NEURAL INFORMATION PROCESSING SYSTEMS

香 港 大 學
THE UNIVERSITY OF HONG KONG

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†

Motivation

香 港 大 學
THE UNIVERSITY OF HONG KONG

1. Embodied AI needs robust activity understanding beyond vision-only.

2. Real environments are sensor-rich (RF, IMU, thermal, audio).

Missing: large, unified, aligned benchmarks across heterogeneous modalities.

What is OctoNet?

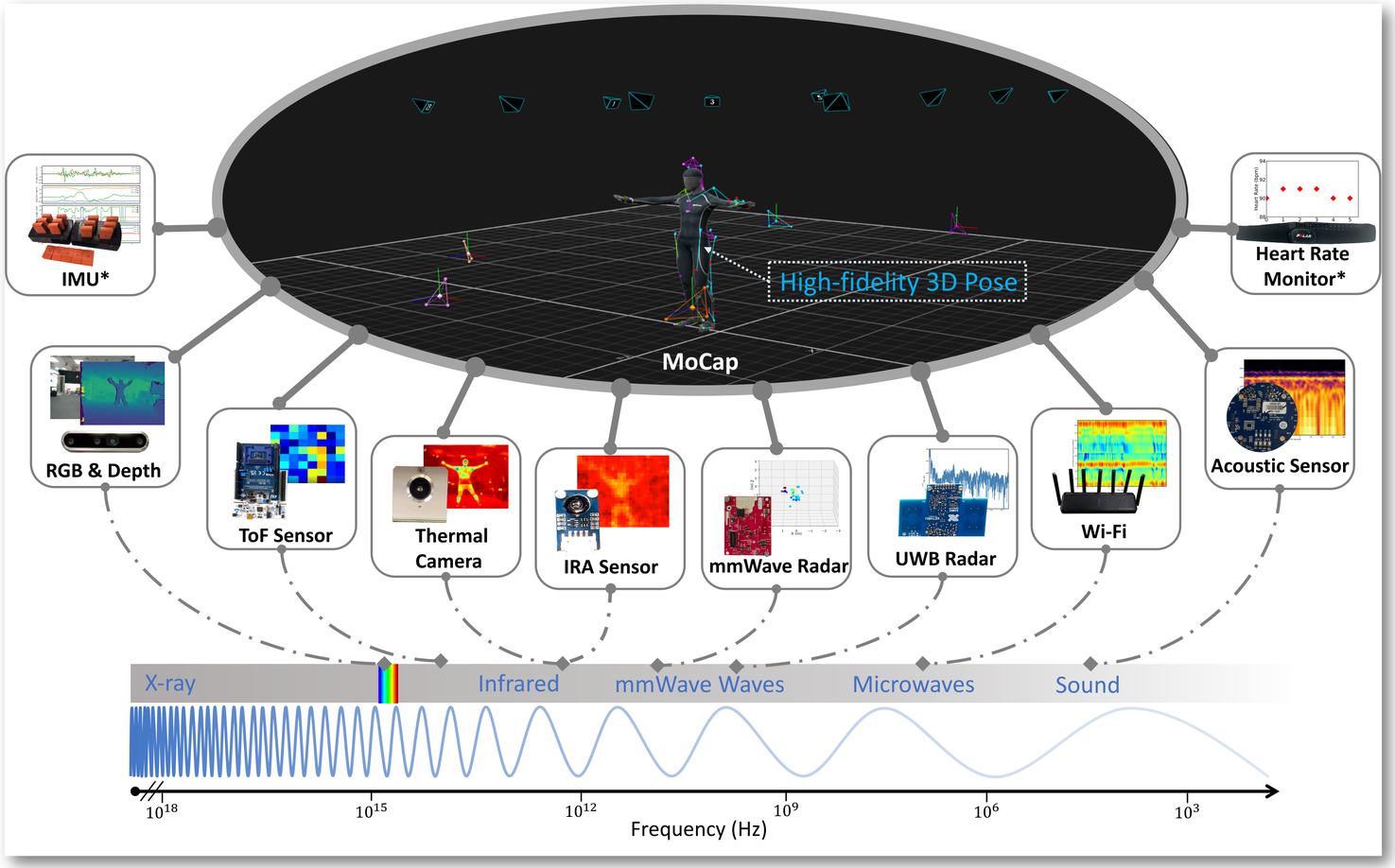
香 港 大 學
THE UNIVERSITY OF HONG KONG

12 modalities, multi-view; 41 participants; 62 activities;

67.72M synchronized frames;

High-fidelity 3D pose labels from

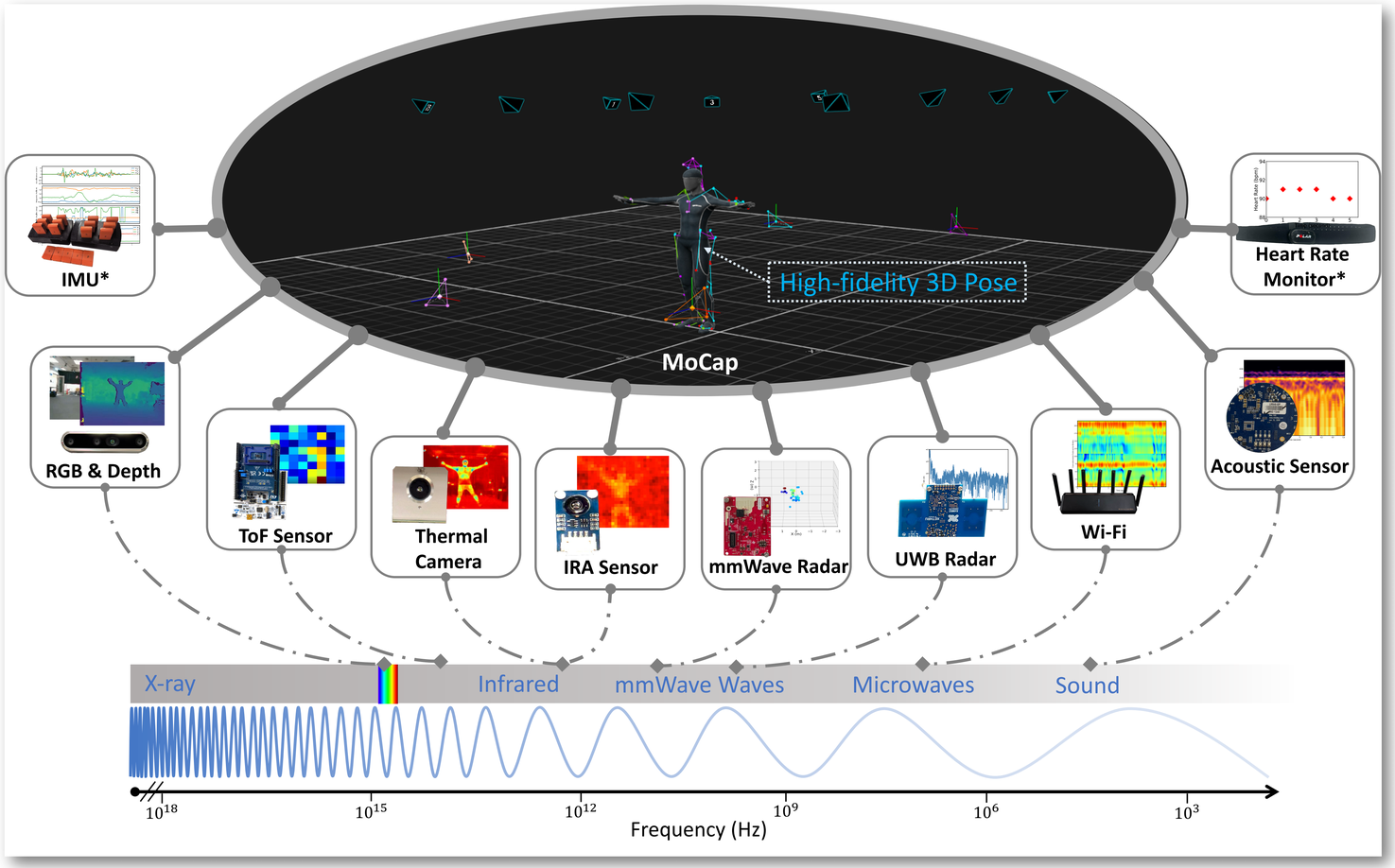
OptiTrack for alignment/supervision;



Supports HAR, 3D HPE, fusion, cross-modal learning, sensor foundation models.

Modalities

香 港 大 學
THE UNIVERSITY OF HONG KONG



Visual

RGB-D (×3), ToF, Thermal (×2), Infrared Array (×5)

RF

mmWave (FMCW×5, SFCW×1), UWB, Wi-Fi

Others

Audio microphones, active acoustic emitter, IMU (17 nodes), Heart Rate sensor

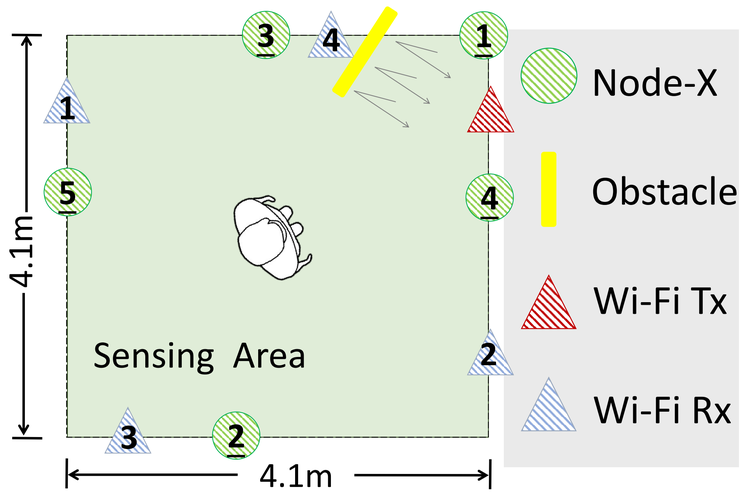
Labels

3D skeletal keypoints (50 markers).

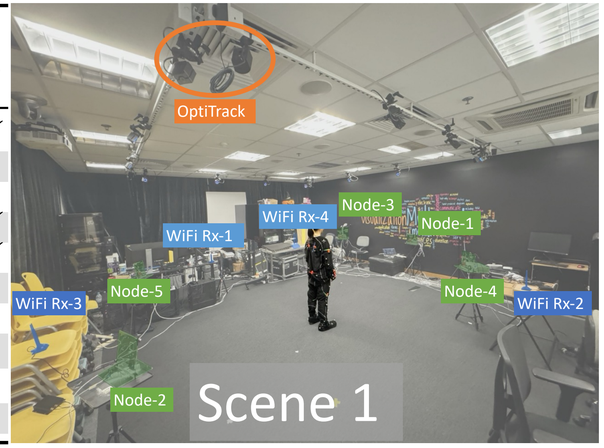
Collection, scenes & sync

香 港 大 學
THE UNIVERSITY OF HONG KONG

- 3 scenes (office, lab, living room); 5 sensing nodes.
- Wi-Fi Rx rectangle; one deliberate NLOS link.
- NTP-based global time; single broadcast start timestamp.
- Activities cover body-motion, H-O/H-C/H-H interactions, medical + aerobics + freestyle.



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



(a) Data collection setup.

Scale & examples

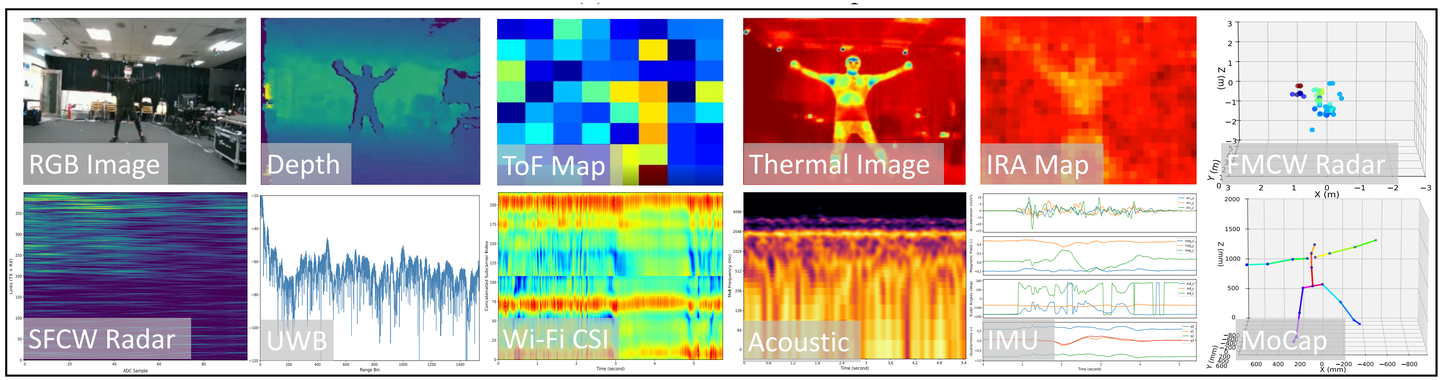
香 港 大 學
THE UNIVERSITY OF HONG KONG

Modality	Total Frames	Sampling Rate (Hz)	Number of Nodes	Data Dimension (per frame)	Storage Size (GB)
RGB-D	7.82M	29.95	3	480×640 $(\times 3)^{\dagger}$	**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**

Scale & examples

香 港 大 學
THE UNIVERSITY OF HONG KONG

Modality	Total Frames	Sampling Rate (Hz)	Number of Nodes	Data Dimension (per frame)	Storage Size (GB)
RGB-D	7.82M	29.95	3	480×640 $(\times 3)^{\dagger}$	**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**

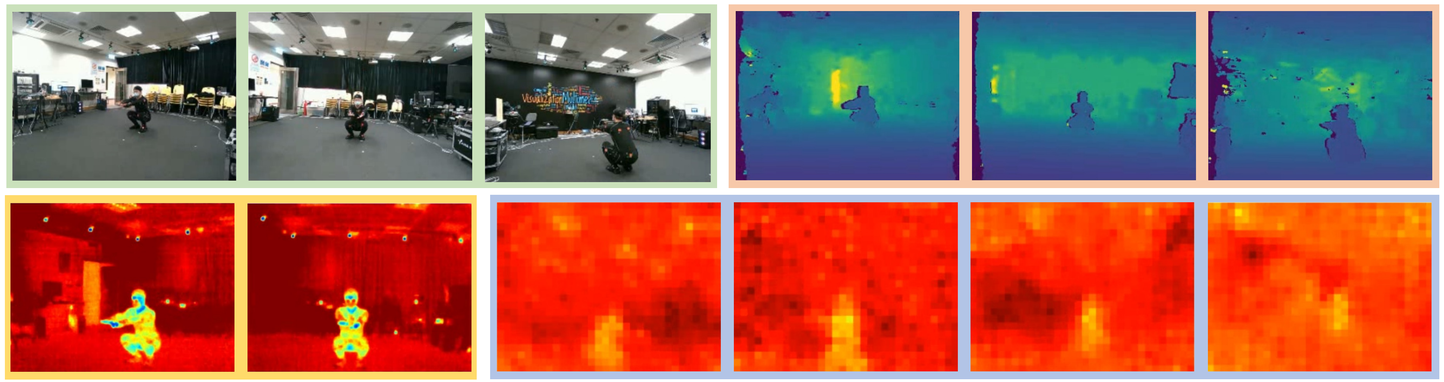


(b) Multi-modality samples.

Scale & examples

香 港 大 學
THE UNIVERSITY OF HONG KONG

Modality	Total Frames	Sampling Rate (Hz)	Number of Nodes	Data Dimension (per frame)	Storage Size (GB)
RGB-D	7.82M	29.95	3	480×640 $(\times 3)^{\dagger}$	**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**



(c) Multi-view samples.

Evaluation results

香 港 大 學
THE UNIVERSITY OF HONG KONG

Human Activity Recognition (HAR): 62 daily activities across five categories (body-only, human–object, human–computer, human–human, medical), class-balanced.

Table 1: 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	Model			
RGB	91.5±2.6 / 93.4±0.9	93.2±2.3 / 91.2±1.0	94.9±2.0 / 93.1±0.9	89.7±2.8 / 60.9±1.8
Depth	89.7±2.8 / 86.6±1.2	90.6±2.7 / 83.2±1.3	86.3±3.2 / 81.7±1.4	87.2±3.1 / 40.0±1.8
ToF	86.8±3.1 / 70.3±1.6	N/A	82.6±3.5 / 51.8±1.8	89.3±2.8 / 75.9±1.5
Thermal	90.1±2.7 / 85.0±1.3	91.7±2.5 / 85.4±1.3	85.1±3.2 / 79.2±1.5	47.1±4.6 / 28.6±1.6
IRA	25.6±4.0 / 1.8±0.5	N/A	14.0±3.2 / 3.7±0.7	19.0±3.6 / 4.2±0.7
FMCW	39.3±4.5 / 24.0±1.6	74.4±4.1 / 46.3±1.8	36.8±4.5 / 5.0±0.8	38.5±4.5 / 12.6±1.2
SFCW	30.6±4.2 / 9.0±1.0	59.5±4.5 / 13.0±1.2	26.4±4.0 / 0.9±0.3	28.1±4.1 / 5.1±0.8
UWB	98.3±1.2 / 93.8±0.9	88.4±2.9 / 80.1±1.4	100.0±0.0 / 90.4±1.1	94.2±2.1 / 75.8±1.5
Wi-Fi	93.3±2.3 / 91.1±1.0	90.8±2.6 / 91.0±1.0	91.7±2.5 / 92.3±1.0	81.7±3.5 / 60.5±1.8
IMU	96.6±1.7 / 96.5±0.7	97.4±1.5 / 95.7±0.7	98.3±1.2 / 95.7±0.7	94.0±2.2 / 35.8±1.8

Evaluation results

香 港 大 學
THE UNIVERSITY OF HONG KONG

Human Activity Recognition (HAR): 62 daily activities across five categories (body-only, human–object, human–computer, human–human, medical), class-balanced.

Table 1: 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	Model			
	ResNet	DenseNet	Swin-T	RFNet
RGB	91.5±2.6 / 93.4±0.9	93.2±2.3 / 91.2±1.0	94.9±2.0 / 93.1±0.9	89.7±2.8 / 60.9±1.8
Depth	89.7±2.8 / 86.6±1.2	90.6±2.7 / 83.2±1.3	86.3±3.2 / 81.7±1.4	87.2±3.1 / 40.0±1.8
ToF	86.8±3.1 / 70.3±1.6	N/A	82.6±3.5 / 51.8±1.8	89.3±2.8 / 75.9±1.5
Thermal	90.1±2.7 / 85.0±1.3	91.7±2.5 / 85.4±1.3	85.1±3.2 / 79.2±1.5	47.1±4.6 / 28.6±1.6
FMCW	39.3±4.5 / 24.0±1.6	74.4±4.1 / 46.3±1.8	36.8±4.5 / 5.0±0.8	38.5±4.5 / 12.6±1.2
SFCW	30.6±4.2 / 9.0±1.0	59.5±4.5 / 13.0±1.2	26.4±4.0 / 0.9±0.3	28.1±4.1 / 5.1±0.8
UWB	98.3±1.2 / 93.8±0.9	88.4±2.9 / 80.1±1.4	100.0±0.0 / 90.4±1.1	94.2±2.1 / 75.8±1.5
Wi-Fi	93.3±2.3 / 91.1±1.0	90.8±2.6 / 91.0±1.0	91.7±2.5 / 92.3±1.0	81.7±3.5 / 60.5±1.8
Acoustic	40.8±4.5 / 45.5±1.8	60.0±4.5 / 54.6±1.8	36.7±4.4 / 32.1±1.7	29.2±4.2 / 19.1±1.4
IMU	96.6±1.7 / 96.5±0.7	97.4±1.5 / 95.7±0.7	98.3±1.2 / 95.7±0.7	94.0±2.2 / 35.8±1.8

Evaluation results

香 港 大 學
THE UNIVERSITY OF HONG KONG

Human Activity Recognition (HAR): 62 daily activities across five categories (body-only, human–object, human–computer, human–human, medical), class-balanced.

Table 1: 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	Model			
	ResNet	DenseNet	Swin-T	RFNet
RGB	91.5±2.6 / 93.4±0.9	93.2±2.3 / 91.2±1.0	94.9±2.0 / 93.1±0.9	89.7±2.8 / 60.9±1.8
Depth	89.7±2.8 / 86.6±1.2	90.6±2.7 / 83.2±1.3	86.3±3.2 / 81.7±1.4	87.2±3.1 / 40.0±1.8
ToF	86.8±3.1 / 70.3±1.6	N/A	82.6±3.5 / 51.8±1.8	89.3±2.8 / 75.9±1.5
Thermal	90.1±2.7 / 85.0±1.3	91.7±2.5 / 85.4±1.3	85.1±3.2 / 79.2±1.5	47.1±4.6 / 28.6±1.6
IRA	25.6±4.0 / 1.8±0.5	N/A	14.0±3.2 / 3.7±0.7	19.0±3.6 / 4.2±0.7
FMCW	39.3±4.5 / 24.0±1.6	74.4±4.1 / 46.3±1.8	36.8±4.5 / 5.0±0.8	38.5±4.5 / 12.6±1.2
UWB	98.3±1.2 / 93.8±0.9	88.4±2.9 / 80.1±1.4	100.0±0.0 / 90.4±1.1	94.2±2.1 / 75.8±1.5
Wi-Fi	93.3±2.3 / 91.1±1.0	90.8±2.6 / 91.0±1.0	91.7±2.5 / 92.3±1.0	81.7±3.5 / 60.5±1.8
Acoustic	40.8±4.5 / 45.5±1.8	60.0±4.5 / 54.6±1.8	36.7±4.4 / 32.1±1.7	29.2±4.2 / 19.1±1.4
IMU	96.6±1.7 / 96.5±0.7	97.4±1.5 / 95.7±0.7	98.3±1.2 / 95.7±0.7	94.0±2.2 / 35.8±1.8

Evaluation results (cont.)

香 港 大 學
THE UNIVERSITY OF HONG KONG

Human Pose Estimation (HPE): Multi-view, multi-modal data with motion-capture ground truth (3D skeletons) enabling precise cross-view/cross-sensor training and evaluation on both structured aerobics and freestyle sequences.

Table 2: HPE results (MPJPE in millimeters; lower is better) across sensing modalities. "N/A" denotes model incompatibility. Values are to one decimal with standard error of the mean as x.x.

Modality	Model			
	ResNet	DenseNet	Swin-T	RFNet
RGB	133.3±4.4	147.2±5.1	269.6±6.2	162.8±4.6
Depth	131.4±4.5	147.4±4.6	248.2±6.5	194.8±5.7
Thermal	142.8±4.7	147.0±4.9	259.9±5.9	254.3±6.1
IRA	244.4±6.8	N/A	261.1±6.4	265.1±6.7
FMCW	198.5±5.7	185.4±5.4	272.5±7.3	220.9±6.0
SFCW	206.7±6.2	202.9±6.2	264.2±6.4	270.7±6.6
UWB	142.4±4.8	158.0±5.2	260.5±6.1	159.5±5.0
Wi-Fi	147.3±4.7	147.4±4.9	262.2±6.0	186.8±5.3
Acoustic	258.8±6.9	256.8±6.7	271.2±6.7	243.6±6.8
IMU	147.9±5.0	159.3±5.5	251.6±6.4	180.9±5.3

Evaluation results (cont.)

香 港 大 學
THE UNIVERSITY OF HONG KONG

Human Pose Estimation (HPE): Multi-view, multi-modal data with motion-capture ground truth (3D skeletons) enabling precise cross-view/cross-sensor training and evaluation on both structured aerobics and freestyle sequences.

Table 2: HPE results (MPJPE in millimeters; lower is better) across sensing modalities. "N/A" denotes model incompatibility. Values are to one decimal with standard error of the mean as x.x.

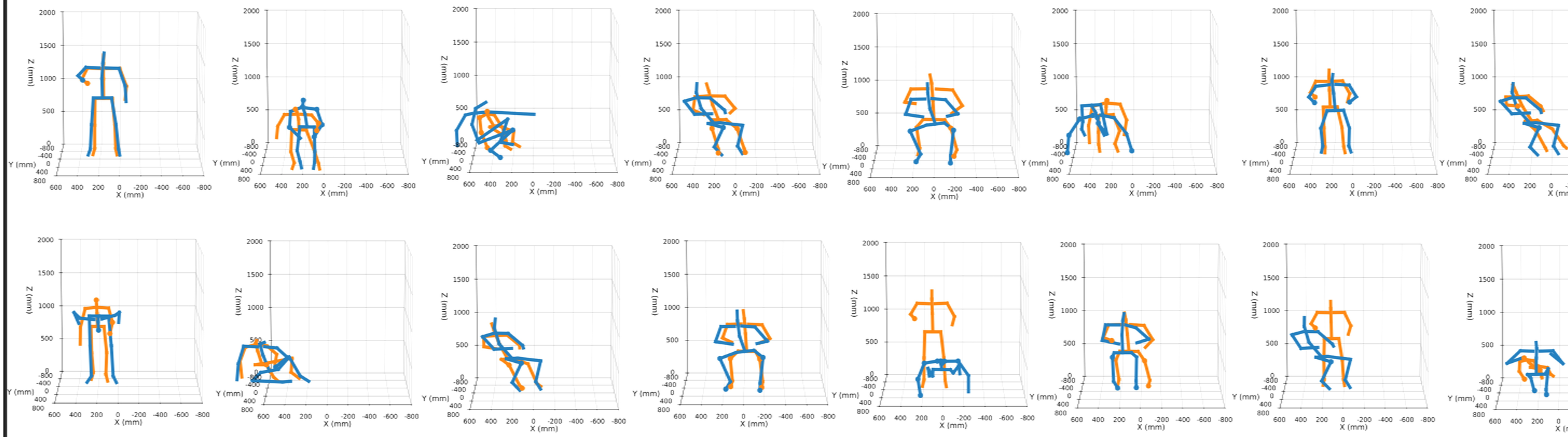
Modality	Model			
	ResNet	DenseNet	Swin-T	RFNet
RGB	133.3±4.4	147.2±5.1	269.6±6.2	162.8±4.6
ToF	152.5±5.2	N/A	252.1±6.0	162.2±5.0
Thermal	142.8±4.7	147.0±4.9	259.9±5.9	254.3±6.1
FMCW	198.5±5.7	185.4±5.4	272.5±7.3	220.9±6.0
UWB	142.4±4.8	158.0±5.2	260.5±6.1	159.5±5.0
Wi-Fi	147.3±4.7	147.4±4.9	262.2±6.0	186.8±5.3
IMU	147.9±5.0	159.3±5.5	251.6±6.4	180.9±5.3

Visualization of human pose estimation

香 港 大 學
THE UNIVERSITY OF HONG KONG

Ground truth skeletons: Predicted skeletons:

RGB Image | Depth Image | ToF Map | Thermal Image | IRA Map | FMCW Radar | SFCW Radar | UW

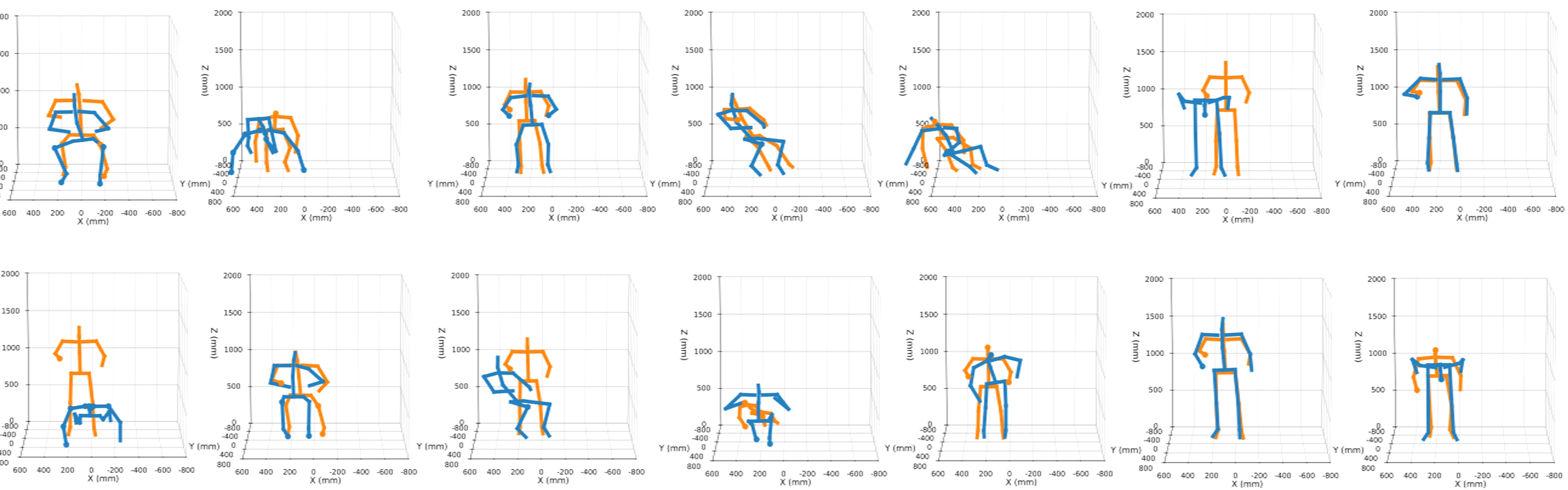


Visualization of human pose estimation

香 港 大 學
THE UNIVERSITY OF HONG KONG

Ground truth skeletons: Predicted skeletons:

IRA Map | FMCW Radar | SFCW Radar | UWB | Wi-Fi | Acoustic | IMU



NEURAL INFORMATION PROCESSING SYSTEMS

香 港 大 學
THE UNIVERSITY OF HONG KONG

THANK YOU