IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 September 2026, accepted 30 September 2026, date of publication 19 October 2026, date of current version 29 October 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3745276

RESEARCH ARTICLE

Bringing Modern Multi-Object Tracking to Millimeter-Wave Radar Target Lists for Indoor Human Tracking

LUKAS ENGEL[1], (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)

[1]Institute of Microwaves and Photonics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany
[2]Machine Learning and Data Analytics Laboratory, 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 (lukas.le.engel@fau.de)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG), German Research Foundation–Sonderforschungsbereich (SFB) 1483 EmpkinS, under Project 442419336; and in part by the Bavarian Ministry of Economic Affairs, Regional Development and Energy, as a part of the BayVFP Fund Program–Funding Line Digitalization–Funding Section Information and Communication Technology, under Grant DIK-0310/02.

ABSTRACT Accurate indoor multi-person tracking is a key enabling technology for smart-home automation and human–machine interaction. While modern deep multi-object tracking (MOT) has revolutionized computer vision through tracking-by-detection and tracking-by-attention paradigms, it remains unclear whether these concepts transfer to sparse millimeter-wave radar target lists, whose unordered and point-based measurements fundamentally differ from image data. This work closes this gap by introducing the first comprehensive study of modern deep MOT paradigms for radar target lists. Two end-to-end architectures are proposed, representing the tracking-by-detection and tracking-by-attention paradigms, and adapted specifically to object-centric radar representations. Furthermore, a large-scale benchmark comprising more than 405,000 annotated radar frames acquired with our custom *RadarBox* platform is presented and released together with high-fidelity ground-truth annotations generated using an Azure Kinect system and refined by a bidirectional Kalman filter. Extensive experiments compare both paradigms against optimized DBSCAN–Kalman tracking pipelines and representative deep-learning baselines while analyzing the role of temporal modeling for radar-based MOT. Both proposed architectures consistently outperform conventional and existing deep-learning approaches across localization, detection, and tracking metrics. The results demonstrate that learning object-centric spatial-temporal representations directly from radar target lists effectively overcomes the characteristic limitations of conventional density-based clustering, particularly under sparse detections, multipath propagation, and occlusions. The presented results further demonstrate that modern end-to-end multi-object tracking paradigms can be successfully transferred to millimeter-wave radar target lists, establishing learned tracking as a superior alternative to conventional clustering-based radar tracking pipelines. Code and dataset are publicly available.

INDEX TERMS Attention, deep learning, DBSCAN, detection, indoor, Kalman filter, localization, radar, remote sensing, remote health care, smart home, tracking.

I. INTRODUCTION

Human localization and motion analysis in indoor environments are central components of intelligent smart-home and health-monitoring systems. Accurate position estimates enable context-aware automation and provide the basis for downstream tasks such as activity recognition and pose estimation [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12].

Wearable and camera-based systems can provide such functionality but often suffer from usability, privacy, or

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

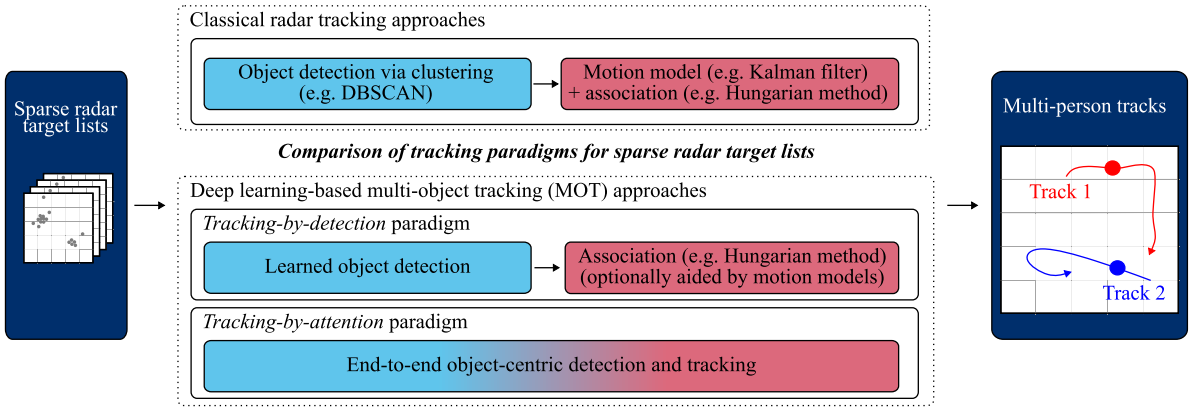


FIGURE 1. Content of this work: Comparison of classical and deep learning-based multi-object tracking paradigms for sparse radar target lists. Classical pipelines first perform object detection by clustering and subsequently associate detections over time using motion models. In contrast, the tracking-by-detection paradigm replaces handcrafted detection with learned object detection while retaining a separate association stage, whereas the tracking-by-attention paradigm performs detection and temporal association jointly in an end-to-end object-centric framework. The color coding is semantic and is used consistently throughout the paper: blue denotes detection-related processing, red denotes tracking and temporal association, and the blue-to-red gradient indicates their joint optimization within a single end-to-end model.

line-of-sight limitations. Radar sensors constitute a robust alternative, as they operate reliably under diverse indoor conditions, penetrate many occlusions, and avoid capturing identifiable imagery [13], [14]. In addition, their fine velocity resolution enables micro-Doppler-based motion analysis, revealing gait and activity patterns for patient monitoring [15], [16], [17].

Despite these advantages, deriving accurate indoor location estimates from radar data remains challenging due to the inherent sparsity of radar data [18], multipath propagation [19], and measurement noise [20]. These factors complicate the subsequent detection, clustering, and tracking stages.

Human tracking using indoor radar data has traditionally relied on handcrafted signal-processing and tracking pipelines [21], [22], [23]. Typically, radar detections were clustered into candidate targets using methods such as DBSCAN and subsequently associated over time using matching algorithms and Kalman filtering [24], [25], [26]. While several works introduced learning-based components for identity association through micro-Doppler signatures or motion prediction using deep sequential models [27], [28], the overall tracking pipeline remained largely based on clustering-based detection and handcrafted association rules, limiting robustness in multi-person scenarios.

In parallel, the computer vision community has achieved substantial progress in multi-object tracking (MOT) through deep learning-based approaches. *Tracking-by-detection* methods first generate frame-wise object detections using powerful (pretrained) object detectors, such as [29], [30], [31], [32], and subsequently associate them across time using motion models, appearance embeddings, or optimization-based data association techniques [33], [34], [35]. In contrast, *tracking-by-attention* approaches integrate temporal association directly into end-to-end neural architectures by propagating object queries across frames and jointly learning detection and identity preservation [36], [37], [38], [39], [40] (see Fig. 1). Both paradigms remain actively investigated, with neither consistently outperforming the other across all tracking scenarios.

However, transferring these paradigms to radar-based localization is non-trivial. In contrast to dense image features, radar target lists are sparse, unordered, noisy, and affected by multipath propagation and ghost detections. Consequently, identities must be inferred primarily from sparse spatial and temporal observations rather than appearance cues, requiring adapted object representations and association mechanisms.

Recent radar-based tracking approaches have adapted modern deep MOT paradigms, originally developed in computer vision and often built upon powerful pretrained object detectors, to image-like radar representations such as range–Doppler maps [41], [42]. In contrast, their applicability to sparse radar target-list data remains largely unexplored.

Therefore, this work investigates the applicability of modern MOT paradigms to indoor radar-based multi-person tracking. As illustrated in Fig. 1, classical tracking pipelines are compared with tracking-by-detection and tracking-by-attention approaches based on Slot Attention [43] and transformer-based query propagation to analyze suitable representation and association strategies for sparse target lists.

Our earlier study [44] demonstrated that deep learning applied to radar target lists enables reliable indoor people localization under occlusion and multipath. Building on this baseline, the present work extends the task from localization to multi-person tracking by incorporating object-centric representations and temporal trajectory estimation.

The remainder of this paper is organized as follows. Section II and Section III describe the hardware setup and radar signal-processing pipeline. Section IV introduces the dataset, while Section VI and Section VII present the classical baseline and the proposed deep-learning approach, respectively. Experimental results are presented in Section VIII, discussed in Section IX, and concluded in Section X.

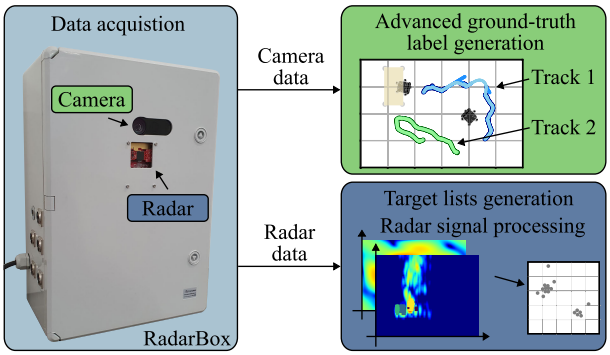


FIGURE 2. Measurement setup: Customized *RadarBox* platform combining calibrated radar sensing with synchronized camera-based ground-truth acquisition.

II. MEASUREMENT SETUP

To acquire diverse radar measurements for training and evaluation, data were recorded using the custom-developed *RadarBox* platform described in detail in [44]. As hardware design is not the focus of this work, only a brief overview is provided here. The system, shown in Fig. 2, consists of a compact and modular enclosure for synchronized multimodal data acquisition. Its main sensing units are a Texas Instruments radar evaluation module and a Microsoft Azure Kinect RGB-D camera[1] for ground-truth generation. Both sensors are rigidly mounted to maintain fixed relative geometry.

The Azure Kinect records hardware-synchronized RGB, depth, and infrared images and, together with the Azure Kinect Body Tracking SDK, provides 3D skeletal joint coordinates for position estimation. Positional accuracy is further refined using a bidirectional Kalman filter (BKF) [44]. The radar subsystem is based on the TI IWR6843AOPEVM,[2] featuring an L-shaped virtual antenna array. The main radar device specifications, antenna characteristics, and employed measurement configuration are summarized in Table 1. The wide field of view enables room-scale coverage using a single radar sensor, making the platform well suited for practical smart-home deployment. However, the compact antenna array provides only limited angular resolution, increasing the ambiguity of closely spaced individuals and thereby making multi-person localization and tracking considerably more challenging. The 15 mm × 15 mm antennas-on-package radar device was mounted centrally approximately 3 mm behind a 45 mm × 45 mm opening in the *RadarBox* enclosure. Consequently, no dielectric housing material covers the antennas, and the opening does not geometrically obstruct the specified ±60° field of view. All measurements reported in this work were acquired using this unobstructed configuration. Raw radar data are captured using the MMWAVEICBOOST[3] and DCA1000EVM[4] extension boards under time-division multiplexed MIMO operation. An update rate of 15 Hz with 128 chirps per transmitter antenna is employed to enable reliable motion tracking and fine-grained Doppler analysis.

[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 device specifications and measurement configuration. The measurement settings and antenna characteristics are adapted in part from [44], while the RF device specifications are taken from [45].

Parameter	Value
Device and antenna specifications	
Radar device	TI IWR6843AOP
Radar chip package dimensions	15 × 15 mm
Supported RF frequency range	60–64 GHz
Single-transmitter effective isotropic radiated power (EIRP)	16 dBm
Maximum IF bandwidth	10 MHz
Number of transmitter antennas	3
Number of receiver antennas	4
Nominal field of view in azimuth and elevation	±60°
Measurement configuration	
Configured chirp start frequency	60 GHz
Configured chirp 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	128
Approximate azimuth/elevation resolution	≈ 30°
Range resolution	≈ 14.8 cm
Unambiguous range	≈ 9.49 m
Doppler resolution	≈ 0.078 m/s
Unambiguous Doppler velocity	≈ ±5.02 m/s

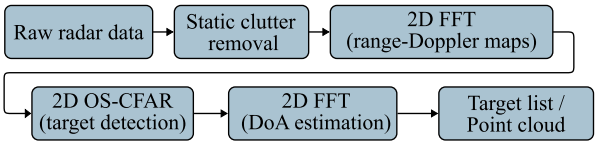


FIGURE 3. Developed radar signal-processing chain for indoor human localization used to extract target lists.

III. RADAR SIGNAL PROCESSING

The radar system employs a frequency-modulated continuous-wave (FMCW) chirp-sequence waveform with a time-division-multiplexed (TDM) MIMO antenna configuration. The signal-processing chain is illustrated in Fig. 3. Raw complex I/Q samples from all virtual transmit-receive antenna pairs are arranged into a radar data cube spanning range, Doppler, spatial antenna dimensions, and temporal frames. Static clutter is suppressed using frame-wise mean subtraction, followed by Hann windowing and a 2D FFT to obtain range–Doppler maps. Target candidates are extracted using a 2D ordered-statistics constant false alarm rate (2D OS-CFAR) detector [46]. Direction-of-arrival estimation is subsequently performed via a 2D spatial FFT across the virtual antenna array, while the CLEAN algorithm [47] is applied to reduce sidelobe interference and improve angular separation of closely spaced targets. Detected radar points are transformed into Cartesian coordinates and augmented with radial velocity, signal-to-noise ratio (SNR), noise power, and reflection intensity. The resulting target-list representation was selected to reflect practical edge-deployment scenarios, as modern radar chipsets increasingly generate target

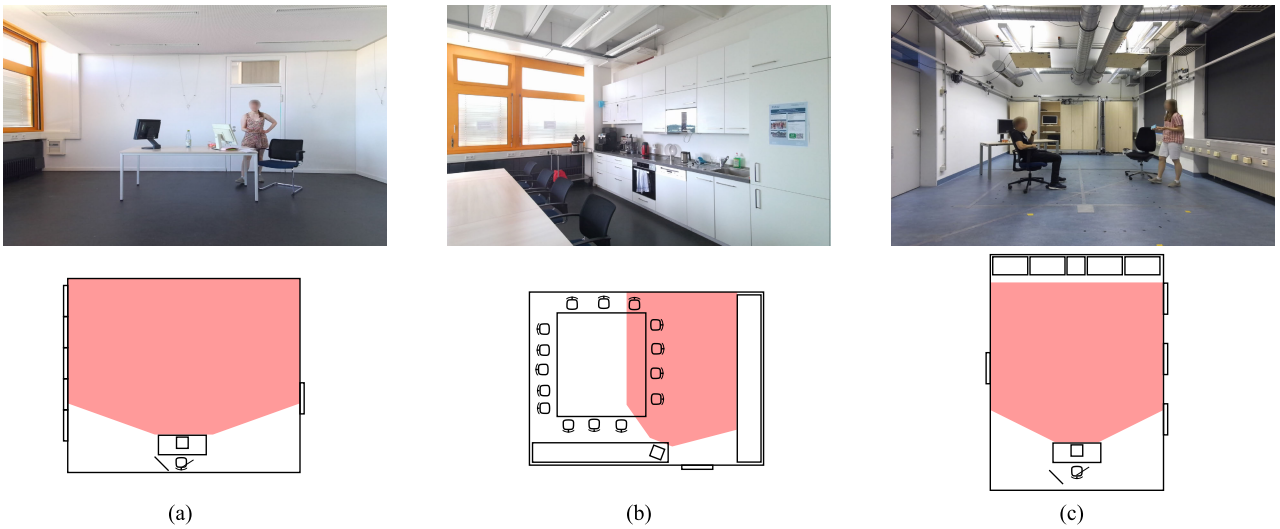


FIGURE 4. Photographs of the rooms utilized in this study: (a) office, (b) kitchen, and (c) laboratory configured as an office. The photographs are reproduced from [44]. The corresponding floor plans highlight the areas in which the participants moved during data acquisition.

detections directly on-chip. Compared to processing full range–Doppler or range–azimuth tensors, this representation reduces data transfer and computational requirements while retaining the information required for localization and tracking. To ensure consistent tensor dimensions for downstream processing, each frame is normalized to a fixed cardinality of 64 detections using zero-padding or intensity-based truncation when required. These processed detections form the common input representation for both the conventional baseline and the proposed deep-learning approach.

IV. ADVANCED LOCALIZATION DATASET

Data were collected in three indoor environments with different layouts and furnishing: a kitchen, an office, and a laboratory configured as an office. Across the three environments, the walls were plastered, while the ceilings comprised plastered or exposed concrete surfaces and, in some areas, exposed pipes. The furniture and other interior elements comprised wood, metal, and glass, with metal components present in chairs and tables, exposed pipes, appliances, cladding, and window structures. For each environment, recordings were conducted under three occupancy conditions: no person, one person, and two persons. The *RadarBox* sensor was placed on a table facing the relevant area of the room to ensure consistent coverage geometry, as highlighted in red in Fig. 4.

Several uninterrupted 15-minute recordings were acquired for each occupancy condition. To increase environmental variability, the arrangement of furniture such as tables and chairs was modified after each recording session, resulting in a substantially larger set of room configurations than the three base environments alone. Participants performed unconstrained office-related activities such as walking, sitting, standing, and phone usage to approximate realistic behavior. In total, the dataset comprises approximately 405,000 radar frames and is balanced across all occupancy levels.

All procedures were approved by the Ethics Committee of Friedrich-Alexander-Universität Erlangen–Nürnberg (Protocol #22-437-B), and written informed consent was obtained from all participants. Ground-truth annotations were generated from Azure Kinect pelvis-root joint coordinates refined using the BKF pipeline proposed in [44]. By combining forward and backward state estimation, the BKF produces temporally consistent and noise-suppressed trajectories, reducing occlusion-induced discontinuities and short-term tracking errors. The advantages of the BKF-based annotation strategy are illustrated in Fig. 5, which compares pelvis trajectories obtained from Azure Kinect Body Tracking, a conventional forward Kalman filter, and the BKF. During periods without measurements, the forward filter exhibits substantial drift because the state estimate evolves solely under the constant-velocity prediction model. In contrast, the BKF performs estimation bidirectionally over the complete temporal sequence, yielding smoother and dynamically more consistent trajectories across occluded intervals. Since ground-truth generation is performed offline, the full sequence can be exploited during filtering, resulting in stable and reliable reference trajectories. To support reproducible research, the dataset and source code are publicly available.[5]

V. EVALUATION METRICS

A predicted person was considered a true positive only if it could be matched to a ground-truth person within a Euclidean distance threshold τ. In this work, τ was set to 1 m. Predictions without a corresponding ground-truth person within this threshold were counted as false positives, while unmatched ground-truth persons were counted as false negatives. The same matching criterion was used for all localization, detection, and tracking metrics.

[5]Dataset: https://doi.org/10.5281/zenodo.17526861; source code: https://github.com/lukeng-ai/mmwave-radar-mot

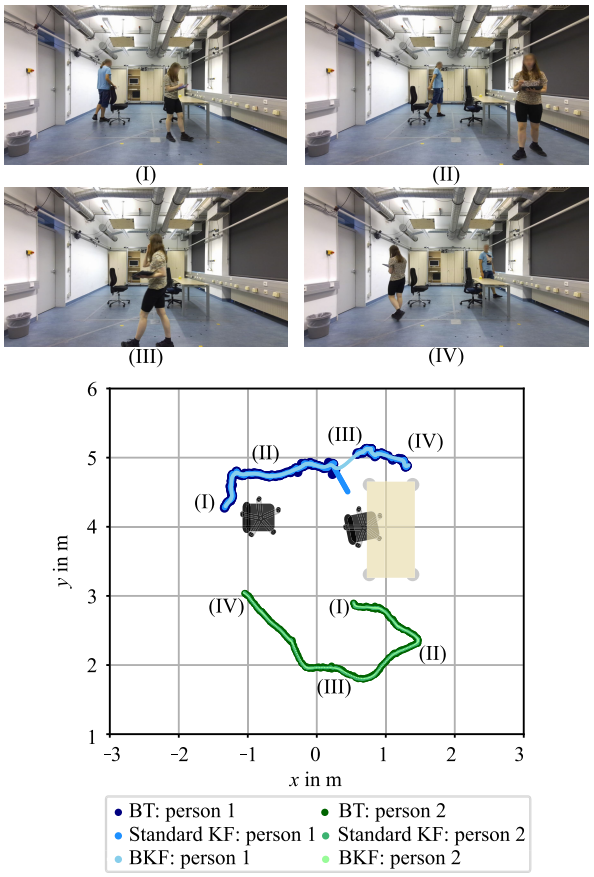


FIGURE 5. Visualization of ground-truth label generation in challenging scenes. Top: Photographs (I)–(IV) depict representative moments from the Azure Kinect camera recordings, showing the trajectories of two subjects and highlighting instances with partial or full visual occlusion. Bottom: Tracking comparison between the standard Body Tracking SDK (BT) and the proposed bidirectional Kalman filter (BKF). The BKF yields smoother and more accurate trajectories, particularly noticeable in Region III. The scene originates from a test-set sequence. The figure was reproduced from [44].

Localization accuracy was evaluated using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of the estimated person positions, computed over all matched detections:

$$\mathrm{MAE} = \frac{1}{N_{TP}} \sum_{i=1}^{N_{TP}} d_i, \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{TP}} \sum_{i=1}^{N_{TP}} d_i^2}. \tag{2}$$

Detection performance was evaluated using Detection Accuracy (DetA), precision, recall, and F1-score:

$$\mathrm{DetA} = \frac{TP}{TP + FP + FN}, \tag{3}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{6}$$

Tracking performance was additionally assessed using the Multiple Object Tracking Accuracy (MOTA) [48] and the Higher Order Tracking Accuracy (HOTA) [49]. MOTA jointly accounts for false positives, false negatives, and identity switches:

$$\mathrm{MOTA} = 1 - \frac{FN + FP + \mathrm{IDS}}{GT}, \tag{7}$$

where GT denotes the total number of ground-truth objects. HOTA combines detection and association quality as

$$\mathrm{HOTA} = \sqrt{\mathrm{DetA} \cdot \mathrm{AssA}}, \tag{8}$$

where AssA quantifies the temporal consistency of identity assignments [49]. Consequently, HOTA provides a balanced assessment of both detection and identity preservation.

VI. CLASSICAL TRACKING APPROACH USING POINT CLOUD-CLUSTERING AND KALMAN FILTERING

We implemented a classical processing pipeline following established radar-based methodologies for human tracking [22], [27], [50], [51]. The developed processing chain is depicted in Fig. 6.

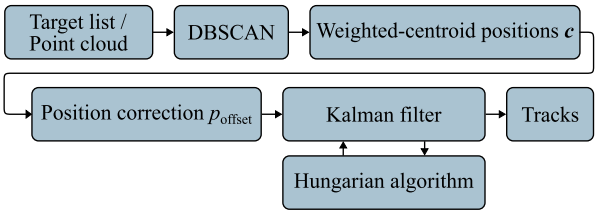


FIGURE 6. Developed classical signal processing chain for target list-based human indoor tracking.

DBSCAN [24] was applied to the xy-projected radar point clouds using a neighborhood radius ε and a minimum cluster size $n_{\min}$. For each cluster, an intensity-weighted centroid $\boldsymbol{c} = \frac{\sum_{i=1}^{N} a_i \boldsymbol{r}_i}{\sum_{i=1}^{N} a_i}$ was computed similarly to [22], [27], [51], where a_i denotes the radar intensity of radar target $\boldsymbol{r}_i$. Since radar reflections mainly originate from the body surface whereas the anatomical pelvis center lies deeper within the body volume, a radial offset correction p_{offset} was applied to approximate the pelvis-centered reference position. Each cluster was interpreted as one potential human target.

Temporal association across frames was performed using a constant-velocity Kalman filter [26] and Hungarian matching [25]. Assignments were accepted only if the distance between predicted and measured positions remained below a threshold t. Short-term occlusions and missed detections were handled using a counter-based gap mechanism, while tracks exceeding the allowed gap duration were terminated.

To maintain comparability with the proposed deep learning approaches, the number of simultaneously tracked targets was limited to two. If more than two clusters were detected, the two largest clusters were retained.

The algorithmic parameters were systematically optimized and evaluated using the same validation split that was applied

for tuning the deep learning frameworks, ensuring a fair and consistent comparison between the approaches.

The DBSCAN parameters were optimized on the validation set, yielding $\varepsilon = 0.3$ m, consistent with [27], [51], and a positional offset of $p_{\text{offset}} = 0.08$ m. The association threshold was set to $t = 0.4$ m, and tracks were terminated after six consecutive missed updates. While ε and p_{offset} require tuning, the minimum-samples parameter $n_{\min}$ has the strongest influence on clustering and tracking performance.

Overall, the performance of the classical tracking pipeline is highly dependent on careful parameter tuning. This sensitivity motivates the investigation of the deep learning–based approaches presented in the following section, which aim to learn robust detection and tracking representations directly from data.

VII. DEEP LEARNING-BASED MULTI-OBJECT TRACKING FROM RADAR TARGET LISTS

Motivated by recent advances in MOT, we investigated two complementary deep learning paradigms for radar-based multi-person tracking: tracking-by-detection and tracking-by-attention. As illustrated in Fig. 1, the former first estimates object detections and subsequently performs temporal association, whereas the latter integrates temporal association directly into the network through propagated object queries. Both approaches operate on the same radar-derived target-list representation and are designed to address the limitations of handcrafted clustering and association rules discussed in the previous section. The following subsections describe the corresponding radar-adapted architectures and their training procedures. To ensure a fair comparison, both approaches use the same input representation and feature extraction pipeline, which are introduced first.

A. INPUT REPRESENTATION AND FEATURE EXTRACTION BACKBONE

Each of the n radar points per frame is represented by an eight-dimensional feature vector comprising its 3D position (x, y, z), radial velocity, SNR, noise estimate, intensity, and a binary indicator distinguishing valid detections from zero-padded points. No constraints based on the physical room dimensions were applied to the point-cloud data.

Feature extraction is performed by two complementary branches (see Fig. 7(a)). The first branch follows the PointNet design principle [52] and consists of three shared multi-layer perceptron blocks, each comprising a linear layer, batch normalization, and a ReLU activation function. Since these shared operations process each point independently, they primarily enrich the point-wise feature representation. The feature dimensionality is progressively increased from 8 to 384. The second branch employs a two-stage Kernel Point Convolution (KPConv) network [53], which models local interactions between neighboring points through convolution-like operations on irregular point sets, as illustrated in Fig. 7(b). Seven kernel points are initialized on a sphere with radius 0.5 m and projected onto the

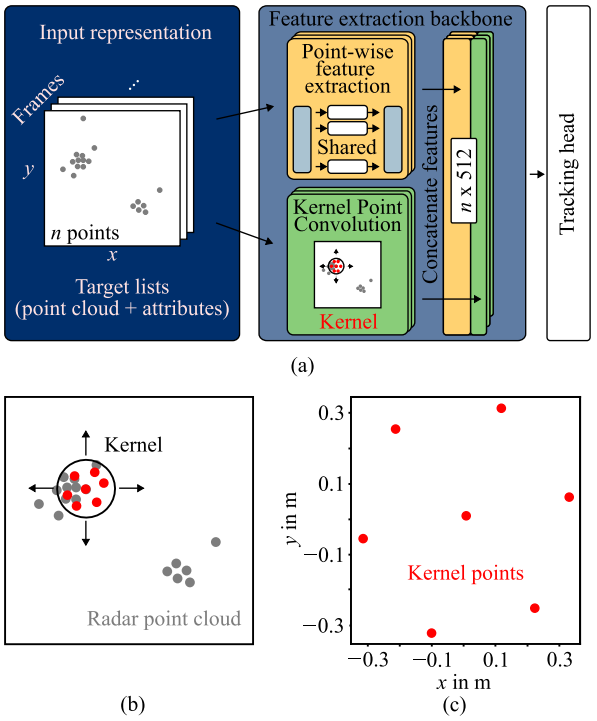


FIGURE 7. Feature extraction backbone: (a) input representation and dual-branch feature extractor, (b) schematic illustration of the KPConv module, in which kernel points are convolved with the discrete input cloud to capture local spatial relationships, and (c) kernel point distribution used in our implementation.

xy-plane, resulting in the kernel configuration shown in Fig. 7(c). Neighboring points are weighted by a Gaussian influence function ($\sigma = 0.24$ m), while trainable kernel weights learn localized spatial filters. The binary indicator channel provides additional information about local point density and padding regions. Across two KPConv stages, the feature dimensionality is increased from 8 to 128. Finally, the outputs of both branches are concatenated into an $n \times 512$ feature representation, batch-normalized, and used as input to the tracking-specific heads described in the following subsections.

B. PROPOSED TRACKING-BY-DETECTION APPROACH

To realize the tracking-by-detection paradigm, we designed a deep neural architecture that estimates person-centric detections from radar-derived target lists, as illustrated in Fig. 8. Following the general principle of tracking-by-detection frameworks such as SORT [33], the approach first generates frame-wise object detections and subsequently associates them over time. Extracted point-wise features are transformed into object-centric representations, refined using short-term temporal context, and finally mapped to occupancy and position estimates. During inference, Hungarian matching and Kalman filtering are employed to obtain temporally consistent trajectories.

1) SLOT ATTENTION MODULE

A Slot Attention module [43] is employed to transform the extracted point-wise feature embeddings into a fixed

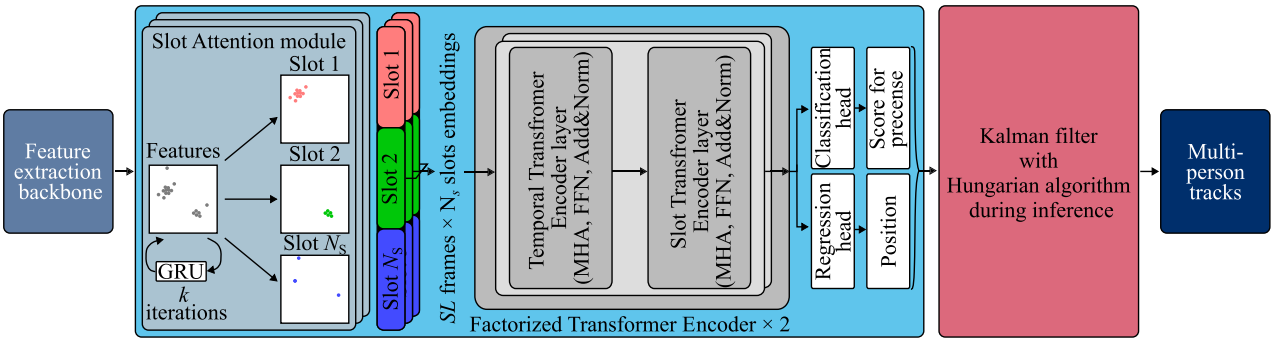


FIGURE 8. Proposed tracking-by-detection architecture: A Slot Attention module transforms the variable-sized point set into a fixed number of object-centric slots. A factorized Transformer encoder models short-term temporal dependencies, and task-specific prediction heads estimate occupancy scores and Cartesian positions. During inference, Hungarian matching and a Kalman filter provide long-term trajectory association and identity stabilization.

number of latent slots. Slot Attention inherits the core principles of the attention mechanism [54] by establishing soft associations between input features and latent slots. It performs k refinement iterations in which each slot aggregates information from the inputs according to the learned attention weights, while the slot states are updated through a gated recurrent unit (GRU) cell [55] to ensure stable and consistent refinement. The mechanism is inherently permutation invariant with respect to the ordering of the input target list.

In this work, Slot Attention serves as a learnable clustering mechanism that replaces the handcrafted DBSCAN stage by transforming point-level radar measurements into object-centric representations. In contrast to conventional clustering, point assignments emerge from learned feature similarity rather than fixed spatial distance thresholds. To accommodate clutter and outliers, the number of slots was fixed to $N_S = 5$, exceeding the maximum number of simultaneously present persons. Example assignments are shown in Fig. 9. Single- and two-person scenarios are represented by one and two dedicated slots, respectively, whereas clutter is assigned to the remaining slots. Although incomplete separation may occur in challenging cases, subsequent network stages further refine the resulting object representations.

2) SHORT-TERM TEMPORAL REFINEMENT USING A TRANSFORMER ENCODER AND PREDICTION HEADS

The proposed architecture follows the tracking-by-detection paradigm by first generating object-centric detections and subsequently associating them over time. In this context, the Slot Attention module serves as a learnable object discovery mechanism that transforms point-level radar measurements into object-centric representations. Previous analyses on the employed radar dataset indicated that exploiting temporal context across multiple frames substantially improves person localization compared to processing individual frames independently. Therefore, rather than directly predicting detections from the Slot Attention outputs, the object-centric representations are further refined using a factorized Transformer encoder operating on short temporal sequences. During both training and inference, the Transformer encoder processes fixed-length sequences of SL consecutive frames. At inference time, a sliding window is evaluated for every new frame; however, each window is processed independently and the Transformer state is reinitialized after every sequence. Consequently, temporal information is exploited only within the current sequence and is not propagated across sequence boundaries. This design enables the model to incorporate motion cues and temporal consistency into the detection process while preserving the slot-wise object representation, such that frame-wise person detections can subsequently be processed by a conventional tracking stage.

The object-centric slot representations obtained from Slot Attention are first projected into compact per-frame embeddings. Before temporal processing, the slots are temporally aligned by selecting the sequence of one-to-one permutations that minimizes the cumulative squared Euclidean distance between slot embeddings of consecutive frames. The factorized Transformer encoder then alternates between temporal attention, which models the motion history of each slot independently, and slot attention, which enables information exchange between slots within each frame to reduce duplicate detections and resolve ambiguous object hypotheses. Each block consists of one temporal and one slot Transformer encoder layer with standard self-attention and feed-forward operations. Finally, the refined slot embeddings of the last frame are processed by shared classification and regression heads to predict person presence and Cartesian positions.

3) LONG-TERM TEMPORAL REFINEMENT USING A KALMAN FILTER

Although the Transformer encoder exploits temporal context within windows of length SL, each sequence is processed independently and the internal network state is reinitialized at the beginning of every sequence. Consequently, temporal information is only propagated within the observed sequence window and is not preserved across longer time horizons. In particular, the employed Slot Attention mechanism is identity-agnostic and does not enforce a consistent

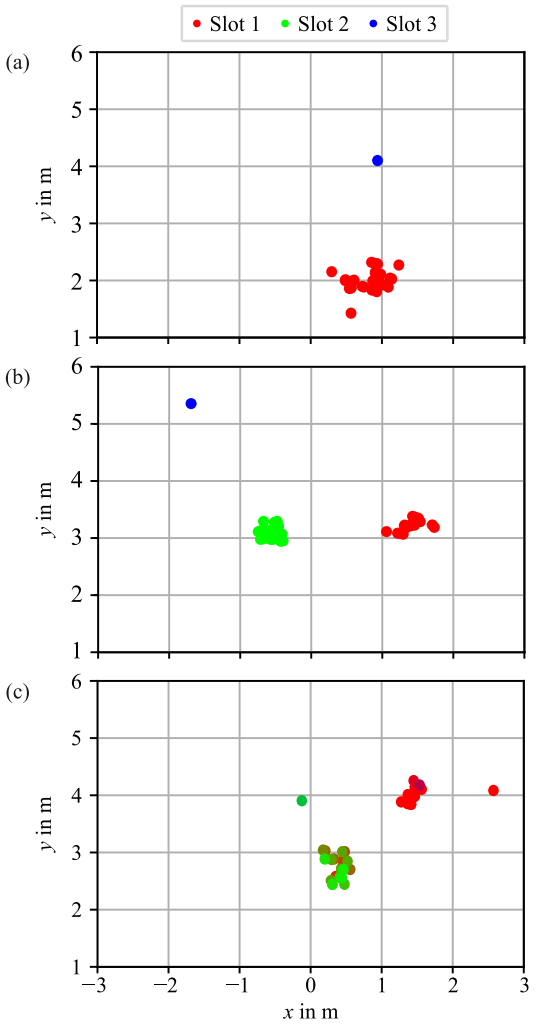


FIGURE 9. Results of the Slot Attention module for a visualization example with three slots: (a) In the single-person scenario, all person-related points are consistently assigned to slot 1. (b) In the two-person scenario, points are cleanly separated, with each individual captured by slot 1 and slot 2, respectively. (c) In the failure case, two individuals are only partially separated, resulting in mixed assignments across slot 1 and slot 2. Nonetheless, subsequent processing stages may further refine these assignments. A brighter color indicates a higher assignment weight for the corresponding slot.

correspondence between slots over time. Consequently, the assignment of individuals to slots may change between consecutive windows, preventing the direct use of slot indices as persistent track identities.

To obtain temporally consistent trajectories, an explicit data association stage is therefore required during inference. We employed Hungarian matching to associate detections with existing tracks. In addition, a Kalman filter based on the previously described constant-velocity motion model is integrated into the tracking framework. While Hungarian matching alone would be sufficient for frame-to-frame association, the Kalman filter provides motion predictions that improve robustness to noisy detections and short-term missed observations while introducing only negligible computational overhead.

Alternative approaches could incorporate identity information directly into the learning framework, for example through cross-window slot propagation, spectrogram-based identity cues [51], or learnable filtering methods [28]. Similar to the appearance embeddings used in DeepSORT [34], spectrogram-based cues provide discriminative identity information for data association. However, these approaches increase model complexity and supervision requirements, whereas the proposed Kalman-based refinement provides a simple and effective mechanism for long-term identity stabilization.

C. PROPOSED TRACKING-BY-ATTENTION APPROACH

To realize the tracking-by-attention paradigm, we designed a deep neural architecture that jointly performs object localization and temporal association from radar-derived target lists. In contrast to the tracking-by-detection approach, temporal association is integrated directly into the network through persistent track queries that are propagated across consecutive frames. These queries maintain track-specific information over time while simultaneously refining object states, thereby unifying detection and tracking within a single end-to-end framework, as illustrated in Fig. 10. During training, the model is optimized on short sequences of length SL to learn motion patterns and temporal query propagation. During inference, however, the architecture operates in a streaming manner by processing one frame at a time while continuously propagating the internal track state between consecutive frames. Consequently, the inference horizon is not limited by the training sequence length SL.

The proposed tracking-by-attention architecture is based on MOTR [37], which was selected among existing transformer-based MOT frameworks due to its explicit query management mechanism. Unlike earlier approaches such as TransTrack [40] and TrackFormer [36], MOTR explicitly models the complete track lifecycle, including initialization, propagation, temporary persistence, and termination, within a unified end-to-end framework. These properties are particularly attractive for radar-based tracking, where measurements are often sparse or intermittent and appearance information is largely unavailable. To better exploit the temporal characteristics of radar target lists, we extend MOTR with an explicit temporal feature memory that stores encoded radar representations from multiple recent frames. While propagated track queries already capture the tracking history, they provide only a compact latent representation. The proposed feature memory additionally enables the decoder to attend directly to radar observations from recent frames, increasing the measurement information available for the current prediction. In contrast, MOTRv2 [38] and MOTRv3 [39] mainly introduce training refinements for large-scale vision datasets without modifying the underlying query-propagation mechanism. Since their benefits for sparse radar target-list data remain unclear, the original MOTR framework was adopted and extended with the proposed temporal feature memory.

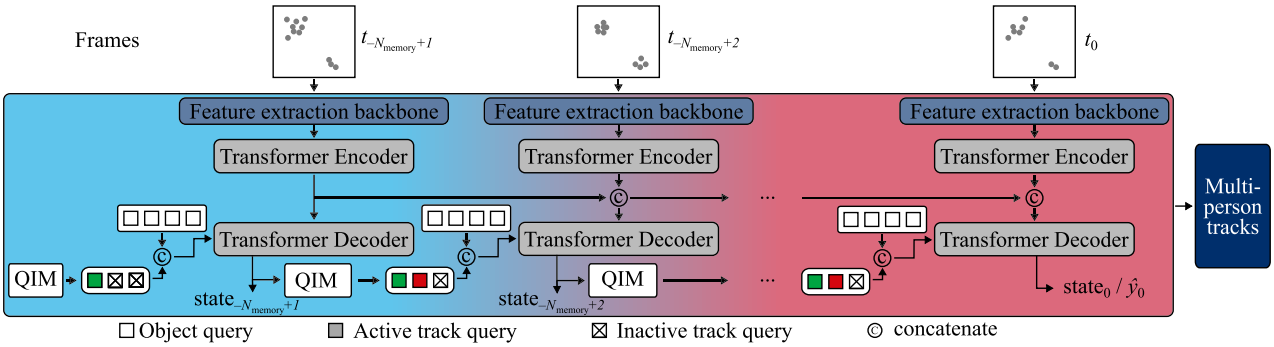


FIGURE 10. Proposed tracking-by-attention-based architecture. At each time step t, the previous track state state_{t-1} is processed by the QIM (query interaction module) to obtain propagated track queries, which are concatenated with object queries and used as decoder targets. The decoder memory is formed from encoded radar point-feature embeddings of the current frame and a FIFO memory of recent N_{memory} previous frames. After decoding, the resulting query representations are mapped to predictions $\hat{y}_t$ and used to update the track state state_t for the next frame. Prediction heads are omitted for visual clarity.

1) QUERY PROPAGATION AND TEMPORAL ASSOCIATION

Temporal tracking is realized through two types of learnable queries. Track queries represent already initialized person tracks, whereas object queries are used to detect newly appearing persons. In all experiments, the number of track queries was fixed to 5, matching the number of slots used by the proposed tracking-by-detection architecture to provide comparable representational capacity. In contrast, 3 object queries were employed, as they are solely required for the initialization of newly appearing tracks, making a small number of object queries sufficient. At each time step t, the track state from the previous frame, denoted as state_{t-1}, is processed by a query interaction module (QIM) [37] before decoding the current frame. The QIM updates the track queries using their latent representations, previous position estimates, confidence scores, and active/inactive states, thereby propagating track-specific information across time. During training, following the original MOTR approach [37], the QIM randomly drops 10% of active track queries and injects false-positive queries before applying self-attention, improving robustness to missing and erroneous tracks during inference. The self-attention operation within the QIM enables information exchange between the remaining active and injected false-positive queries, helping to resolve ambiguities arising from nearby or closely interacting persons.

The propagated track queries are concatenated with object queries and used as the target sequence of a transformer decoder. In contrast to the original MOTR formulation, we additionally equipped the decoder with an explicit temporal feature memory. The decoder memory contains encoded radar feature embeddings from the current frame together with a fixed-size FIFO buffer of recent frames. Learned temporal embeddings encode the relative age of stored features, enabling temporal attention across multiple observations. During training, the model is optimized on sequences of length SL to learn temporal query propagation. During inference, however, the propagated track queries are continuously updated and carried forward from one frame to the next rather than being reinitialized after fixed-length temporal windows, enabling streaming operation over arbitrarily long sequences. While the explicit feature memory is limited to the most recent N_{memory} frames, the propagated track queries preserve temporal identity information beyond this horizon.

Prediction heads are omitted from the figure for visual clarity. At each time step, the decoder outputs are mapped to localization and confidence predictions $\hat{y}_t$. Together with the decoded track-query embeddings, these predictions form the updated track state state_t, which contains the track-query embeddings, predicted positions, confidence scores, and active/inactive flags and is propagated to the next frame.

2) TRACK STATE UPDATE AND QUERY MANAGEMENT

After each frame, the internal track state state_t is updated online to maintain a consistent set of active tracks. Active track queries are propagated together with their latent representation, position estimate, confidence score, and activity status. Tracks whose confidence falls below a predefined threshold are retained for a limited number of consecutive frames to improve robustness against temporary occlusions, sparse radar returns, and missed detections. Conversely, object-query predictions exceeding a new-track threshold and sufficiently separated from existing tracks initialize new track queries. Track queries that remain inactive beyond the permitted duration are reset to their learned initialization state and become available for future track creation.

This query management strategy enables the network to continuously initialize, propagate, and terminate tracks without requiring a separate frame-to-frame association module during inference.

Since identity preservation is learned implicitly from the propagated query representations, tracking performance may degrade after prolonged target absence or ambiguous interactions between nearby persons. Moreover, once incorrect identity information has been incorporated into a track query, it may propagate across subsequent frames. These limitations are addressed implicitly during end-to-end training rather than through explicit motion models or post-processing.

D. TRAINING, OPTIMIZATION, AND EVALUATION

The dataset was divided by room into training, validation, and test sets following an 80:10:10 split. To evaluate cross-room generalization, one environment (Fig. 4(c)) was reserved exclusively for testing. For direct comparability, all deep learning approaches were evaluated on the same validation and test sets as the classical tracking pipeline. Models were trained end-to-end in PyTorch using the Adam optimizer with a batch size of 64, weight decay of 10^{-4}, and an initial learning rate of 10^{-4} decayed exponentially by 0.98 per epoch. Training employed the localization objective

$$\mathcal{L}_{\text{loc}} = \lambda_{\text{pos}} \cdot \mathcal{L}_{\text{MSE}} + \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{BCE}} + \lambda_{\text{count}} \cdot \mathcal{L}_{\text{count}}, \quad (9)$$

where $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{BCE}}$ denote the position regression and occupancy classification losses, respectively, while $\mathcal{L}_{\text{count}}$ penalizes overestimation of the number of present persons. The temporal input sequence was used as context for all models. For the tracking-by-detection model, the loss was computed only on the final supervised frame of each input window. For the tracking-by-attention model, supervision was applied over the complete sequence: query outputs were assigned to ground-truth tracks using a sequence-level order-invariant matching, and the loss terms were then evaluated over all matched frames. For the tracking-by-detection model, the MSE and BCE losses were applied to matched slots after order-invariant assignment. For the tracking-by-attention model, they were applied to the assigned query outputs, including both propagated track queries and newly initialized object queries. In addition, the tracking-by-attention model employed the differentiable detection objective

$$\mathcal{L}_{\text{tracking-by-attention}} = \mathcal{L}_{\text{loc}} + \lambda_{\text{det}} \cdot \mathcal{L}_{\text{softDetA}}, \quad (10)$$

where $\mathcal{L}_{\text{softDetA}}$ penalizes false positives and false negatives.

The models were optimized using the differentiable losses described above. However, these losses do not directly reflect the final tracking objective, while HOTA, which does, depends on discrete detection and association decisions and is therefore non-differentiable. Consequently, model selection was based on the validation score

$$S_{\text{val}} = \text{HOTA} - \lambda_{\text{MAE}} \cdot \text{MAE}. \quad (11)$$

This score rewards high tracking quality while penalizing localization error. The model achieving the highest validation score S_{val} was used for the final evaluation.

VIII. RESULTS

In this section, we report the test-set results for both the deep-learning models and the classical processing chain and directly compare their performance. In the zero-person case, both approaches yielded identical outcomes, with neither producing false detections.

A. QUALITATIVE RESULTS

Representative qualitative results are shown in Fig. 11. Columns (a)–(c) show one- and two-person scenarios, while the rows compare DBSCAN with $n_{\text{min}} = 6$ and 9 to the proposed tracking-by-detection and tracking-by-attention approaches. Frame-wise detections are shown as points and ground-truth trajectories as blue lines. For the DBSCAN-based pipeline and the tracking-by-detection approach, Kalman-filtered trajectories are shown as lines, whereas the tracking-by-attention approach directly predicts identity-specific trajectories without requiring Kalman filtering. Overall, the DBSCAN-based pipeline provides reliable tracking when sufficiently dense radar detections are available. However, multipath reflections generate spurious clusters and false tracks, while sparse detections caused by increased distance, occlusions, or nearby persons often fail to satisfy the DBSCAN density criterion, resulting in fragmented trajectories.

Both proposed deep-learning approaches are considerably more robust to these effects. Instead of relying on local frame-wise point density, they infer person locations directly from the spatial and temporal distribution of the radar measurements, allowing stable localization even from sparse or partially corrupted observations. As a result, they generate denser frame-wise predictions, maintain more cohesive trajectories throughout challenging motion sequences, and are less affected by multipath-induced clutter, which frequently produces spurious detections in the conventional DBSCAN-based pipeline. Nevertheless, both methods are limited by the spatial domain represented in the training data. Since the training data contain no samples beyond approximately 4.9 m, localization accuracy gradually degrades outside this range. Therefore, the quantitative evaluation is restricted to the region in which all methods operate reliably ($y < 4.9$ m).

B. QUANTITATIVE RESULTS

Table 2 summarizes the quantitative comparison between the proposed architectures, classical tracking pipelines, and state-of-the-art deep-learning approaches using tracking, detection, and localization metrics. For a fair comparison, all methods are reported using the same training seed.

Among the classical approaches, DBSCAN with $n_{\text{min}} = 9$ achieved higher tracking performance than $n_{\text{min}} = 6$, whereas $n_{\text{min}} = 6$ yielded a higher recall. Within the tracking-by-detection approaches, the adapted MARS [56] model showed the weakest performance, while PointNet [52], PointNet++ [57], and our previous baseline [44] progressively improved the results. Building on this baseline, the proposed tracking-by-detection architecture achieved the highest HOTA (91.72 %), highest AssA (93.35 %), and lowest MAE (0.104 m) of all evaluated methods.

Within the tracking-by-attention approaches, MOTR consistently outperformed TrackFormer, while the proposed tracking-by-attention architecture achieved the best overall performance within this category. Compared with the proposed tracking-by-detection approach, it obtained a slightly higher MOTA (89.75 % vs. 89.14 %), precision (92.63 % vs. 91.49 %), and the lowest RMSE (0.138 m), whereas the

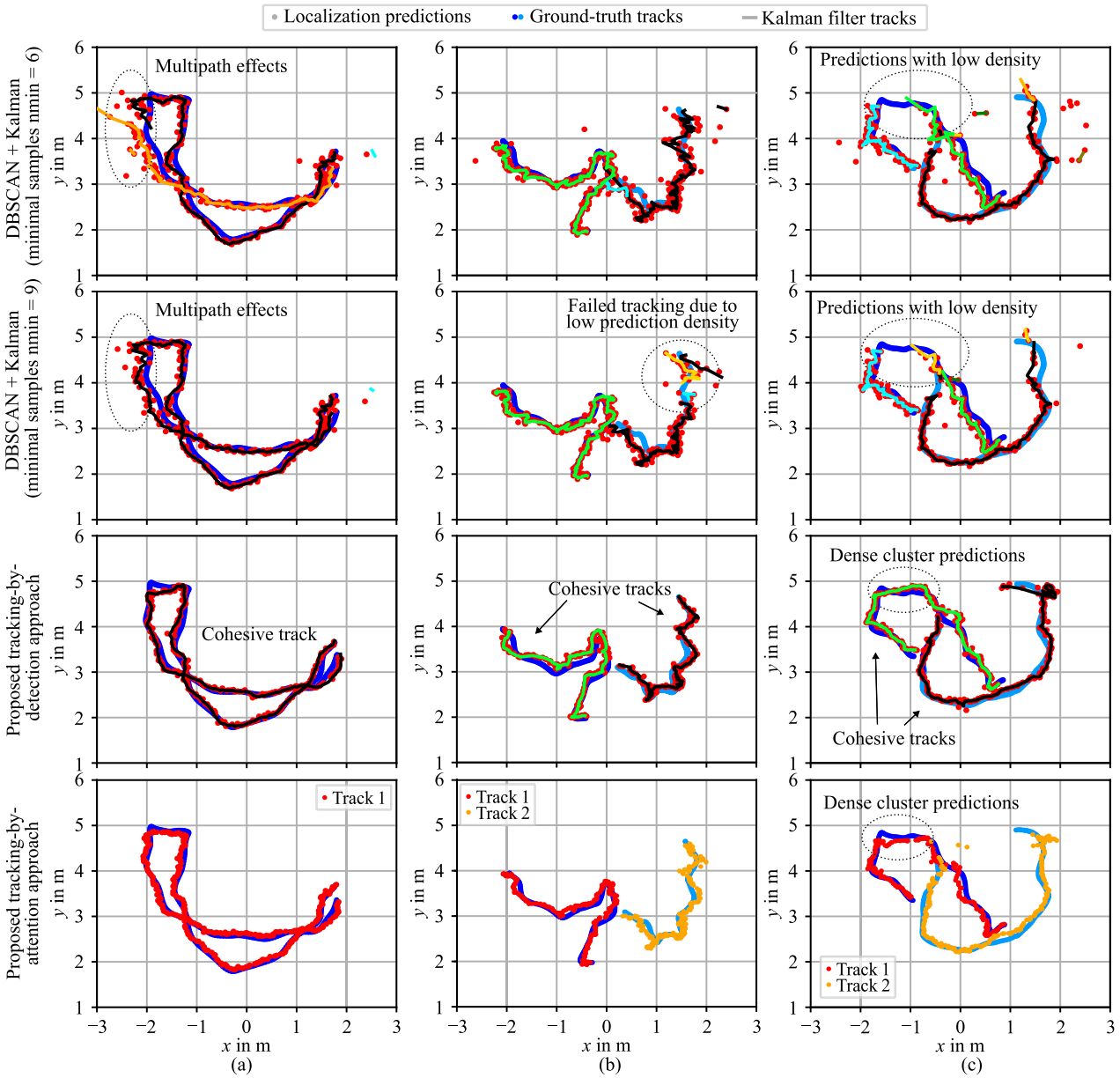


FIGURE 11. Qualitative comparison of representative tracking results for the conventional DBSCAN-based processing chain and the proposed deep-learning approaches in one- and two-person scenarios. Red dots denote frame-wise localization predictions, blue lines the ground-truth trajectories, and gray lines the Kalman-filtered tracks. The highlighted examples illustrate typical failure modes of the DBSCAN-based pipeline, including multipath-induced false tracks and fragmented trajectories caused by low-density detections, whereas both deep-learning approaches generally produce denser predictions and more cohesive tracks.

tracking-by-detection approach remained superior in HOTA, AssA, and MAE.

For TrackFormer, we disabled the original ground-truth query propagation during training and instead propagated model-predicted track queries. In our streaming setting, ground-truth propagation introduced a pronounced train–test mismatch because the ground-truth trajectories used during training are unavailable at inference. This resulted in unstable optimization and numerous false-positive track predictions. We therefore evaluated TrackFormer using the same closed-loop query propagation during training and inference.

Overall, both proposed architectures consistently outperformed the classical and state-of-the-art baselines across the evaluated tracking, detection, and localization metrics.

C. ABLATION STUDY

The influence of temporal context on the proposed tracking-by-detection architecture is illustrated in Fig. 12 and Fig. 13. The reported results correspond to the mean over three independent training runs with different random seeds. In both figures, the reported values are shown relative to the single-frame baseline ($SL = 1$), thereby highlighting

TABLE 2. Quantitative comparison of the proposed deep-learning architectures with classical and state-of-the-art tracking methods. Best values are highlighted in bold. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance. TE: Transformer Encoder, *SL*: sequence length, KF: Kalman filter.

	Tracking in %				Detection in %			Localization in m	
	HOTA ↑	MOTA ↑	DetA ↑	AssA ↑	Precision ↑	Recall ↑	F1 ↑	MAE ↓	RMSE ↓
Classical tracking approaches									
DBSCAN ($n_{\min}$=6) + KF	76.70	69.05	75.44	77.96	79.61	93.51	86.00	0.121	0.154
DBSCAN ($n_{\min}$=9) + KF	78.89	80.37	82.20	75.72	91.65	88.84	90.22	0.117	0.147
Tracking-by-detection approaches									
MARS [56] + KF	44.59	40.08	54.32	36.60	72.13	68.74	70.39	0.352	0.409
PointNet [52] + KF	81.66	80.97	84.21	79.19	85.80	97.84	91.43	0.154	0.199
PointNet++ [57] + KF	81.41	78.22	82.18	80.66	83.82	97.67	90.22	0.138	0.192
Our baseline [44] (SL=25 from paper) + KF	89.79	89.25	90.33	89.25	91.79	98.27	94.92	0.181	0.225
Our baseline [44] (SL=18, best) + KF	90.07	89.50	**90.46**	89.67	91.78	**98.44**	**95.00**	0.175	0.216
Our proposed tracking-by-detection approach	**91.72**	89.14	90.11	**93.35**	91.49	98.35	94.80	**0.104**	0.141
Tracking-by-attention approaches									
TrackFormer [36]	64.14	48.63	56.75	72.50	84.18	63.53	72.41	0.127	0.156
MOTR [37] (trained with SL=10)	89.88	86.73	88.21	91.59	90.43	97.30	93.74	0.128	0.154
Our proposed tracking-by-attention approach	91.25	**89.75**	90.40	92.02	**92.63**	97.51	**95.00**	0.113	**0.138**

the benefit of incorporating temporal information. As shown in Fig. 12, extending the sequence length substantially improves both HOTA and MOTA, confirming that exploiting multiple consecutive frames is considerably more effective than processing each frame independently. The performance increases rapidly for short sequences before gradually saturating beyond approximately 8–10 frames, indicating that most relevant temporal information has already been captured. A similar trend can be observed for localization accuracy in Fig. 13. Increasing the sequence length progressively reduces both the MAE and RMSE, although the improvement continues over slightly longer sequences before converging. Based on these results, a sequence length of 14 frames was selected for the final model.

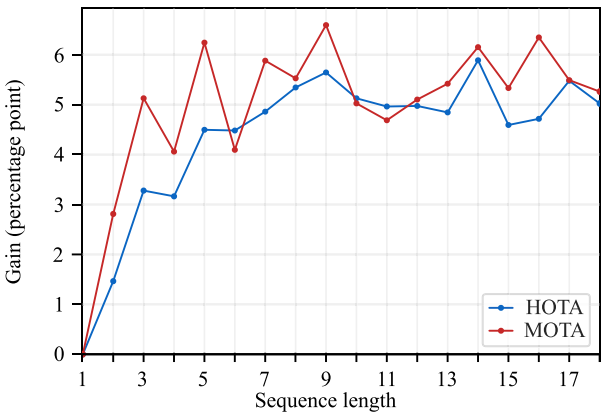


FIGURE 12. Influence of the Transformer sequence length *SL* on the tracking performance of the proposed tracking-by-detection architecture. Performance is reported as the gain relative to the single-frame baseline.

The corresponding ablation study for the proposed tracking-by-attention architecture is presented in Fig. 14 and Fig. 15. The reported results correspond to the mean over three independent training runs with different random seeds. In contrast to the tracking-by-detection approach, the sequence length only affects the training procedure by defining the temporal context available to the network during optimization, whereas the memory size N_{memory} determines how many previous frames are retained and influence the prediction of the current frame during inference. As shown in Fig. 14, increasing the training sequence length improves the HOTA score compared with short sequences, after which the performance gradually saturates. A similar trend is observed in Fig. 15, where longer training sequences reduce the localization error until the improvement levels off. The results further indicate that the frame memory primarily acts as a robustness factor. Very small memory sizes ($N_{\text{memory}} = 1$ or $N_{\text{memory}} = 2$) can reach competitive performance in individual runs, but exhibit substantially larger variance and occasional performance collapses caused by false-positive tracks. Increasing the memory size to moderate values stabilizes tracking, while additional memory yields smaller gains once sufficient temporal context is available. The best overall performance is obtained when the memory size is matched to the sequence length, with the strongest configuration using both a sequence length and memory size of 10. Overall, these results indicate that the two parameters should be chosen jointly: sufficiently long training sequences are essential for learning temporal dependencies, whereas an adequately sized memory is required to robustly exploit temporal information during inference.

Table 3 summarizes the influence of additional architectural and training design choices for both proposed approaches. For the tracking-by-detection architecture, even the single-frame variants ($SL = 1$) outperform the PointNet- and PointNet++-based baselines reported in Table 2, demonstrating the effectiveness of the proposed architecture independently of temporal modeling. As discussed previously, exploiting temporal information provides an additional performance gain. Replacing the proposed factorized Transformer encoder with a conventional Transformer encoder does not improve the overall tracking performance

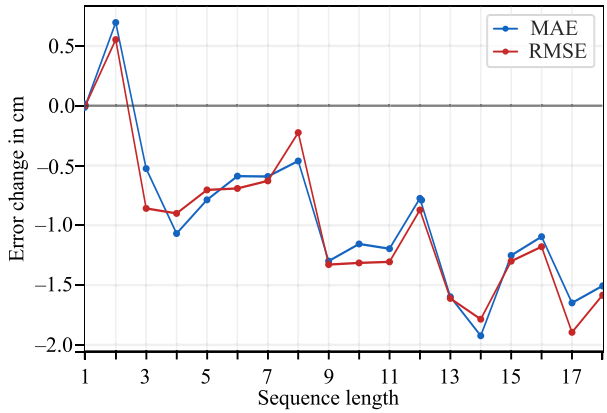


FIGURE 13. Influence of the Transformer sequence length *SL* on the localization accuracy of the proposed tracking-by-detection architecture. Error values are reported relative to the single-frame baseline.

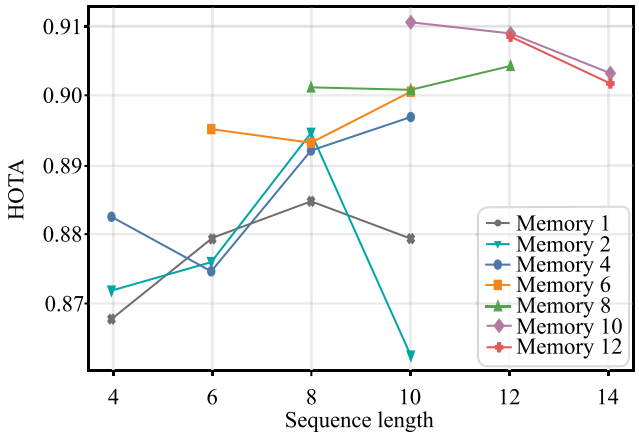


FIGURE 14. Tracking performance (HOTA) of the proposed tracking-by-attention architecture for different sequence lengths *SL* and memory sizes N_{memory}.

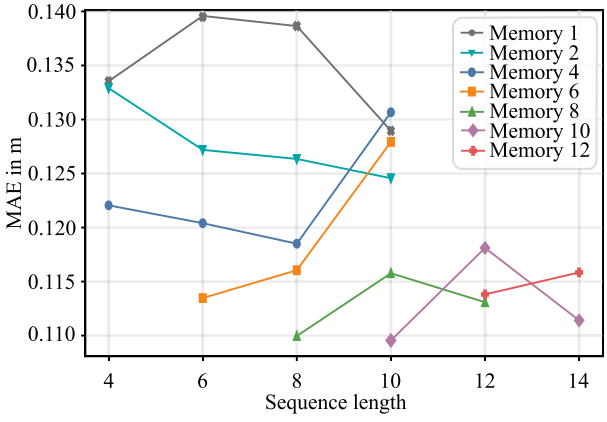


FIGURE 15. Localization performance (MAE) of the proposed tracking-by-attention architecture for different sequence lengths *SL* and memory sizes N_{memory}.

but substantially degrades the localization accuracy while increasing the computational complexity, motivating the proposed factorized design. Furthermore, temporal frame sampling [37], [42], in which consecutive training frames are randomly sampled with a temporal stride between one and three frames, and input point dropout as a robustness-oriented data augmentation strategy to simulate degraded radar measurements have only a minor influence on the final performance. A moderate input point dropout rate of 10 % yields the best overall results, whereas stronger dropout provides no further benefit. Initializing the object queries with DBSCAN-based hints, as suggested in [42], does not improve the overall performance. Although precision slightly increases, the DBSCAN-hint variant yields lower tracking, detection, and localization performance.

D. COMPUTATIONAL PERFORMANCE

The computational performance of the evaluated models is summarized in Table 4. Inference time was measured on an NVIDIA GeForce RTX 3080 in evaluation mode with a batch size of one. After 50 warm-up iterations, the execution time was averaged over five runs of 300 forward passes each. The reported runtime corresponds to a single sliding-window update ($SL = 14$) for the proposed tracking-by-detection approach and a single streaming update ($SL = 10$, memory size $= 10$) for the proposed tracking-by-attention approach. The proposed tracking-by-detection and tracking-by-attention architectures exhibit comparable model complexity and inference time, enabling real-time operation. The proposed baseline contains slightly more parameters but requires the fewest FLOPs. Replacing the factorized Transformer encoder with a conventional Transformer encoder increases the parameter count by more than a factor of six, as its size scales with the product of the number of slots and the embedding dimension, while yielding only comparable performance.

IX. DISCUSSION

The results demonstrate that modern deep multi-object tracking paradigms are not limited to dense image data but can also be successfully applied to sparse millimeter-wave radar target lists. This is remarkable because radar measurements differ fundamentally from images in terms of sparsity, measurement uncertainty, and the absence of appearance information. Nevertheless, both proposed architectures consistently outperform the conventional DBSCAN–Kalman-filter pipeline, demonstrating that learned object-centric representations are more effective than handcrafted clustering and association rules.

The comparison with the evaluated deep-learning baselines indicates that the network architecture must be adapted to sparse radar target lists. Unlike image-based MOT, where convolutional feature maps preserve spatial structure and object hypotheses emerge as localized feature activations, radar target lists consist of unordered point-wise measurements without an underlying spatial feature grid. Consequently, object evidence cannot be inferred from activation locations but must instead be encoded in the learned point representations themselves. The network therefore has to implicitly learn object formation by grouping sparse radar reflections, suppressing clutter and multipath,

TABLE 3. Results of the ablation study for the proposed tracking-by-detection and tracking-by-attention architectures. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance. TE: Transformer Encoder, *SL*: sequence length, KF: Kalman filter.

	Tracking in %				Detection in %			Localization in m	
	HOTA ↑	MOTA ↑	DetA ↑	AssA ↑	Precision ↑	Recall ↑	F1 ↑	MAE ↓	RMSE ↓
Proposed tracking-by-detection approach									
Ours with factorized TE layers and sequence length SL=1	85.68	83.56	85.89	85.46	87.65	97.72	92.41	0.124	0.163
Ours w/o factorized TE layers and sequence length SL=1	85.69	83.87	85.97	85.20	88.56	96.72	92.46	0.110	0.147
Ours with a conventional TE layer	90.95	87.96	89.23	92.72	90.26	98.73	94.30	0.181	0.225
Data augmentation with temporal frame sampling	90.40	87.65	88.95	91.88	90.08	98.61	94.15	0.124	0.161
Input point dropout (10 %)	91.20	89.76	90.65	91.76	91.92	98.47	95.09	0.102	0.138
Input point dropout (20 %)	90.76	87.81	89.07	92.48	90.28	98.52	94.22	0.105	0.140
Input point dropout (50 %)	90.40	87.65	88.95	91.88	90.08	98.61	93.15	0.124	0.161
Proposed tracking-by-attention approach									
Our approach with DBSCAN hints for object queries	90.85	88.64	89.86	91.84	91.72	97.78	94.66	0.122	0.145
Data augmentation with temporal frame sampling	90.29	88.29	89.24	91.35	91.69	97.09	94.31	0.102	0.131
Input point dropout (10 %)	90.89	88.72	89.72	92.09	92.11	97.18	94.58	0.108	0.131
Input point dropout (20 %)	90.44	88.24	89.22	91.68	91.64	97.12	94.30	0.117	0.143
Input point dropout (50 %)	90.87	89.52	90.29	91.45	92.51	97.42	94.90	0.115	0.143

TABLE 4. Computational performance of the evaluated architectures.

Model	Params (M)	FLOPs (G)	Time (ms)
Proposed tracking-by-detection approach	9.93	3.63	6.44
Proposed tracking-by-attention approach	10.78	1.79	5.78
Our baseline ($T = 25$)	14.52	0.66	7.41
Our baseline ($T = 18$)	14.52	0.48	6.47
Tracking-by-detection (ours) but with a conventional Transformer encoder layer	61.37	5.08	6.95

and extracting object-centric representations before localization and tracking become possible. This explains why MARS performs poorly, whereas PointNet and PointNet++ benefit from permutation-invariant processing. The proposed Slot Attention architecture further improves performance by replacing handcrafted clustering with learned object-centric representations.

The small performance differences between tracking-by-detection and tracking-by-attention indicate that both paradigms are well suited for radar-based MOT. Rather than identifying a universally superior approach, the results reveal complementary strengths. Tracking-by-detection provides a modular processing pipeline that facilitates independent optimization of localization and tracking, whereas tracking-by-attention integrates both tasks into a unified streaming architecture without requiring explicit post-processing.

A fundamental difference to vision-based MOT is the absence of discriminative appearance information. Radar target lists provide almost no identity-specific cues such as clothing, color, or texture. Consequently, maintaining temporal continuity becomes substantially more important than in vision-based MOT, as reliable re-identification after track loss is difficult. Future work could address this limitation by incorporating micro-Doppler or spectrogram-based identity features analogous to the appearance embeddings employed by DeepSORT.

The ablation studies further show that temporal information is one of the most important cues for radar-based tracking. Longer temporal contexts consistently improve both tracking and localization performance, while performance saturates for moderate sequence lengths and memory sizes. At the employed frame rate of 15 Hz, approximately one second of motion history appears sufficient to resolve most ambiguities. The explicit feature memory further improves the tracking-by-attention architecture by preserving recent radar observations that cannot be fully retained in propagated track queries alone.

TrackFormer transferred considerably less successfully to the proposed radar setting than MOTR. Our experiments indicate that its propagated-query formulation is more sensitive to the train–test mismatch introduced by closed-loop query propagation, leading to a stronger trade-off between recall and identity stability. In contrast, the MOTR query interaction mechanism proved substantially more robust under sparse radar observations.

The present dataset is limited to two simultaneously present persons and mainly contains trajectories within predefined indoor areas. Consequently, scalability to denser multi-person scenarios, frequent room entries and exits, and larger environments remains to be investigated. Extending the dataset is expected to further improve the proposed deep-learning approaches, as they can learn richer spatial-temporal representations from more diverse data. In contrast, deterministic methods such as DBSCAN are governed by fixed density thresholds and cannot systematically benefit from additional training data. At the same time, classical clustering approaches retain practical advantages, as they require no training data, are directly applicable to previously unseen environments, and are not inherently limited to the spatial extent represented in the training set. In particular, DBSCAN can naturally process detections beyond the approximately 4.9 m range covered by the present dataset, whereas the generalization of learned models outside the training distribution cannot be guaranteed. Overall, the presented results suggest that future radar-based multi-object tracking systems should increasingly adopt learned object-centric representations and temporal modeling instead of handcrafted clustering and association rules.

X. CONCLUSION

In this work, we investigated the applicability of modern deep multi-object tracking paradigms to sparse millimeter-wave radar target-list data. To this end, tracking-by-detection and tracking-by-attention architectures were proposed and evaluated on a large-scale dataset comprising more than 405,000 annotated radar frames. The results demonstrate that deep learning-based approaches consistently outperform a conventional DBSCAN- and Kalman filter-based tracking pipeline in terms of detection, tracking, and localization performance. Furthermore, the comparison of both paradigms indicates that explicitly exploiting temporal context is highly beneficial for radar-based tracking, where object identities must be inferred primarily from spatial and motion cues rather than appearance information. Overall, the presented results demonstrate that modern multi-object tracking paradigms originally developed for computer vision can be successfully transferred to sparse radar target-list data, establishing deep end-to-end tracking as a promising alternative to conventional radar tracking pipelines.

ACKNOWLEDGMENT

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).

REFERENCES

[1] V. B. Vales, T. Domínguez-Bolaño, C. J. Escudero, and J. A. García-Naya, "An IoT system for smart building combining multiple mmWave FMCW radars applied to people counting," *IEEE Internet Things J.*, vol. 11, no. 21, pp. 35306–35316, Nov. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10612821/

[2] P. Nallabolu, L. Zhang, H. Hong, and C. Li, "Human presence sensing and gesture recognition for smart home applications with moving and stationary clutter suppression using a 60-GHz digital beamforming FMCW radar," *IEEE Access*, vol. 9, pp. 72857–72866, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9430867/

[3] 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. (BHI)*, Nov. 2024, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/10913821/?arnumber=10913821&tag=1

[4] 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. (WACV)*. Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 5704–5713. [Online]. Available: https://ieeexplore.ieee.org/document/10030786/

[5] A.-C. Froehlich, D. Mejdani, L. Engel, J. Braeunig, C. Kammel, M. Vossiek, and I. Ullmann, "A millimeter-wave MIMO radar network for human activity recognition and fall detection," in *Proc. IEEE Radar Conf. (RadarConf24)*, Denver, CO, USA, May 2024, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10548702/

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

[7] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "RF-based 3D skeletons," in *Proc. Conf. ACM Special Interest Group Data Commun.* Budapest Hungary: ACM, Aug. 2018, pp. 267–281. [Online]. Available: https://dl.acm.org/doi/10.1145/3230543.3230579

[8] 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. Vehicles Symp. (IV)*, Jun. 2024, pp. 2616–2621. [Online]. Available: https://ieeexplore.ieee.org/document/10588669/

[9] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "MmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave," in *Proc. 19th Annu. Int. Conf. Mobile Syst., Appl., Services*. New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 269–282. [Online]. Available: https://dl.acm.org/doi/10.1145/3458864.3467679

[10] L. Engel, J. Mueller, E. J. F. Rendon, E. Dorschky, D. Krauss, I. Ullmann, B. M. Eskofier, and M. Vossiek, "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. Microw.*, vol. 5, no. 2, pp. 373–387, Mar. 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10904474/?arnumber=10904474

[11] J. L. Mueller, L. Engel, E. Dorschky, D. Krauss, I. Ullmann, M. Vossiek, and B. M. Eskofier, "RadProPoser: Probabilistic radar tensor human pose estimation that knows its limits," 2025, *arXiv:2508.03578*.

[12] J. Mueller, L. Engel, D. Krauss, I. Ullmann, M. Vossiek, B. M. Eskofier, and E. Dorschky, "End-to-end learning for human pose estimation from raw millimeter wave radar data," in *Proc. 58th Asilomar Conf. Signals, Syst., Comput.*, Grove, CA, USA, Oct. 2024, pp. 1495–1499.

[13] Y.-H. Ho, J.-H. Cheng, S. Y. Kuan, Z. Jiang, W. Chai, H.-W. Huang, C.-L. Lin, and J.-N. Hwang, "RT-Pose: A 4D radar tensor-based 3D human pose estimation and localization benchmark," in *Proc. Computer Vision ECCV*, in lecture notes in computer science, vol. 15121, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham, Switzerland: Springer, 2025, pp. 107–125. [Online]. Available: https://link.springer.com/10.1007/978-3-031-73036-87

[14] M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7356–7365. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8578866

[15] 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 (SAM)*, Trondheim, Norway, Jun. 2022, pp. 345–349. [Online]. Available: https://ieeexplore.ieee.org/document/9827886/

[16] 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. [Online]. Available: https://ieeexplore.ieee.org/document/10546280/

[17] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward unobtrusive in-home gait analysis based on radar micro-Doppler signatures," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 9, pp. 2629–2640, Sep. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8613848/

[18] A. D. Singh, Y. Ba, A. Sarker, H. Zhang, A. Kadambi, S. Soatto, M. Srivastava, and A. Wong, "Depth estimation from camera image and mmWave radar point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 9275–9285. [Online]. Available: https://ieeexplore.ieee.org/document/10204897/

[19] 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.[Online]. Available: https://ieeexplore.ieee.org/document/8355715

[20] P. Tschapek, G. Körner, C. Carlowitz, and M. Vossiek, "Detailed analysis and modeling of phase noise and systematic phase distortions in FMCW radar systems," *IEEE J. Microw.*, vol. 2, no. 4, pp. 648–659, Oct. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9875949/

[21] N. Knudde, B. Vandersmissen, K. Parashar, I. Couckuyt, A. Jalalvand, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor tracking of multiple persons with a 77 GHz MIMO FMCW radar," in *Proc. Eur. Radar Conf. (EURAD)*, Oct. 2017, pp. 61–64. [Online]. Available: https://ieeexplore.ieee.org/document/8249147/

[22] M. Jiang, S. Guo, H. Luo, and G. Cui, "Continuous tracking of indoor human targets based on millimeter wave radar," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Chiang Mai, Thailand, Nov. 2022, pp. 2071–2076. [Online]. Available: https://ieeexplore.ieee.org/document/9979904/

[23] X. Huang, H. Cheena, A. Thomas, and J. K. P. Tsoi, "Indoor detection and tracking of people using mmWave sensor," *J. Sensors*, vol. 2021, no. 1, pp. 1–14, Jan. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1155/2021/6657709

[24] M. Ester, H.-P. Kriegel, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Knowl. Discovery Databases (KDD)*, vol. 96, no. 34, pp. 226–231, 1996.

[25] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/nav.3800020109

[26] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960. [Online]. Available: https://asmedigitalcollection.asme.org/fluidsengineering/article/82/1/35/397706/A-New-Approach-to-Linear-Filtering-and-Prediction

[27] J. Pegoraro and M. Rossi, "Real-time people tracking and identification from sparse mm-Wave radar point-clouds," *IEEE Access*, vol. 9, pp. 78504–78520, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9440989/

[28] 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. (RadarConf20)*, Sep. 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9266400/

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 779–788. [Online]. Available: http://ieeexplore.ieee.org/document/7780460/

[30] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448. [Online]. Available: http://ieeexplore.ieee.org/document/7410526/

[31] 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. [Online]. Available: http://ieeexplore.ieee.org/document/7485869/

[32] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Comput. Vision ECCV*, 2020, pp. 213–229. [Online]. Available: https://link.springer.com/10.1007/978-3-030-58452-813

[33] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468. [Online]. Available: https://ieeexplore.ieee.org/document/7533003/

[34] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649. [Online]. Available: http://ieeexplore.ieee.org/document/8296962/

[35] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," in *Proc. Comput. Vision ECCV*, Switzerland, 2022, pp. 1–21, doi: 10.1007/978-3-031-20047-2_1.

[36] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: Multi-object tracking with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8834–8844. [Online]. Available: https://ieeexplore.ieee.org/document/9879668/

[37] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," 2021, *arXiv:2105.03247*.

[38] Y. Zhang, T. Wang, and X. Zhang, "MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 22056–22065. [Online]. Available: https://ieeexplore.ieee.org/document/10204828/

[39] E. Yu, T. Wang, Z. Li, Y. Zhang, X. Zhang, and W. Tao, "MOTRv3: Release-fetch supervision for end-to-end multi-object tracking," 2023, *arXiv:2305.14298*.

[40] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: Multiple object tracking with transformer," 2020, *arXiv:2012.15460*.

[41] K. Fatseas and M. J. G. Bekooij, "Neural network based multiple object tracking for automotive FMCW radar," in *Proc. Int. Radar Conf. (RADAR)*, Toulon, France, Sep. 2019, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9078996/

[42] M. Dell, W. Bradfisch, S. Schober, and C. Klöck, "RadarMOTR: Multi-object tracking with transformers on range-Doppler maps," in *Proc. Int. Radar Conf. (RADAR)*, Rennes, France, Oct. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10994166/

[43] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 11525–11538. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html

[44] L. Engel, M. Bergmann, C. Kammel, I. Ullmann, B. M. Eskofier, and M. Vossiek, "Fully automated, AI-driven multimodal annotation framework for human centric radar applied to deep learning-based people localization," *IEEE J. Microw.*, vol. 6, no. 2, pp. 376–390, Mar. 2025.

[45] *Single-Chip 60- To 64-GHz MmWave Sensor Antennas-On-Package (AOP)*, document IWR6843AOP, Texas Instruments, Jun. 2025.

[46] H. Rohling, "Ordered statistic CFAR technique–an overview," in *Proc. 12th Int. Radar Symp. (IRS)*, Sep. 2011, pp. 631–638. [Online]. Available: https://ieeexplore.ieee.org/document/6042197/?arnumber=6042197

[47] J. A. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines," *Astron. Astrophys. Suppl. Ser.*, vol. 15, pp. 417–426, Jun. 1974. [Online]. Available: https://ui.adsabs.harvard.edu/abs/1974A&AS.15.417H

[48] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, May 2008. [Online]. Available: http://jivp.eurasipjournals.com/content/2008/1/246309

[49] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "HOTA: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 548–578, Feb. 2021. [Online]. Available: https://link.springer.com/10.1007/s11263-020-01375-2

[50] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "MID: Tracking and identifying people with millimeter wave radar," in *Proc. 15th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, May 2019, pp. 33–40, [Online]. Available: https://ieeexplore.ieee.org/document/8804831/

[51] J. Pegoraro, F. Meneghello, and M. Rossi, "Multiperson continuous tracking and identification from mm-wave micro-Doppler signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 2994–3009, Apr. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9194046/

[52] 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. (CVPR)*, Jul. 2017, pp. 77–85. [Online]. Available: http://ieeexplore.ieee.org/document/8099499/

[53] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6410–6419. [Online]. Available: https://ieeexplore.ieee.org/document/9010002/

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Information Processing Systems*, 2026, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[55] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[56] S. An and U. Y. Ogras, "MARS: MmWave-based assistive rehabilitation system for smart healthcare," *ACM Trans. Embedded Comput. Syst.*, vol. 20, no. 5s, pp. 1–22, Oct. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3477003

[57] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*.



LUKAS ENGEL (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 pursuing 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 (FAU), 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, and 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 focused on radar imaging and signal processing in the space sector, and radar-based motion tracking in medical applications. He received the Semikron Promotion Prize, in 2016, the Rohde and Schwarz Prize, in 2018, the Baumüller Master's 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 Postdoctoral Researcher and the Head of the Research Group "Wave-Based Sensing, Signal Processing, and AI," Institute of Microwaves and Photonics, FAU. In 2022, she spent one month as a Visiting Researcher with the Microwave Sensing, Signals and Systems Group, Delft University of Technology, Delft, The Netherlands. Her research interests include radar and ultrasound imaging and AI-based smart sensing for nondestructive testing, security, geophysics, and medical applications. She received the EuRAD Conference Prize, in 2019 and the IEEE SENSORS Best Paper Award, in 2023. She serves as a Reviewer for the European Radar Conference (EuRAD), the IEEE Radar Conference, and various journals in the field of microwaves and sensing. Since 2022, she has been an Associate Editor of IEEE Transactions on Radar Systems.



BJOERN M. ESKOFIER (Senior Member, IEEE) received the Ph.D. degree in biomechanics from the University of Calgary, Canada, under the supervision of Prof. Dr. Benno Nigg. He heads the Machine Learning and Data Analytics (MaD) Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany. He is also the Founding Spokesperson of the Department of Artificial Intelligence in Biomedical Engineering, FAU, the German Ministry of Economic Affairs and Climate Action GAIA-X use case project TEAM-X, and the Co-Spokesperson of the German Research Foundation Collaborative Research Center EmpkinS (www.empkins.de). Since April 2023, he has been an Associate Principal Investigator and a Leader of the Research Group Translational Digital Health, Helmholtz Zentrum Munich. From April to August 2023, he was a Visiting Professor with Prof. Scott Delps NMBL Laboratory that is part of Stanford University's Schools of Engineering and Medicine. He studied electrical engineering with FAU and graduated, in 2006. 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 a recipient of several medical-technical research awards, including the Curious Minds Award 2021 in Life Sciences by Manager Magazin and Merck. In 2016, he was a Visiting Professor with the Motion Analysis Laboratory, Harvard Medical School, from February to March, with Prof. Paolo Bonato. In 2018, he was a Visiting Professor with, the Human Dynamics Group, MIT Media Laboratory, from March to August, with Prof. Alex Sandy Pentland. In addition, he was the Area Editor for IEEE Open Access Journal of Engineering in Medicine and Biology and Associate Editor for IEEE Journal of Biomedical and Health Informatics. He is also active in the organization of several IEEE and ACM meetings (e.g., BSN, BHI, EMBC, IJCAI, ISWC, and UbiComp), most recently serving as the 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 (FAU), Erlangen, Germany. He has authored or co-authored more than 400 publications. His research has led to more than 100 granted patents. His current research interests include radar, microwave systems, wave-based imaging, transponders, RF identification, communications, and 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. He 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 also been serving as the Spokesperson for the DFG Review Board 4.42 Electrical Engineering and Information Technology, since April 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 and Localization. He has received 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. Since October 2022, he has been an Associate Editor-in-Chief for IEEE Transactions on Radar Systems. From 2013 to 2019, he was an Associate Editor for IEEE Transactions on Microwave Theory and Techniques.