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

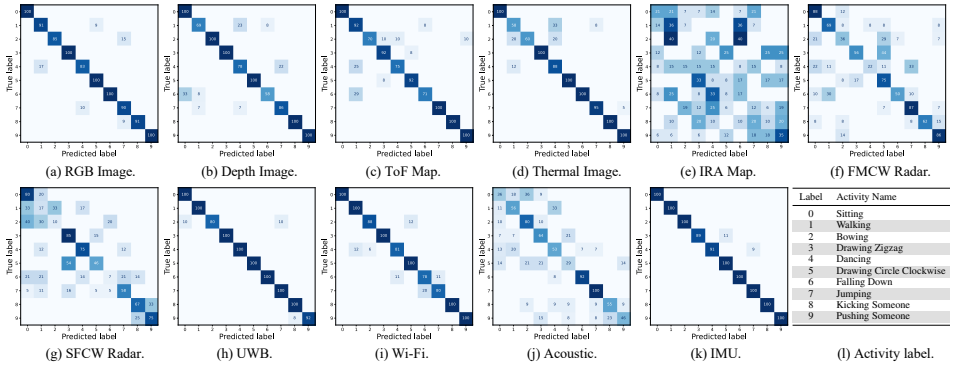
Dongsheng Yuan*, **Xie Zhang***, **Weiying Hou, Sheng Lyu, Yuemin Yu,**
Luca Jiang-Tao Yu, Chengxiao Li, Chenshu Wu†
Department of Computer Science, The University of Hong Kong
Project Website: `https://aiot-lab.github.io/OctoNet/`
Dataset: `https://huggingface.co/datasets/hku-aiot/OctoNet`

A Results visualization

To intuitively compare model performance across modalities, we visualize results for **Human Activity Recognition (HAR)** and **3D Human Pose Estimation (HPE)**. Figure 1 shows the confusion matrix for a representative subset of 10 HAR activities, while Figure 2 displays results for all 62 activities. For HPE, Figure 3 illustrates performance across modalities. All visualizations compare the top-performing baseline models across each of the 11 tested modalities, quantifying their individual contributions to task performance.

A.1 HAR task details

We evaluate OctoNet using both a representative 10-activity subset (covering body-motion, object/computer/human interactions, and medical conditions) and the full 62 activities. Figures 1(l) and 2(l) show the activity labels and color-coded categorization respectively. The consistent performance of baseline models across modalities demonstrates OctoNet's utility as: (1) a comprehensive multi-modal HAR benchmark, and (2) a testbed for identifying modality-specific challenges in cross-modal generalization and fine-grained activity recognition.



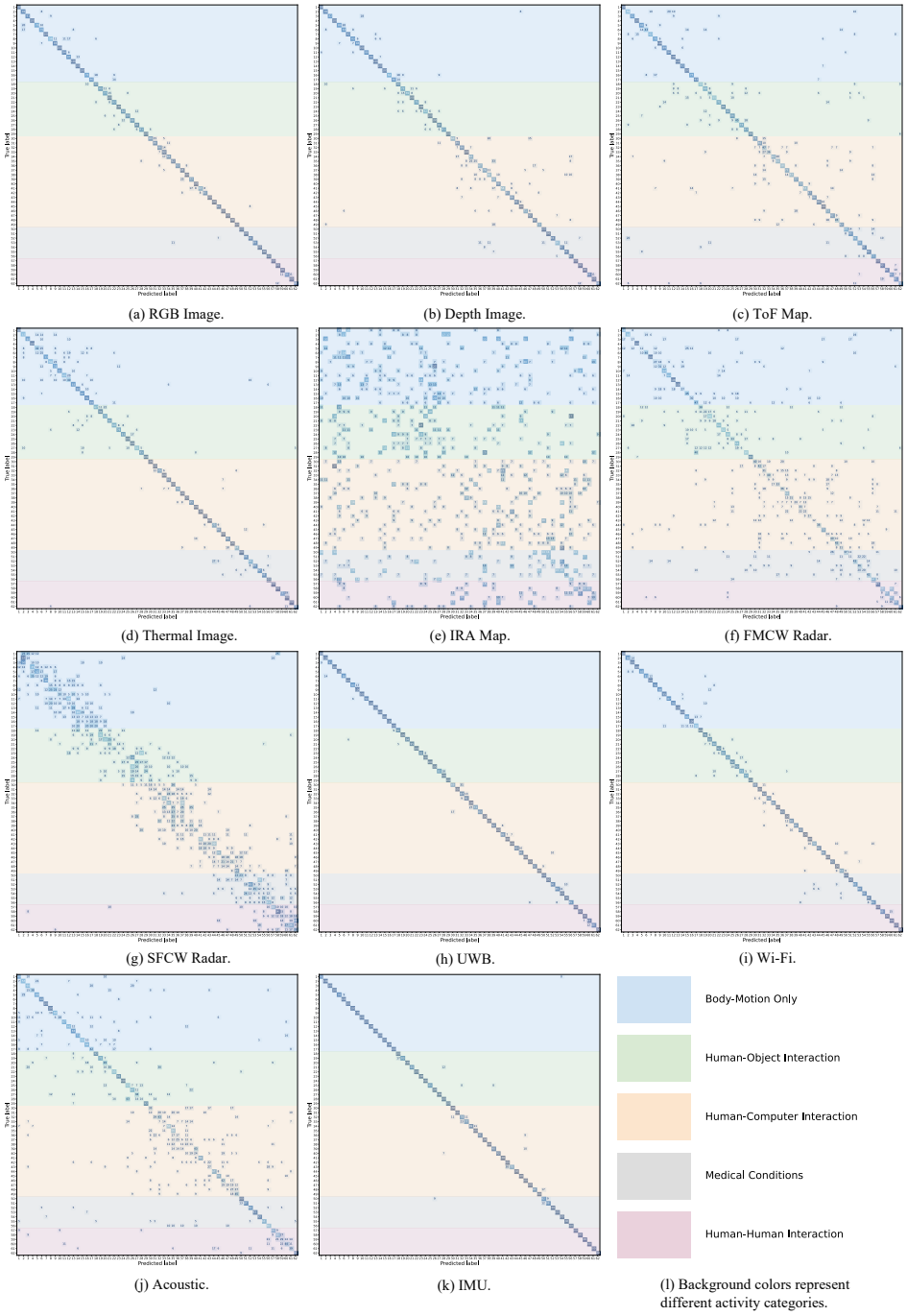
(a) RGB Image. (b) Depth Image. (c) ToF Map. (d) Thermal Image. (e) IRA Map. (f) FMCW Radar.
(g) SFCW Radar. (h) UWB. (i) Wi-Fi. (j) Acoustic. (k) IMU.

Label	Activity Name
0	Sitting
1	Walking
2	Bowing
3	Drawing Zigzag
4	Dancing
5	Drawing Circle Clockwise
6	Falling Down
7	Jumping
8	Kicking Someone
9	Pushing Someone

(l) Activity label.

Figure 1: Confusion matrices for 10 representative activities across 11 modalities. Rows represent ground truth labels, columns show predictions, with color intensity indicating classification accuracy.

*Equal contribution.
†Corresponding author.



(a) RGB Image. (b) Depth Image. (c) ToF Map.

(d) Thermal Image. (e) IRA Map. (f) FMCW Radar.

(g) SFCW Radar. (h) UWB. (i) Wi-Fi.

(j) Acoustic. (k) IMU. (l) Background colors represent different activity categories.

Figure 2: Confusion matrices for all 62 activities across 11 modalities. Activities are color-coded by category, revealing modality-specific recognition patterns.

A.2 HPE task details

Figure 3 demonstrates representative examples of ground truth versus predicted 3D human poses from the best-performing baseline model across different input modalities. The selected samples showcase diverse poses that effectively highlight model performance characteristics. Due to varying original sampling rates across modalities, the results exhibit slight temporal misalignment.

Notably, modalities with lower spatial resolution (IRA, SFCW, and Acoustic) show reduced HPE accuracy, quantitatively demonstrating the importance of spatial information for pose estimation. These systematic performance variations across modalities establish OctoNet as a rigorous benchmark for evaluating sensor-specific capabilities in 3D human pose estimation.

Figure 3: 3D human pose estimation results comparing ground truth (blue) and predicted (orange) skeletons across modalities. Each row shows predictions from models trained with a different input modality, demonstrating both reconstruction accuracy and modality-specific performance characteristics.

B Data annotation

Temporal segment annotation process. During each experimental session, participants repeated the 62 fixed activities continuously (excluding relatively static activities such as sleeping and conversation), with brief pauses of approximately 3 seconds. Before each set of repetitions, the specific activity label was recorded to ensure correct annotation. Although we initially considered using existing activity recognition models to automate pause detection and segmentation, these models lacked sufficient coverage for our diverse, untrimmed multi-action scenarios with subtle behavioral differences [3]. Consequently, we used the motion-capture system's speed profiles of body markers to identify the 3-second pause intervals, resulting in over 8.76k segmented samples.

C Sensor details and data processing

Figure 4 shows the physical sensor units used in our multimodal data collection. The following subsections document each sensor type's (1) hardware specifications, including form factor and operating ranges; (2) data acquisition settings such as sampling rates and resolutions; and (3) preprocessing steps covering calibration and noise reduction procedures.

Figure 4: All sensor hardware used for data collection.

C.1 RGB-D camera details

We employ three Intel RealSense Depth Cameras D455C to record synchronized RGB and depth frames at a resolution of 640×480 pixels and an average frame rate of 29.95 Hz. The D455C cameras utilize stereoscopic depth technology, featuring a depth field of view (FOV) of 87° horizontal by 58° vertical and an RGB FOV of 90° horizontal by 65°, which provides comprehensive environmental coverage. The minimum depth distance is approximately $52\,\mathrm{cm}$, and the depth accuracy is less than 2% at $4\,\mathrm{m}$ [1], ensuring precise depth measurements within the operational range. To optimize data collection efficiency, the RGB data are compressed using the `.mp4` format, which offers high compression ratios while maintaining visual quality, and the depth data are stored in `.png` format, utilizing its near-lossless compression capabilities to preserve the integrity of the depth measurements for accurate post-processing and analysis. The RGB data are collected into a 5D tensor structured as $(N_{\text{node}}, F, H, W, C)$, where $N_{\text{node}} = 3$ is the number of camera nodes, F is the number of frames, $H = 480$ is the height of the pixel array, $W = 640$ is the width of the pixel array, and $C = 3$ is the number of color channels. This results in a tensor of size $3 \times F \times 480 \times 640 \times 3$. The depth data are collected into a 4D tensor structured as $(N_{\text{node}}, F, H, W)$, with the same $N_{\text{node}} = 3$, F is the number of frames, $H = 480$, and $W = 640$, resulting in a tensor of size $3 \times F \times 480 \times 640$.

C.2 Time-of-Flight (ToF) sensor details

We employ a commercially available Single Photon Avalanche Diode (SPAD) sensor (STMicroelectronics VL53L8CH) and refer to the same setting as ToFace [4] to capture depth information. This sensor operates by emitting modulated infrared pulses and measuring the time taken for these pulses to reflect from objects back to the sensor, enabling accurate distance measurements. The VL53L8CH provides a FOV of approximately $45^\circ \times 45^\circ$, allowing focused distance sensing within the area of interest.

Data acquisition from the ToF sensor is managed by an STM32 microcontroller, which records depth measurements at an average frame rate of 7.32 Hz. The raw sensor output comprises impulse-response

histograms organized into 18 discrete bins per spatial location. The ToF data are collected into a 5D tensor structured as $(N_{\text{node}}, F, H, W, C)$, where $N_{\text{node}} = 1$ is the number of sensor nodes, F is the number of frames, $H = 8$ is the height of the spatial array, $W = 8$ is the width of the spatial array, and $C = 18$ is the number of histogram bins. This results in a tensor of size $1 \times F \times 8 \times 8 \times 18$.

C.3 Thermal camera details

We employ two Seek Thermal S304SP Mosaic Core thermal cameras to capture high-precision thermal data. Each camera features an uncooled vanadium oxide microbolometer with a pixel pitch of 12 μm and operates within a spectral response range of 7.8 to 14 μm. The sensors provide a resolution of 320×240 pixels, totaling 76,800 pixels per frame, which allows for detailed thermal imaging of the subjects.

The cameras operate at a frame rate of approximately 9 Hz, complying with export regulations for thermal imaging devices. They have a FOV of 56° horizontal by 42° vertical, enabling comprehensive coverage of the subjects. With a thermal sensitivity of less than 100 mK at 25°C, the cameras ensure accurate temperature measurements within an imaging range of -40°C to $+330^\circ$C. They perform automatic non-uniformity correction (NUC) using an internal shutter, enhancing image quality by compensating for sensor artifacts.

For data acquisition, the cameras interface via USB and provide 16-bit thermal data before automatic gain control (AGC). We store these raw images as `.png` files, applying near-lossless compression to reduce file size while preserving temperature accuracy. Over a recording period of t seconds, the thermal data are collected into a 4D tensor structured as $(N_{\text{node}}, F, H, W)$, where $N_{\text{node}} = 2$ is the number of camera nodes, F is the number of frames, $H = 240$ is the height of the pixel array, and $W = 320$ is the width of the pixel array. This results in a tensor of size $2 \times F \times 240 \times 320$, reflecting the capture of thermal data over time from two cameras.

C.4 Infrared array sensor details

We employ five MLX90640 infrared array sensors to capture two-dimensional thermal maps of the environment. Each sensor comprises a grid of thermopile detectors arranged in a 32×24 pixel matrix, generating temperature data of objects within its FOV. The MLX90640 sensors feature an FOV of $110^\circ \times 75^\circ$, enabling wide-area coverage for thermal sensing. With a temperature precision of $\pm 1.5^\circ$C, these sensors exceed the clinically relevant limit of $\pm 0.5^\circ$C [20], thereby mitigating concerns about user privacy related to precise temperature measurements.

Data acquisition is performed by interfacing the sensors with an ESP32, which collects the temperature data from each pixel at an average frame rate of 6.91 Hz and transfers it to a PC node for storage and processing. The emissivity setting of the sensors is configurable and is set to $\varepsilon = 1$. The collected data are organized into a 4D tensor structured as $(N_{\text{node}}, F, H, W)$, where $N_{\text{node}} = 5$ is the number of sensor nodes, F is the number of frames, $H = 24$ is the height of the pixel array, and $W = 32$ is the width of the pixel array. This results in a tensor of size $5 \times F \times 24 \times 32$, reflecting the capture of thermal data over time from five strategically placed sensors providing comprehensive thermal coverage of the environment.

C.5 FMCW millimeter-wave radar details

We employ Texas Instruments IWR1843Boost mmWave radar [17] with frequency-modulated continuous-wave (FMCW) sensing with multiple transmit and receive antennas, which is widely used in micro motion detection [13, 22], and macro human activity recognition [5]. It transmits linearly frequency-swept chirps, then applies standard range-FFT, Doppler-FFT, and angle-estimation pipelines to extract target range, velocity, and azimuth/elevation. Combined with the known antenna array geometry, these estimates form a three-dimensional point cloud that captures object position and motion. A constant-false-alarm-rate (CFAR) stage suppresses noise and spurious detections, yielding a cleaner point cloud. Data acquisition and processing are performed by connecting the radar to a PC node via a UART interface. The radar data are collected at an average frame rate of 8.81 Hz, generating a set of point-cloud points $\{(x, y, z, v)\}$ per frame. For uniform batch processing, each frame's points are either padded or truncated to a fixed number P. These frames are then collected into a 4D tensor structured as $(N_{\text{node}}, F, N_{\text{point}}, S)$, where F is the number of frames, $N_{\text{node}} = 5$

is the number of radar nodes, $N_{\text{point}} = P$ is the number of points, and $S = 4$ is the number of features, representing P points each with 4 coordinates (x, y, z, v). This results in a tensor of size $5 \times F \times P \times 4$, reflecting the capture of point-cloud data.

C.6 SFCW millimeter-wave radar details

We use a Vayyar IMAGEVK-74 mmWave radar to acquire 3D imagery of object positions and motion. The sensor operates in Stepped Frequency Continuous Wave (SFCW) mode, essentially a discrete form of FMCW, in which the channel response is sampled at N uniformly spaced continuous wave tones rather than along a continuous chirp, yielding precise frequency-domain measurements [2]. Data capture and preprocessing are carried out on a ThinkPad T14 laptop (Intel Core i7-1260P, Windows 11) via MATLAB, using the official SDK to interface with the radar and log raw data.

The key configurations are as follows: the start frequency is 63 GHz, and the stop frequency is 67 GHz, resulting in a bandwidth of 4 GHz. We utilize $N = 100$ frequency steps between the start and stop frequencies, offering fine resolution in the frequency domain. The radar is equipped with 20 transmit (Tx) antennas and 20 receiver (Rx) antennas, creating a virtual antenna array of 400 unique Tx-Rx pairs (i.e., 20×20). This extensive antenna configuration enhances spatial resolution and enables detailed imaging capabilities.

The radar operates at an average frame rate of 3.20 Hz, sufficient for capturing human movements and gestures. Over a recording period of t seconds, the radar data are collected into a 4D tensor structured as $(N_{\text{node}}, F, T_x \cdot R_x, S)$, where $N_{\text{node}} = 1$ is the number of radar nodes, F is the number of frames, $T_x \cdot R_x = 400$ is the number of virtual antenna pairs (with $T_x = 20$ transmit antennas and $R_x = 20$ receiver antennas), and $S = 100$ is the number of frequency steps (ADC samples). This results in a tensor of size $1 \times F \times 400 \times 100$, reflecting the capture of radar responses. We then convert this array into a PyTorch tensor for downstream processing.

C.7 Ultra-Wideband radar details

For ultra-wideband sensing, we employ Novelda XeThru X4M200 radar with co-located transmitter and receiver antennas. The radar operates at a center frequency of 7.29 GHz with a bandwidth of around 2.5 GHz, and with a sampling rate of 23.328 GS/s [9]. The maximum detection range of the radar is approximately 9.9 m, with a range resolution (bin-to-bin distance) of 6.4 mm, enabling high temporal and spatial resolution that facilitates fine-grained motion detection.

We utilize the official Python wrapper to acquire low-level control of the hardware. We disable the on-chip downconversion of received data, thus real-valued Channel Impulse Response (CIR) is obtained. The Digital-to-Analog Converter (DAC) settings are configured with a minimum sweeping threshold of 900 and a maximum threshold of 1150 Hz, and each DAC sweep consists of 16 iterations. At each timestamp, the radar outputs a real vector of size 1535, representing the CIR across different range bins. With an average frame rate of 17.07 Hz, the radar data are collected into a 3D time-domain tensor structured as (N_{node}, F, S), where $N_{\text{node}} = 1$ is the number of radar nodes, F is the number of frames, and $S = 1535$ is the number of CIR range bins. This results in a tensor of size $1 \times F \times 1535$.

C.8 Wi-Fi details

We use one Xiaomi Router AX6000 as the Wi-Fi transmitter and four Raspberry Pi Compute Module 4 devices, each equipped with an Intel AX200 network interface card (NIC), as the Wi-Fi receivers. Inspired by the sensor placement methodology of Widar3.0 [23], the receiver units are strategically positioned at the four corners of the sensing area. This arrangement creates a larger rectangular sensing area conducive to capturing comprehensive movement data.

Each receiver sends ping packets independently to the transmitter, resulting in an average packet rate of 75.62 packets per second. We utilize a modified driver for CSI recording on the AX200 NICs to conduct channel estimation and obtain the CSI of 114 subcarriers. Consequently, over a recording period of t seconds, the CSI data are collected into a 4D tensor structured as $(N_{\text{node}}, F, T_x \cdot R_x, S)$, where F is the number of frames, $N_{\text{node}} = 4$ is the number of receiver nodes, $T_x = 1$ is the number of transmitter antennas, $R_x = 2$ is the number of receiver antennas per node, and $S = 114$ is the number of subcarriers. This results in a tensor of size $4 \times F \times 2 \times 114$.

C.9 Audio details

We use a sampling rate of 48 kHz for audio recording with two microphones. Simultaneously, we play back the probing signals using one speaker. The rationale behind it is that inaudible ultrasonic signals can provide additional sensing abilities, including speech enhancement [15, 19], localization [7], pose estimation [18], gesture recognition [16], and speed estimation [6, 21]. Specifically, we leverage Kasami, a pseudo-noise signal that provides good orthogonality. We adhere to the preprocessing methodology outlined in ASE [6] to modulate these signals, ensuring they remain inaudible to humans while avoiding interference with low-frequency features.

All recordings are saved as `.wav` files in our dataset. For baseline training, we convert the raw waveforms into Mel-spectrograms, which are then organized into $(N_{\text{node}}, C, M, T)$, where $N_{\text{node}} = 2$ represents the number of sensor nodes (microphones), $C = 1$ is the number of channels per node, $M = 128$ is the number of Mel-frequency bands used, and T denotes the temporal dimension. This results in a tensor of size $2 \times 1 \times 128 \times T$.

C.10 Inertial-magnetic measurement unit (IMU) details

To capture detailed motion data, the Xsens Awinda Research Kit is utilized, comprising 17 MTw Awinda wireless motion trackers and an Awinda Station serving as the master interface. Each IMU integrates a 3-axis accelerometer, 3-axis gyroscope, and 3-axis magnetometer, enabling comprehensive sensing of linear accelerations, angular velocities, and magnetic field vectors. The sensors provide raw output data, including sensor free acceleration, magnetic field vectors, quaternion representation of orientation, and Euler representation of orientation, which are essential for calculating joint quaternions and fully specifying body pose and movement. The MTw trackers wirelessly connect to the Awinda Station, which interfaces with the host PC running the Xsens MVN Analyze software (version 2024.2) [8]. Data are sampled at a rate of 60 Hz, ensuring high temporal resolution for capturing dynamic movements. The Awinda Station ensures that data from each MTw are synchronized within $10\,\mu$s, crucial for accurate temporal alignment across all sensors. Sensors are securely affixed to the subjects using the provided straps and calibrated following the procedures outlined in the Xsens whitepaper [14]. The IMU data are collected into a 3D tensor structured as (F, D, N_{IMU}), where F is the number of frames, $N_{\text{IMU}} = 17$ is the number of IMU sensors, and $D = 13$ is the dimension of the feature vector, comprising: (1) sensor free acceleration (x, y, z), (2) magnetic field vectors (x, y, z), (3) Euler angle representation of orientation (x, y, z), and (4) quaternion representation of orientation (q_0, q_1, q_2, q_3). This results in a tensor of size $F \times 13 \times 17$.

C.11 OptiTrack motion-capture system details

We use the OptiTrack motion-capture system [11] to generate high-precision 3D human skeletal data, which serves both as ground truth for our HPE baselines and as a foundation for future research exploring multimodal data relationships. The system includes 12 Prime$^{\text{x}}$ 13 cameras that emit and receive infrared light for precise motion capture. During data collection, 50 reflective markers are placed on each subject to define their skeletal structure. The marker placement strategy is detailed in [10], and additional skeletal information (*e.g.*, 4D rotations) is recorded in the released dataset's `.csv` files. According to the manufacturer, each camera achieves positional errors within $\pm 0.20\,\text{mm}$ and rotational errors within 0.5° [12]. Motion data is captured at a sampling rate of 120 Hz, ensuring high temporal resolution in Scenarios 1-3. For the benchmarks, we extract a subset of 20 key skeletal markers stored in `.csv` format, ultimately forming an $(F \times 20 \times 3)$ tensor of 3D coordinates, which we saved in `.npy` format.

References

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

[2] Fangqiang Ding, Zhen Luo, Peijun Zhao, and Chris Xiaoxuan Lu. milliflow: Scene flow estimation on mmwave radar point cloud for human motion sensing. In *European Conference on Computer Vision*, pages 202–221. Springer, 2024.

[3] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 1011–1030, 2023.

[4] 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.

[5] Kun Liang, Anfu Zhou, Zhan Zhang, Hao Zhou, Huadong Ma, and Chenshu Wu. mmstress: Distilling human stress from daily activities via contact-less millimeter-wave sensing. In *Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2023.

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

[7] Wenguang Mao, Zaiwei Zhang, Lili Qiu, Jian He, Yuchen Cui, and Sangki Yun. Indoor follow me drone. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*, pages 345–358, 2017.

[8] Movella Inc. *Xsens MVN Analyze Software, version 2024.2*. Movella Inc., 2023. URL https://www.movella.com/products/mvn-analyze. Accessed: 2023-10-15.

[9] Novelda AS. XeThru X4 Radar User Guide. https://github.com/novelda/Legacy-Documentation/blob/master/Application-Notes/XTAN-13_XeThruX4RadarUserGuide_rev_a.pdf, 2018. Application Note XTAN-13, Rev. A.

[10] OptiTrack. OptiTrack Core (50) Full-Body Marker Set—Online Documentation. https://docs.optitrack.com/markersets/full-body/core-50, 2025. Accessed: 2025-04-27.

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

[12] OptiTrack. PrimeX 13 Camera. https://www.optitrack.com/cameras/primex-13/, n.d..

[13] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, Min Wu, and KJ Ray Liu. Radioses: mmwave-based audioradio speech enhancement and separation system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, undefined(undefined), 2023.

[14] 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.

[15] Ke Sun and Xinyu Zhang. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th annual international conference on mobile computing and networking*, pages 160–173, 2021.

[16] 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.

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

[18] Zhijian Yang, Xiaoran Fan, Volkan Isler, and Hyun Soo Park. Posekernellifter: Metric lifting of 3d human pose using sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13179–13189, 2022.

[19] Luca Jiang-Tao Yu, Running Zhao, Sijie Ji, Edith CH Ngai, and Chenshu Wu. Uspeech: Ultrasound-enhanced speech with minimal human effort via cross-modal synthesis. *arXiv preprint arXiv:2410.22076*, 2024.

[20] Xie Zhang and Chenshu Wu. Tadar: Thermal array-based detection and ranging for privacy-preserving human sensing. *arXiv preprint arXiv:2409.17742*, 2024.

[21] 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.

[22] Running Zhao, Jiangtao Yu, Hang Zhao, and Edith CH Ngai. Radio2text: Streaming speech recognition using mmwave radio signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–28, 2023.

[23] 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.