IEEE Journal of Microwaves

Received 1 June 2025; revised 9 October 2025; accepted 6 December 2025.

Digital Object Identifier 10.1109/JMW.2025.3642570

Fully Automated, AI-Driven Multimodal Annotation Framework for Human-Centric Radar Applied to Deep Learning-Based People Localization

LUKAS ENGEL [1] (Graduate Student Member, IEEE), MARKUS BERGMANN [1], CHRISTOPH KAMMEL [1], INGRID ULLMANN [1] (Member, IEEE), BJOERN M. ESKOFIER [2,3] (Senior Member, IEEE), AND MARTIN VOSSIEK [1] (Fellow, IEEE)

(Regular Paper)

[1]Institute of Microwaves and Photonics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany
[2]Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Germany
[3]Translational Digital Health Group, Institute of AI for Health, Helmholtz Zentrum München—German Research Center for Environmental Health, 85764 Neuherberg, Germany

CORRESPONDING AUTHOR: Lukas Engel (e-mail: lukas.le.engel@fau.de).

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1483 – under Project 442419336, EmpkinS; and in part by the Bavarian Ministry of Economic Affairs, Regional Development and Energy, as a part of the BayVFP Funding Program – funding line digitalization – funding section information and communication technology under Grant DIK-0310/02.

ABSTRACT Radar-based machine learning pipelines require extensive annotated datasets. However, producing large volumes of precise labels remains prohibitively laborious and prone to inconsistency, as radar signals lack a direct visual correspondence. To address this limitation, we introduce a fully automated, multi-modal annotation pipeline built around our custom *RadarBox* that co-registers a FMCW MIMO radar with an Azure Kinect RGB-D camera. Precise spatial calibration and hardware-level synchronization yield exact pixel-to-radar alignment. RGB images undergo panoptic segmentation to generate per-pixel human masks, which are fused with depth measurements to reconstruct a voxelized surface mesh. We extract 3D joint positions from the Kinect Body Tracking SDK and apply a bidirectional Kalman filter to derive precise per-joint positions and velocity vectors free from sudden, non-physiological fluctuations. These enhanced labels are projected into 5D radar cube slices and target lists through robust spatio-temporal association. As a demonstration, we train a deep neural network on annotated radar target lists for indoor people localization, achieving a mean positional error of 0.31 m and 91.8% occupancy accuracy, even under occlusion. Unlike prior semi-automatic or heuristic-based methods, our approach delivers consistent 5D labels at scale, bridging spatial, temporal, and Doppler dimensions, and thus paves the way for large-scale, learning-based radar sensing in human-centered applications.

INDEX TERMS Automatic labeling, AI-driven, deep learning, human-centric, people localization, radar.

I. INTRODUCTION

Radar sensors are recently gaining increased attention in human-centered applications, such as activity recognition [1], [2], [3], [4], [5], [6], fall detection [7], [8], and pose estimation [9], [10], [11], [12]. These diverse use cases are enabled by radar's inherent advantages: privacy-preserving operation, robustness to adverse lighting [9], [10], [13] and weather conditions [14], the ability to penetrate certain materials [15], [16], [17], [18], and its capacity to deliver dynamic and spatial information such as range, velocity, and angle.

These sensing capabilities make radar an attractive modality for intelligent systems. However, fully exploiting radar data requires models capable of interpreting its complex and multi-dimensional structure. This is where machine learning

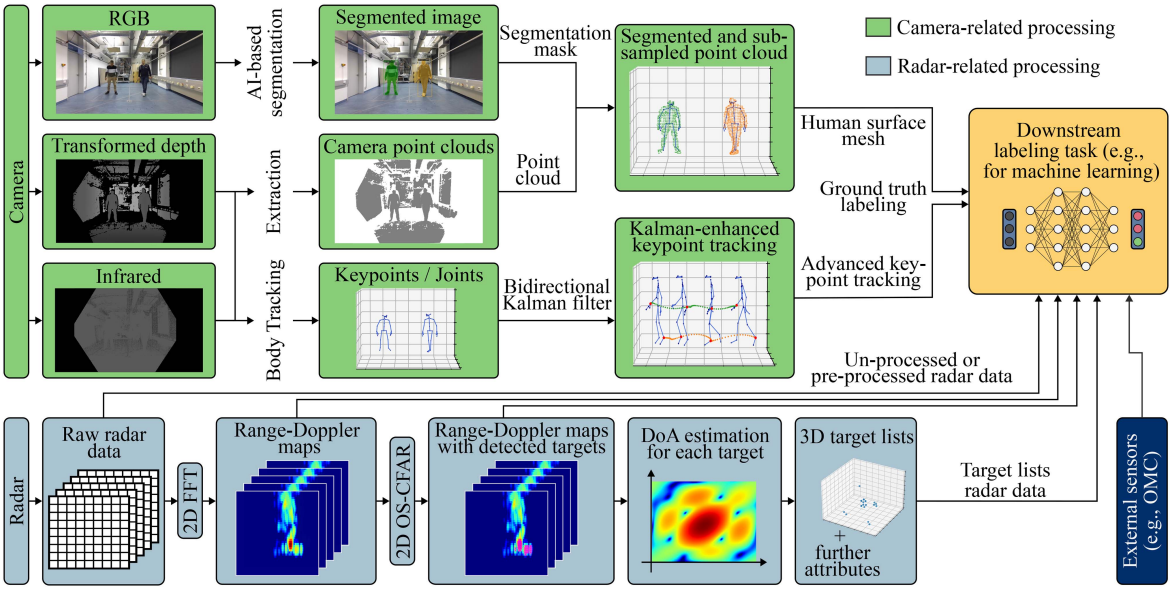


FIGURE 1. Paper content overview: The block diagram illustrates the collected data and developed processing chains for the various sensor modalities utilized in our *RadarBox*. The processing chain delivers calibrated and synchronized, high-precision human-centric labeling information for downstream labeling tasks, e.g., for machine learning. OS-CFAR: ordered-statistic constant false-alarm rate, DoA: direction of arrival, OMC: optical motion capture.

plays a critical role, as artificial intelligence continues to progress in diverse fields [6], [19], [20], [21]. In contrast to the human perceptual system, which is primarily trained on two-dimensional visual stimuli, machine learning models offer the ability to process and learn from the high-dimensional, non-intuitive representations characteristic of radar sensors. The effectiveness of supervised learning critically depends on the availability of large volumes of high-quality labeled data [22], [23], [24]. Obtaining such labels is particularly challenging: manual annotation is labor-intensive, time-consuming, costly, and requires domain-specific expertise. Moreover, the abstract nature of radar data makes labeling ambiguous and prone to inconsistency. Even for experienced annotators, distinguishing targets from clutter or interpreting overlapping reflections often leads to subjective decisions.

Various approaches have been proposed in the literature to automate the annotation of sensor data. In recent years, many solutions for automated radar data annotation have originated from automotive applications. Early methods often relied on manual or semi-automatic procedures, such as drawing bounding boxes in 3D point clouds [25]. Subsequent work used ground-truth information from complementary sensors to enable automatic radar labeling; for example, [26] employed vision-based ground truth to annotate radar cubes and [27] proposed a semi-automatic approach for labeling radar point clouds. However, such vision-based methods are inherently limited by their lack of precise depth information. In [28], computer vision was used to annotate lidar data, while [29] applied lidar-based object detectors for radar data annotation. Multi-modal data fusion has proven particularly effective in this context. [30] combined lidar and camera data for automated radar annotation, and [31] integrated lidar, camera, and odometry for semi-automatic labeling of radar point clouds. Yet, these often omit motion dynamics such as tracking or velocity estimation. In contrast, [32] demonstrated the utility of multi-sensor fusion to generate labels for tracking vulnerable road users, incorporating both spatial and temporal aspects.

Automotive-focused methods typically prioritize object-level detection, limiting their suitability for indoor applications requiring finer motion detail and articulated pose estimation. Several studies have targeted indoor human-centric monitoring: For instance, [33] utilized camera-based computer vision for person detection, and [34] transferred camera-derived annotations onto human-oriented radar point clouds. In [35], Kinect-derived skeletal data was used to label motion over time, such as spectrograms, highlighting the potential of cross-modal skeleton-based supervision. Additionally, some approaches in human pose estimation [9], [10], [11], activity recognition [36], or gait analysis [37] extract skeletons from camera data to train radar-based pose estimators. These pipelines, however, often require manual post-labeling, lack depth cues, or depend on multi-camera setups.

To address the shortcomings of prior work, we introduce a powerful automatic radar data labeling framework driven by multi-modal sensor fusion, which enables the capture of precise depth information as well as detailed motion data, as illustrated in Fig. 1. In contrast to previous works relying on heuristic rules or manual annotation, our method enables consistent and reproducible label generation across varying scenes and subject motions. At its core, our system fuses a radar sensor embedded in a custom-engineered *RadarBox* with an Azure Kinect camera system. The Kinect delivers precisely aligned RGB, depth, and skeletal streams, serving as high-fidelity ground truth for the radar domain. Through spatial calibration and precise temporal synchronization, we enable fully automated label generation via robust association

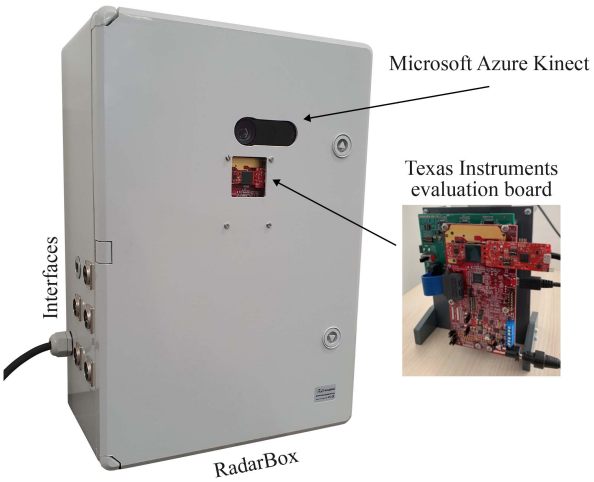


FIGURE 2. Developed measurement setup: Our *RadarBox* with an integrated multiple-input multiple-output frequency-modulated continuous-wave radar (Texas Instruments evaluation board) and RGB-D camera (Microsoft Azure Kinect). The box provides interfaces to synchronize with other sensor systems, such as optical motion capture. The figure is reproduced from [12].

— entirely removing the need for manual annotation. In addition to skeletal tracking, the Kinect's depth modality yields dense spatial point clouds that capture the surface geometry of objects and human bodies. These point clouds offer an additional source of spatial structure, which can be leveraged for enriching radar annotation with surface-level context or non-skeletal reference targets. Crucially, by employing a bidirectional Kalman filter, we further refine skeletal tracking accuracy, yielding high-precision joint positions. Most notably, this allows for the reliable estimation of joint-wise velocity vectors — a key advancement that is otherwise nontrivial to derive from position data alone. This enriched motion representation substantially broadens the labeling scope, enabling targeted annotation within the Doppler domain.

Our framework supports the creation of accurately labeled radar datasets at scale, enabling downstream tasks such as object detection, activity recognition, and pose estimation. The modular design also allows adaptation to diverse radar configurations and environments.

The remainder of this paper is structured as follows: Section II describes the *RadarBox* hardware and sensor setup. Section III details the camera-based data processing pipeline used to extract keypoints and surface meshes. Section IV outlines the radar signal processing chain. Section V summarizes hardware requirements and processing time. Section VI presents the automatic labeling methods for radar data. Section VII demonstrates the application of the proposed labeling pipeline to a deep learning-based indoor people localization task. Section VIII concludes the paper.

II. *RADARBOX* – MEASUREMENT AND SENSOR SETUP

In this section, we describe the measurement and sensor setup. The developed sensor system, *RadarBox*, is illustrated in Fig. 2. *RadarBox* is a compact and robust sensing platform that integrates several hardware components to enable synchronized multi-modal data acquisition. At its core, the system includes two primary sensing devices: a Texas Instruments multiple-input multiple-output (MIMO) radar evaluation board for capturing radar raw data and a Microsoft Azure Kinect depth camera for ground truth labeling. These sensors are enclosed in a durable plastic casing (35 cm × 50 cm × 19 cm) to ensure mechanical stability and fixed relative positions of the sensors. Additional supporting components include a mini PC, which manages sensor operation, data storage, and software-based synchronization, and an Arduino board that provides precise hardware-level time synchronization. Additionally, several external interfaces are provided for system control and external triggering, such as optical motion capturing as demonstrated in [12], making *RadarBox* a highly versatile and extensible platform for advanced experimental setups.

A. AZURE KINECT CAMERA SYSTEM

We employed the Microsoft Azure Kinect camera[1] to capture the visual scene for downstream ground-truth labeling. The Azure Kinect delivers a comprehensive set of synchronized data streams tailored for multimodal perception tasks. It provides high-resolution RGB imagery, precise depth sensing via a time-of-flight sensor, and infrared imaging for robust performance under low-light conditions. Additionally, the device includes an integrated microphone array and an inertial measurement unit, enabling the acquisition of spatial audio and motion data. When combined with the Azure Kinect Body Tracking SDK, the system supports real-time skeletal tracking using advanced algorithms to estimate 3D joint positions and orientations across a detailed 32-joint model. This model encompasses key anatomical landmarks — such as the head, neck, spine, shoulders, elbows, wrists, hips, knees, and ankles — facilitating high-fidelity skeletal motion analysis. Furthermore, the Azure Kinect provides external synchronization interfaces intended for multi-device configurations.

B. RADAR SYSTEM SETUP

The radar subsystem was centered on the Texas Instruments IWR6843AOPEVM[2] evaluation board, which integrates on-package antennas. The board includes four receiving and three transmitting antennas, configured in an L-shaped virtual array to enable two-dimensional angle estimation. This antenna configuration offers a wide field of view in both azimuth and elevation, supporting comprehensive spatial coverage. To facilitate raw radar data acquisition, the MMWAVEICBOOST[3] and DCA1000EVM[4] extension boards were employed, forming a flexible and high-performance radar signal capture setup. The system operates in frequency-modulated continuous-wave (FMCW) mode and employs time-division multiplexing

[1] https://azure.microsoft.com/products/kinect-dk
[2] https://ti.com/tool/IWR6843AOPEVM
[3] https://ti.com/tool/MMWAVEICBOOST
[4] https://ti.com/tool/DCA1000EVM

TABLE 1. Radar System Parameter Settings

Parameter	Value
Frequency	60 GHz
Radio-frequency bandwidth	≈ 1.02 GHz
Update rate	15 Hz
Chirp duration	≈ 17 µs
Samples per chirp	64
ADC sampling frequency	3.8 MHz
Chirps per frame per transmitter antenna	128
Measurement duration per frame	≈ 32 ms
Number of transmitter antennas	3
Number of receiver antennas	4
TDM-MIMO virtual data channels	12
Azimuth/Elevation resolution	≈ 30 °
Field of view (azimuth and elevation)	±60 °
Range resolution	≈ 14.8 cm
Unambiguous range	≈ 9.49 m
Doppler resolution	≈ 0.078 m s^{-1}
Unambiguous Doppler velocity	≈ ±5.02 m s^{-1}

TDM: time-division multiplexing, MIMO: multiple-input multiple-output.

(TDM) MIMO techniques to achieve effective channel separation. Table 1 summarizes the key radar configuration parameters. An update rate of 15 Hz was selected to reliably capture human motion dynamics, while an increased number of chirps per frame was configured to enhance Doppler resolution. This setup enables fine-grained discrimination of velocity components associated with individual body parts in the range-Doppler domain, thereby supporting detailed motion analysis in dynamic human-centered scenarios.

C. CALIBRATION AND SYNCHRONIZATION OF SENSOR SYSTEMS

To enable synchronized data fusion between the two sensing modalities, which operate in distinct coordinate systems, a precise spatial registration within a common Cartesian reference frame and accurate time synchronization are required. As a preliminary step, the radar and Azure Kinect systems were individually calibrated within their respective coordinate frames. Subsequently, a joint calibration procedure was conducted using a shared reference target that is simultaneously perceivable by both sensors. The measurement setup employed for this purpose is depicted in Fig. 3. To ensure high measurement fidelity and suppress multi-path effects, all calibration experiments were performed in an anechoic chamber at the Institute of Microwaves and Photonics (FAU). A metallic trihedral corner reflector was utilized to generate a strong radar target, while an infrared-reflective marker placed precisely at the reflector's apex facilitated accurate localization in the Kinect's depth image. We conducted radar measurements at varying distances and viewing angles, obtaining a diverse set of observations. The reflector's phase center was meticulously identified in both sensor frames across all poses.

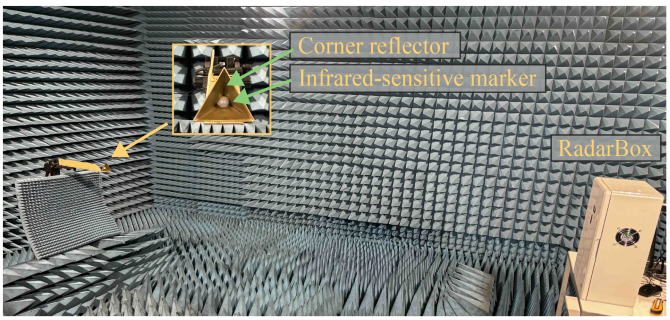


FIGURE 3. Calibration measurement setup in an anechoic chamber with a shared reference target. The image shows a metallic corner reflector with an infrared-sensitive marker attached at its apex. This configuration enables precise co-localization in both the radar and Azure Kinect frames, enabling accurate spatial registration between the two coordinate systems.

Based on these correspondences, a spatial transformation was estimated by solving a constrained optimization problem, minimizing the registration error in a least-squares sense [38]:

$$\boldsymbol{R}_{\text{opt}}, \boldsymbol{T}_{\text{opt}} = \arg\min_{\boldsymbol{R},\boldsymbol{T}} \|\boldsymbol{X}_{\text{radar}} - (\boldsymbol{R}\boldsymbol{X}_{\text{azure}} + \boldsymbol{T})\|_F^2, \quad (1)$$

where $\boldsymbol{R}$ is an orthogonal rotation matrix ($\boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}$), $\boldsymbol{T}$ is a translation matrix, $\boldsymbol{X}$ represents the coordinates in their respective systems, and $\|\cdot\|_F^2$ denotes the Frobenius norm. We applied Procrustes analysis [38], [39], [40] to estimate the optimal rotation $\boldsymbol{R}_{\text{opt}}$ and translation $\boldsymbol{T}_{\text{opt}}$ parameters.

In our setup, accurate time synchronization was achieved by repurposing the external output signal of the Azure Kinect as a hardware trigger for the radar system, ensuring precise alignment with the radar data acquisition pipeline and maintaining consistent, low-latency synchronization across all sensing modalities.

III. CAMERA DATA PROCESSING FOR LABEL EXTRACTION

This section details the processing pipeline applied to camera data to generate groundtruth labels for downstream tasks. The overall pipeline, shown in the upper part of Fig. 1, comprises two stages. First, we perform panoptic segmentation on the RGB images to obtain perpixel scene labels (Section III-A). These labels serve as masks for the point cloud reconstructed from transformed depth images (i.e. depth images that have been resampled and spatially aligned to match the RGB images in resolution and dimensions), yielding a segmented point-cloud mesh (Section III-B). Second, we employ the Azure Kinect SDK Body Tracking module to extract human keypoints and refine their positions and velocities using a bidirectional Kalman filter (Section III-C). Each of these stages is described in detail below.

A. OBJECT DETECTION FROM RGB CAMERA IMAGES USING DEEP LEARNING

For the purpose of generating ground-truth labels, relevant objects are automatically detected and segmented using the camera data. Owing to the rich visual information in RGB images, modern computer vision techniques — particularly

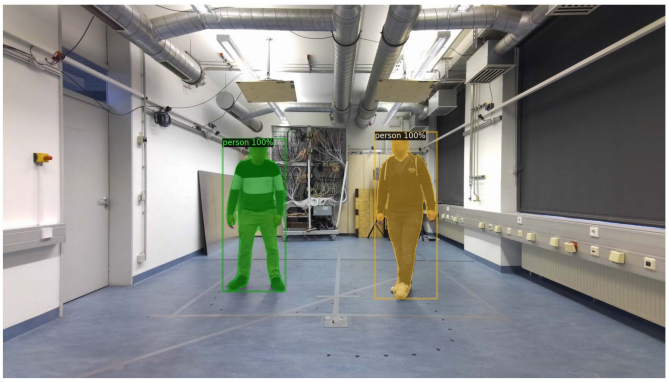


FIGURE 4. Exemplary panoptic segmentation results on an RGB image, with all non-human classes filtered out to highlight detected person instances.

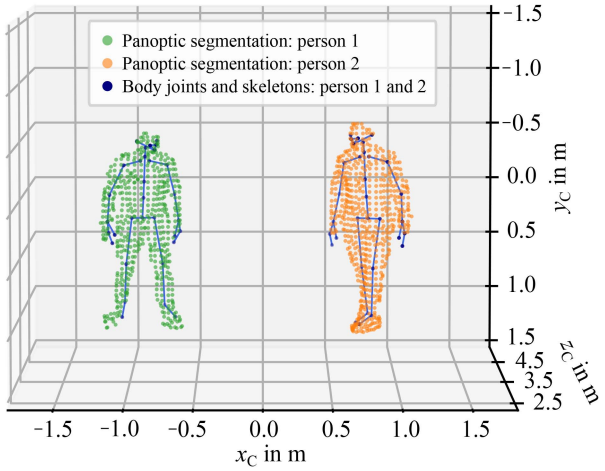


FIGURE 5. 3D point cloud with overlaid skeletal keypoints for two individuals. The figure shows the voxelized surface point clouds derived from panoptic segmentation, with body joint positions extracted using the Azure Kinect Body Tracking SDK.

deep neural networks — have shown strong performance in object recognition and scene understanding tasks [41], [42], [43]. In this work, we leverage panoptic segmentation, a technique that integrates semantic and instance segmentation into a unified framework. This approach yields pixel-wise semantic labels along with instance-specific object masks, enabling both class-level understanding and precise object delineation. The resulting segmentation masks are directly aligned with the image's pixel grid, allowing accurate spatial correspondence. We adopt the panoptic feature pyramid network (FPN) model from [44], implemented within the open-source Detectron2 framework developed by Meta Research [45]. This model provides strong out-of-the-box performance on diverse datasets without requiring modification for our camera system. It supports segmentation across a broad range of object categories, such as person, bicycle, car, dog, cat, and umbrella, making it adaptable to various use cases. In our application, only selected classes relevant to human-centered and indoor scenarios are retained for downstream processing. An example output of the segmentation process is illustrated in Fig. 4, where all non-human objects have been filtered out.

B. FUSION OF OBJECT DETECTION AND DEPTH SENSING FOR SURFACE MESH EXTRACTION

To extract dense 3D representations of human surfaces, we combined RGB-based object detection with depth-based point cloud generation from the Azure Kinect camera system. Depth images were converted into 3D point clouds using the intrinsic parameters of the camera, resulting in a dense set of spatial points projected into the camera coordinate frame. These point clouds were then associated with the corresponding semantic information obtained from panoptic segmentation of the RGB images, allowing for per-pixel classification within the 3D space. Given the high resolution of RGB-D imagery, each object yields a large number of 3D points, resulting in high memory and computational demands. To mitigate this, a voxel grid filter was applied for spatial downsampling. The 3D space was discretized into voxels of edge length 5 cm, and only a single representative point — the centroid of the voxel — was retained if multiple points occupied the same voxel. This voxelization significantly reduces the point cloud density while preserving geometric structure. In the example shown in Fig. 5, the point cloud contains only one-seventieth of the points of the denser representation.

Segmentation masks often contain mislabeled pixels at object boundaries, leading to depth values inconsistent with the actual object geometry — typically corresponding to background surfaces. These artifacts project into 3D space as outliers with large deviations in the z-direction. To remove these, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [46] was applied to each segmented object, using a minimum cluster size of 10 points and a maximum neighbor distance of 30 cm. Only the largest identified cluster was retained, as it is assumed to correspond to the correctly segmented object region, which holds true in the majority of near-camera applications. This strategy effectively filters out spurious outliers caused by misclassified edge pixels without discarding the primary structure of the object. Since DBSCAN may separate parts of the same object (e.g., upper and lower body occluded by a table), clustering was constrained to the depth dimension only. This ensures that spatially disjoint but depth-consistent body segments are preserved, while isolated outliers with divergent depth values are removed.

To enable consistent tracking of segmented individuals across frames, an inter-frame association algorithm was implemented. Each segmented object was assigned an index number based on the minimum distance between centroid positions of segments across consecutive frames, using a threshold of 1 m to avoid incorrect associations due to subject occlusion or scene transitions. Unmatched segments received new identifiers. The result of the body surface extraction is illustraded in Fig. 5. For enhanced interpretation, skeletal keypoints were also overlaid on the point clouds.

C. BIDIRECTIONAL KALMAN FILTER-ENHANCED KEYPOINT-WISE TRACKING

The Azure Kinect Body Tracking SDK is employed to extract 3D human joint positions from the captured depth data. Using a deep learning-based approach, the system estimates the spatial locations and orientations of 32 anatomical joints. The SDK is capable of tracking multiple individuals simultaneously, making it suitable for multi-person interaction scenarios. The resulting skeletal representations exhibit temporal continuity and spatial accuracy, making them suitable as structured ground-truth annotations for motion-related downstream tasks. An example of the two extracted skeletons from the scene shown in Fig. 4 is visualized in blue in Fig. 5.

While the SDK provides discrete joint positions x_c, y_c, z_c per frame, applications such as labeling radar-based human motion scenarios can benefit from the additional estimation of joint velocities $\dot{x}_c$, $\dot{y}_c$, $\dot{z}_c$ to enrich the temporal dynamics of the data. To this end, we applied a Kalman filter to not only smooth the position trajectories but also infer the velocity vector components. As real-time operation is not necessary for the offline labeling process, we adopt a bidirectional Kalman filter (BKF) to further optimize the estimates by incorporating both past and future measurements [47], [48], [49]. In particular, we employ the Rauch–Tung–Striebel smoother [47], a classical fixed-interval implementation of the BKF. This is particularly advantageous in the presence of occlusions, missing detections, or temporally inconsistent keypoints, which are common in depth-based body tracking. The BKF can retrospectively correct such errors and interpolate through short-term gaps, thereby providing more accurate and stable keypoint trajectories, including both position and velocity components, for use in downstream radar labeling.

Throughout the manuscript, the index k denotes the current time step, with $k-1$ and $k+1$ referring to the previous and next time steps, respectively. We define the Gaussian distributed system state vector at time step k as:

$$\underline{\boldsymbol{x}}_k = \left[\underline{x}_c, \underline{y}_c, \underline{z}_c, \underline{\dot{x}}_c, \underline{\dot{y}}_c, \underline{\dot{z}}_c\right]^\top \text{ with } \underline{\boldsymbol{x}}_k \sim \mathcal{N}(\boldsymbol{x}_k, \boldsymbol{P}_k). \quad (2)$$

where x_c, y_c, z_c denote the keypoint position vector elements and $\dot{x}_c, \dot{y}_c, \dot{z}_c$ the associated velocity components. $\boldsymbol{x}$ denotes the mean, and $\boldsymbol{P}$ denotes the covariance of that state's probability distribution. In the forward pass, the standard Kalman filter [50] is applied. The prediction step is formulated as:

$$\boldsymbol{x}_{k|k-1} = \boldsymbol{F}\boldsymbol{x}_{k-1|k-1}, \quad (3)$$

$$\boldsymbol{P}_{k|k-1} = \boldsymbol{F}\boldsymbol{P}_{k-1|k-1}\boldsymbol{F}^\top + \boldsymbol{Q}, \quad (4)$$

where $\boldsymbol{F}$ is the state transition matrix (constant velocity model) and $\boldsymbol{Q}$ is the process noise covariance matrix, which is initialized with values described in [51]. The measurement vector $\boldsymbol{y}_k$ at each time step contains the observed position:

$$\boldsymbol{y}_k = \left[x_c, y_c, z_c\right]^\top. \quad (5)$$

The update step of the standard Kalman filter is then computed as [50]:

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{P}_{k|k-1}\boldsymbol{H}^\top + \boldsymbol{R}_k)^{-1}, \quad (6)$$

$$\boldsymbol{x}_{k|k} = \boldsymbol{x}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{H}\boldsymbol{x}_{k|k-1}), \quad (7)$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{H}\boldsymbol{P}_{k|k-1}, \quad (8)$$

where $\boldsymbol{R}$ is the empirically determined measurement noise covariance matrix, $\boldsymbol{K}$ is the Kalman gain, and $\boldsymbol{H}$ is the observation matrix. Since only the position is observed, $\boldsymbol{H}$ is a 3×6 matrix that extracts the position components from the state vector. As the measurement space is a subset of the state space, the transformation from state to measurement domain can be performed via a simple matrix multiplication.

To further improve the estimates, we applied a backward pass using the BKF equations. This refinement step relies on future measurements to enhance the state estimation at each time step. As a prerequisite, we (again) first computed the predicted state and covariance for the $(k+1)$-th time step based on the information up to the k-th time step using the standard Kalman filter prediction formulas [47], [50]:

$$\boldsymbol{x}_{k+1|k} = \boldsymbol{F}\boldsymbol{x}_{k|k}. \quad (9)$$

$$\boldsymbol{P}_{k+1|k} = \boldsymbol{F}\boldsymbol{P}_{k|k}\boldsymbol{F}^\top + \boldsymbol{Q}. \quad (10)$$

The predicted values were subsequently employed in the backward equations to compute the backward-corrected state and covariance estimates [47]:

$$\tilde{\boldsymbol{K}}_k = \boldsymbol{P}_{k|k}\boldsymbol{F}^\top\boldsymbol{P}_{k+1|k}^{-1}, \quad (11)$$

$$\tilde{\boldsymbol{x}}_k = \boldsymbol{x}_{k|k} + \tilde{\boldsymbol{K}}_k(\tilde{\boldsymbol{x}}_{k+1} - \boldsymbol{x}_{k+1|k}), \quad (12)$$

$$\tilde{\boldsymbol{P}}_k = \boldsymbol{P}_{k|k} + \tilde{\boldsymbol{K}}_k(\tilde{\boldsymbol{P}}_{k+1} - \boldsymbol{P}_{k+1|k})\tilde{\boldsymbol{K}}_k^\top, \quad (13)$$

where $\tilde{\boldsymbol{x}}$ represents the backward-corrected state estimate and $\tilde{\boldsymbol{P}}$ the corresponding covariance matrix. The matrix $\tilde{\boldsymbol{K}}$ functions similarly to the Kalman gain in the standard Kalman filter with the difference that the predicted system state vector is no longer compared with a measured vector but rather with the corrected system state vector of the subsequent time step. Thus, the BKF propagates the trajectory from the final to the initial time step, ensuring that the state estimate at each time instant incorporates all past and future measurements, thereby yielding an optimal estimate.

To demonstrate the performance of the BKF, we provide an exemplary measurement of a person walking frontally toward the camera. Fig. 6 shows the estimated position and its corresponding velocity. For clarity, we focus solely on the z-component, which represents the depth dimension in the Azure Kinect system, as it exhibits the most significant variation in this scenario. The top plot illustrates the z-position over time for selected joints that exhibit the most pronounced motion, including a comparison with measurement data from an Azure Kinect sensor. The bottom plot compares the velocity estimated by the BKF with a reference velocity obtained via numerical differentiation using a central difference approximation. To asess the BKF's performance,

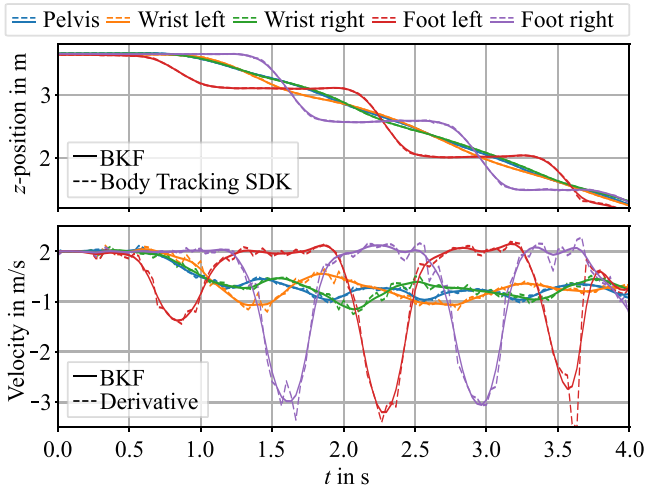


FIGURE 6. An exemplary measurement of a person walking frontally toward the camera. The plots compare the bidirectional Kalman filter (BKF) results with the Body Tracking SDK z-position outputs (top) and the corresponding velocity obtained by differentiating the position (bottom).

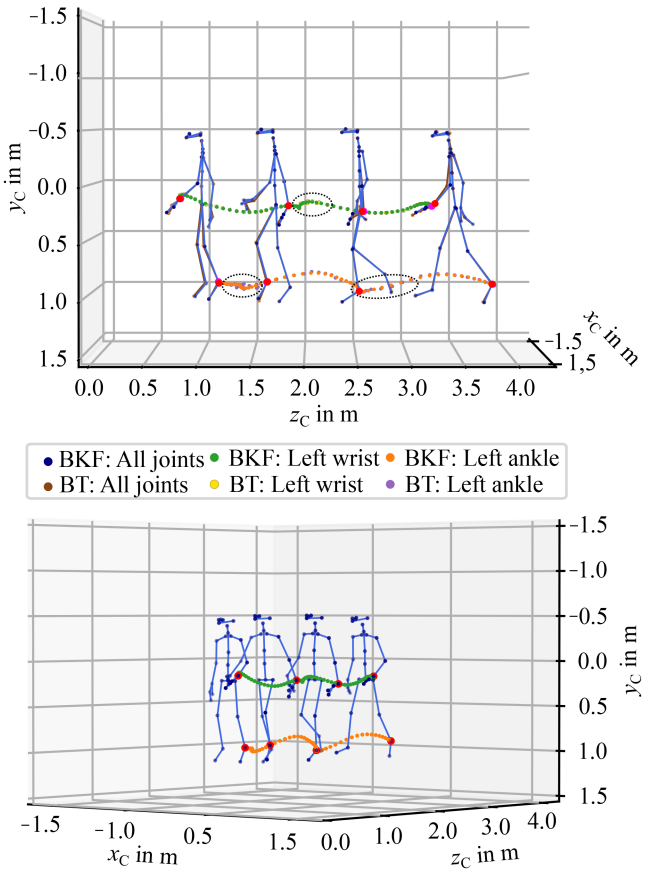


FIGURE 7. Visualization of selected joints (wrist and ankle): The top panel compares outputs from the bidirectional Kalman filter (BKF) and Body Tracking SDK (BT), while the bottom panel shows the BKF-smoothed trajectory from an alternative perspective.

we compare its velocity estimates with those generated by the differential method. Notably, the plot highlights opposing velocity components for corresponding left and right body parts, such as the wrists and feet, which align with the natural gait pattern. Moreover, the BKF effectively smooths sudden, non-physiological fluctuations likely caused by measurement noise, resulting in more stable and consistent velocity estimates. Fig. 7 illustrates a comparison of tracking performance between the BKF and the Body Tracking SDK using selected joints, such as the wrist and ankle. In the upper plot, position estimates from both methods are presented in a single frame to facilitate a direct comparison of their tracking performance, with key regions exhibiting decisive differences clearly highlighted. The lower plot offers an alternative perspective on the BKF-smoothed trajectory, underscoring its enhanced consistency and robustness in accurately capturing movement.

This BKF process yields temporally consistent keypoint trajectories along with their velocity vectors, and improves robustness to short-term occlusions, misdetections, and noise. The advantages of this enhanced occlusion handling will be furhter demonstrated in the application example in Section VII below.

IV. RADAR DATA PROCESSING

The lower panel of Fig. 1 illustrates the radar signalprocessing chain; a more detailed view is presented in Fig. 8. The radar system operates based on a time-division multiplexed MIMO FMCW waveform, processed using a chirp-sequence approach tailored to indoor motion capture scenarios [52], [53]. Raw I/Q data was collected from each virtual channel across all frames and organized for every frame into a 4D complex-valued data structure, $\boldsymbol{X}[i, j, k, l]$. The indices i and j represent fast-time and slow-time samples, respectively, while k and l correspond to the spatial dimensions of the virtual antenna array. Due to the L-shaped array configuration, some entries in the spatial grid were unassigned and thus filled with zeros. To suppress static background reflections and isolate dynamic scene elements, a frame-wise mean subtraction was performed along the slow-time dimension, resulting in a radar cube free from statical reflections (see Fig. 8(a)):

$$\boldsymbol{X}_{\mathrm{SCR}}[i, j, k, l] = \boldsymbol{X}[i, j, k, l] - \bar{\boldsymbol{X}}[i, j, k, l], \qquad (14)$$

where $\bar{\boldsymbol{X}}[i, j, k, l]$ denotes the mean of the slow-time samples and is broadcast across the full cube. This step enhances sensitivity to fine movements and reduces sensitivity to environment-specific static reflections (Fig. 8(b)). Following this, the Hann window [54] $\boldsymbol{W}[i, j, k, l]$ was applied to the fast and slowtime dimensions to suppress spectral leakage before transforming the data into the range–Doppler domain via a 2D FFT:

$$\boldsymbol{Y}[m, n, k, l] = \mathrm{FFT}\{\boldsymbol{W}[i, j, k, l] \cdot \boldsymbol{X}_{\mathrm{SCR}}[i, j, k, l]\}. \qquad (15)$$

Here, m and n denote discrete range and Doppler bins, respectively. A 2D ordered-statistics CFAR detector [55], [56], [57] was applied to $\boldsymbol{Y}$ to identify prominent scatterers in the range-Doppler plane. An example result is shown in Fig. 8(c), which highlights dynamic reflections suitable for further processing.

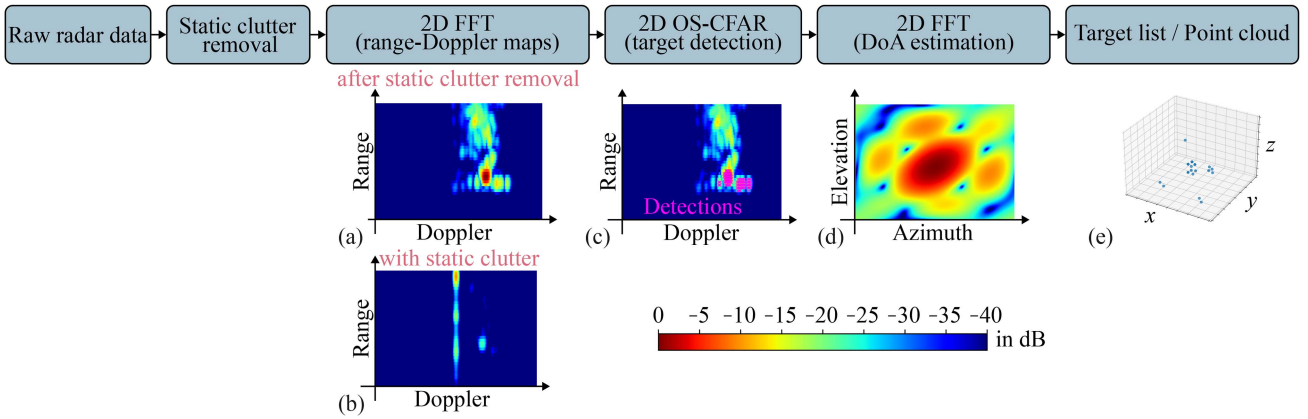


FIGURE 8. Applied radar signal processing: The raw radar data was processed by a 2D FFT to convert it to more descriptive range-Doppler information. Subsequently, a 2D ordered-statistic constant false-alarm rate (OS-CFAR) was applied to extract the relevant scatterer in the range-Doppler domain and apply a 2D FFT beamforming to infer the angular information of the scatterer. The figure is similar to that in [12].

To estimate the direction-of-arrival (DoA), a 2D spatial FFT was then performed over the virtual antenna dimensions:

$$\boldsymbol{Z}[m, n, o, p] = \mathrm{FFT}\{\boldsymbol{Y}[m, n, k, l]\}, \tag{16}$$

where o and p represent the elevation and azimuth angle bins, respectively. To reduce computational load, DoA estimation was limited to the range-Doppler bins where CFAR had detected a target. In high-density scenes, multiple reflections may appear within the same bin. To resolve such overlaps, we applied the CLEAN algorithm [58], enabling separation of multiple sources from within a single angular spectrum. A representative elevation-azimuth slice is visualized in Fig. 8(d). The resulting detections were then mapped into Cartesian coordinates using the estimated range, radial velocity, and angular information. Each detection was further enriched with radar-specific metrics such as signal-to-noise ratio (SNR), noise floor, and intensity. These processed detections were compiled into structured target lists, each representing a set of radar-detected objects in a given frame, as illustrated in Fig. 8(e). To standardize input dimensionality for the learning model, the number of targets per frame was capped at 64. If fewer detections were present, zero-padding was applied to maintain fixed-size input tensors. These radar target lists formed the final processing step for potential subsequent downstream processing.

V. HARDWARE REQUIREMENTS AND PROCESSING TIME

For the operation of the *RadarBox*, a Dell OptiPlex Micro 7010 equipped with an Intel Core™ i5-13500 T processor and 16 GB of RAM was employed. The solid-state drive was expanded to 1 TB to accommodate the storage of extensive raw datasets. To preserve the compact design and minimize thermal stress during continuous operation, the system was configured without a dedicated GPU. Raw data from the camera and radar sensors were stored locally as RGB, infrared, and depth images, along with corresponding radar raw data files. These data were subsequently transferred to an external workstation for post-processing.

For each acquisition sequence, the recording duration was fixed at 15 min to ensure coherent motion segments while maintaining manageable dataset block sizes. The radar operated at 15 Hz while the camera system recorded at 30 Hz, yielding 27,000 frames per recording. Although the higher frame rate increases data volume and computational demands, it provides a finer temporal resolution, which improves the accuracy and continuity of motion representation. After processing, the camera frames were downsampled to match the radar frame rate. The resulting data volumes amounted to 11 GB for RGB, 7.7 GB for infrared, and 5.2 GB for depth images, complemented by 5.3 GB of radar raw data, corresponding to a total of 29.2 GB per recording.

We parallelized the raw data processing across multiple workstations to reduce overall computation time. For benchmarking purposes, however, the complete labeling pipeline was evaluated on a single workstation equipped with an Intel Xeon W-1350 processor, 128 GB RAM, and an Nvidia GeForce RTX3080 GPU with 10 GB VRAM. The current implementation is not yet fully optimized. All reported timings refer to a 15-min recording involving two persons. The body joint extraction using the Body Tracking SDK required 40.6 min when processing four recordings in parallel, corresponding to an effective time of 10.2 min per recording. The panoptic image segmentation, which is computationally intensive, required 180.1 min for three recordings processed in parallel, resulting in an effective time of 60.0 min per recording to extract all scene classes, including people, furniture, and small objects. The subsequent Kalman-based temporal processing, comprising both the standard forward Kalman filter and the BKF, required 100.5 min for six recordings processed in parallel, resulting in an effective time of 16.7 min per recording. The radar signal processing, from raw data to target list generation, took approximately 6.6 min per recording.

Although the overall data volume and processing duration are considerable, the proposed processing framework exhibits a high degree of scalability. Each module of the pipeline operates largely independently and can thus be parallelized

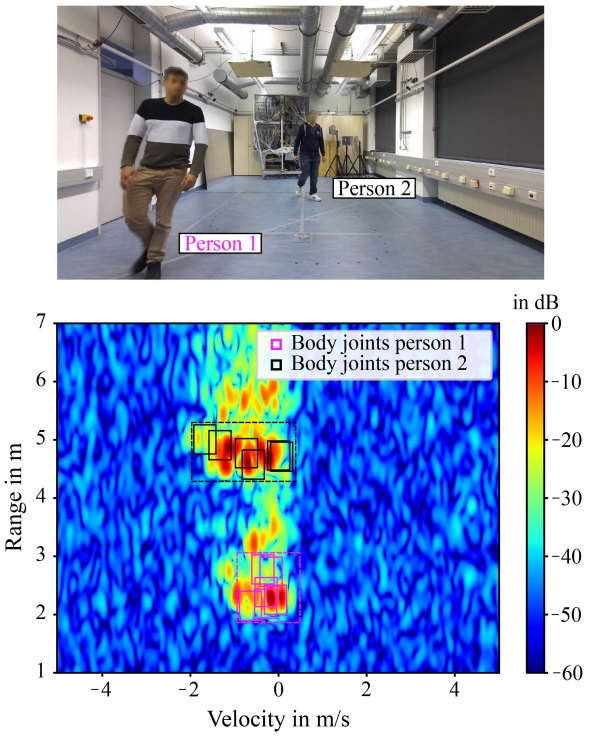


FIGURE 9. Labeling of a range–Doppler slice using the bidirectional Kalman filter (BKF) approach. Strong reflections at ≈ 2 m and ≈ 5 m are automatically annotated with a subset of keypoints (pelvis, hands, feet, and nose shown for clarity) and enclosed by dashed contours indicating each person.

across multiple workstations or computing nodes with minimal interdependence. Furthermore, future improvements in algorithmic efficiency and hardware performance, such as advances in GPU architectures and optimized deep learning models, are expected to further reduce computational demands. Accordingly, the present configuration establishes a robust foundation for future extensions, particularly since the labeling process operates offline and does not require real-time execution.

VI. LABELING RESULTS

In this section, we demonstrate two exemplary labeling possibilities made feasible by the automatic labeling process described above.

A. LABELING OF 5D RADAR CUBE

Here, we demonstrate the feasibility of labeling the 5D radar cube (time, range, Doppler, elevation, and azimuth) with an example in range-Doppler domain. In Fig. 9, the results of labeling a range-Doppler slice are shown alongside the corresponding RGB image. By employing our bidirectional Kalman filter labeling approach, each human keypoint is assigned a velocity, which can be readily converted into the radial velocity relative to the radar, thereby enabling labeling in the Doppler domain. As an example, two individuals move freely in a room, producing a range-Doppler plot with multiple strong targets at approximately 2 m and 5 m. These can be associated with the individuals based on the RGB image. Additional, weaker reflections around 3 m and 6 m do not originate directly or via line-of-sight from the persons. As shown, both individuals' body joints have been automatically annotated. For clarity, only the pelvis, left and right hands, both feet, and the nose (head) are displayed. Windows sizes were set with tolerances of ± 0.25 m in the range dimension and ± 0.25 m in the velocity dimension. Dashed lines illustrate the outlines of each person. It is immediately evident that the marked body joints reliably cover the strong reflections of both individuals, while echoes caused by multi-path propagation and other scatterers are ignored. On this basis, machine learning can be applied to the labeled radar data to derive further insights. Since the x-, y-, and z-coordinates of body joints can also be used to compute azimuth and elevation angles, labeling in these dimensions is feasible and especially advantageous when using radar systems with high angular resolution.

B. POINT CLOUD-BASED LABELING

The signal processing chain described in Section III-B provides the foundation for labeling radar point clouds. Each radar target extracted through the processing chain in Fig. 8 is checked for proximity (≤ 30 cm) to a camera-based reference point in Fig. 5 and, if within range, inherits that reference point's label and instance number. If multiple reference points fall within the threshold, the closest one is chosen to avoid misassignments, especially when two individuals move in close proximity. An important advantage of this approach is the utilization of human surface data derived from camera and depth sensors, as these surfaces directly correspond to the regions responsible for radar reflections, thus significantly enhancing labeling accuracy. Since the panoptically segmented point cloud may contain non-human objects, corresponding object categories can also be assigned. In Fig. 10, however, only the "person" category was considered. We present skeletons derived from the Azure Kinect SDK and refined by Kalman filter–based position estimation, shown alongside the radar point clouds. As illustrated, point clouds assigned to humans appear in color green and orange, while unassigned points are rendered in blue.

VII. APPLICATION EXAMPLE: DEEP LEARNING-ENABLED PEOPLE INDOOR LOCALIZATION USING RADAR TARGET LISTS

The direct labeling of radar measurements in the 5D datacube domain and of further processed target lists has proven effective for many applications. However, for machinelearning tasks that rely exclusively on radar inputs, applying labels derived from camera data enables a model to reproduce the same outcome using radar data alone. In the example below, we demonstrate how camerabased labels are used to train a machinelearning model on radar measurements for room-scale people localization.

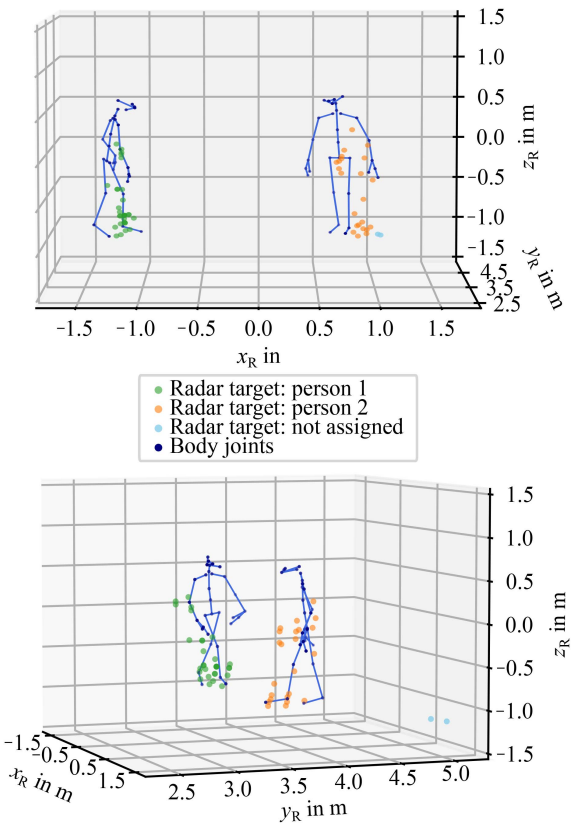


FIGURE 10. Labeling of radar point clouds through proximity matching to camera-based reference points. Radar targets extracted by the signal processing chain (Fig. 8) are assigned the label and instance ID of the nearest panoptically segmented reference point within 30 cm.

Accurately localizing and counting individuals indoors, which is essential for tasks such as home-habit monitoring, shopper-behavior analysis, navigation in museums, or human–machine coordination in industrial halls, remains difficult. This is because multipath propagation can mask true reflections and produce spurious ghost targets, and because occlusions occur when one person blocks another's radar return [59], [60]. Machine learning models may overcome these issues by learning to distinguish genuine targets and by using temporal context to track occluded individuals.

To address this, we present a deep learning model that detects and localizes multiple people within a room. By operating on radar target lists rather than the full 5D data cube, we reduce computational cost while preserving the positional information needed. We conducted a study in an office-like environment, recording people moving in front of our *RadarBox*. Using the outputs of our labeling process, the model was trained using only radar target lists to predict both the number of people and their spatial coordinates.

A. BIDIRECTIONAL-KALMAN-FILTER-ENHANCED DATASET

We deployed the system in three distinct indoor environments with varying layouts and furniture: (a) an office, (b) a kitchen, and (c) a laboratory temporarily configured as an office (see Fig. 11). In each room, we conducted recordings under controlled occupancy levels of 0, 1, and 2 individuals to evaluate the system's performance across different usage scenarios. Our *RadarBox* was placed on a table facing toward the center of the room. The study was approved by the Ethics Committee of Friedrich-Alexander-Universität Erlangen-Nürnberg (Protocol #22-437-B), and all participants provided written informed consent. For each room and occupancy level, we recorded several continuous 15-minute sessions, yielding a total of 405,000 radar samples. Participants performed unscripted office-style activities (walking, standing, sitting, eating, telephoning, or any combination thereof) to emulate natural behavior. The dataset contains equal representation of each occupancy level. Pelvis-root joint coordinates extracted via our presented processing method described in Section II-I-C served as ground truth. We encoded occupancy as a count vector and each individual's pelvis position as their location.

To demonstrate the application-specific benefits of our BKF labeling approach, Fig. 12 compares pelvis trajectories from Azure Kinect Body Tracking, a standard Kalman filter, and the BKF. The standard filter exhibits pronounced deviations during measurement gaps (e.g., of person 1 in region III due to occlusion on the blue trajectory) because it simply forward-propagates the last known velocity without correction. In contrast, the BKF — by processing the entire sequence bidirectionally in time — maintains a smooth, physically plausible trajectory even in the absence of observations. As we now possess accurate ground-truth trajectories throughout these gaps, our radar-based deep learning model can be trained to infer both occupancy and individual positions even in the presence of occlusions.

B. DEEP LEARNING ARCHITECTURE

To accurately infer room occupancy and individual positions from radar-derived target lists, we developed a deep neural network, illustrated in Fig. 13. The architecture employs a multi-stage pipeline to exploit spatial and temporal correlations inherent in the data. As an input to the deep learning model, we used a tensor featuring seven parameters including position (x-, y-, z-components from point clouds), corresponding velocity, SNR, noise, and intensity $\in \mathbb{R}^{7\times 1}$, processed by the radar processing presented in Section IV. The prediction of our network was a hard-coded vector for person or not, and a regression for the the Cartesian positions of the persons.

First, a PointNet [61] backbone extracts spatial features from each radar frame, mapping unordered point sets into fixed-dimensional global features. By cascading successive multi-layer perceptron (MLP) layers and applying learned geometric transformations, the network captures the intrinsic spatial structure of the input. These per-frame embeddings are then sequenced and processed by a Transformer encoder, whose multi-head self-attention mechanism [62] discerns temporal dependencies critical for modeling human motion dynamics. The encoder's output is subsequently fed into two distinct fully connected heads: one implements a classification task to predict occupancy states as discrete labels, and the

FIGURE 11. Photographs of the rooms utilized in this study: (a) Office, (b) kitchen, (c) laboratory configured as an office.

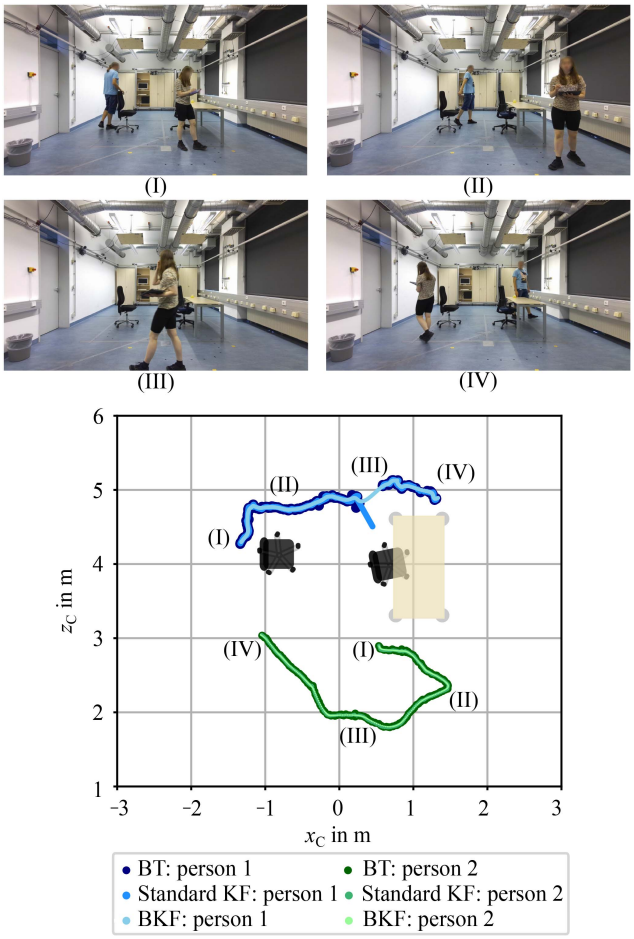


FIGURE 12. Illustration of tracking performance in challenging scenarios: (Top) Photographs (I)–(IV) show selected time frames from the recorded scenarios, illustrating the motion paths of two individuals and highlighting challenging situations involving visual occlusion. (Bottom) Comparison of tracking results: the proposed bidirectional Kalman filter (BKF) improves trajectory continuity and accuracy compared to the standard Body Tracking SDK (BT) output, as demonstrated in Region III.

other performs regression to estimate continuous positional coordinates.

C. TRAINING, OPTIMIZATION, AND EVALUATION

For training our deep learning models, we partitioned the dataset room-wise into three subsets: training, validation, and test in an 80 : 10 : 10 ratio. For assessing cross-room generalization, we trained on data from the environments shown in Fig. 11(a) and (b) and held out the environment in (c) exclusively for testing, ensuring the test set was never seen during training. We trained the model using the Adam optimizer, with a batch size of 32, a weight decay of 0.0001, and an initial learning rate of 0.0001, which was decayed exponentially by a factor of 0.98 per epoch. To mitigate overfitting, we applied early stopping based on validation performance, capping training at a maximum of 120 epochs. All methods were implemented in PyTorch. The network is trained end-to-end using a composite loss — cross-entropy for occupancy classification and mean-squared error for position regression, so both objectives are optimized simultaneously, as we assume that the tasks are interdependent and that learning one supports the other. The optimized parameters for the Transformer encoder are listed in Table 2.

TABLE 2. Transformer Encoder Parameter Overview of the Optimized Model

Parameter	Value
Sequence length	25
Positional encoding	sinusoidal
Embedding dimension	128
Number of heads	16
Number of layers	2
Dropout	0.1

D. EXPERIMENTAL RESULTS AND DISCUSSION

Qualitative top-down localization results are shown in Fig. 14. In panel (a), which features a single occupant, the model's predicted trajectory (red) aligns almost perfectly with the ground truth (blue), and the system correctly identifies exactly one person with no misclassification of count. In panel (b), two people appear and the network correctly identifies both and reconstructs their paths. Panel (c) illustrates the occlusion scenario from Fig. 12: despite a temporary loss of the Body Tracking data from Azure Kinect, the network benefits from the BKF labeling and leverages its learned spatiotemporal context to infer and maintain each person's position, seamlessly bridging the gap with radar-based predictions.

Quantitative results are listed in Table 3. Our model demonstrates strong overall performance: On the regression side, the mean positional error is 0.313 m and the median positional error is 0.270 m; their closeness indicates that the

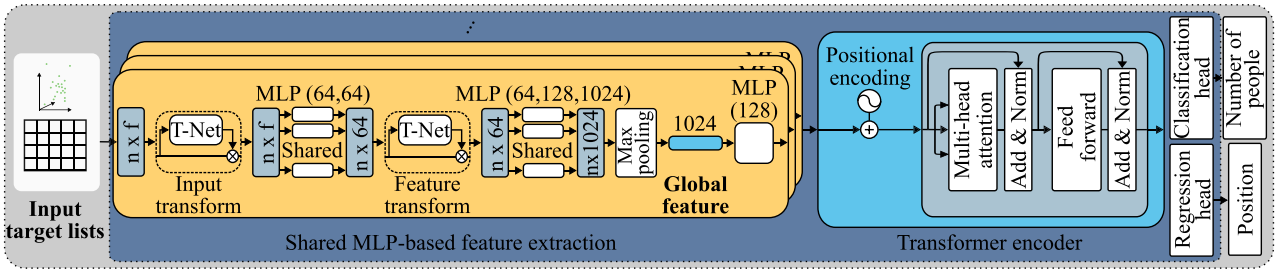


FIGURE 13. Proposed localization network architecture: A shared multi-layer perceptron (MLP)-based feature extractor first encodes spatial correlations from the input sensor data, and a Transformer encoder then models temporal dependencies for the classification and regression heads to produce robust trajectory estimates.

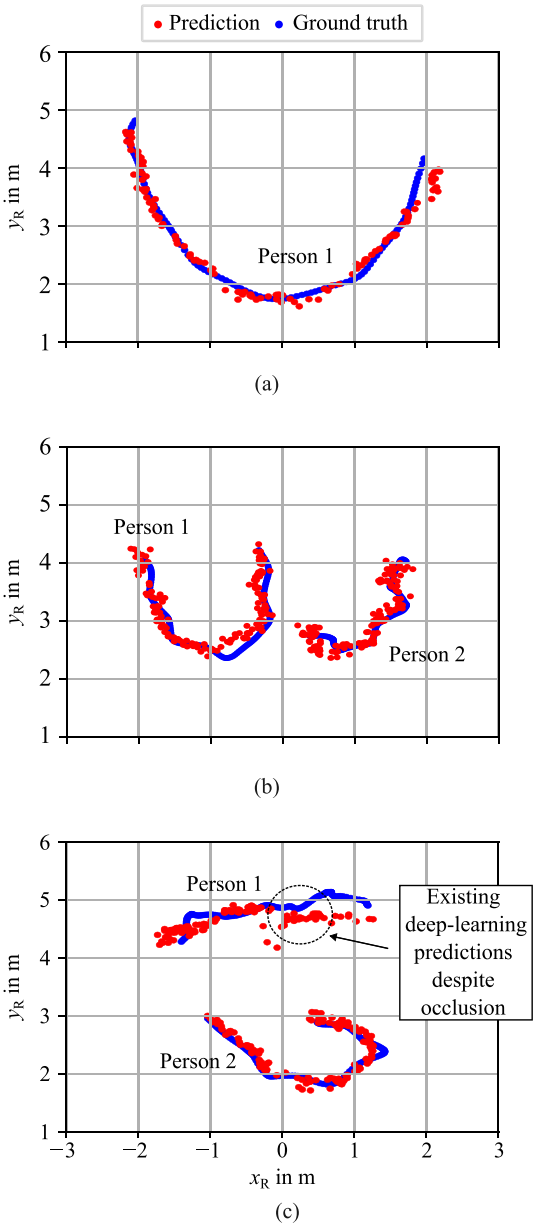


FIGURE 14. Top–down localization results from our proposed deep learning localization network: (a) Single-person scenario, (b) two-person scenario, (c) Occlusion scenario (see Fig. 12) with existing predictions despite occlusion in the original Body Tracking data.

TABLE 3. Quantitative Results for Deep Learning-Based Localization Task

Metric	Value
Mean positional error	0.313 m
Median positional error	0.270 m
Accuracy of occupancy classification	91.78%
Precision of occupancy classification	89.98%
Recall of occupancy classification	93.33%
F1-score of occupancy classification	91.63%

positional predictions exhibit low variance. Achieving positional accuracy below 32 cm demonstrates the approach's strong potential for indoor positioning tasks. Moreover, it correctly classifies about 92% of all persons. With a recall of 93.3%, it successfully identifies nearly all true positives, missing fewer than 7 out of every 100. Its precision of 90.0% indicates that most of the instances it flags as positive are indeed correct, though roughly 10% are false alarms, meaning the model predicted a person where none was actually present. The F1-score of 91.6% reflects a well-balanced trade-off between precision and recall, suggesting that the chosen decision threshold is effective for the objectives. These results demonstrate the correct functionality and high precision of the developed approach. They confirm that deep learning–enabled, radar-based indoor people localization can achieve high performance in future applications.

Benchmarking the localization performance against existing approaches remains challenging, as most state-of-the-art methods do not provide publicly available datasets with precise ground-truth annotations [63], [64]. While a few studies on indoor localization include ground-truth references [65], [66], a direct comparison is further complicated by substantial differences in radar hardware and experimental setups. For instance, some systems employ highly directional antennas that concentrate transmitted power within a narrow sector, thereby reducing multipath effects but restricting spatial coverage. In contrast, our *RadarBox* utilizes a wide field of view that enables full-room monitoring and simultaneous multi-person observation, yet inherently increases the influence of multipath propagation, making the localization task more complex. Moreover, our framework not only estimates

individual positions but also infers the number of persons present, whereas related approaches such as [65], [66] assume this information to be known a priori. Reported localization accuracies in these studies typically range between 20 cm and 30 cm, which is comparable to our achieved results, considering the broader coverage and more challenging conditions of our setup. Nevertheless, we expect that future work employing more sophisticated network architectures and larger training datasets could further enhance localization accuracy and robustness, particularly in multi-person and high-clutter indoor environments.

To extend this work, we will integrate a Kalman filter to fuse sequential model outputs into smooth, robust trajectory estimates and enable continuous target tracking. At the same time, we will expand our dataset to span a broader range of room geometries and layouts, thereby improving the model's generalizability across diverse indoor environments. Finally, by deploying multiple radar–camera station pairs, we intend to increase spatial coverage, reduce occlusions, and achieve higher multi-view localization precision in complex settings.

VIII. CONCLUSION

In this work, we present a fully automated, multi-modal pipeline for high-precision ground-truth labeling of human-centric millimeter-wave radar data. Our *RadarBox* precisely aligns FMCW MIMO radar with Azure Kinect streams and can integrate further modalities, such as optical motion capture. AI-driven panoptic segmentation, voxelised surface reconstruction, and a bidirectional Kalman filter together yield temporally smooth labels for the 5D radar cube. We also transfer those labels to radar point clouds. Because labeling is performed offline, we employ a BKF to refine keypoint trajectories: it smooths and interpolates across frames, bridges occlusion or missed-detection gaps, and outputs per-frame velocity estimates that serve directly as ground-truth Doppler labels.

Using these extracted labels, we trained a deep neural network on BKF-derived labels projected into radar target lists for room-scale people localization. Operating solely on radar inputs, the model achieves an occupancy-estimation accuracy of 91.8% and precise spatial positioning, with a mean positional error of 0.31 m, even under occlusion. These results underscore the high potential of our automatic labeling process to accelerate the creation of large, high-quality radar datasets for downstream learning-based tasks, thereby enabling rigorous development and benchmarking of learning-based radar sensing methods.

ACKNOWLEDGMENT

We gratefully acknowledge Victoria Koch and everyone who helped collect the dataset for their time and enthusiasm. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

REFERENCES

[1] M. M. Rahman, D. Martelli, and S. Z. Gurbuz, "Gait variability analysis with multi-channel FMCW radar for fall risk assessment," in *Proc. IEEE 12th Sensor Array Multichannel Signal Process. Workshop.*, Trondheim, Norway: IEEE, Jun. 2022, pp. 345–349.

[2] E. Kurtoğlu, S. Salehin, M. G. Amin, I. P. Kan, M. A. McKay, and S. Z. Gurbuz, "Ethogram-based personalization of human activity and agility from radar μD signatures," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat.*, Nov. 2024, pp. 1–8.

[3] D. Mejdani et al., "Radar-based tremor quantification using deep learning for improved Parkinson's and palliative care assessment," *IEEE Trans. Radar Syst.*, vol. 2, pp. 1174–1185, 2024.

[4] A.-C. Froehlich et al., "A millimeter-wave MIMO radar network for human activity recognition and fall detection," in *Proc. IEEE Radar Conf.* Denver, CO, USA: IEEE, May 2024, pp. 1–5.

[5] I. Ullmann, R. G. Guendel, N. C. Kruse, F. Fioranelli, and A. Yarovoy, "A survey on radar-based continuous human activity recognition," *IEEE J. Microwaves*, vol. 3, no. 3, pp. 938–950, Jul. 2023.

[6] D. Krauss et al., "A review and tutorial on machine learning-enabled radar-based biomedical monitoring," *IEEE Open J. Eng. Med. Biol.*, vol. 5, pp. 680–699, 2024.

[7] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. C. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 71–80, Mar. 2016.

[8] B. Jokanović and M. Amin, "Fall detection using deep learning in range-Doppler radars," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 1, pp. 180–189, Feb. 2018.

[9] Y.-H. Ho et al., "RT-Pose: A 4D radar tensor-based 3D human pose estimation and localization benchmark," in *Proc. Eur. Conf. Comput. Vis.*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham, Switzerland: Springer Nature, 2025, vol. 15121, pp. 107–125.

[10] S.-P. Lee, N. P. Kini, W.-H. Peng, C.-W. Ma, and J.-N. Hwang, "HuPR: A benchmark for human pose estimation using millimeter wave radar," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 5704–5713.

[11] M. Zhao et al., "RF-based 3D skeletons," in *Proc. Conf. ACM Special Int. Group Data Commun.* Budapest, Hungary: ACM, Aug. 2018, pp. 267–281.

[12] L. Engel et al., "Advanced millimeter wave radar-based human pose estimation enabled by a deep learning neural network trained with optical motion capture ground truth data," *IEEE J. Microwaves*, vol. 5, no. 2, pp. 373–387, Mar. 2025.

[13] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the human figure through a wall," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–13, Nov. 2015.

[14] C. Waldschmidt, J. Hasch, and W. Menzel, "Automotive radar—From first efforts to future systems," *IEEE J. Microwaves*, vol. 1, no. 1, pp. 135–148, Jan. 2021.

[15] M. Zhao et al., "Through-wall human pose estimation using radio signals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7356–7365.

[16] S. Ahmed, A. Schiessl, F. Gumbmann, M. Tiebout, S. Methfessel, and L.-P. Schmidt, "Advanced microwave imaging," *IEEE Microw. Mag.*, vol. 13, no. 6, pp. 26–43, Sep./Oct. 2012.

[17] N. C. Albrecht, J. P. Weiland, D. Langer, M. Wenzel, and A. Koelpin, "Characterization of the influence of clothing and other materials on human vital sign sensing using mmWave radar," in *Proc. 53rd Eur. Microw. Conf.*, Berlin, Germany: IEEE, Sep. 2023, pp. 428–431.

[18] D. M. Sheen, D. L. McMakin, and T. E. Hall, "Three-dimensional millimeter-wave imaging for concealed weapon detection," *IEEE Trans. Microw. Theory Techn.*, vol. 49, no. 9, pp. 1581–1592, Sep. 2001.

[19] Z. Geng, H. Yan, J. Zhang, and D. Zhu, "Deep-learning for radar: A survey," *IEEE Access*, vol. 9, pp. 141800–141818, 2021.

[20] A. S. Martey, A. Ali, and E. Ebenezer, "AI-based palm print recognition system for high-security applications," in *Proc. IEEE AFRICON.*, Nairobi, Kenya: IEEE, Sep. 2023, pp. 1–6.

[21] J. Kipongo, T. G. Swart, and E. Esenogho, "Artificial intelligence-based intrusion detection and prevention in edge-assisted SDWSN with modified honeycomb structure," *IEEE Access*, vol. 12, pp. 3140–3175, 2024.

[22] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 843–852.

[23] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar./Apr. 2009.

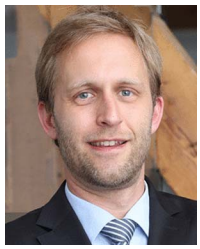
[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[25] M. Meyer and G. Kuschk, "Automotive radar dataset for deep learning based 3D object detection," in *Proc. 16th Eur. Radar Conf.*, Oct. 2019, pp. 129–132.
[26] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "CARRADA dataset: Camera and automotive radar with range- angle- doppler annotations," in *Proc. 25th Int. Conf. Pattern Recognit.*, Jan. 2021, pp. 5068–5075.
[27] S. Agrawal, S. Bhanderi, and G. Elger, "Semi-automatic annotation of 3D radar and camera for smart infrastructure-based perception," *IEEE Access*, vol. 12, pp. 34325–34341, 2024.
[28] F. Piewak et al., "Boosting LiDAR-based semantic labeling by cross-modal training data generation," in *Proc. Eur. Conf. Comput. Vis. Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham, Switzerland: Springer International Publishing, 2019, vol. 11134, pp. 497–513.
[29] M.-H. Sun, D.-H. Paek, S.-H. Song, and S.-H. Kong, "Efficient 4D radar data auto-labeling method using LiDAR-based object detection network," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2024, pp. 2616–2621.
[30] M. Heidbrink, O. Sura, V. K. Rangaraj, M. Reinecke, M. Hoffmann, and M. Vossiek, "Concept for automatic annotation of automotive radar data using AI-Segmented camera and LiDAR reference data," in *Proc. 21st Eur. Radar Conf.*, Sep. 2024, pp. 292–295.
[31] S. Isele, M. Schilling, F. Klein, S. Saralajew, and J. Zoellner, "Radar artifact labeling framework (RALF): Method for plausible radar detections in datasets," in *Proc. 7th Int. Conf. Veh. Technol. Intell. Transp. Syst.*, 2021, pp. 22–33.
[32] M. Dimitrievski, I. Shopovska, D. V. Hamme, P. Veelaert, and W. Philips, "Automatic labeling of vulnerable road users in multi-sensor data," in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, Indianapolis, IN, USA: IEEE Press, Sep. 2021, pp. 2623–2630.
[33] J. Gabsteiger, T. Maiwald, S. Wünsche, R. Weigel, and F. Lurz, "Automated radar data labeling through computer vision," in *Proc. IEEE Wireless Microw. Technol. Conf.*, Apr. 2023, pp. 77–80.
[34] A. Sengupta, A. Yoshizawa, and S. Cao, "Automatic radar-camera dataset generation for sensor-fusion applications," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 2875–2882, Apr. 2022.
[35] K. Ishak, Z. Zafar, M. Steiner, N. Appenrodt, J. Dickmann, and C. Waldschmidt, "A radar measurement setup with a ground truth system for micro-Doppler human movements," in *Proc. IEEE MTT-S Int. Conf. Microw. Intell. Mobility*, Apr. 2019, pp. 1–4.
[36] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16–28, Jul. 2019.
[37] S. Z. Gurbuz, M. M. Rahman, Z. Bassiri, and D. Martelli, "Overview of radar-based gait parameter estimation techniques for fall risk assessment," *IEEE Open J. Eng. Med. Biol.*, vol. 5, pp. 735–749, 2024.
[38] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, Mar. 1966.
[39] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, Mar. 1975.
[40] L. Li, R. Wang, and X. Zhang, "A tutorial review on point cloud registrations: Principle, classification, comparison, and technology challenges," *Math. Problems Eng.*, vol. 2021, pp. 1–32, Jul. 2021.
[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[42] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.
[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[44] A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 6392–6401.
[45] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019. [Online]. Available: https://github.com/facebookresearch/detectron2
[46] M. Ester, H.-P. Kriegel, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise" *kkd*, vol. 96, no. 34, pp. 226–231, 1996.
[47] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA J.*, vol. 3, no. 8, pp. 1445–1450, Aug. 1965.
[48] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering: With MATLAB Exercises*, 4th ed. Hoboken, NJ, USA: Wiley, 2012.
[49] R. R. Labbe, "Kalman and Bayesian filters in Python" Accessed Dec. 20, 2025. [Online]. Available: https://github.com/rlabbe/Kalman-and-Bayesian-Filters-in-Python
[50] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.
[51] S. Bogatin and D. Kogoj, "Processing kinematic geodetic measurements using Kalman filtering," *Acta Geodaetica et Geophysica Hungarica*, vol. 43, no. 1, pp. 53–74, Mar. 2008.
[52] X. Li, X. Wang, Q. Yang, and S. Fu, "Signal processing for TDM MIMO FMCW millimeter-wave radar sensors," *IEEE Access*, vol. 9, pp. 167959–167971, 2021.
[53] A. Wojtkiewicz, J. Misiurewicz, M. Nałecz, K. Jedrzejewski, and K. Kulpa, "Two-dimensional signal processing in FMCW radars," in *Proc. Nat. Conf. Circuit Theory Electron. Netw.*, Oct. 1997, pp. 475–480.
[54] R. B. Blackman and J. W. Tukey, "The measurement of power spectra from the point of view of communications engineering — Part I," *Bell Syst. Techn. J.*, vol. 37, no. 1, pp. 185–282, Jan. 1958.
[55] M. A. Richards, *Fundamentals of Radar Signal Processing*, 2nd ed. New York, NY, USA: McGraw-Hill Education, 2014.
[56] H. Rohling, "Ordered statistic CFAR technique—An overview," in *Proc. 12th Int. Radar Symp.*, Sep. 2011, pp. 631–638.
[57] M. Kronauge and H. Rohling, "Fast two-dimensional CFAR procedure," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 3, pp. 1817–1823, Jul. 2013.
[58] J. A. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines," *Astron. Astrophys. Suppl. Ser.*, vol. 15, Jun. 1974, Art. no. 417.
[59] R. Mautz, "Indoor positioning technologies," 1 Band. Habilitation thesis, Inst. Geodesy and Photogrammetry, Dept. Civil, Environmental and Geomatic Eng., ETH Zurich, Zurich, Germany, 2012.
[60] S. Aditya, A. F. Molisch, and H. M. Behairy, "A survey on the impact of multipath on wideband time-of-arrival based localization," *Proc. IEEE*, vol. 106, no. 7, pp. 1183–1203, Jul. 2018.
[61] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA: IEEE, Jul. 2017, pp. 77–85.
[62] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.
[63] P. Zhao et al., "mID: Tracking and identifying people with millimeter wave radar," in *Proc. 15th Int. Conf. Distrib. Comput. Sensor Syst.*, May 2019, pp. 33–40.
[64] N. Knudde et al., "Indoor tracking of multiple persons with a 77 GHz MIMO FMCW radar," in *Proc. Eur. Radar Conf.*, Oct. 2017, pp. 61–64.
[65] J. Pegoraro, D. Solimini, F. Matteo, E. Bashirov, F. Meneghello, and M. Rossi, "Deep learning for accurate indoor human tracking with a mm-Wave radar," in *Proc. IEEE Radar Conf.*, Sep. 2020, pp. 1–6.
[66] J. Pegoraro and M. Rossi, "Human tracking with mmWave radars: A deep learning approach with uncertainty estimation," in *Proc. IEEE 23rd Int. Workshop Signal Process. Adv. Wireless Commun.*, May 2022, pp. 1–5.

LUKAS ENGEL (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, in 2017 and 2019, respectively, where he is currently working toward the Ph.D. degree. In 2019, he joined the Institute of Microwaves and Photonics (LHFT), FAU. His research interests include radar-based human motion analysis using machine learning and deep learning, radar signal processing, antenna design, radar hardware, and 3D-printed millimeter-wave components.

MARKUS BERGMANN received the B.Sc. and M.Sc. degrees in electrical engineering from Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, in 2020 and 2025, respectively, with a specialization in information and communication technology. His research interests include signal processing, with particular emphasis on radar, image and video applications, high-frequency technologies and data-driven methods, such as machine learning and deep learning.

CHRISTOPH KAMMEL received the M.Sc. degree in electrical engineering from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, in 2020, and the Ph.D. degree from the Institute of Microwaves and Photonics, FAU, in 2025. He is currently with fiveD GmbH, where he is involved in the development of realistic radar simulation. His research interests include radar imaging and signal processing in the space sector, and radar-based motion tracking in medical applications. He was the recipient of the Semikron Promotion Prize in 2016, Rohde & Schwarz Prize in 2018, Baumüller Master Prize in 2021, and the Student Paper Award at the IEEE Radar Conference in 2023.

INGRID ULLMANN (Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, in 2016 and 2021, respectively. She is currently a Postdoc and Head of the research group "Wave-Based Sensing Techniques" with the Institute of Microwaves and Photonics, FAU. In 2022, she spent one month as a Visiting Researcher with Microwave Sensing, Signals and Systems Group, Delft University of Technology, Delft, The Netherlands. Her research interests include radar sensing, signal processing and imaging for nondestructive testing, security screening, and medical applications. She is a reviewer for the European Radar Conference, IEEE Radar Conference and various journals in the field of microwaves. Since 2022, she has been an Associate Editor for IEEE TRANSACTIONS ON RADAR SYSTEMS. Dr. Ullmann was the recipient of the EuRAD Conference Prize in 2019 and the IEEE Sensors Best Paper Award in 2023.



BJOERN M. ESKOFIER (Senior Member, IEEE) studied electrical engineering with FAU, Erlangen, Germany, in 2006. He received the Ph.D. degree in biomechanics under the supervision of Prof. Dr. Benno Nigg from the University of Calgary, Calgary, AB, Canada. In 2016, he was a Visiting Professor with Prof. Paolo Bonato's Motion Analysis Lab, Harvard Medical School, and a Visiting Professor with Prof. Alex Sandy Pentland's Human Dynamics Group, MIT Media Lab, in 2018. Since 2023, he has been an Associate Principal Investigator and Leader of the Research Group Translational Digital Health with Helmholtz Zentrum Munich. From April 2023 to August 2023, he was a Visiting Professor with Prof. Scott Delp's NMBL Lab that is part of Stanford University's Schools of Engineering and Medicine. He is currently the Head of the Machine Learning and Data Analytics Lab, Friedrich-Alexander-University Erlangen-Nuernberg, Erlangen, Germany. He is also the Founding Spokesperson with FAU's Department Artificial Intelligence in Biomedical Engineering, German Ministry of Economic Affairs and Climate Action GAIA-X Usecase Project TEAM-X, and co-spokesperson of the German Research Foundation Collaborative Research Center EmpkinS (www.empkins.de). He authored more than 400 peer reviewed articles, holds five patents, started three spinoff startup companies, and is in a supporting role for further startups. He was the recipient of several medical-technical research awards, including the Curious Minds Award 2021 in Life Sciences by Manager Magazin and Merck. He was the area editor of the IEEE OPEN ACCESS JOURNAL OF ENGINEERING IN MEDICINE AND BIOLOGY and Associate Editor for the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS. He is also active in the organization of several IEEE and ACM meetings (such as BSN, BHI, EMBC, IJCAI, ISWC, and UbiComp), most recently serving as General Chair of BHI 2023.

MARTIN VOSSIEK (Fellow, IEEE) received the Ph.D. degree from Ruhr-Universität Bochum, Bochum, Germany, in 1996. In 1996, he joined Siemens Corporate Technology, Munich, Germany, where he was the Head of the Microwave Systems Group from 2000 to 2003. Since 2003, he has been a Full Professor with Clausthal University, Clausthal-Zellerfeld, Germany. Since 2011, he has been the Chair of the Institute of Microwaves and Photonics (LHFT), Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. He has authored or coauthored more than 400 publications. His research has led to more than 100 granted patents. His research interests include radar, microwave systems, wave-based imaging, transponders, RF identification, communication, wireless sensor and locating systems. He is the spokesman of the Collaborative Research Centre (CRC 1483) EmpkinS, where more than 80 researchers aim to open up innovative wireless and wave-based sensor technologies for medicine and psychology. Dr. Vossiek is a member of the German National Academy of Science and Engineering (acatech) and of the German Research Foundation (DFG) Review Board 4.42-02 Communication Technology and Networks, Microwave Technology and Photonic Systems, Signal Processing and Machine Learning for Information Technology. He has been the spokesperson for the DFG Review Board 4.42 Electrical Engineering and Information Technology, since 2024. He is a member of the IEEE Microwave Theory and Technology (MTT) Technical Committees for MTT-24 Microwave/mm-Wave Radar, Sensing, and Array Systems; MTT-27 Connected and Autonomous Systems (as founding chair); and MTT-29 Microwave Aerospace Systems. He also serves on the advisory board of the IEEE CRFID Technical Committee on Motion Capture & Localization. He was the recipient of the numerous best paper prizes and other awards. In 2019, he was awarded the Microwave Application Award by the IEEE MTT Society (MTT-S) for Pioneering Research in Wireless Local Positioning Systems. He has been a member of organizing committees and technical program committees for many international conferences and has served on the review boards of numerous technical journals. From 2013 to 2019, he was an Associate Editor for IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES. Since 2022, he has been an Associate Editor-in-Chief for IEEE TRANSACTIONS ON RADAR SYSTEM.