TI IWR1443	76–81	✗	✓	✗	✓	✓
*M4Human [18]	50	20	999	661k	Vayyar vTrigB	62–69	✗	✓	✓	✓	✗
UniMM-HAR	33	**62**	**40494**	**1295k**	TI IWR1443 & TI IWR6843	76–81 & 60–64	✓	✓	✓	✓	✓

- We propose DAP-Net for cross-source adaptation of heterogeneous mmWave point clouds via intra-modal and cross-modal enhancement. The Dual-space Doppler Reparameterization (D²R) module performs sample-adaptive geometric densification and feature recalibration to mitigate source-specific biases, while the Text Alignment Module (TAM) provides stable semantic anchors to improve semantic consistency.
- DAP-Net achieves state-of-the-art performance on UniMM-HAR, demonstrating its effectiveness and cross-source robustness. Notably, the D²R module is a parameter-efficient, plug-and-play design that can seamlessly integrate with any point cloud backbone, enhancing mmWave modeling.

2 Related Work

2.1 Human Action Recognition on Point Clouds

Point clouds from LiDAR or depth cameras have led to mature point cloud networks, typically categorized into static and dynamic paradigms. Static point clouds excel in tasks such as object classification and HAR [14,28,39,40,47,48,68], while dynamic point clouds model temporal motion [3,17,31,34]. Although dense point clouds provide fine-grained motion cues, they incur high hardware and computational costs . In contrast, mmWave radar offers a low-cost, privacy-preserving alternative. However, its sparse and noisy nature prevents direct adoption of dense point cloud architectures. DAP-Net addresses this limitation by incorporating Doppler-guided motion priors to compensate for geometric sparsity, enabling dense point cloud backbones to effectively process radar point clouds.

2.2 mmWave-based Human Action Recognition

Millimeter-wave action recognition has witnessed substantial progress with diverse data representations, predominantly spectrum-based [5, 20, 53, 57, 61, 64, 65, 70, 74] and point cloud–based approaches [13, 21, 22, 29, 47, 56, 59, 73, 75]. Spectrum-based methods represent radar signals as 2D spectrum maps and apply image recognition models to capture Doppler cues, but the 2D projection inevitably causes information loss. Point cloud-based methods preserve 3D structure but often overlook Doppler information, a distinctive characteristic of mmWave sensing. Existing studies mainly use Doppler for preprocessing, such as frame selection [23], rather than explicit motion representation learning. In contrast, DAP-Net leverages Doppler to guide feature learning in both geometric and feature spaces.

2.3 Existing mmWave Radar Datasets

The advancement of mmWave-based human understanding has been largely facilitated by public radar datasets [2, 8, 10, 18, 30, 42, 52, 69], with action recognition gaining increasing attention [1, 12, 15, 26, 55, 56, 63, 72, 77–79, 81]. Point cloud datasets are increasingly favored for their intuitive 3D structure and rich geometric semantics. Early mmWave point cloud HAR datasets [2, 56, 63, 78] were limited in scale and relied on random splits. Later datasets [1,12,55,72,81] expanded subjects and actions and introduced cross-subject (C-Sub) and cross-dataset (C-Set) evaluation protocols. Existing datasets remain single-source and limited in scale, diversity, and standardization. UniMM-HAR addresses these gaps by integrating heterogeneous multi-source radar data, significantly increasing samples, subjects, and action diversity, and providing multiple standardized evaluation protocols under heterogeneous distributions.

3 UniMM-HAR Dataset

3.1 Dataset Overview

Existing mmWave point cloud datasets suffer from limitations in scale, subject diversity, and action coverage. They are typically collected using a single radar device at a fixed frequency band and lack standardized evaluation protocols. Inspired by [41], we adopt a large-scale unified paradigm that integrates heterogeneous mmWave point cloud sources under standardized evaluation protocols, resulting in the proposed UniMM-HAR. Table 1 provides a systematic comparison with existing datasets. Specifically, UniMM-HAR consolidates three representative datasets: RadHAR [56], mRI [1], and MM-Fi [72], and standardizes their action categories, subject splits, and sample representations to form a unified benchmark. The heterogeneity in UniMM-HAR mainly arises from differences in radar devices, operating frequency bands, and other recording configs across the sub-datasets. These differences lead to varying data distributions even for the same action, providing a realistic and challenging setting for cross-domain

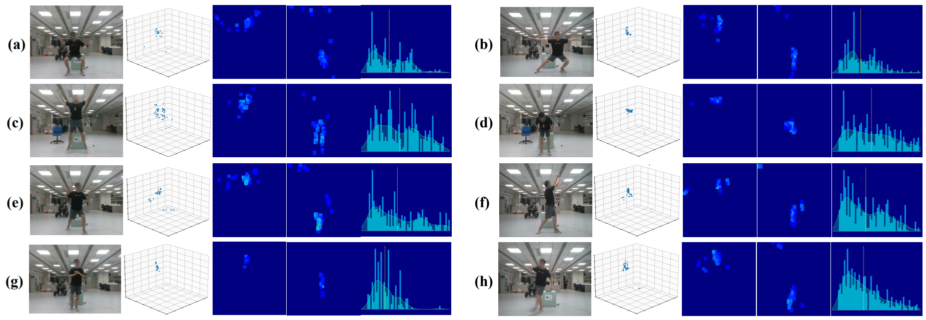

Fig. 2: Visualization of different action samples in UniMM-HAR. Actions (a)–(h) are: *squat, left lunge, jump, bow, left front lunge, throw left, stretch, and kick right.* In each subfigure, the 1st and 2nd columns show the RGB video frame and mmWave point cloud, respectively. The 3rd and 4th columns show the Doppler heatmap in BEV and front view. The 5th column shows the Doppler distribution for the action.

evaluation. UniMM-HAR contains 33 action categories spanning daily and rehabilitation activities, 62 subjects, and over 1.29 million frames. Figure 2 visualizes representative action samples. To the best of our knowledge, UniMM-HAR is the largest mmWave point cloud dataset for HAR to date and the first to systematically support cross-subject and cross-set evaluation under heterogeneous radar hardware and frequency bands.

3.2 Dataset Statistics and Unification Process

UniMM-HAR excludes datasets that are not publicly available, incompletely annotated, or lacking raw point cloud data, and ultimately retains three high-quality datasets that satisfy the unification criteria: RadHAR [56], mRI [1], and MM-Fi [72]. The selected datasets are then systematically unified. Table 2 summarizes the statistics of the source datasets and UniMM-HAR before and after processing. The procedure is detailed as follows.

Action Alignment. After consolidating synonymous actions across the three datasets, UniMM-HAR establishes a unified taxonomy of 33 categories, including 21 daily activities and 12 rehabilitation actions as in Fig.3. The action distribution follows a natural long-tailed pattern, preserving realistic class imbalance while increasing the difficulty of the benchmark.

Dataset-Aware Preprocessing. Given substantial differences among datasets in sequence length, annotation granularity, and structural organization, dataset-aware preprocessing strategies are applied to generate standardized clips. RadHAR is processed using sliding windows (window size 60, stride 10), consistent with its official implementation. mRI adopts sliding windows (window size 32, stride 16). Since MM-Fi sequences are not pre-segmented by action, they are divided into valid action clips according to the provided segmentation files.

Format Standardization. To improve interoperability and reproducibility, UniMM-HAR is released in two formats: CSV and NPZ. The CSV format pre-

Table 2: Dataset statistics before and after processing. The processing strategy (Proc.): Sliding Window (SW), Segmentation (SEG), and Normalization (Norm).

Dataset	Proc.	Format	#Act.	#Sub.	#Seq.	#Frame(k)
	–	TXT	5	2	200	167
RadHAR [56]	SW	CSV	5	2	15,715	942
	SW	NPZ	5	2	15,715	502
	–	CSV	12	20	300	160
mRI [1]	SW	CSV	12	20	8,332	131
	SW	NPZ	12	20	8,332	266
	–	BIN	27	40	1,080	320
MM-Fi [72]	SEG	CSV	27	40	16,447	310
	SEG	NPZ	27	40	16,447	526
UniMM-HAR	–	**CSV**	**33**	**62**	**40,494**	**1,384**
	Norm	**NPZ**	**33**	**62**	**40,494**	**1,295**

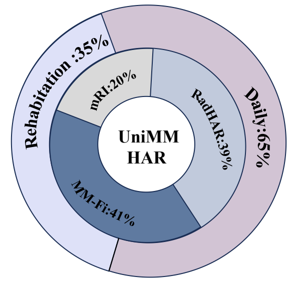

Fig. 3: Action distribution across sources and types.

serves raw annotations for data provenance and traceability, while NPZ is optimized for direct model input. All samples follow a unified naming convention encoding dataset source, action label, subject identity, and scene.

Representation Standardization. Temporal–point normalization converts each clip to a fixed shape $[T, P, C] = [32, 64, 5]$, with channels x, y, z, Doppler, and intensity. When the original sequence length exceeds T, uniform temporal downsampling is applied. Otherwise, zero-padding is used. When the number of points per frame exceeds P, Farthest Point Sampling (FPS) is applied. Otherwise, repeat sampling is performed. Furthermore, UniMM-HAR offers dataset-level and clip-level normalization variants. These normalize x, y, z, Doppler, and intensity using global or per-sample statistics to reduce severe statistical bias, while retaining multi-source distribution gaps as an essential challenge of the benchmark.

3.3 Evaluation Protocols

UniMM-HAR supports three multi-source evaluation protocols: Random Split, Cross-Subject (C-Sub), and Cross-Set (C-Set). 1) Random Split partitions samples 60:40 to simulate seen-subject and seen-action scenarios. 2) Cross-Subject uses mRI and MM-Fi, comprising 19,657 samples across 25 actions and 60 subjects, with a 5:5 split by subjects. 3) Cross-Set covers 28,041 sequences across 27 actions in 6 scenes, following a 4:2 split by scene for training and testing. Importantly, all evaluation protocols are conducted under the heterogeneous radar configuration, which introduces additional domain shifts across sources. As a result, these settings are more challenging than the conventional C-Sub and C-Set protocols defined under single-source radar setups.

4 Methodology

4.1 Task Definition

We consider a mmWave point cloud sequence represented as $X \in \mathbb{R}^{T\times P\times 3} = \{\mathcal{P}_t\}_{t=1}^{T}$, where T denotes the frame number, P denotes the point number per

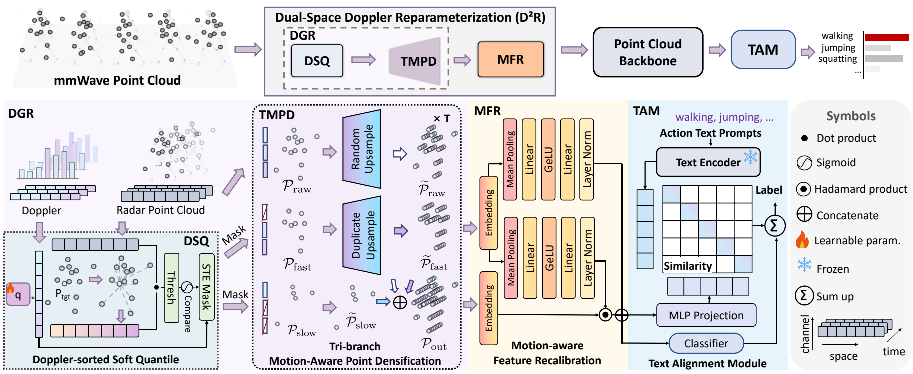

Fig. 4: Overview of DAP-Net: First, the Dual-space Doppler Reparameterization (D²R) converts mmWave point cloud sequences into Doppler-guided dense representations. Within D²R, Doppler-guided Geometry Reparameterization (DGR) performs geometric densification, while Motion-aware Feature Recalibration (MFR) enhances motion-sensitive feature modeling to produce point embeddings. These embeddings are first processed by a point cloud backbone to extract global features and then aligned with textual features through the Text Alignment Module (TAM), with final predictions made based on feature similarity.

frame, and each point is represented by its 3D coordinates (x, y, z). The Doppler velocity of each point is represented by $v \in \mathbb{R}^{T\times P}$. The point number in frame t is denoted by $P_t = |\mathcal{P}_t|$. The objective is to predict the action label y associated with the input sequence. An overview of our DAP-Net is shown in Fig. 4.

4.2 Dual-Space Doppler Reparameterization (D²R)

To effectively address distribution shifts from heterogeneous multi-source data, it is essential to disentangle source-specific representations from action-related ones and extract motion priors that remain stable across different radar sources. Geometric density and absolute coordinates are highly dependent on hardware configurations, whereas Doppler encodes the radial velocity of each point, directly reflecting the underlying physical motion [7,24,38,46]. Although absolute Doppler values may vary across devices, the relative temporal evolution exhibits consistent patterns across sources. For example, arm swinging produces periodic velocity oscillations, whereas squatting shows a regular sign alternation of velocity across the action cycle. Therefore, action-consistent spatio-temporal Doppler patterns can serve as reliable modeling anchors.

Based on this observation, D²R leverages Doppler information as a motion prior and performs sample-adaptive modeling via two complementary modules: Doppler-guided Geometric Reparameterization (DGR), which densifies motion-critical points in the geometric space, and Motion-aware Feature Recalibration (MFR), which conditionally recalibrates dense point cloud features in the feature space to enhance sensitivity to dynamic patterns. Notably, the D²R module is parameter-efficient and plug-and-play, enabling flexible integration with various point cloud backbone networks while enhancing mmWave point cloud modeling with negligible additional computational cost.

4.3 Doppler-guided Geometric Reparameterization (DGR)

Millimeter-wave point clouds are inherently sparse and noisy, with distribution discrepancies further amplified under heterogeneous multi-source settings, rendering fixed-rule densification strategies designed for single-source scenarios difficult to generalize: 1) Repetitive upsampling indiscriminately duplicates both informative motion points and noise, amplifying cross-source bias. 2) Sliding-window aggregation relies on manually defined temporal spans and remains sensitive to variations in frame rate and point density. To better address densification and denoising under heterogeneous distributions, DAP-Net introduces a Doppler-guided Geometry Reparameterization (DGR) module, which performs selective densification of each frame to a fixed target size P_{goal} using per-point Doppler as a motion prior. By adaptively emphasizing discriminative motion regions and suppressing noise, DGR enables heterogeneous point clouds to form aligned spatial structures, thereby improving cross-source generalization.

Doppler-sorted Soft Quantile (DSQ) Unlike methods relying on fixed absolute thresholds, DSQ introduces a learnable quantile parameter $q \in (0, 1)$ to characterize the relative motion intensity boundary of each frame, enabling adaptive partitioning based on the relative distribution. For each frame $\mathcal{P}_t$, let the Doppler magnitudes be $\{v_t^i\}_{i=1}^{P_t}$, sorted in ascending order as $v_t^1 \leq \cdots \leq v_t^{P_t}$.

The target rank position for frame t is defined as $P_t^{\text{tgt}} = q(P_t - 1)$, representing the desired quantile index. To obtain a differentiable approximation of this rank, Gaussian weights are constructed centered at P_t^{tgt}:

$$w_t^i \propto \exp\left(-\frac{\left(i - P_t^{\text{tgt}}\right)^2}{2\sigma^2}\right). \tag{1}$$

The frame-level motion threshold is computed as $\tau_t = \sum_{i=1}^{P_t} w_t^i \, v_t^i$, which can be interpreted as a soft dot product between the sorted Doppler values and the Gaussian weight distribution. Given the threshold τ_t, each Doppler value v_t^i is compared against it to obtain a soft motion indicator:

$$s_t^i = \sigma\left(\frac{v_t^i - \tau_t}{\gamma}\right), \tag{2}$$

where γ controls the sharpness of the transition. The resulting scores are then binarized using a Straight-Through Estimator (STE) [4, 11], maintaining explicit partitioning in the forward pass while propagating gradients through the soft probabilities s_t^i in backpropagation. By modeling relative statistics via quantiles, DSQ achieves stable motion-aware partitioning under heterogeneous multi-source distributions, effectively mitigating source-induced statistical shifts and providing consistent motion priors for subsequent geometric reparameterization.

Tri-branch Motion-Aware Point Densification (TMPD) Given the motion confidence s_i for each point, TMPD partitions the point set into three complementary subsets: $\mathcal{P}_{\text{fast}} = \{\mathbf{p}_i \mid s_i > \delta\}, \quad \mathcal{P}_{\text{slow}} = \{\mathbf{p}_i \mid s_i \leq \delta\}, \quad \mathcal{P}_{\text{raw}} = \mathcal{P}$, where δ is the partition threshold and $\mathcal{P}$ denotes the original point set. The points are assigned to Fast, Slow, and Raw branches, which are processed as:

$$\widetilde{\mathcal{P}}_{\text{fast}} = \bigcup_{k=1}^{r} \mathcal{P}_{\text{fast}}, \quad \widetilde{\mathcal{P}}_{\text{slow}} = \mathcal{P}_{\text{slow}}, \quad \widetilde{\mathcal{P}}_{\text{raw}} = \text{RandSample}\left(\mathcal{P}_{\text{raw}}, P_{\text{raw}}\right), \tag{3}$$

where r is the duplication factor for motion-salient points in the Fast branch, and $\text{RandSample}(\cdot, N)$ uniformly samples N points for the Raw branch. The **Fast branch** focuses on motion-salient regions. Since mmWave point clouds are inherently sparse in high-motion areas and exhibit distribution discrepancies across heterogeneous sources, r-fold controlled duplication increases the sampling density of dynamic structures across sources, improving cross-source representation consistency. Generative synthesis or interpolation is avoided to prevent physically inconsistent noise due to the lack of point-wise supervision and stable inter-frame correspondences. The **Slow branch** handles quasi-static points, which may belong to low-motion human regions or background structures. Their geometry is relatively stable, so the original distribution is preserved without densification to avoid cross-source geometric shifts affecting feature consistency. For the **Raw branch**, the number of points is $\widetilde{P}_{\text{raw}} = P_{\text{goal}} - rP_{\text{fast}} - P_{\text{slow}}$, ensuring the final point cloud has cardinality P_{goal}. The final densified point cloud $\mathcal{P}_{\text{out}}$, with cardinality $|\mathcal{P}_{\text{out}}| = P_{\text{goal}}$, is obtained by aggregating the three branches:

$$\mathcal{P}_{\text{out}} = \widetilde{\mathcal{P}}_{\text{fast}} \cup \widetilde{\mathcal{P}}_{\text{slow}} \cup \widetilde{\mathcal{P}}_{\text{raw}}, \tag{4}$$

This branch-wise adaptive strategy enhances key motion information while mitigating sparsity and noise discrepancies from heterogeneous sources.

4.4 Motion-aware Feature Recalibration (MFR)

Previous works primarily focus on modeling the geometric 3D structure of point clouds, paying limited attention to the motion information embedded in Doppler signals. This results in extracted features with constrained ability to capture dynamic motion patterns. To explicitly enhance motion representation, DAP-Net introduces the Motion-aware Feature Recalibration (MFR) module, which conditionally recalibrates dense point cloud features using Doppler-guided fast points, thereby maintaining consistent motion information across heterogeneous radar sources. The dense point clouds $\mathcal{P}_{\text{out}} \in \mathbb{R}^{T \times P_{\text{out}} \times 3}$ and the fast points $\mathcal{P}_{\text{fast}} \in \mathbb{R}^{T \times P_{\text{fast}} \times 3}$ are embedded to obtain $\mathcal{F}_{\text{out}} \in \mathbb{R}^{T \cdot P_{\text{out}} \times C}$ and $\mathcal{F}_{\text{fast}} \in \mathbb{R}^{T \cdot P_{\text{fast}} \times C}$, respectively. The fast point features $\mathcal{F}_{\text{fast}}$ are then aggregated into a motion summary vector $c \in \mathbb{R}^{C}$ to guide subsequent feature recalibration:

$$c = \frac{1}{|\mathcal{F}_{\text{fast}}|} \sum_{i \in \mathcal{F}_{\text{fast}}} f_i, \tag{5}$$

where f_i denotes the embedding of the i-th fast point. This vector provides a stable motion prior across heterogeneous radar sources. Inspired by [44], the MFR module leverages c to perform channel-wise recalibration of $\mathcal{F}_{\text{out}}$. Specifically, MLPs generate learnable scaling γ and shifting β parameters from c:

$$\gamma = \text{MLP}_\gamma(c), \quad \beta = \text{MLP}_\beta(c), \quad \gamma, \beta \in \mathbb{R}^C. \tag{6}$$

Recalibrated dense features $\mathcal{F}$ are derived via channel-wise affine transformation:

$$\mathcal{F} = \gamma \odot \mathcal{F}_{\text{out}} + \beta, \tag{7}$$

where $\odot$ denotes element-wise multiplication across channels. This operation leverages the motion-guided fast points to modulate point cloud features, effectively enhancing the network's sensitivity to motion patterns while maintaining consistent motion representation across heterogeneous multi-source scenarios.

4.5 Backbone and Text Alignment Module

DAP-Net is backbone-agnostic and can be flexibly integrated into various point cloud backbone networks. Given the input point cloud features $\mathcal{F} \in \mathbb{R}^{T \cdot P_{\text{out}} \times C}$, the backbone network extracts high-level representations, which are aggregated via max pooling to obtain a global feature: $\tilde{\mathcal{F}} = \text{MaxPool}(\text{Backbone}(\mathcal{F})) \in \mathbb{R}^D$. This global feature $\tilde{\mathcal{F}}$ is fed into a classifier to obtain action prediction logits $z \in \mathbb{R}^K$, where K denotes the number of action classes.

mmWave point clouds are sparse and have lower point resolution. To leverage this, DAP-Net introduces the **Text Alignment Module (TAM)**, aligning mmWave features with information-rich textual embeddings. TAM first encodes prompts $T(\cdot)$ for all action categories using a pretrained text encoder $E_{\text{text}}(\cdot)$:

$$F_{\text{text}} = E_{\text{text}}\big([\text{T}(\text{text}_i)]_{i=1}^K\big) \in \mathbb{R}^{K \times C_{\text{text}}}, \tag{8}$$

where C_{text} denotes the dimensionality of the text embeddings. The global mmWave feature $\tilde{\mathcal{F}}$ is projected via an MLP into the shared textual embedding space, yielding the feature $F_{\text{mmw}} \in \mathbb{R}^{K \times C_{\text{text}}}$. Then, TAM computes the similarity between mmWave and text features:

$$S = \text{Softmax}\big(F_{\text{mmw}} F_{\text{text}}^\top\big) \in \mathbb{R}^{K \times K}. \tag{9}$$

Finally, the similarity scores $s = diag(S) \in \mathbb{R}^K$ are fused with the backbone classifier logits $z \in \mathbb{R}^K$ to produce the final prediction $\hat{\mathbf{y}} \in \mathbb{R}^K$. Note that TAM is independent of the specific text encoder and can be seamlessly replaced with more advanced large language models (LLMs) or domain-adapted text encoders.

5 Experiment

5.1 Implementation Details

Experiments are conducted on two NVIDIA RTX 4090 GPUs. DAP-Net is trained for 150 epochs (batch size 128, learning rate 0.01, weight decay $1 \times$

Table 3: Comparison with prior methods on UniMM-HAR C-Sub and C-Set. Original PC type indicates the point cloud type the model was designed for. MPC: mmWave radar point clouds; SPC: static point clouds; TPC: temporal point clouds.

Model	Source	Original PC Type	UniMM-HAR (%) C-Sub	UniMM-HAR (%) C-Set
PointNet [47]	CVPR'17	SPC	59.27	70.96
DGCNN [45]	NN'18	SPC	71.70	52.18
RadHAR [56]	mmNSS'19	MPC	42.49	48.47
PointMLP [40]	ICLR'22	SPC	71.06	78.13
PST-Transformer [17]	TPAMI'22	TPC	63.13	79.70
PointCLIP [76]	CVPR'22	SPC	67.82	69.82
Clip2point [27]	ICCV'23	SPC	64.17	59.56
FastHAR [54]	CIKM'24	MPC	53.72	61.16
3DInAction [3]	CVPR'24	TPC	73.43	57.81
UST-SSM [31]	ICCV'25	TPC	71.50	48.40
DAP-Net	–	MPC	**80.72**	**81.82**

10^{-4}). We set $P_{\text{goal}} = 1024$ for DGR and a 5-fold duplication for TMPD's Fast Branch. Using PointMLP [40] as the backbone, TAM aligns features with a frozen CLIP [49] text encoder via the template *"a mmWave point cloud of a person [CLS]"*. Optimization uses standard cross-entropy loss.

5.2 Compare with SOTA

We evaluate representative methods on the UniMM-HAR C-Sub and C-Set in Table 3, including mmWave radar point cloud methods and LiDAR-based dense point cloud methods (e.g., static point clouds and temporal point clouds). Existing dense point cloud methods fail on mmWave point clouds due to sparsity and noise, and homogeneous mmWave methods perform poorly under heterogeneous settings. In contrast, DAP-Net consistently outperforms all methods, demonstrating its effectiveness, heterogeneous cross-source generalizability.

5.3 Ablation Study

We perform ablation studies on UniMM-HAR C-Sub to systematically investigate the contributions of each component in DAP-Net.

Evaluation of Backbone Generalization As shown in Table 4, integrating the modules of DAP-Net leads to substantial accuracy gains across all backbones. This improvement arises from D^2R's intra-modal enhancement and TAM's cross-modal alignment, which increase the backbone's adaptability to mmWave point cloud characteristics and its robustness to heterogeneous distributions, thereby unlocking the full potential of dense point cloud backbones on mmWave data.

Table 4: Impact of DAP-Net modules across backbones.

Backbone	Module	Acc (%)
PointMLP [40]	–	71.06
	+D^{2}R	$80.00^{\uparrow 8.94}$
	+D^{2}R+TAM	$80.72^{\uparrow 9.66}$
UST-SSM [31]	–	71.50
	+D^{2}R	$74.00^{\uparrow 2.50}$
	+D^{2}R+TAM	$74.94^{\uparrow 3.44}$
PST-Transformer [17]	–	63.13
	+D^{2}R	$76.73^{\uparrow 13.60}$
	+D^{2}R+TAM	$78.21^{\uparrow 15.08}$

Table 5: Impact of D^{2}R modules (DGR and MFR).

Backbone	D^{2}R		Acc (%)
	DGR	MFR	
PointMLP [40]	✗	✗	71.06
	✓	✗	$76.97^{\uparrow 5.91}$
	✗	✓	$78.76^{\uparrow 7.70}$
	✓	✓	$\mathbf{80.00}^{\uparrow 8.94}$
UST-SSM [31]	✗	✗	71.50
	✓	✗	$73.69^{\uparrow 2.19}$
	✓	✓	$\mathbf{74.00}^{\uparrow 2.50}$
PST-Transformer [17]	✗	✗	63.13
	✓	✗	$72.31^{\uparrow 9.18}$
	✓	✓	$\mathbf{76.73}^{\uparrow 13.60}$

Table 6: Impact of different point split strategies and fast branch densification in TMPD.

Backbone	Points Split	Fast Branch Densification	Acc (%)
PointMLP [40]	–	–	71.06
	0.2 quantile	MLP densification	$71.00^{\downarrow 0.06}$
	0.2 quantile	r-fold duplication	$76.46^{\uparrow 5.40}$
	DSQ	r-fold duplication	$76.97^{\uparrow 5.91}$
UST-SSM [31]	–	–	71.50
	0.2 quantile	MLP densification	$71.40^{\downarrow 0.10}$
	0.2 quantile	r-fold duplication	$73.40^{\uparrow 1.90}$
	DSQ	r-fold duplication	$74.00^{\uparrow 2.50}$

Table 7: Impact of different point split quantiles.

Quantile	Split Type	Acc (%)
–	–	71.06
0.2	Fixed	$76.46^{\uparrow 5.40}$
0.3	Fixed	$76.55^{\uparrow 5.40}$
0.5	Fixed	$76.14^{\uparrow 5.08}$
0.8	Fixed	$76.12^{\uparrow 5.06}$
DSQ	Learnable	$76.97^{\uparrow 5.91}$

Evaluation of D^{2}R In Table 5, we ablate D^{2}R on three representative backbones. Both DGR and MFR individually improve performance, and their combination achieves the best results, demonstrating that D^{2}R is backbone-agnostic and its modules provide complementary benefits.

Evaluation of DSQ and TMPD We conduct ablation experiments to evaluate the individual and combined effects of DSQ and TMPD under different point splitting and densification strategies. As shown in Table 6, the proposed combination of DSQ and r-fold duplication of TMPD consistently achieves the best performance across different backbones, demonstrating the effectiveness of jointly modeling motion-aware point selection and densification.

Evaluation of DSQ We compare the proposed learnable soft quantile in DSQ with fixed hard quantiles (0.2, 0.5, 0.8) and a variant without quantile-based splitting. As shown in Table 7, the adaptive soft quantile consistently achieves superior performance, demonstrating the benefit of learning a data-dependent threshold for motion-aware point selection.

Evaluation of DGR In Table 8, we compare DGR with several common point cloud densification strategies under a fixed number of output points. DGR achieves the best performance, substantially outperforming naive duplication

Table 8: Impact of point cloud densification.

Densification	Acc (%)
Repeat Sampling	71.06
Super-frame Fusion	$59.72^{\downarrow 11.34}$
MLP	$43.54^{\downarrow 27.52}$
DGR	$76.97^{\uparrow 5.91}$

Table 9: Impact of different text prompts and encoders.

Method	Action Text Prompt	Text Encoder	Acc (%)
Baseline	–	–	80.00
w/ TAM	[CLS]	CLIP [49]	$80.71^{\uparrow 0.71}$
w/ TAM	A person performing [CLS]	CLIP [49]	$80.65^{\uparrow 0.65}$
w/ TAM	a mmWave point cloud of a person [CLS]	SBERT [50]	$80.67^{\uparrow 0.67}$
w/ TAM	a mmWave point cloud of a person [CLS]	CLIP [49]	$80.72^{\uparrow 0.72}$

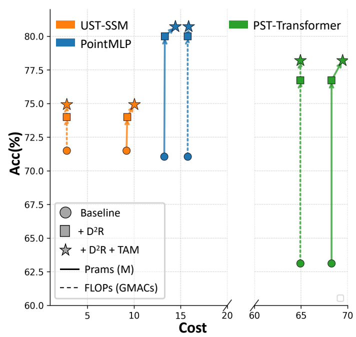

Fig. 5: Accuracy–Cost trade-off across different backbones.

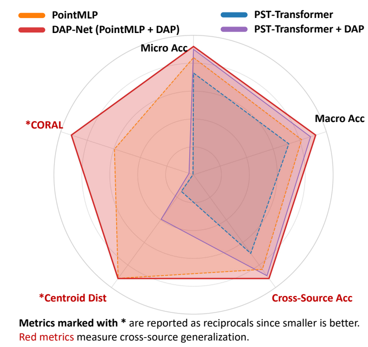

Fig. 6: Accuracy and cross-source generalization comparison.

and MLP-based generation, demonstrating that Doppler-guided geometric reparameterization is more effective for enhancing sparse mmWave representations.

Evaluation of TAM We evaluate TAM with various prompt templates and text encoders in Table 9. All combinations consistently yield performance gains, indicating that TAM consistently benefits from diverse text encoders and prompts, and can easily scale to more advanced text models such as LLMs.

Evaluation of Efficiency We analyze the modules of DAP-Net (D^2R and TAM) in terms of accuracy–cost trade-offs across different baselines in Fig. 5. Both modules improve performance while incurring negligible parameter and FLOPs overhead, with D^2R contributing particularly significant gains.

Evaluation of Cross-Source Performance Fig. 6 shows performance on heterogeneous multi-source data, where *Cross-Source Acc* measures action consistency across sources, *Centroid Distance* quantifies shifts between source feature centroids, and *CORAL* evaluates covariance differences across sources (see Supplementary for metric details). DAP-Net achieves the highest overall accuracy while excelling in cross-source metrics. Compared with the baseline, adding DAP-Net modules significantly improves heterogeneous cross-source metrics.

6 Conclusion

We construct the UniMM-HAR dataset and propose DAP-Net for heterogeneous multi-source mmWave point clouds. UniMM-HAR is the first and largest mmWave point cloud HAR benchmark designed for heterogeneous multi-source scenarios, standardizing three distinct radar configurations and providing C-Sub and C-Set evaluation protocols to rigorously assess cross-source generalization. To mitigate source-specific biases, DAP-Net leverages Doppler as a motion prior, performing sample-adaptive geometric densification and feature recalibration via D^2R, and aligns features with a pretrained textual space through TAM to provide stable semantic anchors. Extensive experiments demonstrate that DAP-Net significantly improves both recognition accuracy and cross-source robustness.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 62473007), Guangdong Outstanding Youth Fund (No. 2026B1515020015), Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

References

1. An, S., Li, Y., Ogras, U.: mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. Adv. Neural Inform. Process. Syst. **35**, 27414–27426 (2022)
2. An, S., Ogras, U.Y.: Mars: mmwave-based assistive rehabilitation system for smart healthcare. ACM Trans. Embed. Comput. Syst. **20**(5s), 1–22 (2021)
3. Ben-Shabat, Y., Shrout, O., Gould, S.: 3dinaction: Understanding human actions in 3d point clouds. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19978–19987 (2024)
4. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
5. Biswas, S., Manavi Alam, A., Gurbuz, A.C.: Hrspecnet: A deep learning-based high-resolution radar micro-doppler signature reconstruction for improved har classification. IEEE Transactions on Radar Systems **2**, 484–497 (2024)
6. Bruce, X., Liu, Y., Zhang, X., Zhong, S.h., Chan, K.C.: Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3522–3538 (2022)
7. Chae, Y., Park, H., Kim, H., Yoon, K.J.: Doppler-aware lidar-radar fusion for weather-robust 3d detection. In: Int. Conf. Comput. Vis. pp. 27197–27208 (2025)
8. Chen, A., Wang, X., Zhu, S., Li, Y., Chen, J., Ye, Q.: mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In: ACM Int. Conf. Multimedia. pp. 3501–3510 (2022)
9. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Int. Conf. Comput. Vis. pp. 13359–13368 (2021)
10. Choi, J., Hor, S., Yang, S., Arbabian, A.: Mvdoppler-pose: Multi-modal multi-view mmwave sensing for long-distance self-occluded human walking pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 27750–27759 (2025)

11. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. arXiv preprint arXiv:1602.02830 (2016)
12. Cui, H., Zhong, S., Wu, J., Shen, Z., Dahnoun, N., Zhao, Y.: Milipoint: A point cloud dataset for mmwave radar. Adv. Neural Inform. Process. Syst. **36**, 62713–62726 (2023)
13. Dang, X., Fan, K., Li, F., Tang, Y., Gao, Y., Wang, Y.: Multi-person action recognition based on millimeter-wave radar point cloud. Applied Sciences **14**(16), 7253 (2024)
14. Deng, Z., Li, X., Li, X., Tong, Y., Zhao, S., Liu, M.: Vg4d: Vision-language model goes 4d video recognition. In: IEEE Int. Conf. Robot. Autom. pp. 5014–5020. IEEE (2024)
15. Ding, F., Luo, Z., Zhao, P., Lu, C.X.: milliflow: Scene flow estimation on mmwave radar point cloud for human motion sensing. In: Eur. Conf. Comput. Vis. pp. 202–221. Springer (2024)
16. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2969–2978 (2022)
17. Fan, H., Yang, Y., Kankanhalli, M.: Point spatio-temporal transformer networks for point cloud video modeling. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 2181–2192 (2022)
18. Fan, J., Zhou, Y., Yang, Y., Cui, X., Zhang, J., Xie, L., Yang, J., Lu, C.X., Ding, F.: M4human: A large-scale multimodal mmwave radar benchmark for human mesh reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 42836–42846 (2026)
19. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Int. Conf. Comput. Vis. pp. 6202–6211 (2019)
20. Fhager, L.O., Heunisch, S., Dahlberg, H., Evertsson, A., Wernersson, L.E.: Pulsed millimeter wave radar for hand gesture sensing and classification. IEEE Sensors Letters **3**(12), 1–4 (2019)
21. Gao, G., Liu, Q., Wang, W., Yu, Z., Liu, X.: Human activity recognition based on 4d millimeter-wave radar. In: Youth Academic Annual Conf. Chinese Assoc. Autom. pp. 1822–1828. IEEE (2025)
22. Gu, Z., He, X., Fang, G., Xu, C., Xia, F., Jia, W.: Millimeter wave radar-based human activity recognition for healthcare monitoring robot. arXiv preprint arXiv:2405.01882 (2024)
23. Guo, Z., Guendel, R.G., Yarovoy, A., Fioranelli, F.: Point transformer-based human activity recognition using high-dimensional radar point clouds. In: IEEE Radar Conference. pp. 1–6. IEEE (2023)
24. Haitman, Y., Bialer, O.: Doppdrive: Doppler-driven temporal aggregation for improved radar object detection. In: Int. Conf. Comput. Vis. pp. 26085–26094 (2025)
25. Hinojosa, C., Marquez, M., Arguello, H., Adeli, E., Fei-Fei, L., Niebles, J.C.: Privhar: Recognizing human actions from privacy-preserving lens. In: Eur. Conf. Comput. Vis. pp. 314–332. Springer (2022)
26. Hor, S., Yang, S., Choi, J., Arbabian, A.: Mvdoppler: Unleashing the power of multi-view doppler for micromotion-based gait classification. Adv. Neural Inform. Process. Syst. **36**, 58064–58074 (2023)
27. Huang, T., Dong, B., Yang, Y., Huang, X., Lau, R.W., Ouyang, W., Zuo, W.: Clip2point: Transfer clip to point cloud classification with image-depth pretraining. In: Int. Conf. Comput. Vis. pp. 22157–22167 (2023)
28. Kim, S., Lee, S., Hwang, D., Lee, J., Hwang, S.J., Kim, H.J.: Point cloud augmentation with weighted local transformations. In: Int. Conf. Comput. Vis. pp. 548–557 (2021)

29. Lai, Z., Yang, J., Xia, S., Lin, L., Sun, L., Wang, R., Liu, J., Wu, Q., Pei, L.: Radarllm: empowering large language models to understand human motion from millimeter-wave point cloud sequence. In: AAAI. vol. 40, pp. 5791–5799 (2026)
30. Lee, S.P., Kini, N.P., Peng, W.H., Ma, C.W., Hwang, J.N.: Hupr: A benchmark for human pose estimation using millimeter wave radar. In: IEEE/CVF Winter Conf. Appl. Comput. Vis. pp. 5715–5724 (2023)
31. Li, P., Wang, Z., Yuan, Y., Liu, H., Meng, X., Yuan, J., Liu, M.: Ust-ssm: Unified spatio-temporal state space models for point cloud video modeling. In: Int. Conf. Comput. Vis. pp. 6738–6747 (2025)
32. Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y., Sun, Z.: Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 29248–29257 (2025)
33. Liu, J., Wang, X., Wang, C., Gao, Y., Liu, M.: Temporal decoupling graph convolutional network for skeleton-based gesture recognition. IEEE Trans. Multimedia **26**, 811–823 (2024)
34. Liu, J., Han, J., Liu, L., Aviles-Rivero, A.I., Jiang, C., Liu, Z., Wang, H.: Mamba4d: Efficient 4d point cloud video understanding with disentangled spatial-temporal state space models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17626–17636 (2025)
35. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition **68**, 346–362 (2017)
36. Liu, M., Liu, J., Jiang, Y., He, B.: Heatmap pooling network for action recognition from rgb videos. IEEE Trans. Pattern Anal. Mach. Intell. **48**(3), 3726–3743 (2026)
37. Liu, Y., Zhou, S., Liu, X., Hao, C., Fan, B., Tian, J.: Unbiased faster r-cnn for single-source domain generalized object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 28838–28847 (2024)
38. Liu, Y., Wang, F., Wang, N., Zhang, Z.X.: Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion. Adv. Neural Inform. Process. Syst. **36**, 53964–53982 (2023)
39. Luo, F., Khan, S., Li, A., Huang, Y., Wu, K.: Edgeactnet: Edge intelligence-enabled human activity recognition using radar point cloud. IEEE Trans. Mobile Comput. **23**(5), 5479–5493 (2023)
40. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In: Int. Conf. Learn. Represent. (2022)
41. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Int. Conf. Comput. Vis. pp. 5442–5451 (2019)
42. Meng, Z., Fu, S., Yan, J., Liang, H., Zhou, A., Zhu, S., Ma, H., Liu, J., Yang, N.: Gait recognition for co-existing multiple people using millimeter wave sensing. In: AAAI. vol. 34, pp. 849–856 (2020)
43. Palipana, S., Salami, D., Leiva, L.A., Sigg, S.: Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(1), 1–27 (2021)
44. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI. vol. 32 (2018)
45. Phan, A.V., Le Nguyen, M., Nguyen, Y.L.H., Bui, L.T.: Dgcnn: A convolutional neural network over large-scale labeled graphs. Neural Networks **108**, 533–543 (2018)

46. Prabhakara, A., Jin, T., Das, A., Bhatt, G., Kumari, L., Soltanaghai, E., Bilmes, J., Kumar, S., Rowe, A.: High resolution point clouds from mmwave radar. In: IEEE Int. Conf. Robot. Autom. pp. 4135–4142. IEEE (2023)
47. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 652–660 (2017)
48. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inform. Process. Syst. **30** (2017)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PmLR (2021)
50. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Conf. Empir. Methods Nat. Lang. Process. pp. 3982–3992 (2019)
51. Salami, D., Hasibi, R., Palipana, S., Popovski, P., Michoel, T., Sigg, S.: Tesla-rapture: A lightweight gesture recognition system from mmwave radar sparse point clouds. IEEE Trans. Mobile Comput. **22**(8), 4946–4960 (2022)
52. Sengupta, A., Cao, S.: mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars. IEEE Trans. Neural Netw. Learn. Syst. **34**(11), 8418–8429 (2022)
53. Seo, H.I., Bae, J.W., Seo, D.H.: Radar-based human activity recognition using adaptive range selection and deep neural network. IEEE Sensors Journal pp. 1–1 (2026)
54. Shao, T., Du, Z., Li, C., Wu, T., Wang, M.: Fast human action recognition via millimeter wave radar point cloud sequences learning. In: ACM Int. Conf. Inf. Knowl. Manag. pp. 2024–2033 (2024)
55. Shao, T., Du, Z., Li, C., Wu, T., Wang, M.: Fast human action recognition via millimeter wave radar point cloud sequences learning. In: ACM Int. Conf. Inf. Knowl. Manag. pp. 2024–2033 (2024)
56. Singh, A.D., Sandha, S.S., Garcia, L., Srivastava, M.: Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In: ACM Workshop Millimeter-wave Netw. Sens. Syst. pp. 51–56 (2019)
57. Tan, T.H., Tian, J.H., Sharma, A.K., Liu, S.H., Huang, Y.F.: Human activity recognition based on deep learning and micro-doppler radar data. Sensors **24**(8) (2024)
58. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1521–1528. IEEE (2011)
59. Tunau, M., Zakka, V.G., Dai, Z.: Enhanced sparse point cloud data processing for privacy-aware human action recognition. In: Int. Conf. AI Healthcare. pp. 142–155. Springer (2025)
60. Wan, Q., Li, Y., Li, C., Pal, R.: Gesture recognition for smart home applications using portable radar sensors. In: IEEE Eng. Med. Biol. Soc. Conf. pp. 6414–6417. IEEE (2014)
61. Wang, S., Song, J., Lien, J., Poupyrev, I., Hilliges, O.: Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In: UIST. pp. 851–860 (2016)
62. Wang, S., Cao, D., Liu, R., Jiang, W., Yao, T., Lu, C.X.: Human parsing with joint learning for dynamic mmwave radar point cloud. ACM Interact. Mob. Wearable Ubiquitous Technol. **7**(1), 1–22 (2023)

63. 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: ICASSP. pp. 8298–8302. IEEE (2021)
64. Wu, R., Li, Z., Wang, J., Xu, X., Zheng, Z., Huang, K., Lu, G.: Diffusion-based mmwave radar point cloud enhancement driven by range images. arXiv preprint arXiv:2503.02300 (2025)
65. Wu, Y., Fioranelli, F., Gao, C.: Radmamba: Efficient human activity recognition through a radar-based micro-doppler-oriented mamba state-space model. IEEE Transactions on Radar Systems **4**, 261–272 (2025)
66. Xia, S., Chu, L., Pei, L., Yang, J., Yu, W., Qiu, R.C.: Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport. IEEE Trans. Mobile Comput. **23**(12), 10734–10751 (2024)
67. Xu, K.: Ai-driven personalized fall prevention for older adults. In: AAAI. vol. 39, pp. 29610–29612 (2025)
68. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. In: Eur. Conf. Comput. Vis. pp. 131–147. Springer (2024)
69. 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: ACM MobiSys. pp. 269–282 (2021)
70. Yan, J., Xu, C., Liu, D.: Og-pcl: Efficient sparse point cloud processing for human activity recognition. arXiv preprint arXiv:2511.08910 (2025)
71. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI. vol. 32 (2018)
72. Yang, J., Huang, H., Zhou, Y., Chen, X., Xu, Y., Yuan, S., Zou, H., Lu, C.X., Xie, L.: Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. Adv. Neural Inform. Process. Syst. **36**, 18756–18768 (2023)
73. Yu, C., Xu, Z., Yan, K., Chien, Y.R., Fang, S.H., Wu, H.C.: Noninvasive human activity recognition using millimeter-wave radar. IEEE Systems Journal **16**(2), 3036–3047 (2022)
74. Yu, J.T., Yen, L., Tseng, P.H.: mmwave radar-based hand gesture recognition using range-angle image. In: IEEE Veh. Technol. Conf. pp. 1–5 (2020)
75. Zeng, X., Shi, Y., Zhou, A.: Multi-har: Human activity recognition in multi-person scenes based on mmwave sensing. In: IEEE Int. Conf. Comput. Commun. pp. 1789–1793. IEEE (2022)
76. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8552–8562 (2022)
77. Zhang, R., Cao, S.: Real-time human motion behavior detection via cnn using mmwave radar. IEEE Sensors Letters **3**(2), 1–4 (2018)
78. Zhao, M., Tian, Y., Zhao, H., Alsheikh, M.A., Li, T., Hristov, R., Kabelac, Z., Katabi, D., Torralba, A.: Rf-based 3d skeletons. In: ACM SIGCOMM. pp. 267–281 (2018)
79. Zhao, P., Lu, C.X., Wang, B., Trigoni, N., Markham, A.: Cubelearn: End-to-end learning for human motion recognition from raw mmwave radar signals. IEEE Internet of Things Journal **10**(12), 10236–10249 (2023)
80. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4396–4415 (2022)

81. Zhou, R., Li, S., Zhang, H., Liu, C., Sun, J.: mmmulti: Multi-person action recognition based on multi-task learning using millimeter waves. ACM Interact. Mob. Wearable Ubiquitous Technol. **9**(2), 1–25 (2025)