RAV4D: A Radar-Audio-Visual Dataset for Indoor Multi-Person Tracking

Yi Zhou
School of Advanced Technology
Xi'an Jiaotong-Liverpool
University
Suzhou, China
zhouyi1023@tju.edu.cn

Ningfei Song
School of Advanced Technology
Xi'an Jiaotong-Liverpool
University
Suzhou, China
Ningfei.Song21@student.xjtlu.edu.cn

Jieming Ma
School of Advanced Technology
Xi'an Jiaotong-Liverpool
University
Suzhou, China
Jieming.Ma@xjtlu.edu.cn

Ka Lok Man
School of Advanced Technology
Xi'an Jiaotong-Liverpool
University
Suzhou, China
Ka.Man@xjtlu.edu.cn

Miguel López-Benítez
Department of Electrical Engineering and Electronics
University of Liverpool
Liverpool, UK
mlopben@liverpool.ac.uk

Limin Yu*
School of Advanced Technology
Xi'an Jiaotong-Liverpool
University
Suzhou, China
limin.yu@xjtlu.edu.cn

Yutao Yue*
Institute of Deep Perception Technology
JITRI
Wuxi, China
ytyue@ustc.edu

***Abstract*—Indoor multi-person tracking is a widely explored area of research. However, publicly available datasets are either oversimplified or provide only visual data. To fill this gap, our paper presents the RAV4D dataset, a novel multimodal dataset that includes data from radar, microphone arrays, and stereo cameras. This dataset is characterised by the provision of 3D positions, Euler angles and Doppler velocities. By integrating these different data types, RAV4D aims to exploit the synergistic and complementary capabilities of these modalities to improve tracking performance. The development of RAV4D addresses two main challenges: sensor calibration and 3D annotation. A novel calibration target is designed to effectively calibrate the radar, stereo camera and microphone array. In addition, a visually guided annotation framework is proposed to address the challenge of annotating radar data. This framework uses head positions, heading orientation and depth information from stereo cameras and radar to establish accurate ground truth for multimodal tracking trajectories. The dataset is publicly available at https://zenodo.org/records/10208199.**

***Index Terms*—Multiple Object Tracking, Sensor Fusion, Speaker Tracking, Radar Tracking**

I. Introduction

Indoor multi-person tracking has emerged as a critical technology in several areas, including video conferencing, human-computer interfaces and virtual reality. This technology enables real-time monitoring and analysis of human movement and behaviour, providing valuable insights for a range of applications. At the heart of indoor multi-person tracking is the ability to effectively detect and smoothly track people in complex indoor environments. This is achieved through the fusion of sensing modalities, each contributing its unique strengths.

Cameras are the most common sensor modality for indoor multi-person tracking. Modern visual detection relies primarily on appearance models for both detection and association [1]. The high accuracy of visual detectors simplifies the tracking process, with many algorithms using a simple Kalman filter for tracking. However, appearance models have limitations and can fail under certain conditions, such as extreme illumination, similar appearances, or occlusion. These scenarios often lead to significant performance degradation in many tracking algorithms. In addition, visual recognition inherently lacks 3D information and raises privacy concerns.

Microphones, known for their cost-effectiveness and maturity, are widely used in indoor environments. By utilising microphone arrays, it is possible to determine the location of sound sources in space [2]. A key aspect of this process is determining the Direction of Arrival (DOA) of sound signals, which is critical to accurately determining where a sound is coming from. Once the DOA is determined, advanced audio enhancement techniques can be implemented in the specified direction, resulting in significant improvements in audio quality.

Millimetre-wave radars are increasingly being used in sensing applications due to their compact size, affordability and mature manufacturing [3]. 4D radar sensors are capable of measuring 3D locations, making them practical for consumer-level personal monitoring and tracking applications. Radar sensors offer advantages over cameras in terms of privacy, 3D measurement and robustness to lighting variations.

Benchmarking the performance of different modalities for human tracking requires a common dataset. However, current indoor human perception datasets present several challenges. First, most are designed for detection tasks and not specifically for multi-object tracking. This design choice results in a lack of crossing trajectories, which are crucial for testing advanced tracking algorithms in dynamic scenarios. In addition, these datasets typically provide only 2D annotations, limiting the

analysis of activities that involve significant spatial movement, such as standing or bending. Another limitation is the limited range of modalities. While visual detections are reliable under standard conditions, they may fall short in scenarios with extreme lighting, similar appearances or occlusions. Such situations highlight the need to incorporate other sensors, such as microphone arrays and radar, which can provide essential data that cannot be captured visually.

In response to these gaps, our work presents the multimodal dataset RAV4D, which includes data from radar, microphone arrays and stereo cameras. This dataset is unique in providing 3D positions, Euler angles and micro-Doppler velocities, with the aim of exploiting the synergy and complementary information of these modalities for robust tracking performance. The creation of RAV4D addresses two main challenges: sensor calibration and 3D annotation. We design a novel calibration target that effectively calibrates the radar, stereo camera and microphone array. In addition, we address the challenge of annotating radar data with a visually guided annotation framework that uses stereo camera detections and depth information to establish ground truth for radar detections and trajectories.

The rest of this article is organised as follows: Section II reviews related datasets in multiple person tracking research. Section III describes the specifications of the sensors used and our data collection methods. In section IV we discuss spatial calibration methods for multi-sensor setups. Section V describes the pre-processing pipeline for each sensor modality. In section VI we present our visually guided annotation pipeline and visualisation tool. section VII analyses example scenarios from our dataset. Finally, in section VIII we summarise the dataset and discuss possible directions for future research.

II. Related Works

In visual detection and tracking, key datasets include the Multiple Object Tracking (MOT) Challenge [4] for tracking multiple individual objects, and DanceTrack [5] for tracking in dynamic environments such as dancing. However, most research has focused on 2D tracking, often ignoring the complexities of 3D human motion.

In audiovisual tracking, which uses audio cues to improve tracking performance, especially in environments with occlusions and unrestricted motion, several key datasets stand out. As detailed in table I, these include the AV 16.3 corpus [6], which is specifically designed for multi-speaker tracking and addresses complex scenarios such as overlapping speech. In addition, SPEVI [7] provides data for multimodal person detection and tracking. AVDIAR [8] provides various multi-speaker scenarios, and CAV3D [9] is recorded on a co-located audio-visual platform for 3D tracking.

Radar sensing for indoor human tracking is limited by a lack of public datasets, with most research centred on pose reconstruction using 4D radar. These studies typically involve stationary subjects and do not fully address tracking challenges such as varying distances, occlusions and viewing angles. Some automotive radar datasets provide tracking annotation of vehicles in outdoor environments. However, in indoor environments, human motion is characterised by a higher degree of flexibility, with increased instances of crossing paths and occlusions. These dynamics present unique challenges to radar-based tracking, requiring more sophisticated algorithms and sensor setups to accurately track human movement indoors.

TABLE I
Multi-Modal Indoor MOT Datasets

Dataset	# Mic	# Cam	Radar	Annot.	# Speakers
AV 16.3 [6]	16	3	no	3D	3
AVDIAR [8]	6	2	no	3D	4
AVTRACK [10]	4	1	no	2D	2
SPEVI [7]	2	1	no	2D	2
CAV3D [9]	8	1	no	3D	3
RAV4D	6	2 (stereo)	yes	3D	3

III. Sensor and Data Recording

A. Data Collection Scenario

The data was collected in a medium-sized meeting room, as shown in fig. 1. The room has a large desk in the centre, surrounded by various other items such as chairs and a whiteboard. Our sensor suite consists of a stereo camera positioned at the centre of the lower edge of the room, a 4D radar sensor in the left corner and a circular microphone array placed on the desk.

To create a dynamic and challenging scenario suitable for MOT tasks, we had one to three people moving around the desk, often crossing paths. This scenario was essential for investigating the problem of identity switching, which is common in tracking applications. To improve sound localisation, participants were instructed to speak loudly, which helped to capture clear audio data.

B. Sensor Modalities

Our dataset contains three sensor modalities: a stereo camera, a 4D FMCW radar and a circular microphone array.

1) Stereo Camera: The stereo camera in our setup captures high-resolution images of 960 x 540 pixels at a frame rate of 30 frames per second. Alongside the visual data, it generates a synchronised depth map covering up to 20 metres, with an accuracy range of 0.5% to 2%. The camera has a field of view of 110 degrees horizontally and 70 degrees vertically.

2) FMCW Radar: The 4D FMCW radar operates at 77 GHz with a bandwidth of 750 MHz. It is a 4-chip cascaded MIMO system with an array of 12 transmit (TX) and 16 receive (RX) antennas, resulting in a virtual 2D array of 192 elements. This radar system offers a range resolution of 0.22m and angular resolutions of 1 degree in azimuth and 2 degrees in elevation. It acquires data at a rate of 20 frames per second.

3) Circular Microphone Array: The microphone array is a 6-element circular design with a diameter of 7 cm. It has an embedded audio-enhanced front end to improve SNR. This array is capable of detecting sound events up to 10 metres away with an angular resolution of approximately 5 degrees.

It records audio in a 6-channel format at a sampling rate of 16 kHz.

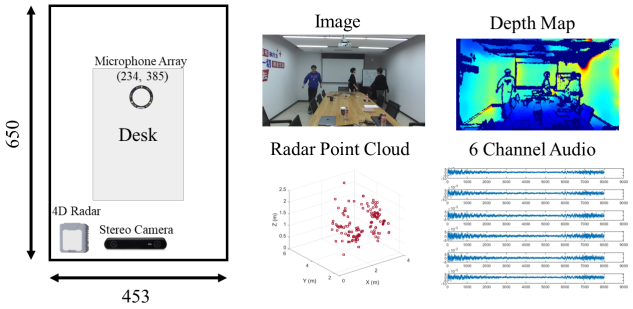


Fig. 1. Room layout and sensor outputs

IV. Multi-sensor Calibration

Calibration was performed using a handmade calibration target consisting of a corner reflector, a checkerboard glued to a clapperboard, as shown in fig. 2. The corner reflector is detectable by radar as a strong point target. The clapperboard, originally used in film production to synchronise audio and video in post-production, is repurposed in our study. We have attached a checkerboard pattern to it to facilitate calibration of the camera and microphone array. The checkerboard's easily identifiable corner features make it ideal for camera calibration. Its attachment to the flat surface of the clapperboard ensures visibility to the camera. In addition, the clapperboard's distinctive clapping sound provides a reference signal for calibrating the DoA angle estimated by the microphone array. The vertical offset between the clapperboard and the corner reflector was measured manually.

The calibration target was placed at seven different locations in the meeting room, varying in height and evenly spaced. The origin of the world coordinate system was set at the left corner of the room. To determine the position of the corner reflector, the ceramic tiles were used as grid units by placing the reflector at a tile corner and counting the coordinates, then multiplying by the tile edge length to obtain the x-y coordinates. The z coordinate was measured directly with a ruler.

After collecting the measurements from the camera, radar and audio sensors, along with their corresponding ground truth values, we computed the camera-to-world and radar-to-world transformation matrices. In addition, we calibrated the audio DoA by aligning the estimated angle from the microphone array with the ground truth angle determined by the spatial relationship between the calibration target and the fixed position of the microphone array.

V. Pre-Processing

A. Radar Data Filtering

In this study, we use a commercial high-resolution 4D radar to capture human motion. This radar can be configured

Fig. 2. Calibration target for microphone, camera and radar

to output the raw radar point cloud produced by the Constant False Alarm Rate (CFAR) detector. Each point is a 5-dimensional vector containing range, azimuth angle, elevation angle, Doppler velocity and RCS. In indoor environments, multi-path propagation often results in significant 'ghost' objects. However, with the room layout available, we can eliminate these artefacts by defining a 3D Region of Interest (ROI) corresponding to the size of the room.

Furthermore, with the radar stationed in a fixed position, we construct an Occupancy Grid Map (OGM) to model the static environment. This OGM, which provides a more robust approach compared to direct point representations, is better suited to tolerate spatial uncertainties in the measurements. We discretise the spatial ROI into a grid with a cell resolution of 0.1 metres. The static OGM is then constructed by the temporal accumulation of static detections, using a fixed count threshold to identify occupied cells. Following these steps, we refine the radar point cloud by removing points that coincide with the static clutter maps. After this filtering process, we use the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to cluster the radar detections into distinct objects and further remove the isolated clutter at each timestamp.

B. Audio DoA Estimation

To estimate the audio DoA, we use the Steered Response Power - Phase Transform (SRP-PHAT) method. The algorithm computes a steering vector for each potential direction through a delay-and-sum beamformer and applies the PHAT weighting function. The PHAT function normalises the magnitude and uses the phase information to calculate the correlation. DoA candidates are then identified from the peaks in the output power spectrum. Due to computational requirements, we implement this process using a grid cell size of one degree. Although originally designed for single source scenarios, SRP-PHAT is capable of handling multiple sources, provided the number of sources is known [11]. Indoor reverberation often leads to noisy results, and periods of silence during motion pose further challenges to accurate DoA estimation. To address these issues, we apply 1D filtering to remove outliers and use interpolation to fill gaps in the DoA trajectories. Finally, we transform the measured DoA into world coordinates and align them with the measured position of the microphone array, as shown in fig. 3.

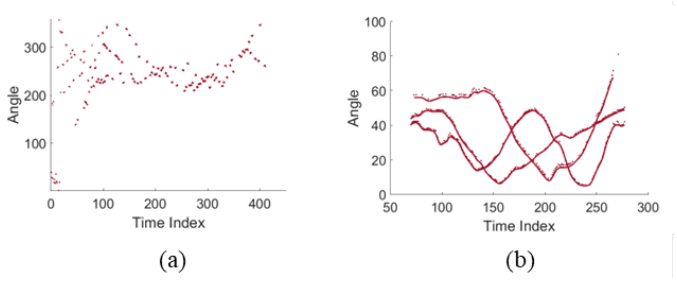


Fig. 3. DoA estimation: (a) raw DoA estimated by SRP-PHAT (b) smoothed DoA trajectories of three speakers in the world coordinate

C. Head Pose Detection through Stereo Camera

In our study, we represent the human head as a point target instead of using traditional bounding box representations. The point target is particularly suitable for tracking tasks. We define the point target as a six-dimensional vector containing the 3D coordinates of the head and its orientation vector, represented by Euler angles. To detect the head, we use a pre-trained YOLOv3 detector [12] to identify the bounding box and calculate the centre of the box as the location of the point target. The depth information for this central point is obtained by projecting it onto the depth map generated by the stereo camera. It is important to note, however, that depth maps can sometimes be incomplete, often producing NaNs (Not a Number) due to missing features or occlusions. Fortunately, due to the high frame rate of our system, depth measurements are continuous over time. This continuity allows us to smoothly refine the depth measurements, thereby improving the accuracy of the head centre depth trajectory. In addition, we predict head orientation using the WHENet [13], where the Euler angles are regressed through an additional MLP layer with the head feature map as input.

VI. Visually Guided Trajectory Annotation

As illustrated in fig. 4, we propose a visually guided annotation framework for trajectories. With respect to trajectory annotation, the sparse nature of the radar point cloud presents a challenge in determining a reliable human point representation. The centre of a radar cluster often results in a zigzag trajectory. To overcome this, we use the heads detected by the stereo camera data as tracked objects. The 3D position of the head can be determined from the pixel coordinate in the image plane, the depth and the intrinsic parameters of the camera. However, there are two main challenges: the lack of strict synchronisation between the stereo camera and the radar sensor, and the inaccuracies in the depth estimation by the stereo camera. Factors such as calibration inaccuracies, lens distortion, image noise and algorithmic limitations can lead to errors in visual depth measurements. In contrast, the radar sensor can directly measure spatial information with high accuracy. Therefore, a fusion algorithm is required to correct visual depth using radar measurements.

To address these challenges, we design a depth calibration module that fuses radar and visual depth. First, we convert radar detections into depth information using extrinsic calibration information. We then accumulate the radar depth trajectory for each individual and interpolate it to match the camera timestamps. Next, we apply an iterative optimisation module to align the visual depth trajectory with the radar depth trajectory, treating the time offset and depth scaling parameter as optimisable variables. After adjusting the visual depth based on these variables, we smooth the trajectory using both the corrected visual depth and the radar depth to obtain the merged depth trajectory. Finally, we compute the 3D trajectory based on the 2D positions and depth. To ensure the accuracy of our dataset, we further develop a GUI interface to visualise the annotations on a frame-by-frame basis, allowing us to check and correct any missed annotations.

The ground truth in our study is generated according to the camera timestamps. For radar-centred applications, such as studying the effect of Doppler information, it is necessary to interpolate the ground truth trajectory to match the slower acquisition frame rate of the radar. Since our analysis includes estimation of head orientation, this interpolation process should be performed in SE(3) space, which takes into account both translation and rotation. To facilitate this, we first convert Euler angles to quaternions and then apply Spherical Linear Interpolation (Slerp) [14] to obtain smooth and continuous trajectories for both head position and heading angles.

VII. Dataset Analysis

A. Calibration Results

Accurate spatial calibration between radar and camera systems is fundamental to our visually guided trajectory annotation pipeline. In this section we present the results of this calibration. First, we project the radar points into the image view, resulting in a reprojection error of 8.6 pixels, given the image resolution of 960 x 540. As our primary focus is on tracking in world coordinates rather than in the image plane, we further assessed the reprojection error in world coordinates.

To do this, we use a transformation matrix to convert radar detections from radar-centred coordinates to world coordinates. Similarly, we use the camera's own matrix and depth information to project image points into world coordinates. The calibration results are quantified by the L2 reprojection error when comparing the radar and image data to ground truth. The reprojection errors are 0.05 metres for the radar data and 0.01 metres for the imagery. These figures demonstrate the accuracy of our calibration process.

B. Dataset Contents

Our dataset is designed to include a range of scenarios to thoroughly test the tracking algorithms under different conditions. It includes three primary cases, categorised according to the number of people involved: one person, two people and three people. For each case, we have developed two different scenarios to simulate different lighting conditions: one with the lights on and one with the lights off. Within each scenario, three classes of trajectories of varying complexity are defined:

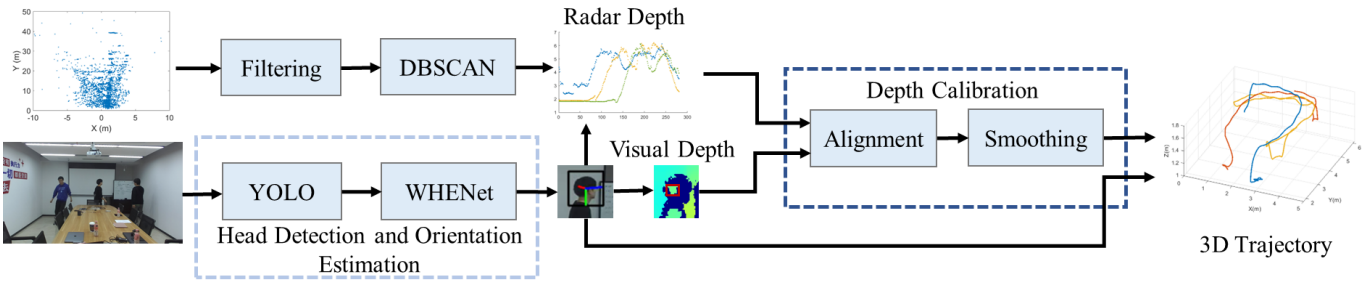


Fig. 4. Visual guided annotation pipeline

- **Simple case** (fig. 5 (a)): In this case, people start from their seats, move to the front of the room, and then return to their seats, all without intersecting paths.
- **Normal case** (fig. 5 (b)): In this case three people cross at the front of the room. The individual following the yellow trajectory crosses the other two individuals twice, for a total of four crossings.
- **Hard Case** (fig. 5 (c)): This case is the most difficult, with each individual crossing the paths of the other two, for a total of six crossings.

It should be noted that the placement of the sensors ensures a suitable field of view to cover the whole of the movement process. Therefore, the number of people is mostly constant during the tracking process. To better evaluate the challenging case where the number of people can vary, the user can define a narrow FoV, such as the upper half of the meeting room. In this case, the person may frequently enter or leave the FoV, making tracking more difficult.

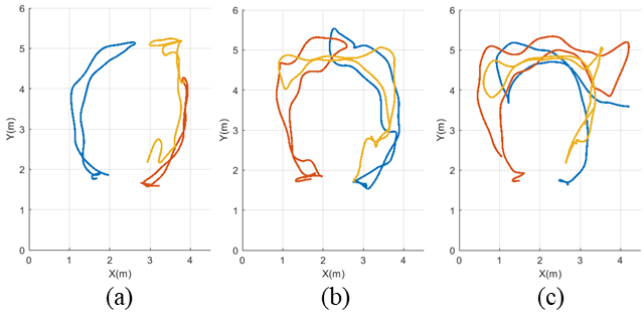


Fig. 5. Three levels of difficluties: (a) easy (b) normal (c) hard

C. Comparing Visual and Radar Tracking in 3D Space

o compare the performance of radar tracking and visual tracking. We implemented a basic MOT tracker using an extended Kalman filter (EKF) for tracking and a global nearest neighbour (GNN) algorithm for classification. The evaluation metrics chosen are MOTA and MOTP as defined by the CLEAR MOT metrics [15]. We chose the 3-person, light, hard case as our test sequence.

As detailed in table table II, radar tracking shows superior performance in both MOTA and MOTP compared to visual tracking. A higher MOTA score for radar tracking indicates better overall tracking accuracy, suggesting that radar tracking is more effective at correctly identifying and tracking objects, with fewer errors such as missing targets or incorrectly tracking irrelevant objects. In addition, radar tracking outperforms visual tracking in terms of MOTP. This suggests that radar tracking not only detects and tracks objects more reliably, but also with greater spatial accuracy, resulting in more precise localisation of tracked objects.

TABLE II
TRACKING PERFORMANCE

Methods	MOTP (%)	MOTA (%)
Visual Tracking	71.412	83.102
Radar Tracking	82.088	87.274

D. Challenges

In our dataset, we address two challenging aspects: variations in illumination and occlusion scenarios.

1) Illumination Condition: A key aspect we investigate in our dataset is the effect of changes in lighting. Figure 6 shows a scenario where the lights in the meeting room are turned off, creating a low-light environment. Despite the reduction in light, the high dynamic range of our camera ensures that the overall image quality is still acceptable. The main challenge comes from extreme lighting contrasts. For example, bright light from a projector in a dark room can significantly reduce the visibility of a person walking in front of it. Our visual detector performs robustly under strong lighting perturbations (as shown in fig. 6 (a)), but struggles when a person's face is obscured by the texture of slides from a projector (as shown in fig. 6 (b)), risking loss of head detection. These conditions highlight the complexity of tracking in different lighting environments and emphasise the need for multi-sensory fusion to achieve reliable tracking.

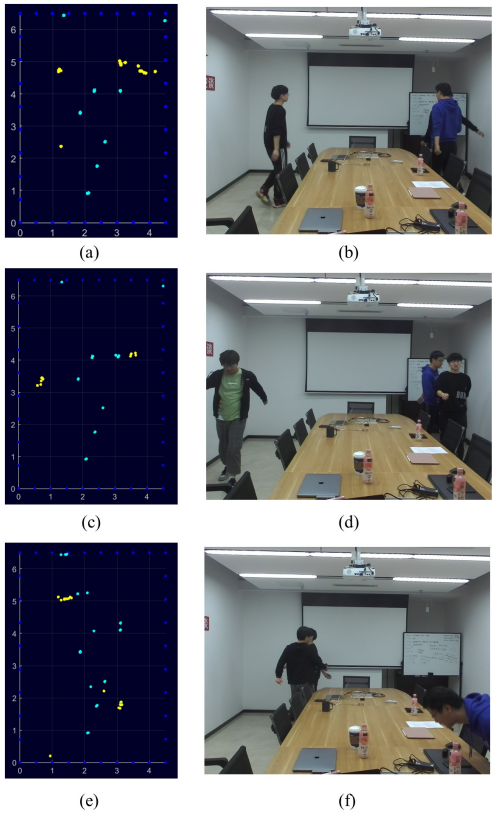
2) Occlusion Scenario: Occlusion poses a significant challenge in our dataset, manifesting itself in two primary forms. The first type of occlusion is environmental, caused by the layout and objects in the room. For example, when the radar is positioned in the left corner, the individual on the left side of the room is more clearly detected, resulting in a denser point cloud. In contrast, the two people on the right are partially

(a) (b)

Fig. 6. Illumination chanllenge in dark scenarios

obscured by the table, resulting in a sparser point cloud. The second type of occlusion is due to trajectory crossing. As the camera and radar are at different positions, occlusions occur at different angles, affecting the visibility of individuals.

In the first row of fig. 7, figure (b) illustrates a scenario where two individuals on the right are visually occluded in the camera's view, while the radar detections in figure (a) clearly identify them. Conversely, the second row shows a situation where two individuals are visible in the camera image (d) but occluded in the radar detections (c). The last row shows a case where occlusions occur simultaneously in both image and radar data. In such cases, the continuous audio DoA becomes crucial and provides an alternative method of confirming the presence of individuals that are occluded in both visual and radar sensors.



(a) (b) (c) (d) (e) (f)

Fig. 7. Occlusion cases: (a) visual occlusion (b) radar occlusion (c) both radar and visual occlusion

VIII. Conclusions

This paper presents the RAV4D dataset, a novel multimodal dataset that integrates 4D radar, audio and visual data to improve multi-person tracking algorithms in challenging indoor environments. RAV4D provides high quality 3D annotations by overcoming key challenges in sensor calibration and radar data interpretation. This involved the creation of a novel calibration target and the development of a visually guided annotation framework. The dataset includes complex trajectories with numerous crossings and a variety of challenging scenarios such as low illumination and occlusion. These features establish RAV4D as an invaluable resource for researchers and practitioners seeking to explore and advance the capabilities of multimodal sensing in complex indoor environments.

References

[1] G. Wang, M. Song, and J.-N. Hwang, "Recent advances in embedding methods for multi-object tracking: a survey," *arXiv preprint arXiv:2205.10766*, 2022.

[2] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, B. Lee, *et al.*, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[3] Y. Zhou, L. Liu, H. Zhao, M. López-Benítez, L. Yu, and Y. Yue, "Towards deep radar perception for autonomous driving: Datasets, methods, and challenges," *Sensors*, vol. 22, no. 11, p. 4208, 2022.

[4] P. Dendorfer, A. Osep, A. Milan, K. Schindler, D. Cremers, I. Reid, S. Roth, and L. Leal-Taixé, "Motchallenge: A benchmark for single-camera multiple target tracking," *International Journal of Computer Vision*, vol. 129, pp. 845–881, 2021.

[5] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, and P. Luo, "Dancetrack: Multi-object tracking in uniform appearance and diverse motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 993–21 002.

[6] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 182–195.

[7] M. Taj, "Surveillance performance evaluation initiative (spevi)—audiovisual people dataset," *Internet: http://www. eecs. qmul. ac. uk/ andrea/spevi. html*, 2007.

[8] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.

[9] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio–visual sensing device," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.

[10] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, "Tracking the active speaker based on a joint audio-visual observation model," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 15–21.

[11] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[12] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[13] Y. Zhou and J. Gregson, "Whenet: Real-time fine-grained estimation for wide range head pose," *arXiv preprint arXiv:2005.10353*, 2020.

[14] M. Zefran and V. Kumar, "Two methods for interpolating rigid body motions," in *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No. 98CH36146)*, vol. 4. IEEE, 1998, pp. 2922–2927.

[15] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.