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

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

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

***Abstract*—Indoor multiple person tracking is a widely explored research field. However, publicly available datasets either are overly simplified or provide solely visual data. To address this gap, our paper introduces the RAV4D dataset, a novel multimodal dataset that encompasses data from radar, microphone arrays, and stereo cameras. This dataset stands out by providing 3D locations, Euler angles, and Doppler velocities. By integrating these diverse data types, RAV4D aims to leverage the synergistic and complementary capabilities of these modalities to enhance tracking performance. The creation of RAV4D tackles two primary challenges: sensor calibration and 3D annotation. A novel calibration target is designed to effectively calibrate the radar, stereo camera, and microphone array. Additionally, a visual-guided annotation framework is proposed to address the challenge of annotating radar data. This framework utilizes head locations, 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 multiple people tracking has emerged as a crucial technology in various domains, including video conferencing, human-computer interfaces, and virtual reality. This technology enables real-time monitoring and analysis of human movement and behavior, providing valuable insights for a range of applications. At the heart of indoor multiple people tracking lies the ability to effectively detect and smoothly track individuals in complex indoor environments. This task is accomplished through a combination of sensor modalities, each contributing its unique strengths.

Cameras stand as the most prevalent sensor modality for indoor multiple people tracking. Modern visual detection primarily relies on appearance models for both detection and association [1]. The high accuracy of visual detectors simplifies the tracking process, with many algorithms employing a simple Kalman filter for tracking [2]. However, appearance models have their limitations and may fail in certain conditions, such as extreme illuminance, similar appearances, or occlusions. These scenarios often lead to significant performance drops in many tracking algorithms. Additionally, visual detection inherently lacks 3D information and raises privacy concerns.

Microphones, known for their cost-effectiveness and maturity, are widely utilized in indoor settings. By utilizing the microphone array, we can determine the location of sound sources in space [3]. A key aspect of this process involves identifying the Direction of Arrival (DOA) of sound signals, which is crucial for accurately detecting where a sound is originating from. Once the DOA is established, advanced audio enhancement techniques can be implemented in the specified direction, leading to significant improvements in audio quality.

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

Benchmarking the performance of different modalities for human tracking requires a common dataset. However, current indoor human perception datasets present several challenges. Firstly, most are designed for detection tasks and not specifically crafted for multi-object tracking. This design choice leads to a scarcity of crossing trajectories, which are crucial for testing advanced tracking algorithms in dynamic scenarios. Moreover, these datasets typically offer only 2D annotations, limiting the analysis of activities that involve significant spatial movement, such as standing up or bending. Another limitation is the restricted range of modalities. While visual detections are reliable in standard conditions, they can fall short in scenarios with extreme illumination, similar appearances, or occlusions. Such situations reveal the necessity of incorporating other sensors like microphone arrays and radar, which can provide essential data not captured visually.

In response to these gaps, our work introduces the multimodal dataset RAV4D, which includes data from radar, microphone arrays, and stereo cameras. This dataset is unique in providing 3D locations, Euler angles, and Micro-Doppler velocities, aiming to leverage 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. Additionally, we tackle the challenge of annotating radar data with a visual-guided annotation framework, using stereo camera detections and depth information to establish ground truth for radar detections and trajectories.

The remainder of this article is organized as follows: Section II reviews related datasets in multiple people tracking research. Section III details the specifications of the sensors used and our data recording methods. In section IV, we discuss the spatial calibration methods for multi-sensor setups. Section V describes the pre-processing pipeline for each sensor modality. In section VI, we introduce our visual guided annotation pipeline and visualization tool. Section VII analyzes example scenarios from our dataset. Finally, in section VIII, we summarize the dataset and discuss potential directions for future research.

II. Related Works

In visual detection and tracking, key datasets include the Multiple Object Tracking (MOT) Challenge [5] for multiple individual tracking and DanceTrack [6] for tracking in dynamic environments like dancing. However, most research focuses on 2D detection, often ignoring the complexities of 3D human motion.

In audio-visual tracking, which leverages audio cues to enhance tracking performance, especially in environments with occlusions and unrestricted movement, several key datasets stand out. As detailed in table I, these include the AV 16.3 corpus [7], specifically designed for multiple speaker tracking and addressing complex scenarios such as overlapping speech. Additionally, SPEVI [8] provides data for multi-modal people detection and tracking. AVDIAR [9] offers various multi-speaker scenarios, and CAV3D [10] is recorded on a co-located audio-visual platform for 3-D tracking.

Radar sensing for indoor people tracking is limited by a lack of public datasets, with most research centered on pose reconstruction using 4D radar. These studies typically involve stationary subjects, not fully addressing tracking challenges like varying distances, occlusion, and viewing angles. Some automotive radar datasets [11], [12] provide the tracking annotation of vehicles in outdoor environments. However, in indoor settings, human movements are characterized by a higher degree of flexibility, with increased instances of crossing paths and occlusions. These dynamics present unique challenges for radar-based tracking, necessitating more sophisticated algorithms and sensor setups to accurately track human movements indoors.

III. Sensor and Data Recording

A. Data Collection Scenario

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

TABLE I
Multi-Modal Indoor MOT Datasets

Dataset	# Mic	# Cam	Radar	Annot.	# Speakers
AV 16.3 [7]	16	3	no	3D	3
AVDIAR [9]	6	2	no	3D	4
AVTRACK [13]	4	1	no	2D	2
SPEVI [8]	2	1	no	2D	2
CAV3D [10]	8	1	no	3D	3
RAV4D	6	2 (stereo)	yes	3D	3

For the purpose of creating a dynamic and challenging scenario suitable for MOT tasks, we had one to three individuals moving around the desk, often crossing paths. This scenario was essential for examining the identity switch issue prevalent in tracking applications. To enhance sound localization, participants were instructed to speak loudly, aiding in the capture of clear audio data.

B. Sensor Modalities

Our dataset features 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 at 960 x 540 pixels resolution, operating at a frame rate of 30 fps. Alongside visual data, it generates a synchronized depth map that covers up to 20 meters, maintaining an accuracy range of 0.5% to 2%. The camera's field of view extends 110 degrees horizontally and 70 degrees vertically.

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

3) Circular Microphone Array: The microphone array is a 6-element circular design with a diameter of 7 cm. It features an embedded audio-enhanced front end to improve the SNR. This array is capable of detecting sound events up to 10 meters away with an angular resolution of approximately 5 degrees. It records audio in a 6-channel format at a 16 kHz sampling rate

IV. Multi-sensor Calibration

Calibration was conducted using a hand-crafted calibration target, consisting of a corner reflector, a checkerboard pasted on a clapperboard, as shown in fig. 2. The corner reflector is detectable by radar as a strong point target. The clapperboard, orifinally used in filmmaking for synchronizing audio and video in post-production, is repurposed in our study. We attached a checkerboard pattern to it, facilitating calibration for both the camera and microphone array. The checkerboard's easily identifiable corner features make it ideal for camera

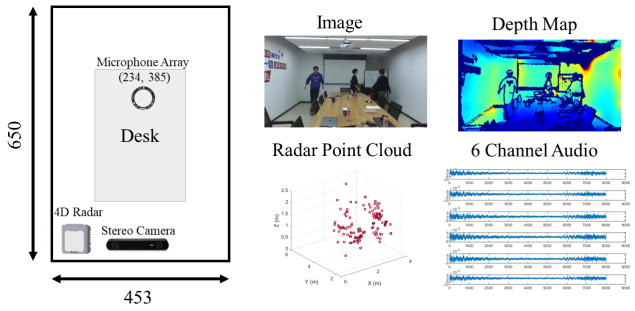

Fig. 1. Room layout and sensor outputs

calibration. Its attachment to the clapperboard's flat surface ensures visibility to the camera. Moreover, 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 manually measured.

The calibration target was positioned in seven different locations across the meeting room, varying in height and uniformly distributed. The world coordinate system's origin was set at the room's left corner. To determine the corner reflector's position, we used the ceramic tiles as grid units, 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 directly measured by the ruler.

Following the collection of 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. Additionally, 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 microphone array's fixed location.

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

V. Pre-Processing

A. Radar Data Filtering

In this study, we utilize a commercial high-resolution 4D radar to capture the human motion. This radar can be configured to output the raw radar point cloud produced by the Constant False Alarm Rate (CFAR) detector. Each point is a 5-dimensional vector, comprising range, azimuth angel, elevation angle, Doppler velocity and RCS. In indoor environments, multi-path propagation often results in significant 'ghost' objects. However, with the room's layout available, we can eliminate these artifacts by defining a 3D Region of Interest (ROI) corresponding to the room's dimensions.

Furthermore, with the radar stationed in a fixed position, we construct an occupancy grid map (OGM) to model the static environment. This OGM, which offers a more robust approach compared to direct point representations, is better suited to tolerate spatial uncertainties in measurements. We discretize the room ROI into a grid with a cell resolution of 0.1 meters. The static OGM is then built through the temporal accumulation of static detections, using a fixed count threshold to identify occupied cells. Subsequent to these steps, we refine the radar point cloud by removing points that coincide with the static clutter maps. After this filtering process, we employ 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

For audio DoA estimation, we employ 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 adopts the PHAT weighting function. The PHAT function normalizes magnitude and leverages phase information for correlation calculation. DoA candidates are then identified from the peaks in the output power spectrum. Due to computational demands, we implement this process using a grid cell size of one degree. Though originally designed for scenarios involving a single sound source, SRP-PHAT is capable of handling multiple sources, provided the number of sources is known [14]. Indoor reverberation often results in noisy outcomes, and silence periods during movement add 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, aligning them with the measured location of the microphone array, as shown in fig. 3.

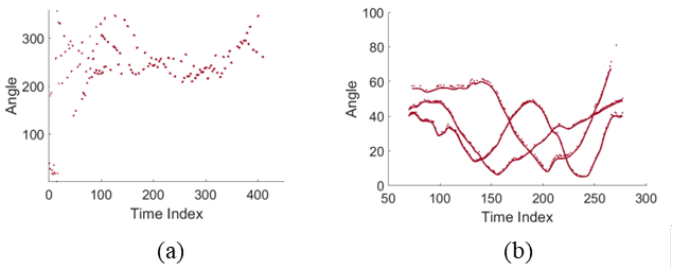

Fig. 3. DoA estimation: (a) raw DoA estimated by SRP-PHAT (b) smoothed DoA trajectories of three speakers in the world coordinate

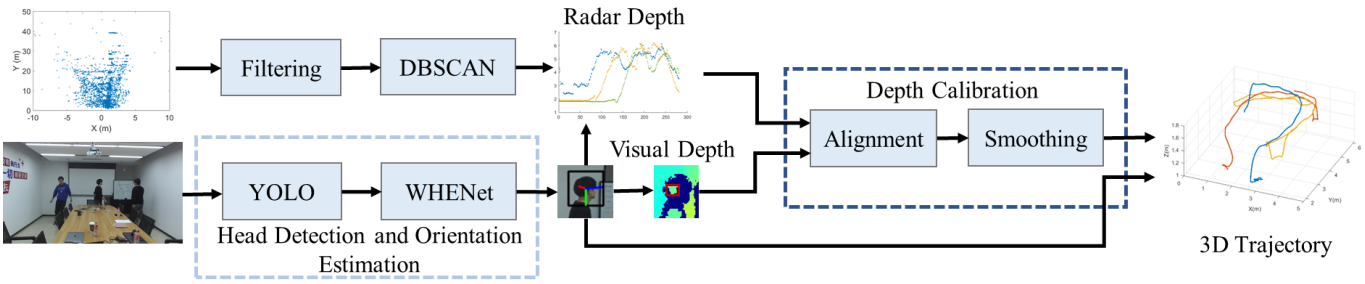

Fig. 4. Visual guided annotation pipeline

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 suited for tracking tasks. This point target model is particularly well-suited for tracking tasks. We define the point target as a six-dimensional vector, which includes 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 [15] to identify the bounding box and calculate the center of 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 a lack of distinctive features or occlusions. Fortunately, given the high frame rate of our system, depth measurements are continuous over time. This continuity allows us to refine the depth measurements smoothly, thus enhancing the accuracy of the head center's depth trajectory. In addition, we predict head orientation using the WHENet [16], where the Euler angles are regressed by an additional MLP layer with the head feature map as input.

VI. Visual Guided Trajectory Annotation

As illustrated in fig. 4, we propose a visual guided annotation framework for trajectories.

Regarding trajectory annotation, the sparse nature of the radar point cloud presents a challenge in determining a reliable point representation of a human. The center of a radar cluster often results in a zig-zag trajectory. To overcome this, we utilize the heads detected by the stereo camera data as the tracked objects. The 3D location of the head can be determined by the pixel coordinate in the image plane, the depth, and the intrinsic parameters of the camera. However, two main challenges arise: the lack of strict synchronization between the stereo camera and the radar sensor, and the inaccuracies in depth estimation by the stereo camera. Factors like 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 necessary to correct the 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 person and interpolate it to match the camera's timestamps. Next, we apply an iterative optimization module to align the visual depth trajectory with the radar depth trajectory, treating the time offset and depth scaling parameter as optimizable 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 fused depth trajectory. Finally, we compute the 3D trajectory based on the 2D locations and the depth.

To ensure the precision of our dataset, we develop a GUI interface to visualize annotations on a frame-by-frame basis, allowing us to review and correct any missed annotations.

The ground truth in our study is generated according to the camera timestamps. For radar-centered applications, for example to study the effect of Doppler information, it becomes necessary to interpolate the ground truth trajectory to align with the radar's slower detection frame rate. Since our analysis also includes the estimation of head orientation, this interpolation process should be conducted in the SE(3) space, which accounts for both translation and rotation. To facilitate this, we first convert Euler angles into quaternions and then apply Spherical Linear Interpolation (Slerp) [17] to achieve smooth and continuous trajectories for both the head's position and its heading angles.

VII. Dataset Analysis

A. Calibration Results

Accurate spatial calibration between radar and camera systems is fundamental to our visual-guided trajectory annotation pipeline. In this section, we present the results of this calibration. Firstly, 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. Since 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 accomplish this, we apply a transformation matrix to convert radar detections from radar-centered coordinates into world coordinates. Similarly, by utilizing the camera's intrinsic matrix and depth information, we can project image points

into world coordinates. The calibration results are quantified by the L2 reprojection error when comparing the radar and image data against the ground truth. The reprojection errors are 0.05 meters for the radar data and 0.01 meters for the camera data, respectively. These figures highlight the precision of our calibration process.

B. Dataset Contents

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

- **Simple Case** (fig. 5 (a)): This case involves individuals starting from their seats, moving to the front of the room, and then returning to their seats, all without any intersections in their paths.
- **Normal Case** (fig. 5 (b)): In this case, three individuals cross paths at the front of the room. The individual following the yellow trajectory intersects with the other two individuals twice, resulting in a total of four crossings.
- **Hard Case** (fig. 5 (c)): This case presents the most challenging setup, with each individual crossing paths with the other two, leading to a total of six crossings.

In addition, our dataset specifically annotates the periods of crossing to provide a more rigorous test for tracking algorithms, especially in scenarios involving occlusions. This comprehensive structure of the dataset is designed to offer a robust testing ground for evaluating the performance and accuracy of various tracking algorithms under different conditions and complexities.

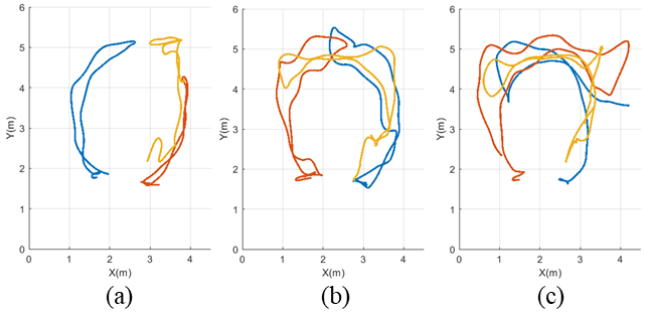

Fig. 5. Three levels of difficluties: (a) easy (b) normal (c) hard

C. Comparing Visual and Radar Tracking in 3D Space

To 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-neighbor (GNN) algorithm for assignment. The evaluation metrics selected are MOTA and MOTP, as defined by the CLEAR MOT metrics [18]. We chose the 3-person, bright, hard case as our test sequence.

As detailed in Table table II, radar tracking demonstrates 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 in correctly identifying and following objects, and encounters fewer errors such as missing targets or incorrectly tracking irrelevant objects. Additionally, radar tracking outperforms visual tracking in terms of MOTP. This indicates that radar tracking not only more reliably detects and tracks objects but also does so with greater spatial accuracy, leading to more precise localization 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: One key aspect we investigate in our dataset is the influence of illumination changes. Figure 6 presents a scenario where the lights in the meeting room are turned off, creating a low illumination 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 arises from extreme lighting contrasts. For example, bright light from a projector in a dark room can significantly obstruct the visibility of a person walking in front of it. Our visual detector performs robustly in strong lighting perturbations (as demonstrated in fig. 6 (a)) but struggles when a person's face is occluded by the texture of slides from a projector (as depicted in fig. 6 (b)), risking loss of head detections. These conditions underscore the complexities of tracking in varied lighting environments and emphasize the necessity for multi-sensory fusion to achieve reliable tracking.

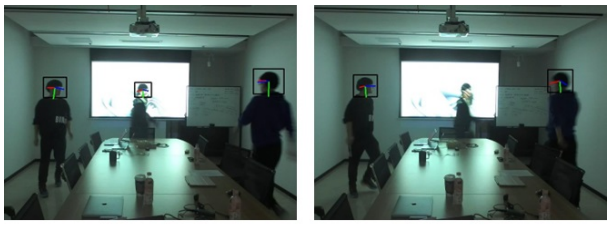

Fig. 6. Illumination chanllenge in dark scenarios

2) Occlusion Scenario: Occlusion presents a significant challenge in our dataset, manifesting in two primary forms. The first type of occlusion is environmental, caused by the layout and objects within the room. For instance, with the

radar positioned in the left corner, the individual on the left side of the room is more clearly detected, returning a denser point cloud. In contrast, the right two individuals, partially obscured by the table, yield a sparser point cloud. The second type of occlusion is due to trajectory crossing. As the camera and radar are located at different positions, occlusions occur at various 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 are occluded in the radar detections (c). The last row demonstrates a case where occlusions occur simultaneously in both the image and radar data. In such instances, the continuous audio DoA becomes instrumental, offering an alternative method to confirm the presence of individuals who are occluded in both visual and radar sensors.

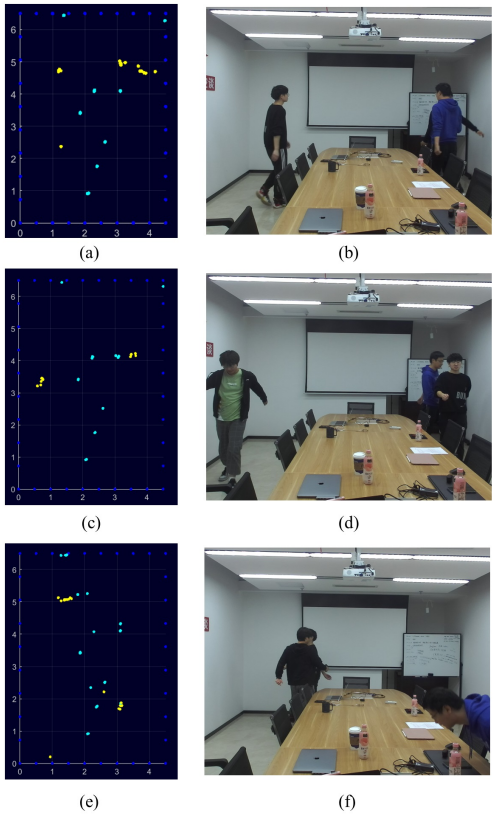

Fig. 7. Occlusion cases: (a) visual occlusion (b) radar occlusion (c) both radar and visual occlusion

VIII. Conclusions

This paper introduces the RAV4D dataset, a novel multi-modal dataset that integrates 4D radar, audio, and visual data to enhance multiple people 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 creating a novel calibration target and devising a visual-guided annotation framework. The dataset encompasses complex trajectories with numerous crossings and a variety of challenging scenarios, such as low illumination conditions and occlusions. These features establish RAV4D as an invaluable resource for researchers and practitioners aiming to explore and advance the capabilities of multi-modal sensing in complex indoor settings.

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] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] O. Schumann, M. Hahn, N. Scheiner, F. Weishaupt, J. F. Tilly, J. Dickmann, and C. Wöhler, "Radarscenes: A real-world radar point cloud data set for automotive applications," in *2021 IEEE 24th International Conference on Information Fusion (FUSION)*. IEEE, 2021, pp. 1–8.

[12] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.