IEEE Journal of Microwaves

Received 30 October 2024; revised 1 December 2024; accepted 18 January 2025. Date of publication 26 February 2025; date of current version 14 March 2025.

Digital Object Identifier 10.1109/JMW.2025.3535525

Advanced Millimeter Wave Radar-Based Human Pose Estimation Enabled by a Deep Learning Neural Network Trained With Optical Motion Capture Ground Truth Data

LUKAS ENGEL [1] (Graduate Student Member, IEEE), JONAS MUELLER[2], EDUARDO JAVIER FERIA RENDON[2], EVA DORSCHKY [2] (Member, IEEE), DANIEL KRAUSS [2] (Graduate Student Member, IEEE), 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 by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1483 – under Project 442419336, EmpkinS.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Ethics Committee of Friedrich-Alexander-Universität Erlangen-Nürnberg under Application No. #22-437-B, and performed in line with the principles of the revised Declaration of the World Medical Association of Helsinki.

ABSTRACT This paper presents a deep learning-enabled method for human pose estimation using radar target lists, obtained through a low-cost radar system with three transmitters and four receivers in a multiple-input multiple-output setup. We address challenges in previous research that often relied on extracting ground truth poses from RGB data, which are constrained by the need for 3D mapping and vulnerability to occlusions. To overcome these limitations, we utilized optical motion capture, which is widely recognized as the gold standard for precise human motion analysis. We conducted an extensive optical motion capture study involving various recorded movement activities, which resulted in *mmRadPose*, a new dataset that enhances existing benchmarks for radar-based pose estimation. This dataset has been made publicly accessible. Building on this approach, we designed an application-tailored radar signal processing chain to generate suitable input for the machine learning algorithm. We further developed an attentional recurrent-based deep learning model, *PntPoseAT*, which predicts 24 keypoints of human poses using radar target lists. We employed cross validation to thoroughly evaluate the model. This model surpasses previous approaches and achieves an average mean per-joint position error of 6.49 cm with a standard deviation of 3.74 cm on totally unseen test data. This excellent accuracy of the reconstructed keypoint positions is particularly remarkable when you consider that a very simple radar was used for the measurements. Additionally, we conducted a comprehensive analysis of the model's performance by exploring aspects such as network architecture, the use of long short-term memory versus gated recurrent units, input data selection, and the integration of multi-head self-attention mechanisms.

INDEX TERMS Human pose estimation, radar, optical motion capture, millimeter wave, target list, point cloud, machine learning, deep learning.

I. INTRODUCTION

Human pose estimation is a highly engaging field [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42] with a wide range of applications across multiple domains. In healthcare and rehabilitation [15], [19], [20], [21], [22], [23], [24], [25], [26], [27], it plays a crucial role in monitoring patients' movements during recovery by ensuring that exercises are performed correctly to optimize outcomes. In sports and fitness, human pose estimation enables quantitative performance analysis, which helps athletes and trainers optimize their routines while minimizing injury risks [28], [29], [30], [31], [32], [33]. By analyzing movement patterns, coaches and athletes can make informed decisions to sustainably improve performance. In human-computer interaction [34], [35], [36], [37], [38], [39], [40], human pose estimation significantly enhances user interfaces by enabling gesture-based commands for more intuitive interactions. This opens up exciting possibilities for virtual and augmented reality experiences [37], [43] as well as innovations in smart home applications, gaming, and entertainment [41], [42].

State-of-the-art human pose estimation techniques primarily rely on RGB(-D) data [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], thereby leveraging advancements in deep learning and computer vision. These methods utilize sophisticated neural networks to analyze and interpret visual information to provide accurate and efficient pose detection.

However, despite their effectiveness and advancements, notable limitations remain. First, they rely on visible light, which makes them sensitive to changes in lighting conditions, such as poor illumination, bright sunlight, or reflective surfaces. Second, privacy concerns arise when visual images are captured in sensitive environments. Third, occlusions present a challenge as RGB-D sensors cannot penetrate through materials, which limits their usability when parts of the body are obscured. Radar technology overcomes some of these shortcomings and makes it a promising alternative as it is unaffected by lighting conditions [17], [19], [22], [44] and poses fewer privacy concerns [22], [45], [46], [47]. Being dependent on the utilized frequency range, radar can even penetrate certain materials, such as clothing, foliage, or thin walls [22], [44], [48], [49], [50]. By exploiting multipath effects, it is possible to achieve a robust human pose estimation, even with partially occluded individuals.

The literature that investigates radar-based human pose estimation offers several approaches. A summary of the key approaches is presented in Table 1.

Zhao et al. [12] employed a multiple-input multiple-output (MIMO) radar to predict 14 keypoints in a multi-person scenario using convolutional neural network (CNN)-driven feature extraction. Their method consisted of two key components: one that detected the presence of a person in the region and another that estimated the corresponding pose. To generate a ground-truth label for the keypoints, they utilized a multi-view camera setup combined with a deep learning-enabled 2D image-based pose extraction. The recorded dataset consisted of 16 hours of data from office and indoor environments with people moving randomly. However, the hardware configuration with an SIMO array was elaborate and quite bulky in the 7 GHz frequency range.

In [13], a target list-based approach to human pose estimation was introduced. The authors transformed target-list information into the channels of an image and applied a CNN. However, the dataset was relatively small, and two separate orthogonal radars are required to be able to measure both solid angles, which makes the setup more complex.

An et al. [14] proposed an mmWave-based assistive rehabilitation system for smart healthcare that processes radar point clouds from a MIMO radar system to provide feedback to the user about the correctness of the movement. Their deep learning model was a CNN-based approach, and an RGB-D camera was utilized to generate ground truth. The dataset comprised rehabilitation movements, such as limb extensions, squats, and lunges. However, the point cloud data was further processed and tailored to the specific approach and neglected additional information or disturbance. Based on this work, the same authors introduced in [15] a new, extended dataset with an increased number of synchronized radar and RGB frames as well as synchronized IMU data.

Yu et al. [16] suggested RFPose-OT. In contrast to previous work, the model was based on an optimal-transport-theory approach, which seeks the most cost-effective way to transform one probability distribution into another. They used two orthogonally mounted radars with a performant MIMO array setup of with 4 transmitting and 16 receiving antennas, which enabled a good spatial resolution but at the same time generated a high data load. They recorded data in an office and an outdoor environment. People moved randomly without instructed actions. In terms of overall precision, it outperformed the model shown in [12].

Lee et al. [19] created a dataset for human pose estimation using two orthogonal radars to receive important information in both the elevation and azimuth dimensions. A cost-effective and simple radar sensor with three transmitters and four receivers was utilized. Simple activity movements were performed in front of the radars, such as walking, standing, and waving hands. However, unlike the other works, they only proposed a 2D human pose estimation. As a reference, they extracted 2D keypoints of human poses from RGB images using the HR-Net [52]. The incorporation of an additional neural network, for instance, such as VideoPose3D [51], is necessary to map 2D information into 3D space, which inherently improves prediction accuracy but consequently reduces comparability.

Ho et al. [17] added another sensor modality to the measurement setupa and exploited lidar data in addition to RGB data to accurately annotate the ground-truth labels. A high-performance state-of-the-art MIMO radar consisting of 12 transmitter and 16 receiver channels was used to record a dataset with indoor and outdoor environments, where simple movements were performed. With a CNN-based backbone,

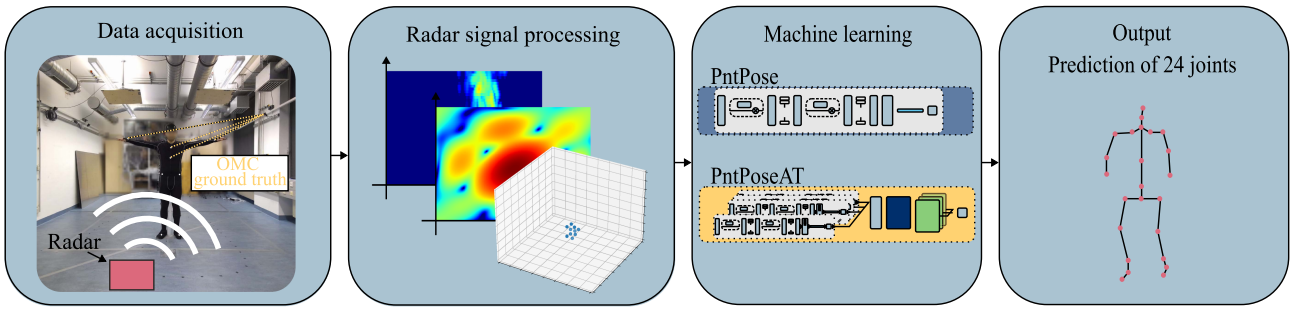

FIGURE 1. Paper content overview: We conducted a motion study with healthy participants performing various activities in front of a radar system using optical motion capture (OMC) markers for ground-truth validation. We developed a radar processing chain to convert raw radar data into target lists containing seven features. Finally, we designed and trained a neural network that used these target lists as input to estimate 24 human skeleton keypoints.

TABLE 1. Overview of Related Work: A Comparison of Relevant Aspects of Previous Work in Human Pose Estimation

Work	FPS[1]	# total frames	# data channels (Tx x Rx channels)	# joints	Ground truth	Input type	Avg. MPJPE in cm
RF-Based 3D Skeleton [12]	30	1,694,440	n.r.	14	RGB (multi view[2])	Heatmap	18.43
mmPose [13]	20	39,700	2 orthogonal radars (each 2 x 4, linear)	17 (25)	RGB-D	Target list[5]	n.r.
MARS [14]	10	40,083	3 x 4 (limited field of view[6])	19	RGB (multi view[2])	Target list	n.r.
mRI [15]	10	160,000	3 x 4 (limited field of view[6])	15	RGB-D, IMU	Target list	9.40
RFPose-OT [16]	20	89,090	2 orthogonal radars (each 12 x 16)	n.r.	RGB (multi view[2])	Heatmap	8.68
HuPR [19]	20	141,000	2 orthogonal radars (each 3 x 4)	14	RGB	Heatmap	6.82[3]
RT-Pose [17]	10	72,000	12 x 16	15	RGB, lidar	Heatmap	9.93
mmMesh [18]	10	480,000	3 x 4 (limited field of view[6])	Mesh	OMC	Target list	2.18[4]
PntPoseAT (Ours)	**15**	**203,149**	**3 x 4**	**24**	**OMC**	**Target list**	**6.49**

[1]from radar system, [2]multiple cameras with different angles toward the scene, [3]performance lifted by VideoPose3D [51], [4]correlated data from test set also included in training set, [5]transformed to images, [6]limited due to antenna beam width; #: number of, n.r.: not reported, FPS: frames per second, Tx: transmitter, Rx: receiver, MPJPE: mean per-joint position error, OMC: optical motion capture.

the deep learning approach generates a body center probability map together with a keypoint offset map. For comparison, the authors trained models from previous human pose estimation work in [13], [18], and [12] with their dataset. They showed a significant improvement in terms of mean per-joint position error (MPJPE) in their work. However, the annotation process was partially hand-crafted and, therefore, very time-consuming.

The aforementioned works primarily relied on ground-truth labeling from RGB images. Despite the ongoing advancements in image-based 2D human pose estimation models, uncertainty persists, especially when projecting 2D poses into 3D space. To overcome this limitation, optical motion capture (OMC) systems can be employed to generate precise ground-truth data for training, which offers several key advantages over traditional RGB-based methods. Unlike RGB systems, OMC inherently provides position data in a 3D coordinate system, thereby eliminating the need for conversion from 2D images to 3D coordinates, which reduces mapping inaccuracy through further processing. Additionally, OMC excels in particular in handling occlusions. It can infer the positions of hidden body parts based on their last known locations and the positions of visible markers, whereas RGB systems struggle with this operation, which often results in lost or inaccurate tracking.

Therefore, Xue et al. [18] introduced a radar-based human pose estimation method based on OMC. In their work, they proposed a dynamic human mesh reconstruction method using mmWave radar combined with deep learning. The dataset included 20 participants performing eight common daily activities. However, the radar only provided a limited field of view due to the use of serial-fed patch antennas, and it suffered from significantly reduced resolution in one angular direction (only two antennas). This resulted in a decreased sensitivity to movements in the elevation direction, particularly when the movements occured near the radar. Furthermore, one significant limitation was that correlated data (data from the same participant) from the training set was also used in the test set, which significantly improved performance but deteriorated the generalizability of the network.

In this article, we propose *PntPoseAT*, a radar-based human pose estimation method that uses deep learning to predict skeleton keypoints, with an overview illustrated in Fig. 1. To enable this approach, we conducted a motion study using OMC with 12 participants performing 12 actions in three different angles toward the radar to form an extensive dataset

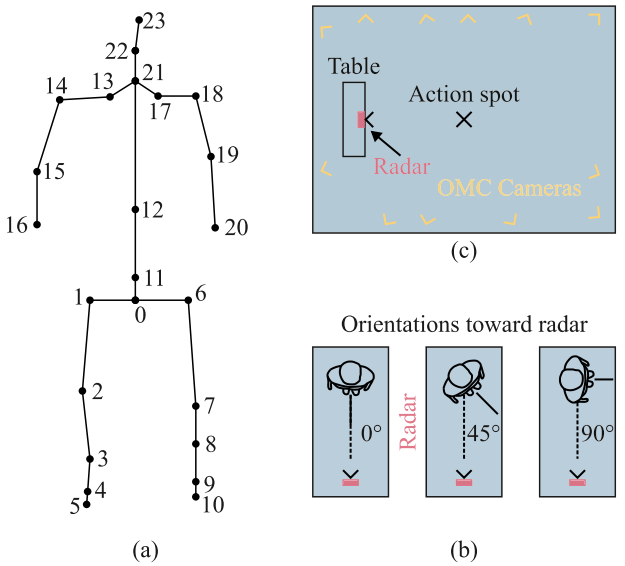

FIGURE 2. (a) Sketch of the applied skeleton pose: We predict a skeleton consisting of 24 joints. The denotation of the joints is shown in Table 2. (b) Layout of the room and sensor setup where the activities were performed. (c) Orientations toward radar: The activities were recorded from different orientations toward the radar: 0°, 45°, and 90°.

named *mmRadPose*. We designed the dataset so that typical human activity movements were performed with the limbs equally including movements of both the upper and lower body. We utilized a human skeleton model with the 24 most relevant keypoints as depicted in Fig. 2(a) and denoted in Table 2. Our approach leverages data from a state-of-the-art, commercially available MIMO radar to generate target lists (point clouds with further attributes, such as velocity, SNR, and intensity) which are then processed by a deep learning network. We employed a relatively simple radar system with a MIMO array with 3 transmit and 4 receive antennas to demonstrate the applicability and performance of low-cost radar hardware, thereby emphasising the potential of its scalability in the application. However, due to the sensor design, the radar provided a wide field of view and a resolution in both solid angles. This approach highlights the potential for widespread deployment across various sectors, including smart homes, healthcare, and industry.

We developed a deep learning model by adapting the PointNet [53] architecture to extract key features. We used a long short-term memory (LSTM) [54] architecture to effectively capture and utilize temporal long-range dependencies over consecutive radar frames in the sequential time data. Furthermore, the multi-head self-attention (MHSA) mechanism was applied, which is commonly used in transformer models [55] and allows the model to focus on different parts of an input sequence simultaneously and capture various relationships by applying multiple attention functions in parallel.

We compared and reported the MPJPE and standard deviation for these different deep learning architectures. A persistent challenge in human pose estimation using deep learning is the ambiguity in many papers regarding which data was used solely for training and which was reserved for testing, which resultes in a lack of comparability. To address this, we explicitly ensured that our models were tested on unseen data.

TABLE 2. Joint Number, Designation, and Abbreviations Referring to the Keypoints of the Skeleton in Fig. 2

Number	Name	Abbreviation
0	Hips	Hips
1	Right upper leg	RUL
2	Right leg	RL
3	Right foot	RF
4	Right toe base	RTB
5	Right toe base end	RTBE
6	Left upper leg	LUL
7	Left leg	LL
8	Left foot	LF
9	Left toe base	LTB
10	Left toe base end	LTBE
11	Spine	Spine
12	Spine upper	SpineU
13	Right shoulder	RS
14	Right arm	RA
15	Right forearm	RFA
16	Right hand	RH
17	Left shoulder	LS
18	Left arm	LA
19	Left forearm	LFA
20	Left hand	LH
21	Neck	Neck
22	Head	Head
23	Head end	HeadE

The remainder of this paper is structured as follows. In Section II the utilized sensor system setup is explained in detail. Section III describes the recorded dataset and explains the movements performed in the study. The radar signal process to yielded target lists used as inputs for deep learning is explained in Section IV. Our proposed deep learning model and results are presented in Section V. We then discuss these findings in Section VI and conclude our study in Section VII.

II. MEASUREMENT AND SENSOR SETUP

In this section, we outline the measurement and sensor setup used to collect the radar and OMC data. Measurements were carried out at the motion capture laboratory at the Institute of Microwaves and Photonics (LHFT), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The room's dimensions were 10.60 m long, 4.72 m wide and 3.27 m high. Our developed sensor system, "RadarBox", (depicted in Fig. 3) was positioned on a table at one end of the room, directed toward the center of the room. A picture of the room is shown in Fig. 4. An OMC system was set up in the room for accurate localization and tracking. To prevent interference with the

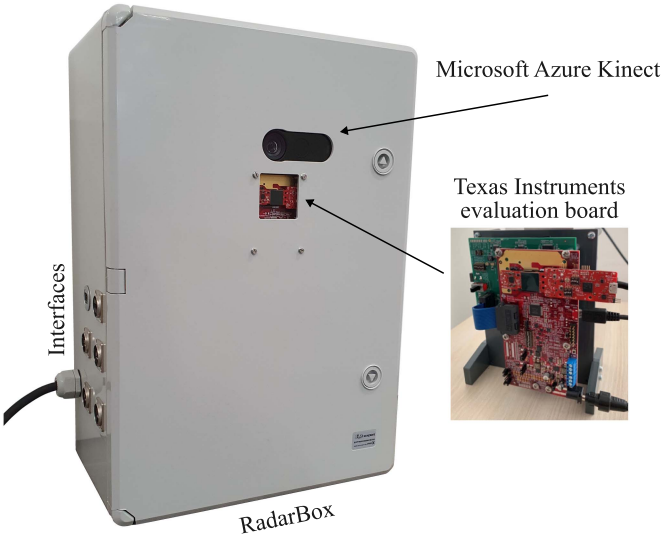

FIGURE 3. The study's measurement setup: Our "RadarBox" with an multipe-input multiple-output integrated frequency-modulated continuous-wave radar (Texas Instruments evaluation board) and RGB-D camera (Microsoft Azure Kinect). The box provides interfaces to synchronize with other sensor systems, such as optical motion capture. (The camera was not used for ground truth in this paper.).

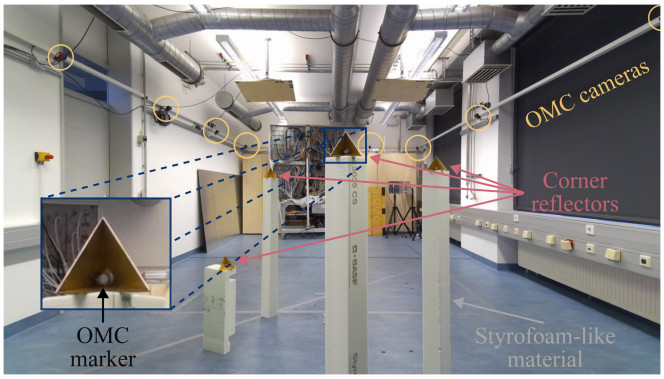

FIGURE 4. A photograph of the measurement setup shows corner reflectors with optical motion capture (OMC) markers attached positioned in front of the radar to align the coordinate systems. The OMC cameras, highlighted by yellow circles, track the markers' positions. For accurate registration of the radar and OMC coordinate systems, a common target must be detectable at the same location by both systems.

OMC system's infrared signals, a shutter on the windows was installed and closed during measurements.

A. OPTICAL MOTION CAPTURE

We utilized an OMC system to precisely determine the joint positions and the poses of the participants to be used as ground-truth data for the study. In this study, the Optitrack Flex 13[1] was utilized. The systems consisted of 12 infrared cameras distributed on the walls in the room enabled different camera perspectives to ensure tracking reliability. A photograph of the room showing the OMC cameras is shown in Fig. 4. The participants were instructed to wear a suit, hat, and shoes equipped with infrared markers sensitive to the OMC system. We used Optitrack's Motive:Body[2] software (version 2.3.1) to track these markers. Motive includes several predefined full-body marker sets that specify biological landmarks where the markers must be attached to the motion capture suit. In this study, we used the Conventional[3] full-body marker set, which consists of 39 markers. When the markers are correctly placed, Motive automatically generates a model-based skeleton. However, due to occlusion, manual hand-crafted post-labeling was required to ensure accurate pose reconstruction. To ensure time synchronization, the system was synchronized with the radar via a hardware input trigger with a frame rate of 15 Hz.

B. RADAR SYSTEM SETUP

The radar system employed during the measurement study is included in our "RadarBox" setup shown in Fig. 3. The RGB-D camera (Microsoft Azure Kinect) was not used in this research. The radar system consisted of the Texas Instruments IWR6843AOPEVM[4], which is an evaluation board with integrated antennas-on-package. It features four receiver antennas and three transmitter antennas arranged in an L-shaped virtual antenna array. The on-package patch antennas deliver a wide field of view in both azimuth and elevation, which results in extensive coverage area. Additionally, we used the MMWAVEICBOOST[5] and DCA1000EVM[6] evaluation boards from Texas Instruments to capture raw radar data. The system operates in a frequency-modulated continuous-wave (FMCW) mode using time-division multiplexing MIMO methods to separate the channels. We installed a hardware trigger output to enable time synchronization with other sensor systems. Table 3 summarizes the key radar system parameters that were utilized. We opted for a high update rate of 15 Hz to accurately track human body movements. Moreover, we chose a high number of chirps within a frame to decrease the Doppler resolution, which allowed for clearer separation in the range-Doppler domain between the Doppler components of the individual body limbs.

C. CALIBRATION OF COORDINATE SYSTEMS

To align the two sensors originating from the different coordinate systems, registration in a joint coordinate system (Cartesian coordinates) is required. First, each sensor system (OMC and radar) is individually calibrated within its own coordinate system space. Moreover, to register the two coordinate systems, a target is required that can be localized properly in both coordinate systems. The registration measurement setup is depicted in Fig. 4. To achieve this, four metallic corner reflectors were mounted on a styrofoam-like bar to maximize

[1] [Online]. Available: https://optitrack.com/cameras/flex-13

[2] [Online]. Available: https://optitrack.com/software/motive

[3] [Online]. Available: https://docs.optitrack.com/markersets/full-body/conventional-39

[4] [Online]. Available: https://ti.com/tool/IWR6843AOPEVM

[5] [Online]. Available: https://ti.com/tool/MMWAVEICBOOST

[6] [Online]. Available: https://ti.com/tool/DCA1000EVM

TABLE 3. Radar System Parameter Settings

Parameter	Value
Frequency	60 GHz
Radio-frequency bandwidth	≈ 1.02 GHz
Frame 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	±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}

ADC: analog-to-digital-converter, TDM: time-divison multiplexing, MIMO: multiple-input multiple-output.

reflected power at the corner reflector's phase center position and minimize clutter from other adjacent static reflectors. A tiny infrared-sensitive marker ball was afficed to the corner of the corner reflectors so that the phase reflection center of both coordinate systems matched. Finally, to spatially register the two systems, we formulated an optimization problem as follows [56]:

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

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 [56], [57], [58] to estimate the optimal rotation $\boldsymbol{R}_{\text{opt}}$ and translation $\boldsymbol{T}_{\text{opt}}$ parameters.

III. DATASET

An extensive study involving OMC and radar motion analysis was conducted. A total of 12 healthy participants (4 female and 8 male aged 28.92 ± 3.15 years with a height of 176.25 ± 6.25 cm and weighing 72.33 ± 12.37 kg) participated in the measurement activity. The study received approval from the ethics committee of Friedrich-Alexander-Universität Erlangen-Nürnberg (Protocol #22-437-B), and all participants provided written informed consent prior to participation.

The participants were instructed to perform 12 specific movements from three different angles toward the radar (see Fig. 2(b)). The test area, a 3.55 m × 3.05 m rectangular space, was marked in the center of the room. This area was within the radar's field of view, and the OMC cameras were strategically positioned to fully capture the movements. All recordings took place within this defined space (see Fig. 2(c) for a top-down view of the setup). All movements started from the initial position that is marked with an "X" (action spot) in Fig. 2(c).

The activity set comprised eleven distinct movements along with a static T-pose and are visualised in Fig. 5:

- T-pose
- Left upper limb extension
- Right upper limb extension
- Bilateral upper limb extension
- Bicep curls
- Front arm rotation
- Torso forward bending
- Left front lunge
- Right front lunge
- Squats
- Side lower limb extension
- Front lower limb extension

Each movement sequence was recorded for a duration of 36 seconds and repeated three times from different angle perspectives (0°, 45°, and 90°). To ensure the extraction of a reliable skeleton from the OMC system, each sequence began with a T-pose followed by the action shortly after the recording started. We cropped the frames to start when the activity commenced after the T-pose. This process resulted in 432 total sequences of varying lengths, depending on when the action began. In total, we extracted 203,149 synchronized frames, which was equivalent to approximately 226 minutes of active movements. Our dataset was published[7] with additional detailed information about the dataset.

IV. RADAR DATA PROCESSING

In this work, we utilized radar target lists to predict human poses. The goal was to drastically reduce the input data size for the downstream machine learning compared to raw radar data while ensuring suitability for applications on edge devices. Therefore, radar point clouds needed to be processed from raw radar data. For radar signal processing, we used the time-division multiplexing MIMO FMCW chirp-sequence processing chain [59], [60] depicted in Fig. 6. This chain is specifically tailored for applications focused on extracting human motion in indoor environments. Raw radar data from each virtual channel per frame was captured and organized into a complex-valued 4D radar cube $\boldsymbol{X}[i, j, k, l]$. The first two axes ($i$ and j) corresponded to the fast-time and slow-time elements, with 64 elements along the fast-time axis (i), and 128 elements along the slow-time axis (j). The axes in direction k and l correspond to the virtual antennas (channels) in spatial direction. Due to the antenna array shape, four elements in this sub-matrix were filled with zeros.

We implemented a static clutter removal algorithm to eliminate non-moving objects from the frame, which is essential for three main reasons. First, it enables the distinction between human movements and static objects. Second, it enhances the system's sensitivity to small movements originating from smaller body parts. This step is particularly important because

[7] [Available]. Online: https://doi.org/10.5281/zenodo.14738837

FIGURE 5. The 11 activities that were performed at an angle of 0° (facing the radar) as depicted in Fig. 2. The static T-pose is omitted.

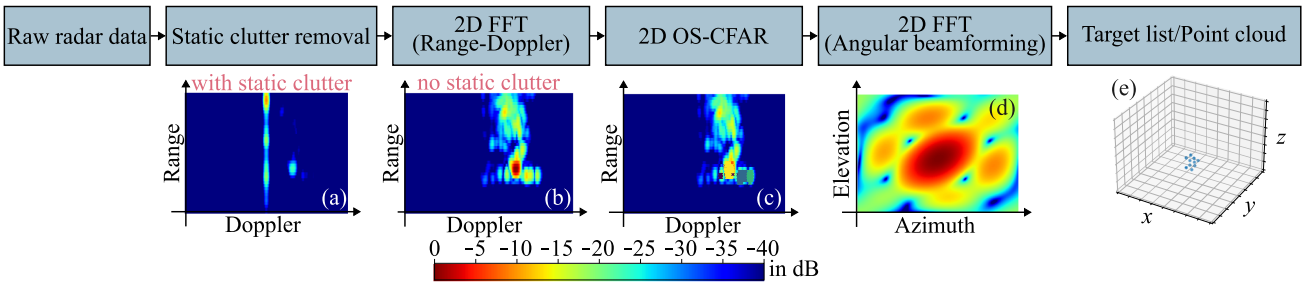

FIGURE 6. Applied radar signal processing: The raw radar data was processed by a 2D fast Fourier transform (FFT) to convert it to more descriptive range-Doppler information. Subsequently, a 2D ordered-statistic constant false-alarm rate (OS-CFAR) was applied to extract the relevant scatterer in the range-Doppler domain and apply a 2D FFT beamforming to infer the position of the scatterer.

strong static reflections from elements like metallic ventilation pipes or door frames can obscure decisive movement-related data. Third, this process reduces the radar data's dependence on the specific static environmental features of a room. To remove these static reflections, we subtracted the mean value within each chirp:

$$\boldsymbol{X}_{\mathrm{SCR}}[i, j, k, l] = \boldsymbol{X}[i, j, k, l] - \bar{\boldsymbol{X}}[i, j, k, l], \tag{2}$$

where $\bar{\boldsymbol{X}}[i, j, k, l]$ represents the mean value across the entries in the j-direction and is expanded to the same dimensions as $\boldsymbol{X}[i, j, k, l]$. Subsequently, we applied a window function $\boldsymbol{W}[i, j, k, l]$ on the fast-time and slow-time axis (e.g., the Hann function) to reduce side-lobes and transform it via 2D fast Fourier transform (FFT) into frequency domain $\boldsymbol{Y}[m, n, k, l]$, where m and n corresponded to range and Doppler (velocity).

$$\boldsymbol{Y}[m, n, k, l] = \mathrm{FFT}\{\boldsymbol{W}[i, j, k, l] \cdot \boldsymbol{X}_{\mathrm{SCR}}[i, j, k, l]\}. \tag{3}$$

An exemplary plot from a range-Doppler matrix $(20 \cdot \log_{10}|\boldsymbol{Y}[m, n, 1, 1]|)$ is illustrated in Fig. 6(a) without clutter removal and in Fig. 6(b) using the clutter removal algorithm. A 2D ordered-statistic constant false-alarm rate (CFAR)CFAR algorithm [61], [62], [63] was applied to the range-Doppler matrix to extract the most dominant scatterer from the range-Doppler domain, as seen in Fig. 6(c). This ensured that the extracted targets belonged to relevant objects in the scene rather than being artifacts of noise. A 2D angular beamforming was applied to estimate the angle of arrival of the reflection. Therefore, another 2D FFT was processed on the radar cube on the dimensions of the virtual array k and l to transform into elevation o and the azimuth p domain:

$$\boldsymbol{Z}[m, n, o, p] = \mathrm{FFT}\{\boldsymbol{Y}[m, n, k, l]\}. \tag{4}$$

To optimize computational efficiency, the second FFT was only applied and evaluated at specific range-Doppler bins where the ordered-statistic CFAR algorithm had detected a target. An exemplary elevation-azimuth plot $(20 \cdot \log_{10}|\boldsymbol{Z}[16, 90, o, p]|)$ is shown in Fig. 6(d). Multi-object beamforming was used to extract multiple targets from a single range-Doppler bin as it is common in such complex human-related scenarios for two or more reflections to occur within the same range bin and at the same velocity. To address this, the CLEAN algorithm [64] was used, which enables the extraction of multiple reflections from a single range-Doppler bin. The processed data was then converted into Cartesian coordinates as seen in Fig. 6(e). Using the range, Doppler, elevation, and azimuth information, the system can determine each target's position and corresponding radial velocity relative to the radar. Additionally, key metrics, such as signal-to-noise ratio (SNR), noise level, and intensity for each target were filled in the target list. These target lists were the input data for the machine learning model. Due to the necessity of a fixed input size, the maximal number of targets per frame was set at 64. Target lists with less than 64 detected targets were extended with zeros.

V. PREDICTION OF HUMAN POSES

To accurately estimate the skeleton keypoints' positions based on radar target lists, we developed a deep learning model which was adapted from Pointnet [53]. We implemented our method and the proposed framework in PyTorch using Python. In the next sections, we describe the architecture and training of our model.

A. POSE ESTIMATION MODEL

Our deep learning model design is depicted in Fig. 7. We used a multi-stage approach for skeleton keypoints prediction using temporal and spatial correlations in our input data. We used the upper section in Fig. 7, PntPose, for single-frame high-level feature extraction and extended the network with a recurrent neural network (RNN) and an MHSA mechanism as illustrated in the lower section of Fig. 7. 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 $\in \mathbb{R}^{7\times 1}$. The prediction of our network was 24 joint keypoints skeleton $\in \mathbb{R}^{24\times 3}$.

The feature extraction of PntPose is based on the well-established PointNet model from [53]. It extracts relevant features from point cloud data while being invariant to the order of the input features, which is achieved through max

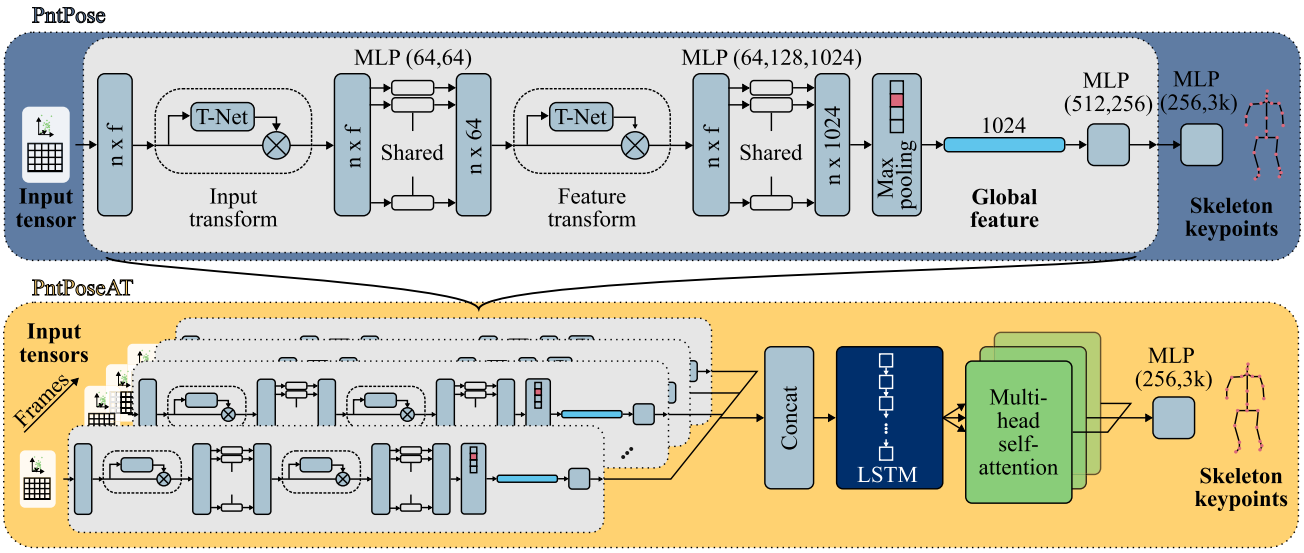

FIGURE 7. Architectures of the proposed keypoints-extracting networks: PntPose (blue) predicts the skeleton keypoints $\in \mathbb{R}^{24\times3}$ from a single-frame input tensor $\in \mathbb{R}^{7\times1}$. *PntPoseAT* (yellow) is a multi-frame input network that predicts the skeleton keypoints from multiple consecutive frames using a long short-term memory (LSTM) and the multi-head self-attention (MHSA) mechanism. MLP: multi-layer perceptron, T-Net: transformation network, LSTM: long short-term memory.

pooling layers. Through multiple linear and transformation layers, the model learns the spatial structure of the input data.

To exploit the temporal correlations in the kinematic tree of the estimated poses, an RNN was used. Two different RNN architectures were compared: an LSTM-based [54] and a gated recurrent unit (GRU)-based [65] network. The GRU cell comprises fewer parameters than an LSTM cell and, therefore, is easier and faster to train. On the other hand, GRU cells can suffer from exploding gradients by design [66], a problem that is solved by the more complex LSTM cell. Thus, the challenge of finding a suitable architecture balances simplicity and performance. Additionally, in line with state-of-the-art models in radar-based human pose estimation, we incorporated an attention mechanism in our network to increase prediction accuracy [13], [19]. In our hybrid solution, we upgraded the LSTM cell with an MHSA [55] mechanism across the hidden states of the RNN, thereby effectively circumventing the problem of information loss by hidden state accumulation with weighted attention processing [67] and linear reprojection to the output size. For example, if we had a tensor $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{B\times T\times H}$ as the output hidden state sequence of the RNN, where B is the batch, T is the sequence length, and H is the hidden state dimension, and we process this tensor with an MHSA layer. The query, key, and value matrices are computed as linear projections of the input sequence $\boldsymbol{X}_{\text{in}}$ by multiplying it with the learned weight matrices: the query matrix $\boldsymbol{Q} = \boldsymbol{X}_{\text{in}}\boldsymbol{W}^{Q}$, the key matrix $\boldsymbol{K} = \boldsymbol{X}_{\text{in}}\boldsymbol{W}^{K}$, and the value matrix $\boldsymbol{V} = \boldsymbol{X}_{\text{in}}\boldsymbol{W}^{V}$, where $\boldsymbol{W}^{Q}$, $\boldsymbol{W}^{K}$, and $\boldsymbol{W}^{V}$ are the learned weight matrices, respectively. MHSA can then be computed as

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_{\text{k}}}}\right)\boldsymbol{V}, \tag{5}$$

where d_{k} is the dimensionality of the key matrix. The attention output is then implemented with an additive residual connection

$$\boldsymbol{X}_{\text{out}} = \text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) + \boldsymbol{X}_{\text{in}}. \tag{6}$$

To obtain the next prediction, this output matrix is then flattened over the time dimension and linearly reprojected to the output dimension for the next keypoint prediction with a linear layer.

B. TRAINING, OPTIMIZATION, AND EVALUATION

For the training and optimization of our deep learning model, we initially split the dataset participant-wise into three subsets: train, validation, and test. The test set, which comprising two participants randomly chosen from the 12 participants (participant 3 (male) and participant 9 (female)), was set aside for the entire development and optimization process. The remaining 10 participants were further divided into training and validation sets, to implement cross validation, thereby optimizing the hyperparameters within every training fold independently. We employed a five-fold cross validation, with eight participants in the training set and two in the validation set for each fold. This ensured that each participant was used exactly once for validation. To speed up the hyperparameters optimization process, we used 30% of the data per activity, thereby ensuring multiple repetitions of each activity movement in the training set. In each fold, we trained the model on the eight participants by employing early stopping to prevent overfitting, with training limited to a maximum of 50 epochs. The initial learning rate was set to 0.001, with a learning rate decay applied after 30 epochs, thus reducing it to 0.0001. We optimized the network weights by minimizing the mean squared error loss using the Adam optimizer and a batch size

TABLE 4. Layer Overview of the Optimized Model Using Cross Validation

Parameter	Quantity
LSTM sequence length	50
LSTM input features	256
LSTM hidden size	72
MHSA total dimensions	72
MHSA number of heads	24
Dropout in backbone	0
Dropout in regression head	0.2

LSTM: long short-term memory, MHSA: multi-head self-attention.

TABLE 5. Lowest Average Mean Per-Joint Position Error (MPJPE) on the Validation Set Per Fold From Cross Validation With the Optimized Hyperparameter Configuration Set

Fold	Avg. MPJPE in cm	Mean standard deviation in cm
Fold 1	8.29	4.53
Fold 2	8.10	4.81
Fold 3	8.90	4.88
Fold 4	8.30	5.13
Fold 5	7.30	5.43
Average	8.18	4.96

MPJPE: mean per-joint position error.

of 64. We selected the best hyperparameters for our final model by averaging the lowest validation loss for each fold within every hyperparameter configuration. This allowed us to identify the configuration that performed best overall. As a result of the cross validation, an overview of the optimized model's layers is provided in Table 4.

We used the MPJPE metric to assess and compare the models' performance. It is defined by:

$$\mathrm{MPJPE}_k = \frac{1}{N_S}\sum_{i=1}^{N_S} \|\mathbf{x}_{\mathrm{pred},i,k} - \mathbf{x}_{\mathrm{gt},i,k}\|_2, \tag{7}$$

where N_S is the number of considered samples (skeletons), $\mathbf{x}_{\mathrm{pred},i,k}$ is the predicted position of keypoint k in sample i, and $\mathbf{x}_{\mathrm{gt},i,k}$ is the ground-truth position of keypoint k in sample i. The average MPJPE is calculated as the mean MPJPE across all keypoints.

The average MPJPE results for every cross validation fold using the optimized hyperparameter configuration are presented in Table 5, with an average MPJPE of 8.18 cm and a standard deviation of 4.96 cm.

Once the optimal hyperparameters were identified, we trained the model using the combined training and validation data and evaluated the performance of the overall model *PntPoseAT* on the unseen test set. Training was performed for up to 60 epochs by utilizing early stopping to prevent overfitting and ensure the selection of the optimal model.

C. EXPERIMENTAL RESULTS

The MPJPE and their corresponding standard deviation for each skeleton keypoint are shown in Fig. 8. While the keypoints on the hips, spine, and neck show small mean position errors in the range of 5.25 cm to 5.69 cm, keypoints on the limbs — especially the hands (RH: right hand, LH: left hand) and elbows (LFA: left forearm, RFA: right forearm) — display higher MPJPEs, from 7.81 cm to 11.33 cm.

To gain deeper insights into our model's performance, we conducted a series of experiments focusing on the evaluation of individual model components. Table 6 presents the performance metrics for various network architectures and input data configurations. Our reference model, *PntPoseAT*, achieved pose predictions with an average MPJPE of 6.49 cm and a standard deviation of 3.74 cm. Comparatively, the basic PntPose framework, which predicts human poses on a frame-by-frame basis, demonstrated a higher average MPJPE of 10.11 cm, which was 56% greater than that of the *PntPoseAT*. Substitution of the LSTM with a GRU in the network architecture resulted in a similar average MPJPE of 6.53 cm. It was observed that reducing the input data components negatively impacted the accuracy across all network variations. We delve into a detailed discussion of these results in the following section.

VI. DISCUSSION

Initially, we concentrate on the results utilizing all seven target list components as an input. As seen in Fig. 8, the limbs showed a higher average MPJPE and standard deviation compared to the pelvis components. This was expected as the limbs anatomically undergo more movement during the activity compared to the hips, spine, and neck, which makes accurate prediction of these keypoints more challenging. This behavior was already reported in previous works, e.g., in Cao et al. [68]. Models enhanced with an LSTM or GRU cell to capture temporal changes showed improved regression results. Incorporating an MHSA mechanism further boosted the model's performance. Additionally, as a GRU does not have a cell state contrary to LSTMs, it reduces the learnable parameters by 25%, thereby making it a plausible alternative for applications with limited processing memory and power. In fact, replacing the GRU with an LSTM in our model led only to a slight reduction of 0.04 cm of average MPJPE, which prompted cost-effectiveness considerations tailored to the specific application.

Next, we examined how the model's performance depended on the characteristics of the input data. The results showed that reducing the input data from a tensor containing x-, y-, z-coordinates, velocity, SNR, noise, and intensity to only x-, y-, z-coordinates led to a drop in performance across all network architecture. This demonstrated that including additional information beyond the basic point cloud data helped the network to better capture the pose. However, the disparity in average MPJPE between the compared networks decreased when the LSTM and MHSA units were applied. This suggests that the recurrence and attention mechanisms can effectively mitigate the impact of reduced input data.

Considering that this study utilized a relatively simple radar system, the results are highly promising. Despite using a MIMO radar array with three transmitters and four receivers,

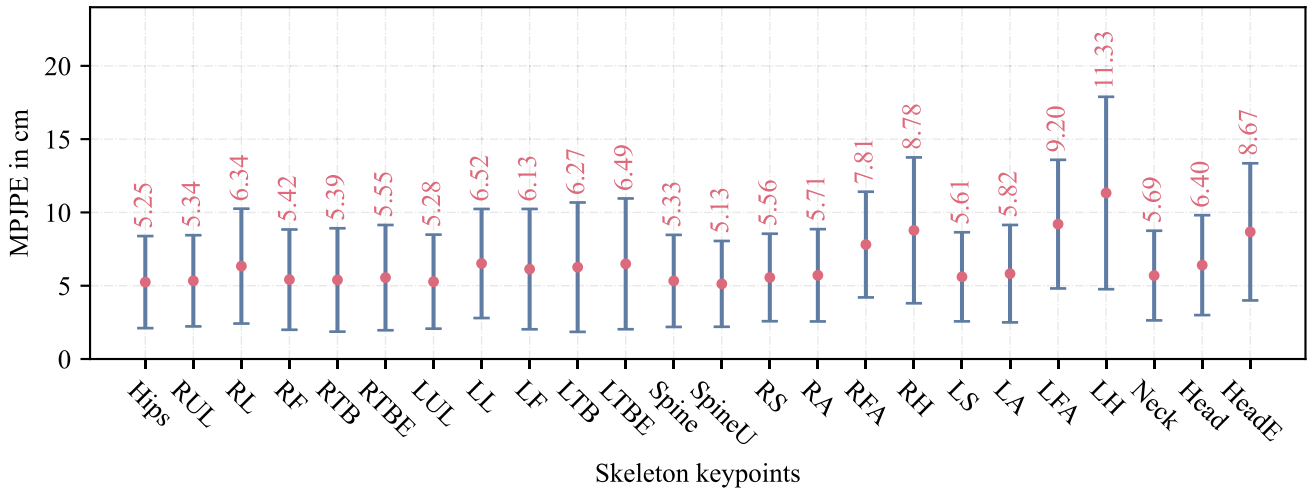

FIGURE 8. Results of the mean per-joint position error (MPJPE) using *PntPoseAT:* The mean value (red) and its standard deviation (blue) of the prediction. The designations of the abbreviations for the respective joints are shown in Table 2.

TABLE 6. Comparison of the Performance of Different Network Components and Input Data

Network components	Input data (target list components)	Avg. MPJPE in cm	Standard deviation in cm
PntPose	[x-,y-,z] of mmRadPose	11.38	9.88
PntPose	[x-,y-,z, v, SNR, noise, intensity] of mmRadPose	10.11	9.02
PntPose+LSTM	[x-,y-,z] of mmRadPose	6.77	4.37
PntPose+LSTM+MHSA	[x-,y-,z] of mmRadPose	6.63	4.32
PntPose+LSTM	[x-,y-,z, v, SNR, noise, intensity] of mmRadPose	6.57	3.94
PntPose+GRU+MHSA	[x-,y-,z, v, SNR, noise, intensity] of mmRadPose	6.53	3.92
PntPose+LSTM+MHSA (***PntPoseAT***)	[x-,y-,z, v, SNR, noise, intensity] of mmRadPose	6.49	3.74

MPJPE: mean per-joint position error, LSTM: long short-term memory, MHSA: multi-head self-attention, SNR: signal-to-noise ratio.

we were able to achieve results comparable to or even better than previous work with the same [14], [15], [18] or even larger antenna arrays [12], [16], [17], thereby resulting in inherently better spatial resolution. However, for a better comparison, it is essential to evaluate the methods using a uniform dataset.

Moreover, our model predicted a very detailed human skeleton with 24 keypoints. To the best of our knowledge, Sengupta et al. [13] is the only work that has proposed a model predicting more keypoints than ours, with a total of 25 keypoints, but, due to inaccuracies, they reduced the count to 17 in their evaluation. A higher number of predicted keypoints, however, is favorable for downstream tasks, such as full-body biomechanical analysis [69].

Unlike other approaches that relied on multiple orthogonal radar units [13], [16], [19], our method only required a single radar, thus reducing the complexity and volume of data to process. This demonstrates the potential of our approach to provide accurate pose estimation with fewer resources, which would make it not only efficient but also scalable for real-world applications. Moreover, the reduced hardware and data requirements enable deployment in more constrained or cost-sensitive environments. Coupled with recent advancements in radar technology, including enhanced on-device processing capabilities, tasks like CFAR calculation can now be executed directly on the device (e.g., TI-IWR2944[8]). This makes our network an ideal solution for running human pose estimation applications on edge devices, thereby drastically reducing the input size of the deep learning network and eliminating the need for extensive raw data transfer as well as reliance on external processing resources [16], [17], [19].

In this study, we focused on scenarios involving a single individual within a constrained area of the room. However, in complex real-world applications, situations often arise where multiple people are randomly distributed throughout the room. There are existing approaches to address this challenge. For instance, Zhao et al. [12] and Ho et al. [17] tackled this issue by dividing the neural network into two components: one component identified a region of interest that encompasses a single person, while the other treated this region as a single-person human pose estimation problem — an approach that is well-established in computer vision [70], [71], [72]. Additionally, by incorporating three different angles facing the radar, we demonstrated that human pose estimation is possible regardless of the person's orientation toward the radar. Consequently, it is crucial to effectively separate individuals in at least one dimension of the radar data. Given the radar's excellent resolution in the range and Doppler domains, along

[8][Available]. Online: www.ti.com/product/IWR2944

with reasonable angular resolution, this separation is assumed to be feasible. We plan to investigate this topic in future work.

Our dataset was specifically designed to capture human movements during activity. To enhance the diversity of body motion patterns, we recorded each activity from three distinct angles. Additionally, we ensured a balanced representation of arm and leg movements, thus preventing the network from favoring predictions for one part of the body over another. This extends existing human motion and rehabilitation datasets, but we recognize the need for further expansion. In the future, we plan to enhance the dataset by incorporating more participants and a wider variety of activities to further enhance deep learning performance in terms of position error and generalizability.

VII. CONCLUSION

This paper presents a novel radar-based human pose estimation approach using deep learning by leveraging accurate OMC data as ground truth. An extensive, published dataset — *mmRadPose* — was generated to develop and train a deep learning model. We hope our dataset will be a valuable resource for researchers working in the field of machine learning-enabled radar-based pose and activity estimation. We present *PntPoseAT*, a multi-stage deep learning model that captures both temporal and spatial correlations of the input data. This approach outperforms comparable existing human pose estimation models in terms of mean per-joint position error with an overall average mean per-joint position error of 6.49 cm. We explored additional aspects of the model design by analyzing the performance gains contributed by individual components within our deep learning model. The results demonstrate the high potential of radar technology as a viable alternative to RGB-based methods that offer enhanced privacy protection and eliminates reliance on lighting conditions. Additionally, we demonstrated that accurate keypoint predictions can be achieved with relatively cost-efficient and unsophisticated radar sensors, thus enabling their versatile use in a wide range of environments, such as smart homes, healthcare, and industry.

ACKNOWLEDGMENT

The authors would like to thank all of the participants in motion study for their valuable time and enthusiasm.

REFERENCES

[1] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1653–1660.

[2] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 44–58, Jan. 2006.

[3] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2019.

[4] Z. Jiang, H. Ji, C.-Y. Yang, and J.-N. Hwang, "2D human pose estimation calibration and keypoint visibility classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2024, pp. 6095–6099.

[5] Z. Zhou, Z. Jiang, W. Chai, C.-Y. Yang, L. Li, and J.-N. Hwang, "Efficient domain adaptation via generative prior for 3D infant pose estimation," Nov. 2023, *arXiv:2311.12043*.

[6] Z. Jiang, Z. Zhou, L. Li, W. Chai, C.-Y. Yang, and J.-N. Hwang, "Back to optimization: Diffusion-based zero-shot 3D human pose estimation," Oct. 2023, *arXiv:2307.03833*.

[7] H. Liu et al., "PoSynDA: Multi-hypothesis pose synthesis domain adaptation for robust 3D human pose estimation," Oct. 2023, *arXiv:2308.09678*. [Online]. Available: https://dl.acm.org/doi/10.1145/3581783.3612368

[8] C.-Y. Yang et al., "CameraPose: Weakly-supervised monocular 3D human pose estimation by leveraging in-the-wild 2D annotations," Jan. 2023, *arXiv:2301.02979*.

[9] Z. Jiang, W. Chai, L. Li, Z. Zhou, C.-Y. Yang, and J.-N. Hwang, "UniHPE: Towards unified human pose estimation via contrastive learning," 2023, *arXiv:2311.16477*.

[10] W. Chai, Z. Jiang, J.-N. Hwang, and G. Wang, "Global adaptation meets local generalization: Unsupervised domain adaptation for 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 14655–14665.

[11] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors," Mar. 2021, *arXiv:2103.17265*.

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

[13] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmWave radars and CNNs," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10032–10044, Sep. 2020.

[14] 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.

[15] S. An, Y. Li, and U. Ogras, "mRI: Multi-modal 3D human pose estimation dataset using mmWave, RGB-D, and inertial sensors," Oct. 2022, *arXiv:2210.08394*. [Online]. Available: https://dl.acm.org/doi/10.5555/3600270.3602258

[16] C. Yu et al., "RFPose-OT: RF-based 3D human pose estimation via optimal transport theory," Dec. 2022, *arXiv:2301.13013*. [Online]. Available: https://link.springer.com/article/10.1631/FITEE.2200550

[17] Y.-H. Ho et al., "RT-pose: A 4D radar tensor-based 3D human pose estimation and localization benchmark," Jul. 2024, *arXiv:2407.13930*.

[18] H. Xue et al., "mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave," in *Proc. 19th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, New York, NY, USA, Jun. 2021, pp. 269–282.

[19] 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, Jan. 2023, pp. 5704–5713.

[20] C.-L. Lin, Y.-H. Ho, W.-C. Chiu, T.-C. Chu, and Y.-H. Liu, "Innovative shoe-integrated system based on time-of-flight range sensors for fall detection on various terrains," *IEEE Sensors Lett.*, vol. 5, no. 10, Oct. 2021, Art. no. 6002404.

[21] T. Tao et al., "Trajectory planning of upper limb rehabilitation robot based on human pose estimation," in *Proc. 17th Int. Conf. Ubiquitous Robots*, Jun. 2020, pp. 333–338.

[22] 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.

[23] B. Jokanovic 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.

[24] K. Chen et al., "Patient-specific pose estimation in clinical environments," *IEEE J. Transl. Eng. Health Med.*, vol. 6, 2018, Art. no. 2101111.

[25] B. Jokanovic, M. Amin, and F. Ahmad, "Radar fall motion detection using deep learning." in *Proc. IEEE Radar Conf.*, Philadelphia, PA, USA, 2016, pp. 1–6.

[26] N. Vysotskaya et al., "Continuous non-invasive blood pressure measurement using 60 GHz-radar—A feasibility study," *Sensors*, vol. 23, no. 8, Apr. 2023, Art. no. 4111.

[27] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB-Depth sensors for fall detection: A review," *IEEE Sensors J.*, vol. 17, no. 12, pp. 3585–3604, Jun. 2017.

[28] A. Raza, A. M. Qadri, I. Akhtar, N. A. Samee, and M. Alabdulhafith, "LogRF: An approach to human pose estimation using skeleton landmarks for physiotherapy fitness exercise correction," *IEEE Access*, vol. 11, pp. 107930–107939, 2023.

[29] A. Badiola-Bengoa and A. Mendez-Zorrilla, "A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise," *Sensors*, vol. 21, no. 18, Jan. 2021, Art. no. 5996.

[30] L. Citraro et al., "Real-time camera pose estimation for sports fields," *Mach. Vis. Appl.*, vol. 31, no. 3, Mar. 2020, Art. no. 16.

[31] J. Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, "AI coach: Deep human pose estimation and analysis for personalized athletic training assistance," in *Proc. 27th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2019, pp. 374–382.

[32] L. Bridgeman, M. Volino, J.-Y. Guillemaut, and A. Hilton, "Multi-person 3D pose estimation and tracking in sports," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 2487–2496.

[33] K. Ludwig, S. Scherer, M. Einfalt, and R. Lienhart, "Self-supervised learning for human pose estimation in sports," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2021, pp. 1–6.

[34] Y. Yuan, S.-E. Wei, T. Simon, K. Kitani, and J. Saragih, "SimPoE: Simulated character control for 3D human pose estimation," Apr. 2021, *arXiv:2104.00683*.

[35] X. Zhou, W. Liang, K. I.-K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for internet of healthcare things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020.

[36] J. Wang, Q. Gao, X. Ma, Y. Zhao, and Y. Fang, "Learning to sense: Deep learning for wireless sensing with less training efforts," *IEEE Wireless Commun.*, vol. 27, no. 3, pp. 156–162, Jun. 2020.

[37] H.-Y. Huang, C.-W. Ning, P.-Y. Wang, J.-H. Cheng, and L.-P. Cheng, "Haptic-go-round: A surrounding platform for encounter-type haptics in virtual reality experiences," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, Apr. 2020, pp. 1–10.

[38] X. Ma, Y. Zhao, L. Zhang, Q. Gao, M. Pan, and J. Wang, "Practical device-free gesture recognition using WiFi signals based on metalearning," *IEEE Trans. Ind. Inform.*, vol. 16, no. 1, pp. 228–237, Jan. 2020.

[39] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "RadHAR: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proc. 3rd ACM Workshop Millimeter-Wave Netw. Sens. Syst.*, New York, NY, USA, Oct. 2019, pp. 51–56.

[40] O. G. Guleryuz and C. Kaeser-Chen, "Fast lifting for 3D hand pose estimation in AR/VR applications," in *Proc. 25th IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 106–110.

[41] S.-R. Ke, L. Zhu, J.-N. Hwang, H.-I. Pai, K.-M. Lan, and C.-P. Liao, "Real-time 3D human pose estimation from monocular view with applications to event detection and video gaming," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2010, pp. 489–496.

[42] T. He, "Human pose-based activity recognition approaches on smart-home devices," in *Proc. Distrib., Ambient Pervasive Interact. Smart Environ., Ecosystems, Cities: 10th Int. Conf. 24th HCI Int. Conf. Virtual Event*, Berlin, Germany, Jun. 2022, pp. 266–277.

[43] H.-Y. Lin and T.-W. Chen, "Augmented reality with human body interaction based on monocular 3D pose estimation," in *Advanced Concepts for Intelligent Vision Systems*, J. Blanc-Talon, D. Bone, W. Philips, D. Popescu, and P. Scheunders, Eds. Berlin, Germany: Springer, 2010, pp. 321–331.

[44] 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.

[45] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023.

[46] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison, "Vid2Doppler: Synthesizing Doppler radar data from videos for training privacy-preserving activity recognition," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2021, pp. 1–10.

[47] 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.

[48] 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.

[49] 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 Sep. 2023, pp. 428–431.

[50] 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.

[51] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 7745–7754.

[52] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 5385–5394.

[53] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," Apr. 2017, *arXiv:1612.00593*.

[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[55] A. Vaswani et al., "Attention is all you need," Dec. 2017, *arXiv:1706.03762*. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295349

[56] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, Mar. 1966.

[57] J. C. Gower, "Generalized Procrustes Analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, Mar. 1975.

[58] 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.

[59] 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.

[60] A. Wojtkiewicz, J. Misiurewicz, M. Nałcz, K. Jedrzejewski, and K. Kulpa, "Two-dimensional signal processing in FMCW radars," in *Proc. XX KKTOiUE*, Kołobrzeg, Poland, Oct. 1997, pp. 475–480. [Online]. Available: https://scholar.google.de/citations?view_op=view_citation&hl=de&user=KZ7kH7wAAAAJ&citation_for_view=KZ7kH7wAAAAJ:Se3iqnhoufwC

[61] M. A. Richards, *Fundamentals of Radar Signal Processing*, 2nd ed. New York, NY, USA: McGraw-Hill, 2014.

[62] H. Rohling, "Ordered statistic CFAR technique - An overview," in *Proc. 12th Int. Radar Symp.*, Sep. 2011, pp. 631–638.

[63] M. Kronauge and H. Rohling, "Fast two-dimensional CFAR procedure," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 3, pp. 1817–1823, Jul. 2013.

[64] J. A. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines," *Astron. Astrophys. Suppl. Ser.*, vol. 15, Jun. 1974, Art. no. 417.

[65] K. Cho et al., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," Sep. 2014, *arXiv:1406.1078*. [Online]. Available: https://aclanthology.org/D14-1179/

[66] S. Kanai, Y. Fujiwara, and S. Iwamura, "Preventing gradient explosions in gated recurrent units," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Dec. 2017, pp. 435–444.

[67] J. Müller, R. Braun, H. P. A. Lensch, and N. Ludwig, "Glacier movement prediction with attention-based recurrent neural networks and satellite data," in *Proc. NeurIPS Workshop Tackling Climate Change Mach. Learn.*, 2023. [Online]. Available: https://www.climatechange.ai/papers/neurips2023/42

[68] Z. Cao, J. Zhang, R. Chen, X. Guo, and G. Wang, "Task-specific feature purifying in radar-based human pose estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 6, pp. 9285–9298, Dec. 2023.

[69] A. Seth et al., "OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement," *PLoS Comput. Biol.*, vol. 14, no. 7, Jul. 2018, Art. no. e1006223.

[70] 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.

[71] 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.
[72] 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.

LUKAS ENGEL (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering in 2017 and 2019, respectively, from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, 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, radar signal processing, radar hardware, antenna design, and 3D-printed mmWave components.

JONAS MUELLER received the bachelors degree in psychology from Alpen-Adria-University, Klagenfurt, Austria, and completed his studies in cognitive science from the University of Tübingen, Tübingen, Germany, strengthening his interest in applications of artificial intelligence from a human inspired perspective. He is currently working toward the Ph.D. degree with the Machine Learning and Data Analytics Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. His research interests include applications of human pose estimation from camera and sensor data and deep learning methods for remote sensing applications.

EDUARDO JAVIER FERIA RENDON received the B.Sc. degree in computer science from the University of Havana, Havana, Cuba, in 2019, and the M.Sc. degree in artificial intelligence from Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, in 2024, where he conducted research which resulted in his thesis, *Human Pose Estimation with mmWave Radar Sensors*. He began his career with the Cuban Neuroscience Center, focusing on the development of audiology testing software. In 2021, he moved to Germany. His research interests include smart home automation, security, and human-computer interaction.

EVA DORSCHKY (Member, IEEE) received the Graduate (with Distinction) degree in electrical engineering with a focus on signal processing and machine learning and the Ph.D. degree in computer science from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, in 2014 and 2022, respectively. She led the Sports Analytics Group with Machine Learning and Data Analytics Laboratory, FAU, as a Postdoc. Since October 2024, she has been a Senior AI Engineer with fiveD GmbH, where she specializes in realistic human modeling and simulation for radar ray tracing applications. She was the recipient of the two Doctoral awards for her dissertation on machine learning and optimal control in biomechanical simulations.

DANIEL KRAUSS (Graduate Student Member, IEEE) was born in 1996, Schwäbisch Hall, Germany. He received the M.Sc. degree in medical engineering in 2022 from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, where he is currently working toward the Ph.D. degree with the Machine Learning and Data Analytics Lab. His research interests include biosignal analysis and the application of machine learning algorithms and focuses on the early detection of Parkinson's disease, with a particular interest in radar-based sleep analysis. He was the recipient of the 2nd place Routledge Young Researcher Award from the International Association of Computer Science in Sport in 2022.

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 the Head of the Research Group "Wave-based Imaging Systems" 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 imaging and radar signal processing for nondestructive testing, security screening, medical radar, and automotive applications. She is also a Reviewer of European Radar Conference (EuRAD) 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 Argus Science Award (sponsored by Airbus Defense and Space, now Hensoldt) in 2016 and the EuRAD Conference Prize in 2019. She is a member of the IEEE CRFID Technical Committee on Motion Capture & Localization.

BJOERN M. ESKOFIER (Senior Member, IEEE) received the Graduation degree in electrical engineering from Friedrich-Alexander-University Erlangen-Nuernberg (FAU), in 2006, and the Ph.D. degree in biomechanics from the University of Calgary, Calgary, AB, Canada, under the supervision of Prof. Dr. Benno Nigg. From February to March 2016, he was a Visiting Professor with Prof. Paolo Bonato's Motion Analysis Laboratory, Harvard Medical School, and he was a Visiting Professor with Prof. Alex Sandy Pentland's Human Dynamics Group, MIT Media Laboratory, from March to August 2018. Since April 2023, he has been an Associate Principal Investigator and a Leader with Research Group Translational Digital Health, Helmholtz Zentrum Munich, Germany. From April to August 2023, he was a Visiting Professor with Prof. Scott Delp's NMBL Laboratory that is part of Stanford University's Schools of Engineering and Medicine. He currently Heads the Machine Learning and Data Analytics (MaD) Laboratory with the Friedrich-Alexander-University Erlangen-Nuernberg (FAU), 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 with 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 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, UbiComp), and is also 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 with the Institute of Microwaves and Photonics (LHFT), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 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, and wireless sensor and locating systems. Since April 2024, he has been the spokesperson for the DFG Review Board 4.42 Electrical Engineering and Information Technology. He is also the Spokesman with 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 is also 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 is on the Advisory Board of the IEEE CRFID Technical Committee on Motion Capture & Localization. Dr. Vossiek was the recipient of numerous Best Paper prizes and other awards, Microwave Application Award in 2019 by the IEEE MTT Society (MTT-S) for Pioneering Research in Wireless Local Positioning Systems. He is a member of Organizing Committees and Technical Program Committees for many international conferences and was 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 October 2022, he has been an Associate Editor-in-Chief of IEEE TRANSACTIONS ON RADAR SYSTEM.