2026 International Conference on Automatic Face and Gesture Recognition (FG)

ARiSE: Efficient Mesh-Based Action Recognition from Wi-Fi Sensing on Edge Devices

Zhankai Ye[1], Shuoqiu Li[1], Bofan Li[1], Yili Ren[2], Bo Mei[3], Shangqian Gao[1,*], Xin Liu[1,*]
[1]Florida State University, [2]University of South Florida, [3]Texas Christian University
{zy22b, sl24cs, bl21n, sg24bi, xliu15}@fsu.edu, yiliren@usf.edu, b.mei@tcu.edu

***Abstract*— Wi-Fi sensing technology has advanced to the point where it can reconstruct 3D human meshes, offering a low-cost and privacy-preserving paradigm for human action recognition. However, existing methods for processing these meshes typically rely on dense vertices spread across the entire body to maintain recognition accuracy. This approach imposes a heavy computational burden, making it difficult to deploy in resource-constrained WiFi edge devices. Our work presents an in-depth analysis of vertex distribution in these approaches. We find that the informational value of vertices is not uniformly distributed across the body. Based on this, we propose the ARiSE framework, which introduces a two-stage optimization strategy. First, we address the inherent complexity of the human anatomy. We recognize that body parts with fine motor details, like the hands and torsos, require a higher density of vertices to be accurately represented. Our selective down-sampling method is therefore designed to strategically preserve more vertices in these critical regions while reducing them elsewhere. Second, we introduce a dynamic, motion-aware weighting mechanism. For any given action, different body parts contribute unequally to the movement energy. Our framework learns to identify the high-energy parts and focus more on them during recognition. We proposes a energy-based mixture of experts (MoE) in which each body part is processed by a lightweight CNN expert. A learned motion-aware gate estimates part energy over time and sets the expert weights so inference focuses on the parts that drive the action. This dual approach ensures that we not only start with a more strategic, anatomy-aware distribution of points, but also dynamically focus on the most expressive body parts as an action unfolds. Extensive experiments validate our approach. Critically, we are the first to demonstrate that ARiSE achieves state-of-the-art accuracy on resource-constrained edge platforms like the NVIDIA Jetson TX2, while reducing computational overhead by more than an order of magnitude. This confirms the practical feasibility of on-device mesh-based action recognition with Wi-Fi sensing.**

I. Introduction

Action recognition is crucial for applications like interactive systems, behavior analysis, and home monitoring, which benefit from detailed 3D motion representation. However, traditional methods relying on cameras, radar, or LiDAR face significant practical limitations. They require expensive, sparsely deployed hardware, are susceptible to occlusions, and often raise serious privacy concerns. Conversely, Wi-Fi sensing directly addresses these limitations. It utilizes existing, low-cost networks for broad coverage, overcoming the hardware and deployment barriers [36]. Its signals penetrate obstacles like walls, thus eliminating reliance on line-of-sight [1]. Furthermore, it enhances privacy by design, as it functions without capturing any identifiable visual data, making it ideal for various environments.

However, translating this detailed 3D motion representation into practical applications on Wi-Fi devices presents a significant challenge. These detailed representations often take the form of high-resolution 3D meshes [5], [37], [43]. Recognizing a single action requires processing a sequence of these meshes over time, and with each mesh typically containing thousands of vertices (e.g., 6,890 in an SMPL model [34]), the computational and memory load quickly becomes impractical for resource-constrained edge devices.

A seemingly natural solution is to first downsample the mesh to a smaller set of vertices, and then feed the reduced data into a model for action recognition. Many existing approaches follow this general strategy, typically using Farthest Point Sampling (FPS) to globally select a few hundred vertices from each mesh [16], [41], [42], [45]. However, despite the reduced input size, the models adopted—such as the hierarchical transformer architecture in STMT [45], which applies inter- and intra-frame attention—remain computationally expensive. As a result, this pipeline still poses significant challenges for deployment on resource-constrained devices.

To address this issue, we propose the ARiSE approach, which reduces computational overhead by more than an order of magnitude while maintaining higher accuracy than state-of-the-art methods at the same frame rate under equal or lower per-frame vertex counts. This enables practical deployment of mesh-based action recognition on resource-constrained devices, including radar and LiDAR systems.

The key insight of ARiSE is that certain body parts, such as the hands and torsos, play a more critical role in achieving high accuracy for action recognition tasks. Conventional mesh down-sampling techniques, such as FPS, treat the mesh globally, often preserving an excessive number of vertices in less informative regions like the abdomen while underrepresenting crucial areas such as the hands (as shown in Figure 1, third row). However, different actions require varying levels of granularity. For example, large-scale motions such as walking or jumping can be effectively recognized with just a few vertices, whereas fine-grained actions like drinking or writing require detailed hand and arm movements for accurate distinction. If these regions are insufficiently represented, the model may fail to capture subtle yet crucial details, leading to degraded performance.

979-8-3315-7231-0/26/$31.00 © 2026 IEEE
* Corresponding Authors

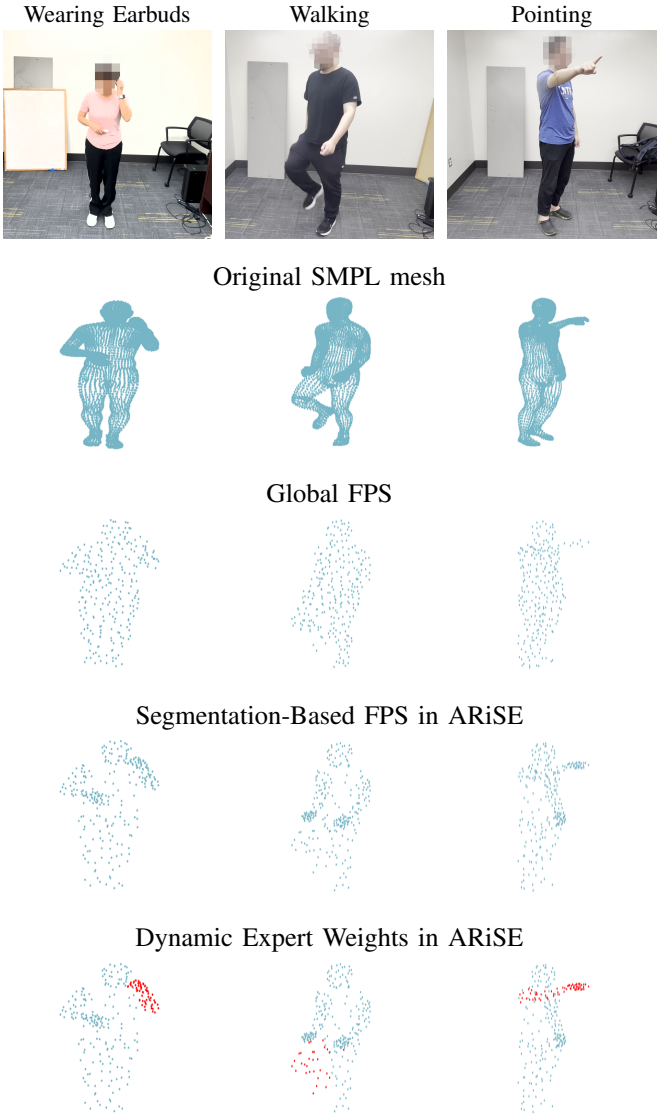

Fig. 1: Comparison of different down-sampling methods on six action sequences: Wearing Earbuds, Walking and Pointing. The second row shows the original SMPL mesh with 6,890 vertices, the third row displays results using global FPS, and the fourth row presents segmentation-based FPS. The fifth row shows segmentation-based FPS with highlight body part with the highest motion energy. ARiSE successfully preserved more vertices in certain body parts (i.e. hands) and accurately highlighted the body part with the highest motion energy for different actions.

To fix this, our first step is a segmentation-based sampling strategy. As illustrated in the fourth row of Figure 1, this method partitions the mesh by body parts and strategically allocates more vertices to inherently complex regions, creating an efficient and structurally-informed baseline.

However, segmentation-based down sampling alone is insufficient. A model might naturally learn to ignore a body part represented by fewer vertices, even if that part is central to the action (e.g., the leg gesture in "walking"). To solve this, our second step is an energy-based MoE [28] framework. In this design, we treat each body part as a specialized "expert," processed by its own lightweight convolutional neural network (CNN). A central, motion-aware gating module then learns to analyze the movement and estimate the motion energy of each part over time. This gate assigns computational weights to each expert, functioning as a dynamic attention mechanism. As visualized by the highlighted regions in the final row of Figure 1, this forces the model to focus its resources on the most expressive parts of the body as the action unfolds. This MoE architecture ensures that even sparsely sampled but critical regions receive the computational focus they need for accurate recognition.

To validate our approach, we conduct extensive experiments on public benchmarks (BABEL, NTU RGB+D 60) and our own self-collected Wi-Fi sensing dataset. The results demonstrate that ARiSE not only achieves state-of-the-art accuracy but does so while reducing computational overhead by more than an order of magnitude. Critically, we are the first to demonstrate this level of performance on severely resource-constrained edge platforms like the NVIDIA Jetson TX2. This confirms the tangible feasibility of deploying high-fidelity, on-device mesh-based action recognition using Wi-Fi sensing.

Our key contributions are three-fold:

- We propose ARiSE, a novel, two-stage framework that solves the critical trade-off between accuracy and efficiency in mesh-based action recognition, making it practical for on-device deployment.
- We design a dual-optimization strategy that combines (1) an anatomy-aware, segmentation-based sampling method to preserve structural details with fewer points, and (2) an MoE architecture where a learned gate dynamically allocates moreon the most expressive body parts during an action.
- We provide the first conclusive evidence of a mesh-based action recognition model achieving state-of-the-art accuracy on real-world, resource-constrained edge hardware, demonstrating a greater than 10x reduction in computational cost.

II. Related Work

With the emergence of IEEE 802.11bf [24], the standardization of RF sensing marks a substantial step toward integrating authentication capabilities into widely deployed wireless infrastructure. Leveraging this existing, pervasive Wi-Fi infrastructure helps overcome the hardware dependency and integration challenges faced by traditional camera and LiDAR-based methods. Additionally, it offers lower cost, resilience to lighting and occlusions, and enhanced privacy [6], [33].

A. RF-Based Skeleton Action Recognition

RF sensing has been widely explored for skeleton-based action recognition [9]–[14], [17], [31], [38], [44], where

human motion is represented as a set of key joints. This approach is particularly attractive due to its low data complexity, computational efficiency, and robustness to occlusions, making it well-suited for privacy-preserving applications and deployment on resource-constrained devices.

Early studies primarily leveraged radar-based RF systems to extract human skeletons. For example, RF-Pose3D [38] introduces the first RF-based system capable of inferring dynamic 3D human skeletons. It efficiently captures spatio-temporal information by decomposing high-dimensional convolutions into low-dimensional operations. RF-Wear [11] is a novel wearable solution that tracks human body movements by leveraging passive RFID tags. WiPose [10] is the first framework that accurately constructs 3D human skeletons using commercial Wi-Fi Channel State Information (CSI).

Despite their efficiency, skeleton-based representations often lack fine-grained details, particularly in regions, for instance, hands, which are critical for recognizing complex human actions.

B. RF-Based Human Mesh Modeling

Mesh-based models provide a detailed representation of human motion by capturing surface dynamics and articulated movement [5], [34], [37], [39], [40], [43]. In this domain, the Skinned Multi-Person Linear (SMPL) model [18] and its extended versions (SMPL-X) [21] are widely used for 3D human mesh generation, producing triangular meshes with at least 6,890 vertices to accurately represent body shape and posture.

For instance, RF-Avatar [43] employs customized FMCW radar devices, leveraging Mask-RCNN and multi-head self-attention mechanisms to extract human motion trajectories from reflected signals and generate an SMPL mesh. *mmMesh* [39] employs a commercial millimeter-wave device and a deep learning pipeline to handle sparse 3D point clouds, producing an SMPL mesh. *WiMesh* [34] exploits Wi-Fi CSI measurements, fusing time-of-flight, amplitude, and phase data through a neural model to reconstruct an SMPL mesh. *3DMesh-GAR* [25] reconstructs meshes for all individuals in a scene without bounding boxes, averaging multiple 3D body meshes into a single, holistic feature set that powerfully represents the entire group's activity.

While RF-based human mesh modeling has advanced, using high-resolution meshes for action recognition on RF devices remains challenging due to high computational costs.

C. Mesh Sequence-Based Action Recognition

To capture dynamic motions over time, action recognition relies on sequences of 3D meshes rather than single-frame shapes. However, processing thousands of vertices per frame (e.g., 6,890 from an SMPL model) quickly overwhelms resource-constrained devices. In order to address this issue, researchers resampled point cloud data by Farthest Point Sampling (FPS) algorithm to select representative subset of points, which helps in reducing computational complexity while preserving geometric information. PointNet++ [23] introduces mini-PointNet to extract features in local sub-regions. SequentialPointNet [16] adapts this and introduces a Hyperpoint-Mixer module to perform feature mixing. PST-Net [42] utilizes another CNN for temporal feature across frames. 3DV [35] converts the point cloud sequence into a voxelized 3D representation and then captures the temporal dynamics within these voxels. P4Transformer [41] uses attention encoder to learn long-range dependencies along temperol features. While these methods effectively model spatio-temporal relationships in irregular data, STMT [45] extends this approach to mesh sequences by incorporating a hierarchical transformer framework with inter- and intra-frame attention layers. However, this design results in a model exceeding 10 million parameters, making deployment on RF devices highly impractical.

TABLE I: A 128-Vertex Mesh Generated via STMT on the NTU RGB+D 60 and BABEL Datasets

Body Part	**Distribution**		**Accuracy Drop w/o Part**	
	Left	Right	NTU RGB+D 60	BABEL
Hands	3	3	-3.38%	-2.45%
Arms	8	8	-20.05%	-3.90%
Legs	23	23	-5.29%	-5.78%
Abdomen	53		-0.1%	-1.50%
Head	7		-1.25%	-3.10%

D. Other Recognition Method Oriented from 3D Point Cloud

Recent advancements in 3D action recognition show a shift away from legacy methods that relied on downsampling point clouds for computational efficiency. Growing computational power now permits the use of richer, more structurally informative full mesh data. 3DInAction [2] introduces a large-scale dataset and a multi-branch, multi-scale network that operates directly on raw point cloud sequences. It choose to implicitly learn surface geometry from unstructured point clouds. However, 3DInAction is computationally demanding, imposing a significant cost on both training time and hardware resources. Consequently, due to this high computational expense, we have chosen not to adopt this approach in our work, favoring a more resource-efficient methodology. This highlights the need for efficient down-sampling and lightweight network to enable mesh-sequence action recognition on resource-constrained devices.

III. Key Insight

Our key insight is that action recognition accuracy is highly dependent on the distribution of vertices across different body parts. Existing approaches rely on FPS for mesh down-sampling, formulated as:

$$S = v_0,\ v_i = \arg\max_{v \in V \setminus S} \min_{u \in S} d(v, u),\ \forall i = 1, 2, \ldots, k \quad (1)$$

where S represents the set of sampled vertices, v_0 is an arbitrarily selected initial vertex from the original set V, $d(v, u)$ denotes the Euclidean distance (or another metric)

TABLE II: NTU RGB+D 60 - Classes With Lowest Accuracy

Action Name	Counter Class	Possible Factor
Reading	Play Phone	Finger Movement
Writing	Typing	Finger Movement
Clapping	N/A	Body Surface
Put on a Shoe	Take off a Shoe	Body Surface
Take off a Shoe	Put on a Shoe	Body Surface
Drinking	N/A	Finger Movement
Eating	N/A	Finger Movement
Taking Selfie	Pointing	Torso Rotation
Sneeze	Headache	Body Surface
Typing	Writing	Finger Movement
Play Phone	Writing	Finger Movement

between vertices v and u, and k is the total number of vertices to be sampled. This algorithm ensures a well-distributed selection across the entire mesh [7].

While effective in preserving overall structure, traditional FPS has two major limitations. First, since the initial vertex is randomly selected, applying FPS to each frame in a mesh sequence results in inconsistent vertex correspondences across frames, as all vertices remain unordered. Second, FPS does not account for the varying importance of different body parts. As a result, critical regions such as the hands receive too few vertices, whereas less informative areas like the abdomen retain an excessive number.

To illustrate this issue, we applied STMT's global FPS-based down-sampling to the NTU RGB+D 60 [27] and BABEL [22] dataset, generating 128-vertex meshes and analyzing the distribution of vertices across body parts. As shown in Table I, different body parts receive varying numbers of vertices: hands (6), arms (16), legs (46), abdomen (53), and head (7). However, this uniform down-sampling fails to prioritize the most informative regions, leading to suboptimal accuracy in action recognition.

To investigate the role of different body parts in action recognition, we systematically removed vertices corresponding to each region and measured the resulting accuracy drop on both the NTU RGB+D 60 and BABEL datasets. As shown in Table I, eliminating abdomen vertices has minimal impact (-0.1% in NTU, -1.5% in BABEL), whereas removing vertices from the arms results in a substantial accuracy drop (-20.05% in NTU, -3.9% in BABEL). The legs, hands, and head also contribute notably. These results highlight that certain body parts play a critical role in action recognition.

As previously shown, using global FPS-based down-sampling results in a significant 20% accuracy drop. This is because the NTU RGB+D 60 dataset contains a large number of fine-grained actions, making action recognition particularly challenging. Table II highlights the action classes with the lowest accuracy, where many misclassifications involve subtle motion cues. For instance, actions like Reading vs. Playing Phone and Writing vs. Typing require fine-grained finger movements, while actions like Taking Selfie vs. Pointing rely on subtle torso rotations. These findings suggest that traditional FPS fails to preserve critical motion details, particularly in the hands and arms, leading to confusion between visually similar actions.

(a) SMPL mesh (6,890 vertices) (b) Existing methods (256 vertices) (c) ARiSE (256 vertices)

Fig. 2: Comparison of different down-sampling methods.

To further illustrate this issue, Figure 1 compares different down-sampling strategies across three action sequences: Wearing Earbuds, Walking and Pointing. The third row shows results using traditional FPS, where Making a Phone Call vs. Wearing Earbuds, Eating vs. Drinking, and Reading vs. Writing appear visually similar due to insufficient preservation of fine-grained motion details. In contrast, the fourth row demonstrates our segmentation-based FPS, which better retains essential motion cues, making these actions more distinguishable. This highlights the importance of allocating more vertices to motion-critical regions—such as the hands and arms—for improved recognition accuracy.

IV. Design

A. Segmentation-Based FPS

In our approach, we first segment the mesh into different body parts. Specifically, each vertex corresponds to a specific body part, such as the hands, arms, abdomen, etc. These vertex indices are fixed, meaning they always belong to their respective body parts for all mesh sequences. Then, we label each vertex and classify it into the corresponding body part, ensuring that each vertex remains consistently assigned to a specific body part, making the segmentation ordered in every frame. Next, we apply the FPS algorithm within each segmented region and record the indexes of sampled vertices. These indexes are feasible for all mesh sequences generated by the SMPL model, which means we only need to perform our down-sampling method once. Also, instead of selecting points purely based on spatial distribution, we adopt a uniform allocation strategy.

Figure 2 presents one single frame of a person performing the same action, which is jumping. Figure 2a is the original SMPL mesh with 6890 vertices. Figure 2b shows the mesh down-sampled by the traditional global FPS algorithm. And Figure 2c shows the outcome of our segmentation-based FPS. As shown, global FPS results in a low vertex density in the hands and arms, while preserving a large number of vertices in low-information regions such as the abdomen. In contrast, our method achieves a more balanced distribution, preserving more vertices in regions such as the hands and arms.

B. Energy-Guided MoE for Spatio-Temporal Learning

In point cloud-based action recognition, which includes mesh-based action recognition, two aspects need handling:

spatial features and temporal features. The spatial features denote the vertex-level relations within one single frame, while the temporal features stand for the frame-wise relations. We have developed a MoE-based neural network that combines spatial learning, temporal feature extraction, and residual networks to ensure efficiency and accuracy. Our neural network consists of an arbitrary number of basic blocks. The base block is illustrated in Figure 3.

In STMT [45], the spatial features are handled by a network structure called Surface Field Convolution. Since the mesh sequences being used in STMT are unordered. It is necessary to form local vertex patches before performing classification tasks. STMT computed both Euclidean distance and Geodesic distances to approximate the connectivity between vertices. However, the mesh sequences in ARiSE are ordered since we sampled the vertices by index instead of by spatial coordinates. Therefore, we could bypass this step to avoid a significant waste of computational resources.

In addition, transformer networks are used in the temporal learning process for STMT. Transformer-based neural network architectures work extraordinarily on long time sequences, such as plain text articles, because of the attention mechanism [30]. However, transformers are not necessarily needed since the mesh sequences in our approach are, at most, 32 frames, and using a transformer architecture could be too much. Instead, we could simply use convolution operations to handle the temporal relationships in our setting.

Consequently, we abandoned the inefficient architecture used by STMT and instead proposed a spatio-temporal residual block along with the MoE strategy that effectively and dynamically captures both sequential and spatial information from the input features. We define the convolution operations of our spatial-temporal residual block by:

$$X_i = \text{Conv-2d}_i\left(\widetilde{X}_{i-1}\right), \quad \text{for } i = 1, 2, 3, 4, \tag{2}$$

where

$$\widetilde{X}_{i-1} = \begin{cases} X_{i-1}, & \text{if } i \neq 3, \\ X_{i-1}^{\top_{(2,1,3)}}, & \text{if } i = 3. \end{cases}$$

Here, $X_0 \in \mathbb{R}^{D\times F\times V}$ is the input tensor (with D as the number of channels, F the number of frames, and V the number of vertices), $X_i \in \mathbb{R}^{C_i\times F\times V}$ for $i \in \{1, 2\}$ (with $C_2 = D$), and $X_i \in \mathbb{R}^{C'_{i-1}\times D\times V}$ for $i \in \{3, 4\}$ (with $C'_4 = F$). The operator $\top_{(i,j,k)}$ permutes the dimensions of a tensor so that the ith dimension becomes the first, the jth becomes the second, and the kth becomes the third. Finally, the output of our spatial-temporal residual block is defined as

$$X_{\text{out}} = X_0 + X_4^{\top_{(2,1,3)}}, \tag{3}$$

with $X_4 \in \mathbb{R}^{F\times D\times V}$. Each Conv-2d$(\cdot)$ is composed of Conv-BN-ReLU operations, which is regularly used in residual networks [8]. Our spatial-temporal blocks enable the CNN to capture sequential relationships when it goes deeper with a larger reception field [19]. Additionally, a key advantage of CNN-based neural networks is their compact size. To deploy our approach on resource-constrained devices, a deep learning model with fewer parameters is essential. Thus, we choose to leverage CNN for temporal learning instead of transformers.

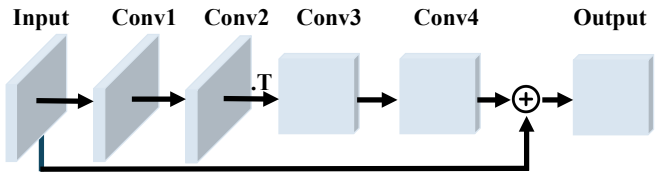

Fig. 3: The Basic Block of Our Residual Spatial-Temporal Neural Network

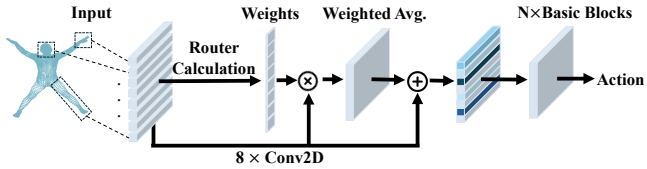

Fig. 4: Our Architecture With Energy-based MoE Inputs

To further enhance the discriminative ability of the network, we introduce an energy-based MoE architecture that reflect the motion intensity of different body regions. For pth body part ($p = 1, \cdots, 8$), we compute its motion energy by accumulating the inter-frame displacement of vertices. Formally, let S_p denote the set of vertices belonging to part p, then the motion energy is defined as

$$E_p = \frac{1}{|S_p|} \sum_{v\in S_p} \sum_{t=1}^{F-1} |V_{t+1,v} - V_{t,v}|_2 \,, \tag{4}$$

where $V_t, v \in \mathbb{R}^3$ denotes the 3D position of vertex v at frame t. The obtained energies are fed into the router, which produces gating weights over a set of Conv2D experts. Specifically, we define:

$$X_{\text{MoE}} = \sum_{p=1}^{8} g_p \cdot f_p(X_p), \quad \text{and} \quad g = \text{softmax}(f_r(E)), \tag{5}$$

where $f_p(\cdot)$ denotes the p-th expert corresponding to each body part, $f_r(\cdot)$ denotes a learnable router to produce motion coefficient given the energy E, and g is the softmax normalized gating coefficient. X_{MoE} is then added back to $X_p = X_p + X_{\text{MoE}}$. These gating coefficients adjust the contributions of different feature groups in the first convolutional block, amplifying responses from informative regions (e.g., moving limbs) while suppressing less discriminative ones (e.g., static torso). This MoE-inspired design enables the network to adaptively focus on body parts exhibiting stronger motion dynamics, thereby improving accuracy.

V. Experiment Setup

To evaluate the performance and adaptability of ARiSE, we conducted a series of experiments comparing it with the state-of-the-art methods, across hardware configurations and multiple datasets. We picked the four most recent related

approaches as baselines, which are PSTNet [42], SequentialPointNet [16], P4Transformer [41], and STMT [45].

A. Hardware Platforms

High-End Platform. Due to the high computational demands of STMT, we evaluated the accuracy of both all approaches, including ARiSE on a high-performance platform equipped with an NVIDIA V100 GPU, which features 32 GB of high-bandwidth memory. The V100 boasts 5,120 CUDA cores and delivers a peak performance of 15.7 TFLOPs. This GPU is designed to handle data-intensive tasks, making it suitable for evaluating computationally demanding models like STMT.

Resource Constrained Platform. To assess the adaptability of ARiSE in environments with limited computational resources, we trained and validated our approach on an NVIDIA Jetson TX2 platform, which only contains 8 GB of LPDDR4 memory, 256 CUDA cores, along with approximately 2 TFLOPs of peak FP32 performance.

Since STMT ran extremely slowly on this hardware due to its high computational demands, we excluded it from our evaluation on this platform.

B. Datasets

To validate the generalizability of ARiSE across different sensing modalities and assess its performance on diverse motion types, we selected the following datasets:

Comprehensive Motion-Capture Dataset. We utilized the BABEL 60 dataset, an extension of the AMASS dataset [20] that integrates semantic labels with high-quality 3D human motion data. BABEL contains approximately 43 hours of motion capture sequences, covering over 150 unique action categories. The training set consists of 23,565 data clips, while the test set includes 8,671 clips, spanning 60 classes, which can be organized and converted into the SMPL mesh. We chose BABEL due to its diverse motion categories, which enable a comprehensive evaluation of our model's ability to capture various movement patterns.

Fine-Grained RGB-Based Dataset. We also utilized the NTU RGB+D 60 dataset, a widely recognized benchmark for human activity recognition using RGB cameras and depth sensors. This dataset comprises 60 action categories with a total of 56,880 video clips. To focus on evaluating the model's capability in recognizing fine-grained actions, we selected 11 specific categories from the NTU dataset (as detailed in Table II) . These categories emphasize detailed limb and body movements, such as finger, arm, leg, and torso actions, where missing fine-grained motion details can significantly impact recognition accuracy. The training set consists of 7,054 RGB video clips, while the test set includes 2,872 clips. By selecting these specific categories, we assess the model's performance in capturing fine-grained motion details. Mesh sequences were generated from these video clips using the open-source ExPose project [3], which is based on the SMPL-X model.

Our Wi-Fi Sensing Dataset was collected using the Wi-Mesh system [34], leveraging commodity Wi-Fi devices for 3D human mesh construction. The data was collected with a transmitter-receiver setup consisting of a Dell LATITUDE laptop as the transmitter with three antennas, and an L-shaped antenna array with nine antennas as the receiver, utilizing Intel 5300 NICs, as shown in Figure 5. A camera was used to capture ground truth for comparison. We recruited 13 volunteers (10 males and 3 females) who performed various everyday activities, including walking, arm movements, leg lifting, and random motions. The Wi-Fi Channel State Information (CSI) data was recorded using Linux 802.11 CSI tools, capturing 30 OFDM subcarriers per packet. To extract meaningful human motion data, we applied CSI preprocessing to remove phase distortions and estimated the 2D Angle of Arrival (AoA) of Wi-Fi reflections using a 4D MUSIC algorithm [26]. By eliminating static object reflections and leveraging deep learning models, we transformed 2D AoA images into 3D human meshes using the SMPL-X model. It supports non-line-of-sight (NLoS) scenarios, clothing variations, and demonstrates high adaptability across different environments. After exploring multiple public datasets, such as MSR Action 3D [15], TOYOTA Smart Home [4], UCF101 [29], and HMDB-51 [32], we selected 16 representative actions for our dataset. In total, we recorded 5,218 Wi-Fi-based data clips, comprising 4,194 training clips and 1,014 test clips.

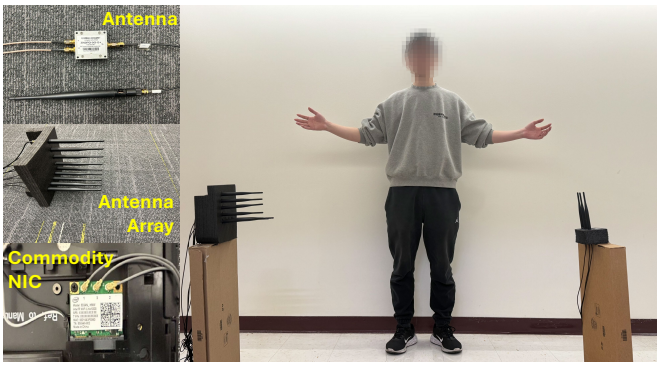

Fig. 5: Our Wi-Fi Sensing Devices and Experiment Scene

VI. Evaluation

A. Recognition Accuracy on Benchmark Datasets

To evaluate ARiSE, we compare it against four state-of-the-art approaches (PSTNet, SequentialPointNet, P4Transformer, STMT) on three datasets: NTU RGB+D 60, BABEL, and our Wi-Fi sensing dataset. The full results (Table III) show that ARiSE consistently delivers the best accuracy with significantly lower resource usage.

On **NTU-RGB+D 60**, ARiSE achieves 81.09% Top-1 accuracy. This is a substantial improvement over the strongest baseline, STMT, which reaches only 68.52%. The margin of over 12% underscores our method's superior capability in capturing fine-grained action patterns from sparse mesh data. On the **BABEL-60** dataset, which features complex motion capture sequences, ARiSE maintains its lead with 57.05% accuracy compared to STMT's 55.05%, confirming its robustness across diverse action categories. Finally, on

TABLE III: Experimental Results on NTU-RGBD 60, BABEL-60 and Our WiFi Sensing Dataset.

Approach	# Vertices	GFLOPs	NTU-RGBD-60		BABEL-60		Wi-Fi Sensing	
			Top-1 (%)	Top-5(%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
PSTNet [42]	2048	19.57	51.81	94.33	48.72	69.23	70.41	88.89
SequentialPointNet [16]	512	2.89	59.84	94.98	54.09	77.86	79.88	96.94
P4Transformer [41]	2048	121.18	49.26	92.31	52.88	77.77	87.08	**100.0**
STMT [45]	512	39.22	68.52	90.93	55.05	78.64	72.01	91.34
ARiSE (Ours)	512	**1.20**	**81.09**	**97.10**	**57.05**	**79.58**	**96.04**	98.50

TABLE IV: Computational Efficiency on NVIDIA V100

Approach	Batch	VRAM (MB)	TPI (s)	Latency (s)
PSTNet	64	26160	2.24	0.76
SequentialPointNet	64	29938	0.57	0.46
P4Transformer	24	30842	1.74	0.57
STMT	32	29028	0.89	0.42
ARiSE (Ours)	64	**3584**	**0.056**	**0.022**

TABLE V: Runtime on Jetson TX2 (Edge Device)

Approach	TPI (s)	Latency (s)
PSTNet	OOM	OOM
SequentialPointNet	OOM	OOM
P4Transformer	OOM	OOM
STMT	OOM	OOM
ARiSE (Ours)	**1.91**	**0.83**

our proprietary **Wi-Fi dataset**, ARiSE sets a new state-of-the-art with 96.04% Top-1 accuracy. This significantly outperforms P4Transformer (87.08%), ensuring reliable and high-fidelity recognition for wireless sensing applications.

B. Computational Efficiency & Edge Performance

Beyond accuracy, ARiSE achieves remarkable computational efficiency across multiple dimensions (Table IV). First, in terms of complexity, ARiSE requires only 1.20 GFLOPs. This is drastically lower than transformer-based methods like P4Transformer (121.18 GFLOPs) and STMT (39.22 GFLOPs), representing a reduction of up to two orders of magnitude. Second, regarding memory usage, ARiSE consumes just 3.5 GB of VRAM with a batch size of 64, whereas baseline methods require approximately 30 GB. This low memory footprint allows ARiSE to train effectively on lower-end hardware where other models would fail. Finally, this efficiency translates into superior runtime performance: ARiSE achieves a training Time-Per-Iteration (TPI) of 0.056s (10× faster than the nearest competitor) and an inference latency of just 0.022s, making it highly suitable for real-time applications.

Crucially, this architecture enables deployment on resource-constrained edge devices, as shown in Table V. We evaluated all methods on an NVIDIA Jetson TX2. While all baseline approaches failed with Out-Of-Memory (OOM) errors due to their excessive resource demands, ARiSE ran successfully with a TPI of 1.91s and an inference latency of 0.83s. This result strictly validates ARiSE as the only currently viable solution for performing on-device, privacy-preserving mesh action recognition directly at the edge.

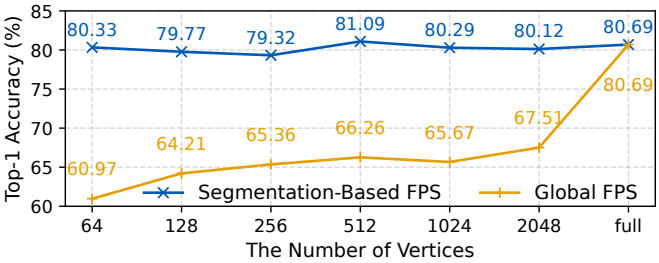

Fig. 6: Top-1 accuracy differences between global FPS and our segmentation-based FPS (NTU RGB+D 60).

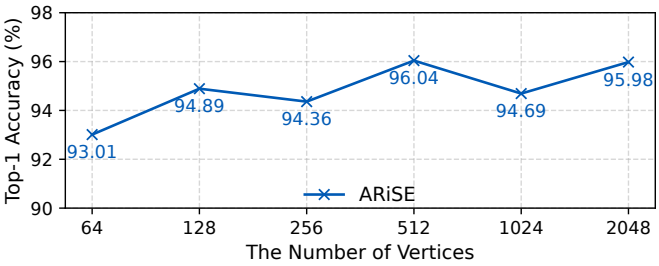

Fig. 7: Recognition accuracy of ARiSE on the Wi-Fi sensing dataset across different vertex counts.

VII. Ablation Study

A. Global FPS VS. Segmentation-Based FPS

To further investigate the impact of our Segmentation-Based FPS strategy, we compare it against the traditional global FPS on the NTU RGB+D 60 dataset, as illustrated in Figure 6. The results indicate that global FPS consistently lags behind, showing a substantial accuracy gap across all vertex settings. For instance, when sampling 512 vertices, Segmentation-Based FPS achieves 81.09% Top-1 accuracy, whereas global FPS only reaches 66.26%, yielding a margin of 14.83%. Similar trends are observed at other sampling densities, confirming the effectiveness of our part-aware sampling mechanism.

Another important observation is that Segmentation-Based FPS exhibits stable accuracy even at very low vertex counts. With only 64 vertices, its accuracy (80.33%) already approaches that of the full mesh (80.69%), demonstrating that our method can preserve the most informative structural cues while discarding redundant points. In contrast, global FPS

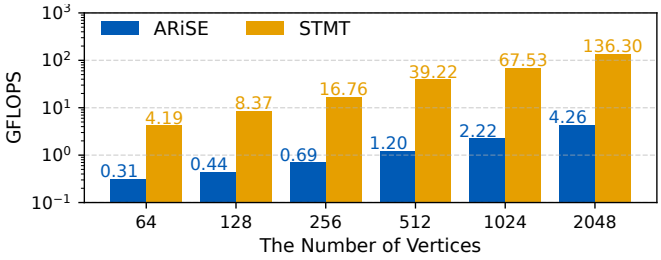

Fig. 8: Computational efficiency comparison between ARiSE and STMT on benchmark datasets.

TABLE VI: Top-1 accuracy VS. number of residual spatial-temporal blocks (NTU RGB+D 60)

# Blocks	Accuracy (%)	# Parameters	GFLOPs
4	81.09	1.86M	1.201
3	79.32	1.58M	1.196
2	78.20	1.35M	1.177
1	74.03	1.32M	1.157

suffers a clear degradation at smaller vertex settings, reflecting its indiscriminate sampling behavior. Because it ignores body-part segmentation and temporal structure, global FPS tends to miss fine-grained motion details in critical regions such as hands, arms, and legs, which are essential for precise action recognition. These results underscore the robustness and efficiency of Segmentation-Based FPS in maintaining accuracy under aggressive downsampling.

B. Performance with Different Numbers of Residual Spatial–Temporal Blocks

To further analyze the contribution of the residual spatial-temporal blocks in ARiSE, we varied their number and evaluated the impact on action recognition accuracy using the NTU RGB+D 60 dataset with 512 vertices. The results, summarized in Table VI, show a clear performance trend: accuracy steadily improves as the number of blocks increases. With a single block, the model achieves 74.03% Top-1 accuracy, which rises to 78.20% with two blocks and 79.32% with three blocks. The best performance is observed when four blocks are stacked, reaching 81.09% Top-1 accuracy.

Importantly, these gains are achieved with only a marginal increase in model complexity. As the number of blocks increases from one to four, the parameter count grows modestly from 1.32M to 1.86M, and GFLOPs remain nearly unchanged (1.157 to 1.201). This demonstrates that the residual spatial-temporal blocks provide substantial improvements in representational capacity without significantly increasing computational cost. Overall, these results confirm the effectiveness of our block design and justify the choice of using four residual spatial-temporal blocks as the default configuration in ARiSE.

C. Performances on Different Number of Vertices

We investigate the influence of the number of vertices on recognition accuracy (Fig. 9, 10, 7) and computational efficiency (Fig. 8). On **NTU RGB+D 60** (Fig. 9), ARiSE consistently outperforms STMT across all vertex settings.

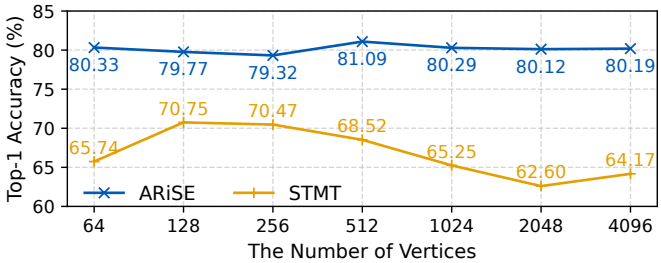

Fig. 9: Top-1 accuracy comparison between ARiSE and STMT on the fine-grained RGB-based dataset (NTU RGB+D 60).

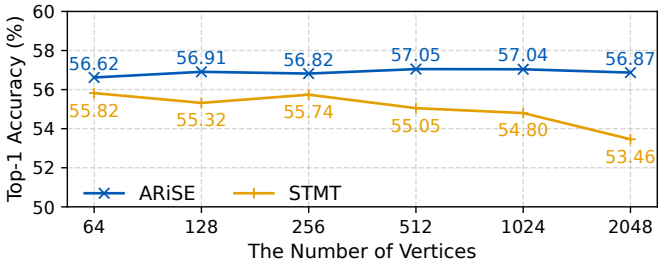

Fig. 10: Top-1 accuracy comparison between ARiSE and STMT on the motion-capture dataset (BABEL).

Even with 64 vertices, it maintains 80.33% accuracy and peaks at 81.09% with 512 vertices. In contrast, STMT fluctuates and degrades with more vertices (dropping to 62.60% at 2048), indicating its transformer architecture struggles with higher resolution inputs without proportional depth increases. Results on **BABEL** (Fig. 10) follow a similar trend. ARiSE maintains a stable lead, peaking at 57.05%, whereas STMT declines to 53.46% at 2048 vertices. On our **Wi-Fi sensing dataset** (Fig. 7), performance remains robust (<93%) even at low vertex counts, peaking at 96.04% with 512 vertices. Regarding **Efficiency** (Fig. 8), ARiSE shows a drastic reduction in computation. At 2048 vertices, it requires only 4.26 GFLOPs—32× fewer than STMT—demonstrating superior scalability.

VIII. Conclusion

In this paper, we introduced ARiSE, a novel framework designed to overcome the critical barrier of deploying mesh-based action recognition on resource-constrained Wi-Fi edge devices. Our dual-optimization approach begins with an anatomy-aware downsampling strategy, which partitions the mesh to strategically allocate more vertices to structurally complex regions, creating an efficient yet detailed baseline representation. Building on this, our energy-wise MoE architecture then learns to dynamically assign greater weight to the outputs of experts corresponding to the most expressive body parts. Extensive experiments confirm that ARiSE achieves state-of-the-art accuracy while reducing computational overhead by more than an order of magnitude. By being the first to validate this level of performance on real-world edge hardware like NVIDIA Jetson TX2, our work establishes the practical feasibility of efficient, high-fidelity action recognition for the next generation of WiFi sensing applications.

REFERENCES

[1] F. Adib and D. Katabi. See through walls with wifi! SIGCOMM '13.

[2] Y. Ben-Shabat, O. Shrout, and S. Gould. 3dinaction: Understanding human actions in 3d point clouds. In *CVPR*, pages 19978–19987, 2024.

[3] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black. Monocular expressive body regression through body-driven attention. In *ECCV*, volume 12355, pages 20–40. Springer, 2020.

[4] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca. Toyota smarthome: Real-world activities of daily living. In *ICCV*, pages 833–842, 2019.

[5] H. Ding, Z. Chen, C. Zhao, F. Wang, G. Wang, W. Xi, and J. Zhao. Mi-mesh: 3d human mesh construction by fusing image and millimeter wave. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):1–24, 2023.

[6] P. Duan, X. Diao, Y. Cao, D. Zhang, B. Zhang, and J. Kong. A comprehensive survey on wi-fi sensing for human identity recognition. *Electronics*, 12(23), 2023.

[7] M. Han, L. Wang, L. Xiao, H. Zhang, C. Zhang, X. Xu, and J. Zhu. Quickfps: Architecture and algorithm co-design for farthest point sampling in large-scale point clouds. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(11):4011–4024, 2023.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su. Towards 3d human pose construction using wifi. In *MobiCom*, 2020.

[10] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su. Towards 3d human pose construction using wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2020.

[11] H. Jin, J. Wang, Z. Yang, S. Kumar, and J. Hong. Rf-wear: Towards wearable everyday skeleton tracking using passive rfids. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 369–372, 2018.

[12] H. Jin, Z. Yang, S. Kumar, and J. I. Hong. Towards wearable everyday body-frame tracking using passive rfids. *UbiComp*, 1(4):1–23, 2018.

[13] L. H. Kim and S. Follmer. Ubiswarm: Ubiquitous robotic interfaces and investigation of abstract motion as a display. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, 2017.

[14] H. Lai, G. Luo, Y. Liu, and M. Zhao. Enabling visual recognition at radio frequency. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 388–403, 2024.

[15] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR Workshops*, pages 9–14, 2010.

[16] X. Li, Q. Huang, Z. Wang, Z. Hou, T. Yang, Q. Li, Z. Huang, T. Wang, and Y. Yang. Sequentialpointnet: A strong parallelized point cloud sequence classification network for 3d action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, pages 1232–1240, 2021.

[17] C. Liu, Z. Dong, L. Huang, W. Yan, X. Wang, D. Fang, and X. Chen. Tagsleep3d: Rf-based 3d sleep posture skeleton recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–28, 2024.

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015.

[19] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[20] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, Oct. 2019.

[21] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10967–10977, 2023.

[22] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black. BABEL: Bodies, action and behavior with english labels. In *CVPR*, pages 722–731, June 2021.

[23] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, pages 5099–5108, 2017.

[24] T. Ropitault, S. Blandino, A. Sahoo, and N. T. Golmie. Ieee 802.11bf: Enabling the widespread adoption of wi-fi sensing. 2023-05-31 04:05:00 2023.

[25] M. Saqlain, D. Kim, J. Cha, C. Lee, S. Lee, and S. Baek. 3dmesh-gar: 3d human body mesh-based method for group activity recognition. *Sensors*, 22(4), 2022.

[26] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[27] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *CVPR*, pages 1010–1019, 2016.

[28] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

[29] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. *arXiv:1212.0402*, 2012.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] C. Wang, J. Liu, Y. Chen, L. Xie, H. B. Liu, and S. Lu. 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.

[32] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015.

[33] S. Wang, L. Mei, R. Liu, W. Jiang, Z. Yin, X. Deng, and T. He. Multi-modal fusion sensing: A comprehensive review of millimeter-wave radar and its integration with other modalities. *IEEE Communications Surveys & Tutorials*, 27(1):322–352, 2025.

[34] Y. Wang, Y. Ren, Y. Chen, and J. Yang. Wi-mesh: A wi-fi vision-based approach for 3d human mesh construction. In *SenSys*, page 362, 2022.

[35] Y. Wang, Y. Zhang, Z. Huang, and L. V. Gool. 3dv: 3d dynamic voxel for action recognition in depth video. In *CVPR*, pages 508–517, 2020.

[36] Y. Xie, J. Xiong, M. Li, and K. Jamieson. md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking. MobiCom '19.

[37] H. Xue, Q. Cao, Y. Ju, H. Hu, H. Wang, A. Zhang, and L. Su. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 391–406, 2022.

[38] H. Xue, Q. Cao, C. Miao, Y. Ju, H. Hu, A. Zhang, and L. Su. Towards generalized mmwave-based human pose estimation through signal augmentation. ACM MobiCom '23, New York, NY, USA, 2023. Association for Computing Machinery.

[39] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In *MobiSys*, pages 14–26, 2021.

[40] J. Zhang, D. Wang, X. An, M. Lv, D. Chen, and A. Sun. A voxel-based 3d reconstruction and action recognition method for construction workers. *Advanced Engineering Informatics*, 65:103203, 2025.

[41] H. Zhao, L. Jiang, J. Jia, P. H. S. Torr, and V. Koltun. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *CVPR*, pages 14149–14158, 2021.

[42] H. Zhao, L. Jiang, J. Jia, P. H. S. Torr, and V. Koltun. Pstnet: Point spatio-temporal convolution on point cloud sequences. *arXiv preprint arXiv:2205.13713*, 2022.

[43] M. Zhao, Y. Liu, A. Raghu, T. Li, H. Zhao, A. Torralba, and D. Katabi. Through-wall human mesh recovery using radio signals. In *ICCV*, 2019.

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

[45] X. Zhu, P.-Y. Huang, J. Liang, C. M. de Melo, and A. Hauptmann. Stmt: A spatial-temporal mesh transformer for mocap-based action recognition. In *CVPR*, pages 1526–1536, 2023.