OctoNet: A Large-Scale Multi-Modal Dataset for Human Activity Understanding Grounded in Motion-Captured 3D Pose Labels

Dongsheng Yuan*, **Xie Zhang***, **Weiying Hou**, **Sheng Lyu**, **Yuemin Yu**,
Luca Jiang-Tao Yu, **Chengxiao Li**, **Chenshu Wu**†
Department of Computer Science, The University of Hong Kong
Project Website: `https://aiot-lab.github.io/OctoNet/`
Dataset: `https://huggingface.co/datasets/hku-aiot/OctoNet`

Abstract

We introduce **OctoNet**, a large-scale, multi-modal, multi-view human activity dataset designed to advance human activity understanding and multi-modal learning. OctoNet comprises 12 heterogeneous modalities (including RGB, depth, thermal cameras, infrared arrays, audio, millimeter-wave radar, Wi-Fi, IMU, and more) recorded from 41 participants under multi-view sensor setups, yielding over 67.72M synchronized frames. The data encompass 62 daily activities spanning structured routines, freestyle behaviors, human-environment interaction, healthcare tasks, etc. All modalities are annotated by high-fidelity 3D pose labels captured via a professional motion-capture system, allowing precise alignment and rich supervision across sensors and views. OctoNet is one of the most comprehensive datasets of its kind, enabling a wide range of learning tasks such as human activity recognition, 3D pose estimation, multi-modal fusion, cross-modal supervision, and sensor foundation models. Extensive experiments have been conducted to demonstrate the sensing capacity using various baselines. OctoNet offers a unique and unified testbed for developing and benchmarking generalizable, robust models for human-centric sensing AI.

1 Introduction

Understanding human activity is fundamental for embodied AI, as it forms the foundation for intelligent systems that can seamlessly interact with and navigate the physical world [56]. Accurate modeling, perception, and interpretation of human behaviors are essential for developing AI agents that collaborate with humans [63], assist in real-world tasks [44, 53], and adapt to practical environments [10, 68].

Despite growing interest in human-centric AI, much of today's embodied and perceptual AI is dominated by vision-first paradigms [13, 32]. However, the physical world is far more sensor-rich than a camera lens, with a rich spectrum of sensing signals available in real-world environments, such as Radio Frequency (RF) (*e.g.*, Wi-Fi, millimeter-wave radars, UWB), inertial measurement units (IMUs), and thermal sensors. These non-visual modalities offer unique and complementary information that is particularly critical in poor-lighting, occluded, and privacy-sensitive scenarios. Despite their proven potential, learning across these diverse modalities remains largely underexplored. This limitation is primarily compounded by the lack of large, unified benchmarks across diverse, heterogeneous modalities, which fundamentally hinders progress in several key areas: 1) Multi-modal

*Equal contribution.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

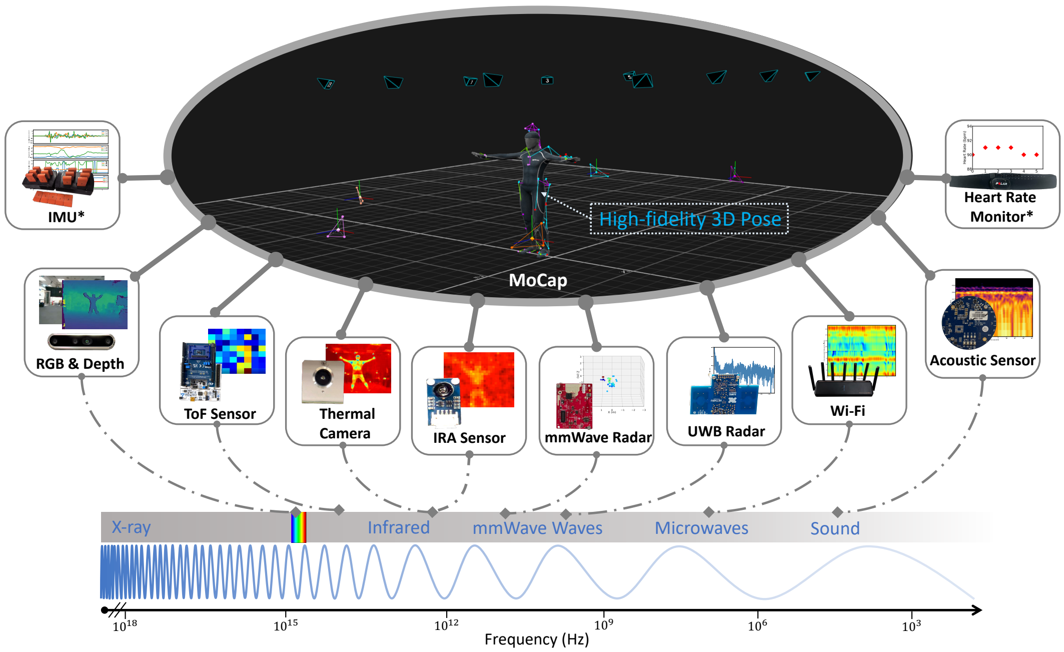


Figure 1: Overview of sensing modalities in **OctoNet**. The system integrates wearable sensors (marked with *) and non-contact modalities spanning the frequency spectrum, unified through high-fidelity 3D poses from a professional motion-capture system. These poses provide an explicit representation to align and correlate multi-modal data streams.

fusion: most current efforts focus on vision-language models, leaving other sensing modalities underutilized; 2) Cross-modal understanding: learning relationships across sensing types is often infeasible without extensive, aligned data collection; 3) Sensing foundation models: Building foundation models for non-visual sensors heavily relies on massive, diverse multi-modal datasets. Moreover, the scarcity of such data also restricts exploration in areas like cross-modal data generation, modality translation, and robust perception under varying environmental and sensory configurations.

To address these challenges, we introduce **OctoNet**, a large-scale, multi-modal dataset for human activity understanding. OctoNet features 62 diverse activity classes, spanning both structured tasks (*e.g.*, falling down, dancing, drawing zigzag, programmed aerobics) and freestyle actions (*e.g.*, impromptu sports, random walk), performed by 41 participants. The data were captured simultaneously using 12 heterogeneous sensing modalities (Figure 1), yielding over 67.72 million synchronized frames. Additionally, OctoNet provides high-fidelity 3D skeletal pose annotations obtained via an OptiTrack motion-capture system [45], offering precise ground truth for human activities. By aligning a diverse range of signals, OctoNet supports a wide spectrum of research tasks, including human activity recognition, 3D pose estimation, multi-modal fusion, cross-modal alignment, and the development of foundation models for physical sensing. The key contributions and features of OctoNet are listed as follows:

❶ **Comprehensive perception modalities:** As shown in Figure 1, to the best of our knowledge, OctoNet is the first dataset that comprehensively covers 12 distinct data modalities encompassing a wide spectrum of electromagnetic (*e.g.*, RGB-D, ToF, thermal, infrared, mmWave, UWB, Wi-Fi) and non-electromagnetic (*e.g.*, acoustic, inertial, physiological) signals to record human activities.

❷ **Precision poses as the label:** Besides the activity labels, we integrate high-fidelity 3D poses captured via a motion-capture system as additional labels for human activities. These pose labels serve as fundamental and explicit common representations, offering deep insights into understanding human activities and enhancing generalizability.

❸ **Large-scale and diverse coverage:** To the best of our knowledge, OctoNet represents the largest human activity dataset to date for several modalities, including thermal, IRA, and ToF, and ranks among the largest for others such as Wi-Fi and UWB/mmWave radars. This scale, combined with

Table 1: Summary of existing single- and multi-modality datasets. OctoNet provides both action labels and high-fidelity 3D whole-body keypoint (3DKP) annotations. #Frames: total frames across all modalities; *: RGB-only datasets (no depth).

Dataset	Modalities											Annotations		#Subj	#Act	#Seq	#Frame
	RGB-D	ToF	Thermal	IRA	mmWave	UWB	Wi-Fi	Acoustic	IMU	HR	MoCap	3DKP	Action				
CMU Panoptic [25]	✓	-	-	-	-	-	-	-	-	-	-	✓	-	8	5	65	154M
NTU RGB+D [58]	✓	-	-	-	-	-	-	-	-	-	-	✓	✓	40	60	56k	4M
Kinetics-700 [5]	✓*	-	-	-	-	-	-	-	-	-	-	-	✓	-	700	650.3k	-
KAIST-MP [19]	✓*	-	✓	-	-	-	-	-	-	-	-	-	-	1.18k	-	-	95.3k
PETS2017 [49]	✓*	-	✓	-	-	-	-	-	-	-	-	-	✓	-	10	36	-
CAMEL [14]	✓*	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	26	44.5k
RF-Pose3D [83]	✓*	-	-	-	✓	-	-	-	-	-	-	✓	✓	>5	5	-	-
mmMesh [77]	✓*	-	-	-	✓	-	-	-	-	-	-	✓	✓	20	8	-	3k
mmBody [6]	✓	-	-	-	✓	-	-	-	-	-	✓	✓	✓	20	100	-	200k
Bocus UWB [3]	✓*	-	-	-	-	✓	-	-	-	-	-	-	✓	1	3	-	2.9M
Widar 3.0 [84]	-	-	-	-	-	-	✓	-	-	-	-	-	✓	17	22	-	17.8k
WiPose [24]	✓*	-	-	-	-	-	✓	-	-	-	-	✓	✓	10	16	-	96k
GoPose [55]	✓*	-	-	-	-	-	✓	-	-	-	-	✓	✓	10	>9	-	676.2k
Ubicoustics [27]	-	-	-	-	-	-	-	✓	-	-	-	-	✓	12	30	-	-
SAMoSA [41]	-	-	-	-	-	-	-	✓	✓	-	-	-	✓	20	26	1560	-
UTD-MHAD [7]	✓	-	-	-	-	-	-	-	✓	-	-	✓	✓	8	27	861	-
USC-HAD [81]	-	-	-	-	-	-	-	-	✓	-	-	-	✓	14	12	840	-
Total Capture [66]	✓*	-	-	-	-	-	-	-	✓	-	✓	✓	✓	5	4	60	1.9M
Stanford-ECM [42]	✓*	-	-	-	-	-	-	-	✓	✓	-	-	✓	10	24	113	-
Opportunity++ [9]	✓*	-	-	-	-	✓	-	-	✓	-	-	-	✓	4	43	24	-
mRI [2]	✓	-	-	-	✓	-	-	-	✓	-	-	✓	✓	20	12	300	160k
OPERAnet [4]	✓	-	-	-	-	✓	✓	-	-	-	-	✓	✓	6	6	61	-
MM-Fi [78]	✓	-	-	-	✓	-	✓	-	-	-	-	✓	✓	40	27	1080	320.8k
XRF55 [70]	✓	-	-	-	✓	-	✓	-	-	-	-	-	✓	39	55	42.9k	-
OctoNet (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	41	62	8.76k	67.72M

its comprehensive modality coverage, enables a wide range of learning paradigms, from supervised to self-supervised and cross-modal learning, and fosters the development of models that generalize across subjects, activities, sensing conditions, and modalities.

2 Related work

Single modality datasets. Many existing datasets for human activity recognition are confined to a single modality, including RGB-D [5, 22, 25, 58], thermal [14, 19, 49], acoustic [12, 15, 27, 71], IMUs [7, 46, 81], and RF radars [3, 83, 84]. While each of these datasets demonstrates strengths within its specific field, they are hampered by inherent limitations of the sensors used, such as privacy concerns, occlusion issues, signal drifting, and constraints in spatial resolutions.

Multi-modality datasets. Recently, an increasing number of datasets have sought to integrate multiple modalities to better understand human behaviors. These datasets often feature core modalities such as RGB-D and IMU, while incorporating additional signals to enrich data representation. For example, Total Capture [66] combines RGB-D and IMU data, Stanford-ECM [42] supplements this with heart rate signals, and mRI [2] includes mmWave data. Some studies further incorporate various RF modalities, such as MM-Fi [78], XRF55 [70], and OPERAnet [4]. As highlighted in Table 1, existing multi-modal datasets typically provide limited subsets of available sensing modalities. Additionally, RF-based datasets [2, 9, 42, 66, 78] suffer from limited participants and limited sets of meaningful activities. Moreover, reliance on RGB-D data for human pose estimation [2, 78] can be susceptible to occlusions and inaccuracies.

To this end, we propose an extensive and highly integrated multi-modal dataset that encompasses a full spectrum of sensing modalities. As detailed in Table 1, we aim to deliver an all-in-one solution that comprehensively captures human activities. We also provide high-fidelity 3D poses as labels,

Table 2: Modality-specific dataset statistics. The data dimension column indicates the shape of data *after* preprocessing for our training pipeline. The raw data are provided in the released dataset. †: The "×3" indicates three color channels (RGB). ✪: For FMCW, the first dimension represents the # points, with 150 as the maximum number. ▲: We preprocess the acoustic data into Mel-Spectrograms.

Modality	Total Frames	Sampling Rate (Hz)	Number of Nodes	Data Dimension (per frame)	Storage Size (GB)
RGB-D	7.82M	29.95	3	$480 \times 640\ (\times 3)^{\dagger}$	**522.45**
ToF	645.08k	7.32	1	$8 \times 8 \times 18$	**6.03**
Thermal	1.50M	8.80	2	240×320	**42.51**
IRA	3.02M	6.91	5	24×32	**18.03**
mmWave (FMCW)✪	3.74M	8.81	5	150×4	**5.52**
mmWave (SFCW)	280.28k	3.20	1	400×100	**167.16**
UWB	1.49M	17.07	1	1×1535	**19.34**
Wi-Fi	27.35M	75.62	4	2×114	**94.85**
Acoustic▲	5.39M	48000	2	1×128	**15.46**
IMU	5.42M	60.01	17	13×17	**9.02**
Heart Rate	90.10k	1.03	1	1	**0.007**
MoCap	10.97M	120	50	20×3	**82.04**

enhancing generalization capabilities and enabling versatile approaches to human activity recognition without reliance on predefined sets. Furthermore, we carefully select the human activities involving body-motion, human-object interaction, human-computer interaction, human-human interactions and medical conditions, which are tailored for broad real-world applications.

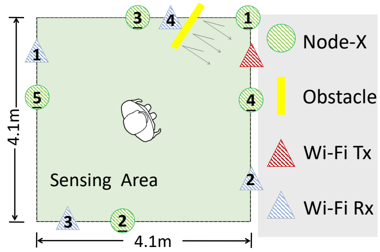
3 Data collection platform

3.1 Modality overview

Visual-related modalities. We adopt RGB-D cameras, time-of-flight sensors (ToF), thermal cameras, and infrared array sensors (IRA). Specifically, we use three Intel RealSense D455C cameras [1] that employ stereoscopic depth sensing to capture RGB and depth frames at an average frame rate of 29.95 Hz. We deploy a Single Photon Avalanche Diode (SPAD) sensor (STMicroelectronics VL53L8CH [61]) that measures distance by emitting modulated infrared pulses and timing their returns [28]. For thermal imaging, two Seek Thermal S304SP Mosaic Core cameras [57] are equipped with uncooled microbolometers to capture thermal images. We also employ five MLX90640 infrared arrays [37] that convert captured infrared radiation into approximate temperature readings.

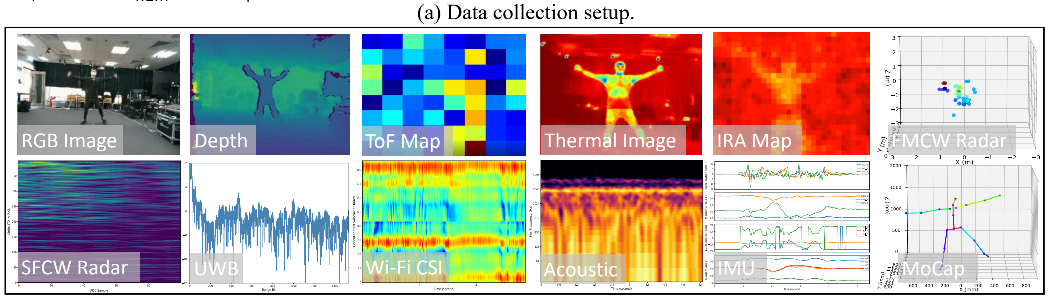
Radio-Frequency (RF) signals. We incorporate two types of millimeter-wave (mmWave) radars with different modulation schemes, *i.e.*, Frequency-Modulated Continuous Wave (FMCW) and Stepped-Frequency Continuous Wave (SFCW). For FMCW, we utilize five Texas Instruments (TI) IWR1843Boost mmWave radars [65] to capture three-dimensional point-cloud data. For SFCW, we use a Vayyar IMAGEVK-74 radar [67] with a bandwidth of 4 GHz and 20 transmitter and 20 receiver antennas. A Novelda XeThru X4M200 Ultra-Wideband (UWB) radar [43] is also employed, which has a bandwidth of 2.5 GHz and provides a maximum detection range of 9.9 m. Moreover, we integrate Wi-Fi sensing for its best ubiquity. We use a Xiaomi AX6000 router [76] as the transmitter and four Raspberry Pi Compute Module 4 devices (with Intel AX200 NICs) [20, 54] as receivers. Packets are sent from the transmitter through one antenna to each of four receivers on channel 36 (5.18 GHz) with a bandwidth of 40 MHz. Each receiver monitors this channel separately using two antennas, resulting in a total of 8 Wi-Fi links.

Others. We also capture acoustic, inertial, and physiological data. We employ a MiniDSP UMA-8-SP USB microphone array [39] and a UMIK-2 microphone [40] to capture the acoustic events in the environment. To enable active acoustic sensing [64, 69, 82], we also deploy a speaker that emits sounds at inaudible frequencies [35] simultaneously. For inertial tracking, we include an Xsens Awinda Research Kit [50], comprising 17 MTw Awinda wireless motion trackers. Besides, heart rate data are collected using a Polar H10 heart rate sensor [51]. The sensor is worn as a chest strap, ensuring reliable and consistent measurements throughout the data collection period.

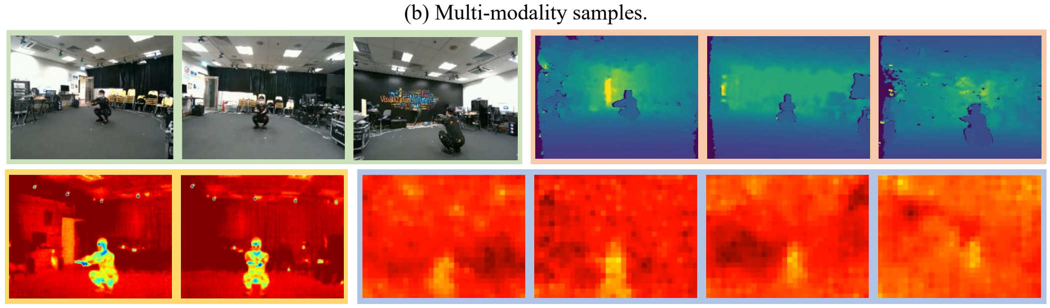


Modality	Node-X 1	2	3	4	5
RGB-D	✓	-	✓	-	✓
ToF	-	-	-	✓	-
Thermal	✓	-	✓	-	-
IRA	✓	✓	✓	✓	✓
FMCW	✓	✓	✓	✓	✓
SFCW	-	-	✓	-	-
UWB	✓	-	-	-	-
Microphone	✓	-	-	✓	-
Speaker	-	-	-	✓	-
HR	✓	-	-	-	-

(a) Data collection setup.



(b) Multi-modality samples.



(c) Multi-view samples.

Figure 2: Overview of our multi-modal data collection setup. (a) Top-down layout, node configuration, and data collection scenes. (b) Representative raw data samples from different modalities. (c) Synchronized multi-view captures illustrating concurrent data collection.

Motion-capture system (MoCap). We employ an OptiTrack motion-capture system [45] with 12 Prime^x 13 cameras. 50 markers are attached to the human body to reconstruct the human skeleton. We summarize the detailed modality information in Table 2.

3.2 Node-wise modality deployment

As illustrated in Figure 2(a), the overall configuration consists of five customized nodes, with specific available modalities provided in the corresponding table. Node-1 is positioned to face participants directly and integrates the most modalities on a single mini-PC platform. The remaining nodes share a similar hardware architecture but are equipped with different sets of modalities. Following the typical data collection procedure for Wi-Fi sensing [84], we arrange four Wi-Fi receivers in a rectangular configuration. Notably, one Wi-Fi receiver (receiver 4) is deliberately obstructed using a 3 cm-thick wooden board, blocking its direct path to the Wi-Fi transmitter and creating a Non-Line-Of-Sight (NLOS) condition. Furthermore, we attach three reflective markers to each node (labeled Node-1–5). These markers are tracked by the motion-capture system to obtain the precise locations of the nodes.

3.3 Time synchronization

To ensure temporal alignment across all nodes and modalities during data collection, we utilize a synchronization mechanism based on the Network Time Protocol (NTP). A master mini-PC serves as the central NTP server, distributing a global timestamp reference for all nodes. Each modality independently aligns its local time with this shared reference, ensuring temporal consistency across data streams. However, constantly sending synchronization signals would strain the network resources and potentially introduce latency. To mitigate this, we implement a one-time broadcast of a reference start timestamp from an NTP server.

4 Dataset construction

Subjects. We recruit 41 participants (22 males, 19 females), aged 19–70 years, with heights ranging from 151 to 192 cm and weights from 41 to 99 kg. The cohort includes individuals of British, Chinese, European, and Indian backgrounds, ensuring broad demographic representation in terms of body types, movement patterns, and ethnicity. Before each session, participants read and sign the consent form in accordance with the approved protocol. We inform them of the research goals, data collection procedures, and any potential risks. Our team assists participants in attaching the necessary markers. An on-site coordinator oversees the sessions, managing the start and end of each experiment. We provide instructional slides if a participant is unfamiliar with a particular activity. Participation is entirely voluntary. Each session lasts approximately 1 hour, and we provide equivalent compensation of 12 USD, which exceeds the local minimum wage. This study is approved by our university's Institutional Review Board (IRB), and detailed IRB approval and user consent are provided in Appendix A.

Multiple scenes. We conduct our data collection sessions in three different settings, as shown in Figure 1(a), designed to reflect different real-world environments: an office (Scene 1), a laboratory (Scene 2), and a living room (Scene 3). We deploy OptiTrack in each scene for motion capture.

Activity categories. We attentively curate the activity sets to select the most representative activities encountered in daily life. We divide the activities into two main categories as follows.

62 Activities. After thoroughly investigating the existing datasets, we select the 62 most representative activities. These activities are further grouped into five subcategories based on interaction contexts: body-motion only [5, 7, 16, 21, 23, 26, 58–60, 70, 72, 73, 75], human-object interaction [16, 26, 27, 29, 58, 62, 70, 73], human-computer interaction [7, 30, 31, 36, 52, 58, 70, 72, 74, 79, 80, 84], human-human interaction [5, 15, 16, 26, 52, 58, 70], and medical conditions [30, 38, 58]. This taxonomy covers a broad range of human actions from prior work, enabling comprehensive evaluation across daily activities, social interactions, human–device engagement, and healthcare scenarios. We also ensure a balanced class distribution, with each class comprising 1.45–1.62% of the samples (average 1.61%, *i.e.*, 1/62). Detailed activity definitions are provided in Appendix D.

Programmed aerobics and freestyle. To capture both structured and spontaneous movements, the second category includes a standardized aerobics routine and freestyle activities. The programmed aerobics sequence consists of synchronized, full-body movements in a five-minute session, providing structured data for evaluating pose estimation across sequential actions. In the freestyle session, participants perform three to five self-chosen movements, yielding diverse and unstructured data. Together, these components support robust and generalizable modeling of complex human dynamics.

Annotation protocol. During each experimental session, participants repeat the 62 activities continuously with specifically assigned activity labels. Furthermore, we adopt human pose as the additional labels for human activities by leveraging the motion-capture system to obtain 3D Skeletal KeyPoints (3DKP) as the ground truth. The rationale behind this is twofold. First, in addition to activity classification, the dataset enables human pose estimation by providing precise pose ground truth. Second, since pose provides a fundamental and interpretable representation of human motion, learning from pose data could facilitate few-shot or zero-shot activity recognition without relying on a predefined set of actions, thereby greatly enhancing generalizability for future studies.

5 Benchmark and evaluation

To demonstrate the practical utility of OctoNet, we establish baseline results on two key tasks: human activity recognition (HAR) and 3D human pose estimation (HPE). We conduct experiments on individual sensing modalities using standard network architectures from both the vision and sensor domains. Our evaluations do not aim for optimized performance, but instead provide baselines to demonstrate the dataset's effectiveness and capacity. To further validate the dataset, we evaluate both multi-modal fusion methods and recent representative approaches for RGB, IMU, and acoustic modalities on HAR. We detail our evaluation protocols, metrics, and models below, and summarize the results in Tables 3, 4, 5, and 6. Further details are provided in the supplementary materials.

5.1 Evaluation protocols

Task setting. To balance the comprehensiveness and computational efficiency, HAR is evaluated under two configurations: (1) a curated 10-class subset and (2) the full 62-class activity set. The 10 selected actions span locomotion, gestures, and interactions (*i.e., sit, walk, bow, dance, fall down, jump, draw zigzag, draw circle clockwise, kick someone, and push someone*), providing a compact yet diverse benchmark for modality comparison. The full 62-class setup introduces greater variability and complexity, approximating real-world recognition scenarios. HPE, in contrast, focuses on fine-grained spatial reconstruction by estimating 3D skeletal keypoints from the same inputs. Together, these two tasks offer complementary views of human motion—semantic recognition and spatial reconstruction.

Test set setting. To comprehensively assess the robustness and generalization of the model, we apply three test set settings across both HAR and HPE: (1) *In-Domain (ID):* Training and testing are conducted on data from the same pool of users and scenes. Specifically, data from Scene 1 and Scene 2 are used with a 7:1:2 split for training, validation, and testing. (2) *Cross-Scene (CS):* Models trained in-domain are evaluated on Scene 3, which remains entirely unseen during training, to assess robustness to environmental variations. (3) *Cross-User (CU):* Models are trained on a subset of users and tested on unseen users from Scene 1, evaluating subject-level generalization. This unified setup enables systematic assessment of how different sensing modalities perform under domain shifts, and highlights key challenges in multi-modal robustness and adaptation.

5.2 Evaluation metrics

HAR. Following the HAR benchmarks [70], we use top-1 classification accuracy as the metric.

HPE. For HPE, we report the Mean Per Joint Position Error (MPJPE)—the average Euclidean distance between predicted and ground-truth 3D joint coordinates. Ground-truth poses are captured by the OptiTrack system, and evaluation is performed on 20 consistently annotated joints, with MPJPE measured in millimeters across all settings.

5.3 Baseline methods

We evaluate HAR and HPE performance using four widely adopted architectures: ResNet [17], DenseNet [18] and Swin-T [33], commonly used in visual tasks, and RFNet [11] tailored for RF signals. All the above models are trained from scratch and adapted as necessary to modality-specific input formats. We intentionally select common architectures rather than SOTA models specific to each modality. This decision is motivated by two factors. First, several modalities, especially IRAs and ToF sensors, are relatively underexplored, with no well-established models available. Second, we aim to ensure a consistent and fair comparison across modalities by using architectures with flexible input handling and minimal modality-specific engineering. This approach provides a unified and interpretable baseline that future work can build upon. For multi-modal fusion, we employ the above architectures as backbones and concatenate intermediate features for fusion. At the same time, we benchmark our dataset on the recent representative approaches used for comparison, including Video Swin [34] (RGB), CALANet [47] (IMU), and HTS-AT [8] (acoustic).

5.4 Results and analysis

Human activity recognition. Table 3 reports HAR accuracy across all sensing modalities, models, and protocols for both the 10-class and 62-class settings. **Vision-based modalities** (RGB and Depth) achieve high accuracy, especially in the 10-class task. RGB with Swin-T reaches 94.9% (10-class) and 93.1% (62-class), while Depth yields 86.3% (10-class) and 81.7%(62-class), confirming the strength of visual information for activity recognition. **IMU** and **UWB** also perform strongly and consistently. IMU achieves 98.3% on the 10-class and 95.7% on the full 62-class, directly capturing motion with minimal environmental interference. UWB attains 98.3% (ResNet) in-domain accuracy and remains robust under cross-scene and cross-user conditions, reflecting its strong sensing capability.

ToF and **Thermal** perform competitively, with ToF (RFNet) reaching 89.3%/75.9% and Thermal (DenseNet) 91.7%/85.4%. Both highlight the potential of privacy-preserving human sensing. In contrast, **IRA**, **Acoustic**, and **SFCW** radar show lower performance—IRA (25.6%) suffers from coarse resolution and garment-induced attenuation, while Acoustic and SFCW achieve 40–60% due to noise and weak discriminative cues. **Cross-domain evaluations** (CU and CS) show consistent performance drops, underscoring domain-shift challenges. In the 10-class setting, RGB falls from 94.9% to 37.0% (CU) and 12.1% (CS) with Swin-T, indicating reliance on scene-specific cues. IMU generalizes better, retaining 82.9% and 47.5% in CU and CS. This suggests motion-coupled sensors are less affected by environment than vision or thermal modalities. Overall, the results reveal a clear modality hierarchy: visual and wearable sensors perform best, RF-based modalities show moderate yet promising results, while low-resolution thermal and acoustic signals are weakest. The consistent domain-shift degradation highlights OctoNet's value as a benchmark for multi-modal robustness, generalization, and fusion.

Comparisons of representative approaches. As shown in Table 4, we evaluate recent representative approaches for three commonly used modalities. For RGB data, we implement **Video-Swin** [34], a state-of-the-art vision transformer for video-based activity recognition, trained from scratch. It achieves 93.2% accuracy on the 10-class subset and 91.3% on the full 62-class setting, confirming the strong visual discriminability of our dataset. For IMU signals, we adopt **CALANet** [47], which attains high in-domain performance of 94.9% and 85.1% on the two settings, respectively, highlighting the reliability of wearable sensing for motion characterization. For acoustic data, we use **HTS-AT** [8], pretrained on AudioSet [15]. While it reaches 80.0% accuracy in-domain, its performance drops notably under cross-user and cross-scene conditions, similar to the other approaches.

Modality fusion. To evaluate the effectiveness of cross-modality fusion, we conduct experiments on the 10-class subset using representative modality combinations. As shown in Table 5, fusion substantially improves in-domain performance across all configurations, with accuracies approaching 99%. The fusion of **Thermal, IRA, and IMU** achieves the highest robustness, maintaining 75.2% under cross-user and 52.7% under cross-scene evaluations, demonstrating the complementary strengths of low-cost thermal sensing and wearable motion data. In contrast, fusions involving **visual modalities** (*e.g.*, RGB, FMCW, Acoustic) perform well in-domain but degrade more sharply under domain shifts, indicating stronger dependence on environmental consistency. These results highlight that integrating heterogeneous sensing modalities, particularly those combining motion, temperature, and spatial cues, can significantly enhance generalization and resilience to domain variation.

3D human pose estimation. Table 6 reports MPJPE results across sensing modalities and evaluation protocols (ID, CU, and CS). **Vision-based modalities** (RGB and Depth) achieve the lowest errors under in-domain conditions, with 133.3 mm and 131.4 mm MPJPE using ResNet, respectively. However, their performance degrades sharply under domain shifts. RGB rises to 473.9 mm in the cross-scene setting, reflecting strong dependence on background appearance and lighting. To ensure fairness across all modalities, no preprocessing (e.g., person cropping) is applied to vision inputs. This decision preserves comparability but causes vision models to overfit to domain-specific context, limiting generalization. **RF-based modalities** (UWB, Wi-Fi, FMCW, and SFCW) exhibit heterogeneous robustness. UWB attains competitive in-domain performance (142.4 mm MPJPE with ResNet) and only moderate deterioration across domains. Wi-Fi shows a large domain gap (147.3 mm to 399.4 mm), reflecting pronounced sensitivity to multipath propagation and environmental entanglement. FMCW and SFCW achieve reasonable accuracy under controlled conditions but degrade under scene shifts, consistent with RF sensing's environmental dependence. **Thermal sensing** exhibits superior cross-scene generalization, with MPJPE rising from 142.8 mm to 308.8 mm using ResNet, indicating reduced sensitivity to background and illumination compared with RGB and Depth. **IMU** maintains stable performance across settings (147.9 mm $\rightarrow$ 252.9 mm $\rightarrow$ 289.7 mm), demonstrating the robustness of body-centric measurements. In contrast, **IRA** and **acoustic** produce substantially higher errors (244.0–398.0 mm and 243.6–441.4 mm, respectively), reflecting the difficulty for general-purpose models in extracting reliable spatial cues from low-resolution or indirect signals. Overall, these results highlight distinct generalization behaviors among sensing modalities: vision excels in-domain but is scene-dependent; RF and thermal modalities offer a trade-off between precision and robustness; and egocentric sensors like IMU generalize best across users and environments. Such observations underscore **OctoNet**'s value as a comprehensive benchmark for studying domain shift and cross-modality robustness in 3D human pose estimation.

Table 3: HAR accuracy (%) across modalities, models, and protocols. Results are shown for the 10-class subset (left) and full 62-class setting (right). “N/A” denotes model incompatibility. Accuracy is given to one decimal with the standard error of the mean as $_{x.x}$.

Modality	Protocol	Model			
		ResNet	**DenseNet**	**Swin-T**	**RFNet**
RGB	**ID**	91.5$_{\pm2.6}$ / 93.4$_{\pm0.9}$	93.2$_{\pm2.3}$ / 91.2$_{\pm1.0}$	94.9$_{\pm2.0}$ / 93.1$_{\pm0.9}$	89.7$_{\pm2.8}$ / 60.9$_{\pm1.8}$
	CU	46.0$_{\pm3.4}$ / 12.3$_{\pm0.9}$	68.2$_{\pm3.2}$ / 24.7$_{\pm1.2}$	37.0$_{\pm3.3}$ / 7.7$_{\pm0.7}$	45.0$_{\pm3.4}$ / 9.2$_{\pm0.8}$
	CS	14.9$_{\pm3.0}$ / 4.1$_{\pm0.7}$	33.3$_{\pm4.0}$ / 11.3$_{\pm1.1}$	12.1$_{\pm2.8}$ / 1.7$_{\pm0.4}$	13.5$_{\pm2.9}$ / 3.1$_{\pm0.6}$
Depth	**ID**	89.7$_{\pm2.8}$ / 86.6$_{\pm1.2}$	90.6$_{\pm2.7}$ / 83.2$_{\pm1.3}$	86.3$_{\pm3.2}$ / 81.7$_{\pm1.4}$	87.2$_{\pm3.1}$ / 40.0$_{\pm1.8}$
	CU	41.2$_{\pm3.4}$ / 11.1$_{\pm0.9}$	64.9$_{\pm3.3}$ / 27.3$_{\pm1.2}$	46.0$_{\pm3.4}$ / 14.4$_{\pm1.0}$	45.0$_{\pm3.4}$ / 11.2$_{\pm0.9}$
	CS	17.7$_{\pm3.2}$ / 3.9$_{\pm0.7}$	22.7$_{\pm3.5}$ / 12.2$_{\pm1.1}$	23.4$_{\pm3.6}$ / 4.3$_{\pm0.7}$	28.4$_{\pm3.8}$ / 4.8$_{\pm0.7}$
ToF	**ID**	86.8$_{\pm3.1}$ / 70.3$_{\pm1.6}$	N/A	82.6$_{\pm3.5}$ / 51.8$_{\pm1.8}$	89.3$_{\pm2.8}$ / 75.9$_{\pm1.5}$
	CU	44.5$_{\pm3.4}$ / 11.8$_{\pm0.9}$	N/A	46.4$_{\pm3.4}$ / 15.3$_{\pm1.0}$	78.7$_{\pm2.8}$ / 28.3$_{\pm1.2}$
	CS	25.5$_{\pm3.7}$ / 8.0$_{\pm0.9}$	N/A	22.7$_{\pm3.5}$ / 4.7$_{\pm0.7}$	44.7$_{\pm4.2}$ / 18.6$_{\pm1.3}$
Thermal	**ID**	90.1$_{\pm2.7}$ / 85.0$_{\pm1.3}$	91.7$_{\pm2.5}$ / 85.4$_{\pm1.3}$	85.1$_{\pm3.2}$ / 79.2$_{\pm1.5}$	47.1$_{\pm4.6}$ / 28.6$_{\pm1.6}$
	CU	50.2$_{\pm3.5}$ / 25.7$_{\pm1.2}$	64.5$_{\pm3.4}$ / 32.5$_{\pm1.3}$	46.8$_{\pm3.5}$ / 15.6$_{\pm1.0}$	15.3$_{\pm2.5}$ / 1.0$_{\pm0.3}$
	CS	36.9$_{\pm4.1}$ / 13.4$_{\pm1.2}$	44.0$_{\pm4.2}$ / 21.0$_{\pm1.4}$	36.2$_{\pm4.1}$ / 10.1$_{\pm1.0}$	17.7$_{\pm3.2}$ / 2.1$_{\pm0.5}$
IRA	**ID**	25.6$_{\pm4.0}$ / 1.8$_{\pm0.5}$	N/A	14.0$_{\pm3.2}$ / 3.7$_{\pm0.7}$	19.0$_{\pm3.6}$ / 4.2$_{\pm0.7}$
	CU	19.9$_{\pm2.8}$ / 2.6$_{\pm0.4}$	N/A	22.3$_{\pm2.9}$ / 2.8$_{\pm0.4}$	21.8$_{\pm2.8}$ / 3.2$_{\pm0.5}$
	CS	18.4$_{\pm3.3}$ / 0.8$_{\pm0.3}$	N/A	20.6$_{\pm3.4}$ / 3.8$_{\pm0.6}$	21.3$_{\pm3.5}$ / 2.7$_{\pm0.6}$
FMCW	**ID**	39.3$_{\pm4.5}$ / 24.0$_{\pm1.6}$	74.4$_{\pm4.1}$ / 46.3$_{\pm1.8}$	36.8$_{\pm4.5}$ / 5.0$_{\pm0.8}$	38.5$_{\pm4.5}$ / 12.6$_{\pm1.2}$
	CU	27.0$_{\pm3.1}$ / 8.9$_{\pm0.8}$	44.1$_{\pm3.4}$ / 16.1$_{\pm1.0}$	24.2$_{\pm3.0}$ / 4.4$_{\pm0.6}$	26.5$_{\pm3.0}$ / 7.2$_{\pm0.7}$
	CS	26.0$_{\pm4.3}$ / 5.3$_{\pm1.0}$	14.4$_{\pm3.5}$ / 7.5$_{\pm1.2}$	14.4$_{\pm3.5}$ / 3.6$_{\pm0.8}$	26.0$_{\pm4.3}$ / 4.3$_{\pm0.9}$
SFCW	**ID**	30.6$_{\pm4.2}$ / 9.0$_{\pm1.0}$	59.5$_{\pm4.5}$ / 13.0$_{\pm1.2}$	26.4$_{\pm4.0}$ / 0.9$_{\pm0.3}$	28.1$_{\pm4.1}$ / 5.1$_{\pm0.8}$
	CU	12.3$_{\pm2.3}$ / 1.6$_{\pm0.3}$	4.3$_{\pm1.4}$ / 1.2$_{\pm0.3}$	7.6$_{\pm1.8}$ / 1.6$_{\pm0.3}$	13.3$_{\pm2.3}$ / 2.2$_{\pm0.4}$
	CS	11.3$_{\pm2.7}$ / 2.5$_{\pm0.5}$	15.6$_{\pm3.1}$ / 1.5$_{\pm0.4}$	7.8$_{\pm2.3}$ / 1.6$_{\pm0.4}$	17.0$_{\pm3.2}$ / 1.5$_{\pm0.4}$
UWB	**ID**	98.3$_{\pm1.2}$ / 93.8$_{\pm0.9}$	88.4$_{\pm2.9}$ / 80.1$_{\pm1.4}$	100.0$_{\pm0.0}$ / 90.4$_{\pm1.1}$	94.2$_{\pm2.1}$ / 75.8$_{\pm1.5}$
	CU	62.6$_{\pm3.3}$ / 21.5$_{\pm1.1}$	59.7$_{\pm3.4}$ / 27.4$_{\pm1.2}$	17.1$_{\pm2.6}$ / 2.7$_{\pm0.4}$	64.5$_{\pm3.3}$ / 13.5$_{\pm0.9}$
	CS	27.0$_{\pm3.7}$ / 6.7$_{\pm0.8}$	20.6$_{\pm3.4}$ / 6.3$_{\pm0.8}$	21.3$_{\pm3.5}$ / 2.4$_{\pm0.5}$	12.1$_{\pm2.8}$ / 1.7$_{\pm0.4}$
Wi-Fi	**ID**	93.3$_{\pm2.3}$ / 91.1$_{\pm1.0}$	90.8$_{\pm2.6}$ / 91.0$_{\pm1.0}$	91.7$_{\pm2.5}$ / 92.3$_{\pm1.0}$	81.7$_{\pm3.5}$ / 60.5$_{\pm1.8}$
	CU	13.3$_{\pm2.3}$ / 3.4$_{\pm0.5}$	11.4$_{\pm2.2}$ / 4.8$_{\pm0.6}$	12.3$_{\pm2.3}$ / 2.3$_{\pm0.4}$	19.9$_{\pm2.8}$ / 4.3$_{\pm0.6}$
	CS	19.1$_{\pm3.3}$ / 2.4$_{\pm0.5}$	11.3$_{\pm2.7}$ / 1.9$_{\pm0.5}$	13.5$_{\pm2.9}$ / 2.8$_{\pm0.6}$	11.3$_{\pm2.7}$ / 1.1$_{\pm0.4}$
Acoustic	**ID**	40.8$_{\pm4.5}$ / 45.5$_{\pm1.8}$	60.0$_{\pm4.5}$ / 54.6$_{\pm1.8}$	36.7$_{\pm4.4}$ / 32.1$_{\pm1.7}$	29.2$_{\pm4.2}$ / 19.1$_{\pm1.4}$
	CU	37.0$_{\pm3.3}$ / 19.9$_{\pm1.1}$	42.7$_{\pm3.4}$ / 16.4$_{\pm1.0}$	27.5$_{\pm3.1}$ / 8.4$_{\pm0.8}$	20.4$_{\pm2.8}$ / 7.1$_{\pm0.7}$
	CS	26.2$_{\pm3.7}$ / 9.3$_{\pm1.0}$	25.5$_{\pm3.7}$ / 8.7$_{\pm1.0}$	12.8$_{\pm2.8}$ / 1.9$_{\pm0.5}$	13.5$_{\pm2.9}$ / 5.1$_{\pm0.7}$
IMU	**ID**	96.6$_{\pm1.7}$ / 96.5$_{\pm0.7}$	97.4$_{\pm1.5}$ / 95.7$_{\pm0.7}$	98.3$_{\pm1.2}$ / 95.7$_{\pm0.7}$	94.0$_{\pm2.2}$ / 35.8$_{\pm1.8}$
	CU	73.5$_{\pm3.0}$ / 43.9$_{\pm1.4}$	74.4$_{\pm3.0}$ / 34.6$_{\pm1.3}$	82.9$_{\pm2.6}$ / 40.8$_{\pm1.3}$	66.4$_{\pm3.3}$ / 13.8$_{\pm0.9}$
	CS	62.4$_{\pm4.1}$ / 43.1$_{\pm1.7}$	62.4$_{\pm4.1}$ / 31.5$_{\pm1.6}$	47.5$_{\pm4.2}$ / 34.4$_{\pm1.6}$	54.6$_{\pm4.2}$ / 12.4$_{\pm1.1}$

Table 4: HAR accuracy (%) of representative models across modalities and protocols, with standard error of the mean shown as $_{x.x}$.

Modality	**Model**	**In-Domain 10/62**	**Cross-User 10/62**	**Cross-Scene 10/62**
RGB	Video-Swin	93.2$_{\pm2.3}$ / 91.3$_{\pm1.0}$	26.5$_{\pm3.0}$ / 5.1$_{\pm0.6}$	11.3$_{\pm2.7}$ / 2.1$_{\pm0.5}$
IMU	CALANet	94.9$_{\pm2.0}$ / 85.1$_{\pm1.3}$	44.5$_{\pm3.4}$ / 17.8$_{\pm1.0}$	31.9$_{\pm3.9}$ / 22.3$_{\pm1.4}$
Acoustic	HTS-AT	80.0$_{\pm3.7}$ / 63.7$_{\pm1.7}$	48.3$_{\pm3.4}$ / 25.1$_{\pm1.2}$	35.5$_{\pm4.0}$ / 22.1$_{\pm1.4}$

6 Limitations and future work

OctoNet introduces a first-of-its-kind comprehensive and richly annotated multi-modal dataset. Yet there are several limitations for improvement: First, to enable high-precision 3D pose tracking with the OptiTrack system, participants were required to wear standardized garments, including a hat, shirt, pants, and shoes that covered their regular clothing. While necessary for motion-capture system, these garments attenuate the body’s natural thermal radiation, potentially reducing the accuracy of readings captured by thermal cameras and infrared arrays. Furthermore, the uniform attire reduces visual variability in the RGB modality, limiting diversity in appearance-based learning tasks. Second, all data in OctoNet were collected in laboratory environments. While this ensures high data quality, it may limit model generalization to real-world environments of varying conditions. Future extensions will consider capturing in-the-wild activities in more variable, dynamic settings. Lastly, OctoNet currently includes 12 diverse sensing modalities spanning a broad portion of the sensing spectrum, but excludes a common modality, LiDAR, as it is uncommon to employ LiDAR for HAR applications.

Table 5: HAR accuracy (%) of different modality fusion configurations on the 10-class subset.

Fused Modalities	Model	In-Domain	Cross-User	Cross-Scene
RGB, FMCW, Acoustic	DenseNet	99.2	66.5	39.8
Depth, ToF, UWB, Wi-Fi	ResNet	99.2	46.9	29.0
Thermal, IRA, IMU	ResNet	99.9	75.2	52.7

Table 6: HPE results (MPJPE in millimeters; lower is better) across sensing modalities under three protocols: In-Domain (ID), Cross-User (CU), and Cross-Scene (CS). "N/A" denotes model incompatibility. Values are to one decimal with standard error of the mean as x.x.

Modality	Protocol	Model			
		ResNet	DenseNet	Swin-T	RFNet
RGB	**ID**	133.3±4.4	147.2±5.1	269.6±6.2	162.8±4.6
	CU	199.8±4.2	204.8±4.8	286.2±5.7	223.7±4.5
	CS	473.9±5.0	524.6±4.3	273.0±7.0	331.8±6.5
Depth	**ID**	131.4±4.5	147.4±4.6	248.2±6.5	194.8±5.7
	CU	197.1±4.6	212.5±4.6	256.2±5.8	230.7±5.3
	CS	363.6±5.6	436.4±5.3	305.1±7.0	444.2±5.8
ToF	**ID**	152.5±5.2	N/A	252.1±6.0	162.2±5.0
	CU	205.7±5.0	N/A	257.3±5.7	193.7±4.8
	CS	361.2±5.4	N/A	303.9±7.0	363.7±4.8
Thermal	**ID**	142.8±4.7	147.0±4.9	259.9±5.9	254.3±6.1
	CU	216.9±4.4	222.4±4.7	259.9±5.8	308.2±5.7
	CS	308.8±5.9	325.4±5.3	313.3±6.8	403.4±7.0
IRA	**ID**	244.4±6.8	N/A	261.1±6.4	265.1±6.7
	CU	373.3±5.8	N/A	261.1±5.9	299.1±5.8
	CS	398.8±7.2	N/A	313.0±6.8	313.4±7.3
FMCW	**ID**	198.5±5.7	185.4±5.4	272.5±7.3	220.9±6.0
	CU	244.0±4.9	236.8±4.7	263.0±6.0	272.0±5.1
	CS	369.4±10.6	389.8±9.8	338.8±10.1	328.3±10.2
SFCW	**ID**	206.7±6.2	202.9±6.2	264.2±6.4	270.7±6.6
	CU	314.6±5.4	334.4±5.4	259.1±5.9	408.1±23.4
	CS	352.2±7.0	408.9±7.0	339.7±8.9	392.7±10.3
UWB	**ID**	142.4±4.8	158.0±5.2	260.5±6.1	159.5±5.0
	CU	241.2±4.8	239.0±4.7	261.3±5.8	241.6±4.6
	CS	310.0±6.5	327.6±6.6	312.5±6.8	295.8±6.8
Wi-Fi	**ID**	147.3±4.7	147.4±4.9	262.2±6.0	186.8±5.3
	CU	270.4±5.6	267.8±5.8	256.3±5.8	274.2±5.6
	CS	399.4±5.7	322.1±6.8	312.8±6.8	400.9±8.3
Acoustic	**ID**	258.8±6.9	256.8±6.7	271.2±6.7	243.6±6.8
	CU	304.1±5.8	312.8±5.8	260.6±5.8	291.8±5.6
	CS	367.2±6.8	441.4±6.9	312.0±6.8	323.3±7.2
IMU	**ID**	147.9±5.0	159.3±5.5	251.6±6.4	180.9±5.3
	CU	252.9±4.9	274.3±5.0	259.9±5.9	266.3±5.0
	CS	289.7±6.8	324.0±6.7	310.8±6.9	328.4±6.9

7 Conclusion

We introduce **OctoNet**, a new benchmark that brings together extensive multi-sensor data, precise annotations, and diverse human activities to support next-generation models for embodied perception. By releasing synchronized recordings across 12 sensing types grounded in high-fidelity 3D pose labels, we aim to facilitate research on fusion, generalization, and cross-modal understanding. We anticipate OctoNet will serve as a valuable asset for the community and lay the groundwork for future progress in human-centric AI.

Acknowledgments and Disclosure of Funding

This work is supported by the NSFC under Grant No. 62222216, Hong Kong RGC GRF under Grant No. 17212224 and No. 17211725, ECS under Grant No. 27204522, HLCA under Grant No. HLCA/E-712/22, and YCRF under Grant No. C5002-23Y.

References

[1] Introducing the intel® RealSense™ depth camera d455. URL `https://www.intelrealsense.com/depth-camera-d455/`.

[2] Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems*, 35: 27414–27426, 2022.

[3] Mohammud J Bocus and Robert Piechocki. A comprehensive ultra-wideband dataset for non-cooperative contextual sensing. *Scientific Data*, 9(1):650, 2022.

[4] Mohammud J Bocus, Wenda Li, Shelly Vishwakarma, Roget Kou, Chong Tang, Karl Woodbridge, Ian Craddock, Ryan McConville, Raul Santos-Rodriguez, Kevin Chetty, et al. Operanet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Scientific data*, 9(1):474, 2022.

[5] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[6] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3501–3510, 2022.

[7] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015.

[8] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022.

[9] Mathias Ciliberto, Vitor Fortes Rey, Alberto Calatroni, Paul Lukowicz, and Daniel Roggen. Opportunity++: A multimodal dataset for video-and wearable, object and ambient sensors-based human activity recognition. *Frontiers in Computer Science*, 3:792065, 2021.

[10] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174, 2020.

[11] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. Rf-net: A unified meta-learning framework for rf-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 517–530, 2020.

[12] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

[13] Yunhao Ge, Yihe Tang, Jiashu Xu, Cem Gokmen, Chengshu Li, Wensi Ai, Benjamin Jose Martinez, Arman Aydin, Mona Anvari, Ayush K Chakravarthy, et al. Behavior vision suite: Customizable dataset generation via simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22401–22412, 2024.

[14] Evan Gebhardt and Marilyn Wolf. Camel dataset for visual and thermal infrared multiple object detection and tracking. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[15] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[16] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[19] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015.

[20] Intel Corporation. Intel® Wi-Fi 6 AX200 network adapter. `https://www.intel.com/content/www/us/en/products/sku/189347/intel-wifi-6-ax200-gig/specifications.html`, 2025. Accessed: 2025-04-27.

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[22] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[23] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. Towards environment independent device free human activity recognition. In *Proceedings of the 24th annual international conference on mobile computing and networking*, pages 289–304, 2018.

[24] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu 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.

[25] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[26] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[27] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st annual ACM symposium on user interface software and technology*, pages 213–224, 2018.

[28] Chengxiao Li, Xie Zhang, and Chenshu Wu. Facial expression recognition with dtof sensing. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887978.

[29] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 872–881, 2019.

[30] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(4):1–28, 2020.

[31] Haipeng Liu, Kening Cui, Kaiyuan Hu, Yuheng Wang, Anfu Zhou, Liang Liu, and Huadong Ma. mtranssee: Enabling environment-independent mmwave sensing based gesture recognition via transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–28, 2022.

[32] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[35] Sheng Lyu and Chenshu Wu. Ase: Practical acoustic speed estimation beyond doppler via sound diffusion field. *arXiv preprint arXiv:2412.20142*, 2024.

[36] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[37] Melexis NV. *MLX90640 32×24 Pixel Infrared Array Sensor—Datasheet*. Melexis NV, 2019. Accessed: 2025-04-27.

[38] Elishiah Miller, Nilanjan Banerjee, and Ting Zhu. Smart homes that detect sneeze, cough, and face touching. *Smart Health*, 19:100170, 2021.

[39] miniDSP. Minidsp uma-8-sp usb microphone array. `https://www.minidsp.com/products/usb-audio-interface/uma-8-sp-detail`, 2025. Accessed: 2025-04-27.

[40] miniDSP. Minidsp umik-2154 microphone. `https://www.minidsp.com/products/acoustic-measurement/umik-2`, 2025. Accessed: 2025-04-27.

[41] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. Samosa: Sensing activities with motion and subsampled audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(3), sep 2022. doi: 10.1145/3550284. URL `https://doi.org/10.1145/3550284`.

[42] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1868–1877, 2017.

[43] Novelda AS. X4M200 Respiration-Sensor Datasheet (rev. e, preliminary). `https://github.com/novelda/Legacy-Documentation/blob/master/Datasheets/X4M200_respiration_sensor_rev_e_preliminary.pdf`, 2018. Accessed: 2025-5-4.

[44] Tønnes F Nygaard, Charles P Martin, Jim Torresen, Kyrre Glette, and David Howard. Real-world embodied ai through a morphologically adaptive quadruped robot. *Nature Machine Intelligence*, 3(5):410–419, 2021.

[45] OptiTrack. OptiTrack Motion Capture Systems. `https://www.optitrack.com/`, n.d. Accessed: 2023-10-15.

[46] Manuel Palermo, Sara Cerqueira, João André, António Pereira, and Cristina P. Santos. Complete inertial pose dataset: from raw measurements to pose with low-cost and high-end marg sensors, 2022. URL `https://arxiv.org/abs/2202.06164`.

[47] Jaegyun Park, Dae-Won Kim, and Jaesung Lee. Calanet: Cheap all-layer aggregation for human activity recognition. *Advances in Neural Information Processing Systems*, 37:69419–69444, 2024.

[48] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[49] Jose Luis Patino, Tahir Nawaz, Tom Cane, and James M. Ferryman. Pets 2017: Dataset and challenge. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2126–2132, 2017. URL `https://api.semanticscholar.org/CorpusID:6904673`.

[50] Monique Paulich, Martin Schepers, Nina Rudigkeit, and Giovanni Bellusci. Xsens mtw awinda: Miniature wireless inertial-magnetic motion tracker for highly accurate 3d kinematic applications. *Xsens: Enschede, The Netherlands*, pages 1–9, 2018.

[51] Polar Electro Oy. Polar h10 heart rate sensor—product page. `https://www.polar.com/us-en/sensors/h10-heart-rate-sensor`, 2025. Accessed: 2025-04-27.

[52] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 27–38, 2013.

[53] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020.

[54] Raspberry Pi Ltd. Raspberry Pi Compute Module 4 datasheet. `https://datasheets.raspberrypi.com/cm4/cm4-datasheet.pdf`, 2024. Accessed: 2025-04-27.

[55] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. Gopose: 3d human pose estimation using wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–25, 2022.

[56] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[57] Seek Thermal Inc. S304SP Mosaic Core Starter Kit (320×240, 57° hfov, 9 hz). `https://shop.thermal.com/S304SP-Mosaic-Core-Starter-Kit-320x240-57HFOV-9HZ`, 2025. Accessed: 2025-04-27.

[58] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[59] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010.

[60] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[61] STMicroelectronics. VL53L8CH Time-of-Flight Sensor. `https://www.st.com/en/imaging-and-photonics-solutions/vl53l8ch.html`, 2025. Accessed: 2025-05-13.

[62] David Strömbäck, Sangxia Huang, and Valentin Radu. Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–22, 2020.

[63] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34: 14502–14515, 2021.

[64] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th annual international conference on mobile computing and networking*, pages 591–605, 2018.

[65] Texas Instruments. IWR1843BOOST mmWave Sensor Evaluation Module. `https://www.ti.com/tool/IWR1843BOOST`, n.d.

[66] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John P Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, volume 2, pages 1–13. London, UK, 2017.

[67] Vayyar Imaging Ltd. Vayyar imaging. `https://vayyar.com/`. Accessed: 2025-05-13.

[68] Andrew Wagenmaker, Kevin Huang, Liyiming Ke, Kevin Jamieson, and Abhishek Gupta. Overcoming the sim-to-real gap: Leveraging simulation to learn to explore for real-world rl. *Advances in Neural Information Processing Systems*, 37:78715–78765, 2024.

[69] Anran Wang, Dan Nguyen, Arun R Sridhar, and Shyamnath Gollakota. Using smart speakers to contactlessly monitor heart rhythms. *Communications biology*, 4(1):319, 2021.

[70] Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. Xrf55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–34, 2024.

[71] Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: A dataset for finding humans using room acoustics. *Advances in Neural Information Processing Systems*, 36:52238–52264, 2023.

[72] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*, pages 65–76, 2015.

[73] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. E-eyes: Device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 617–628, 2014.

[74] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. Push the limit of acoustic gesture recognition. *IEEE Transactions on Mobile Computing*, 21(5):1798–1811, 2020.

[75] Yuheng Wang, Haipeng Liu, Kening Cui, Anfu Zhou, Wensheng Li, and Huadong Ma. m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8298–8302. IEEE, 2021.

[76] Xiaomi Inc. Xiaomi Mi Router AX6000 wi-fi 6 router. `https://www.mi.com/r6000`, 2025. Accessed: 2025-04-27.

[77] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 269–282, 2021.

[78] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems*, 36, 2024.

[79] Zheng Yang, Yi Zhang, Kun Qian, and Chenshu Wu. {SLNet}: A spectrogram learning neural network for deep wireless sensing. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1221–1236, 2023.

[80] Fusang Zhang, Jie Xiong, Zhaoxin Chang, Junqi Ma, and Daqing Zhang. Mobi2sense: empowering wireless sensing with mobility. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 268–281, 2022.

[81] Mi Zhang and Alexander A Sawchuk. Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 1036–1043, 2012.

[82] Yongzhao Zhang, Hao Pan, Yi-Chao Chen, Lili Qiu, Yu Lu, Guangtao Xue, Jiadi Yu, Feng Lyu, and Haonan Wang. Addressing practical challenges in acoustic sensing to enable fast motion tracking. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*, pages 82–95, 2023.

[83] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. Rf-based 3d skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 267–281, 2018.

[84] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *Proceedings of the 17th annual international conference on mobile systems, applications, and services*, pages 313–325, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (*e.g.*, independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, *e.g.*, if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplementary material, but if they appear in the supplementary material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplementary material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (*e.g.*, in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (*e.g.*, a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (*e.g.*, with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (*e.g.*, to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplementary material?

Answer: [Yes]

Justification: In the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (*e.g.*, for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplementary material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (*e.g.*, data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplementary material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (*e.g.*, Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (*e.g.*negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (*e.g.*, preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: In the supplementary material.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (*e.g.*, if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (*e.g.*, disinformation, generating fake profiles, surveillance), fairness considerations (*e.g.*, deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (*e.g.*, gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (*e.g.*, pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: In the supplementary material.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (*e.g.*, code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (*e.g.*, CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (*e.g.*, website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: In the supplementary material.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: See section 4.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplementary material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: See section 4 and the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

Justification: No large language model (LLM) was employed as a core, original, or non-standard component in our research methods.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Ethics statement

All participants provide written informed consent and receive compensation higher than the minimum hourly wage under local labor regulations. We obtained ethics approval from the Institutional Review Board (EA240308). We remove all the personal identifiers and inform participants that de-identified data will be made publicly available for research purposes. All collected data have been carefully examined to ensure the absence of security or safety risks, and the dataset is hosted on HuggingFace with privacy safeguards that preclude the collection of any additional personal information. Overall, this study poses minimal foreseeable harm to participants, and we adhere to all relevant institutional and ethics guidelines throughout data collection, processing, sharing, and publication.

B Dataset toolbox

To facilitate the use of the data, we convert the sensing data from various modalities into open, widely used formats. We also provide a *dataset toolbox* in our public GitHub repository `https://github.com/aiot-lab/OctoNet`, which includes a PyTorch-compatible dataloader. Users may download the data from the provided link and follow step-by-step instructions in the repository to easily load and preprocess the dataset.

For the RGB data containing identifiable attributes, an anonymized version with explicit permission from all participants is available upon request by completing an application form. Please refer to the repository documentation for details on the application process and usage terms. Additionally, a sample dataset, including RGB data with extra approval from the user for distribution, is directly available for immediate exploration.

C Implementation

Our experiments are implemented in PyTorch [48] and trained on an Intel Xeon Gold 5418Y (2 GHz, 96 cores, 512 GB RAM) and eight NVIDIA GeForce RTX 4090 GPUs. We open-source both our code and the datasets under the CC BY-NC 4.0 license for the benefit of the research community.

D Details on 62 activities

We include overall 62 most representative activities in the real world. They are further grouped into five categories: body-motion only, human-object interaction, human-computer interaction, human-human interaction and medical conditions, as illustrated in Table 7. Additionally, we provide the visual illustration of the activities, as displayed in Figure 3.

Table 7: The overview of 62 activities. They are colored by category: Body-Motion Only, Human-Object Interaction, Human-Computer Interaction, Medical Conditions, Human-Human Interaction

ID	Activity Name	ID	Activity Name	ID	Activity Name
1	Sitting	2	Walking	3	Bowing
4	Sleeping	5	Dancing	6	Jogging
7	Falling Down	8	Jumping	9	Jumping Jack
10	Squatting	11	Lunging	12	Turning
13	Push-Up	14	Leg Raising	15	Air Drumming
16	Boxing	17	Shaking Head	18	Answering Phone
19	Eating	20	Drinking	21	Wiping Face
22	Picking Up	23	Jumping Rope	24	Mopping Floor
25	Brushing Hair	26	Bicep Curl	27	Playing Phone
28	Brushing Teeth	29	Typing	30	Thumbs-Up
31	Thumbs-Down	32	Making OK Sign	33	Making Victory Sign
34	Drawing Circle Clockwise	35	Drawing Circle Counterclockwise	36	Stop Sign
37	Pulling Hand In	38	Pushing Hand Away	39	Handwave
40	Sweeping	41	Clapping	42	Sliding
43	Drawing Zigzag	44	Dodging	45	Bowling
46	Lifting Up A Hand	47	Tapping	48	Spreading and Pinching
49	Drawing Triangle	50	Sneezing	51	Coughing
52	Staggering	53	Yawning	54	Blowing Nose
55	Stretching Oneself	56	Touching Face	57	Shaking Hands
58	Hugging	59	Pushing Someone	60	Kicking Someone
61	Punching Someone	62	Conversation		

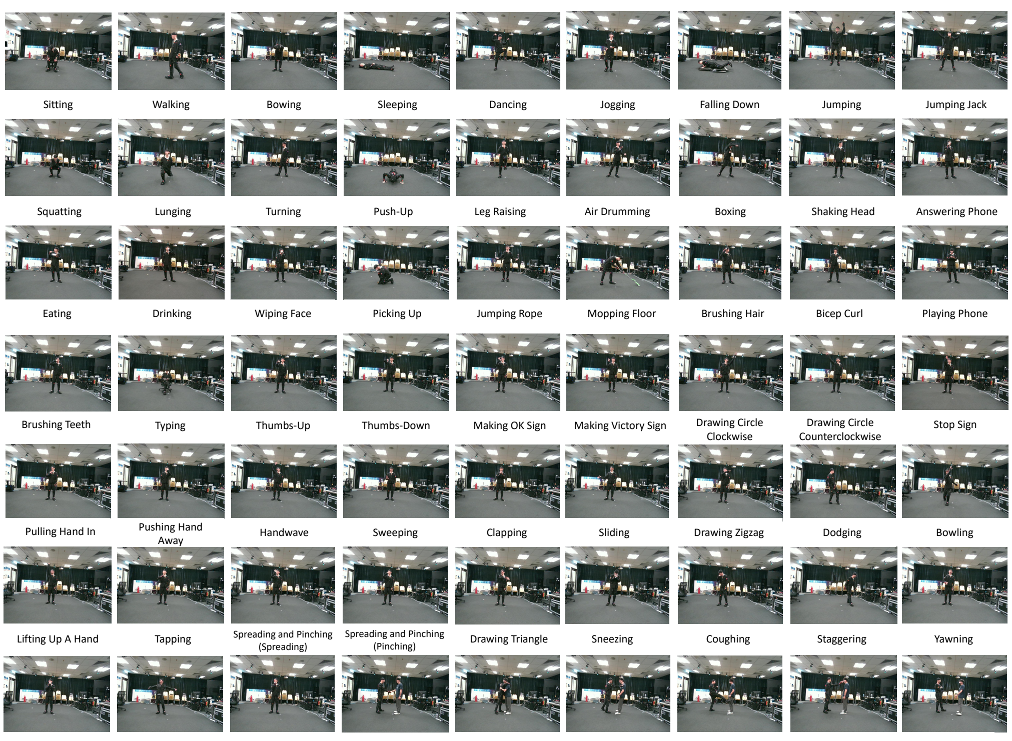


Figure 3: An illustration of the 62 distinct activities, which are further grouped into five subcategories reflecting different interaction contexts. Note that we split spreading and pinching for visualization.