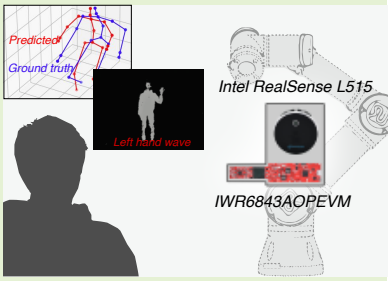
mmPrivPose3D: A RaDAR-based approach to privacy-compliant pose estimation and gesture command recognition in human-robot collaboration

Nima Roshandel, Constantin Scholz, Hoang-Long Cao, Hoang-Giang Cao, Milan Amighi, Hamed Firouzipouyaei, Aleksander Burkiewicz, Sebastien Menet, Felipe Ballen-Moreno, Dylan Warawout Sisavath, Emil Imrith, Antonio Paolillo, Jan Genoe, Bram Vanderborght

***Abstract*— Various sensors are employed in dynamic human-robot collaboration manufacturing environments for real-time human pose estimation to improve safety through collision-avoidance systems and gesture command recognition to enhance human-robot interaction. However, the most widely used sensors – RGBD cameras – often underperform under varying lighting and environmental conditions and raise privacy concerns. This paper introduces mmPrivPose3D, a novel system designed to prioritize privacy while performing human pose estimation and gesture command recognition using a 60 GHz industrial Frequency Modulated Continuous Wave (FMCW) RaDAR with a 10 m maximum range and 29 degrees angular resolution. The system employs a parallel architecture including a 3D Convolutional Neural Network (CNN) for pose estimation, which extracts 19 keypoints of the human skeleton, along with a random forest classifier for recognizing gesture commands. The system was trained on a dataset involving ten individuals performing various movements in a human-robot interaction context, including walking in the workspace and hand-waving gestures. Our model demonstrated a low Mean Per Joint Position Error (MPJPE) of 4.8% across keypoints for pose estimation and, for gesture recognition, an accuracy of 96.3% during k-fold cross-validation and 96.2% during inference. mmPrivPose3D has the potential for application in human workspace localization and human-to-robot communication, particularly in contexts where privacy is a concern.**



***Index Terms*— human-robot collaboration, RaDAR, pose estimation, gesture command recognition, deep learning, privacy**

I. Introduction

The progression towards Industry 5.0, and the increasing need for adaptable human-robot collaboration, has accelerated the advancement of collaborative robots [1]–[3]. These robots can boost productivity and efficiency in various industries. This capability is crucial for addressing the labor shortages experienced globally [4]. Despite being designed to complement human capabilities, collaborative robots struggle with safety at high speeds and are difficult to understand in terms of their intentions [5]. Their rapid operation compromises human safety, but slowing them down impacts their return on investment and acceptance [6].

Sensor-based solutions have been implemented to estimate the 3D position of a human in the workplace relative to the robot's position, ensuring safety in human-robot collaboration and compliance with ISO 15066 regulations [7]. For example, some techniques have used stereo vision or monocular cameras for 3D pose estimation using different algorithms, including CNN-based 2D pose estimation from an RGB image and 3D registration using depth images [8], multi-view 2D pose estimation [9], and a transformer-based approach to learn spatial and temporal correlations between joints [10]. However, such light-based sensors are susceptible to lighting and environmental conditions [11], [12], which can compromise their effectiveness. Additionally, these sensors raise privacy concerns and can cause discomfort among workers and their unions, particularly under GDPR, because the use of cameras can be associated with constant monitoring, especially during

This work was funded by the imec SAFEBOT program and the European Commission Horizon Europe Research and Innovation Program as part of the project euROBIN grant no. 101070596.

N. Roshandel, C. Scholz, H.-L. Cao, M. Amighi, H. Firouzipouyaei, A. Burkiewicz, S. Menet, F. Ballen-Moreno, D. Sisavath, E. Imrith, A. Paolillo, and B. Vanderborght are with BruBotics, Vrije Universiteit Brussel, Pleinlaan 2, Brussels, Belgium (e-mail: hoang.long.cao@vub.be).

N. Roshandel, C. Scholz, M. Amighi, H. Firouzipouyaei, A. Burkiewicz, S. Menet, D. Sisavath, and B. Vanderborght are with imec-IMS-VUB, Belgium.

H.-L. Cao and F. Ballen-Moreno are with Flanders Make, Brussels, Belgium.

H.-G. Cao is with Ming Chi University of Technology, Taiwan, and Can Tho University, Vietnam.

J. Genoe and N. Roshandel are with KU Leuven, Belgium.

J. Genoe is with imec-SAT, Belgium.

A. Paolillo is with SOFT Languages Lab, Vrije Universiteit Brussel, Belgium.

work.[1].

To address the limitations of light-based sensors and the need for non-invasive sensing technologies to mitigate privacy concerns, using RaDAR sensors has been explored as a suitable alternative [13]. These sensors offer reliable detection even under harsh environmental conditions [12], [14] and effectively track obstacles, including humans close to robots. Often, human obstacles are represented by a single point that encapsulates the 3D point-cloud data from the sensor [15]. The use of FMCW RaDARs for 3D human pose estimation has shown higher resolution accuracy than single-representation tracking by allowing the tracking and recognition of multiple keypoints on the human body, which can be used and interpreted for human-robot interaction [16]. Moreover, RaDAR sensors have drawn research interest for gesture recognition; for example, one study used a 60 GHz FMCW RaDAR combined with a Recurrent Neural Network (RNN) [17]. Although many studies have explored the use of multiple RaDAR sensors in human 3D pose estimation, there remains a gap in the literature regarding the use of a single industry-certified 60 GHz mmWave FMCW RaDAR such as IWR6843AOPEVM[2] for real-time 3D human pose estimation and gesture command recognition, reducing the cost and exposure to radiation rather than using multiple sensors.

In this paper, we present mmPrivPose3D, a novel architecture that uses a single IWR6843AOPEVM RaDAR sensor for real-time pose estimation and gesture command recognition through a parallel architecture. This method is suitable for speed and separation monitoring compliant with ISO15066 regulations [7] and tasks that can be facilitated by human-robot interactions such as palletizing. A demonstration video is available at https://youtu.be/vbuGenAu3rE. The system uses a 3D CNN architecture for pose estimation, capturing 19 3D keypoints. Given that 3D pose estimation errors can be significant in certain body parts that have lower RaDAR reflectivity, such as the arms [18], [19], the system employs a random forest classifier in parallel for gesture command recognition to enhance accuracy. The system was trained and validated using our mmPrivPose3D dataset [20] with the human keypoints predicted from the RGBD images of an Intel RealSense L515 camera serving as the ground truth. mmPrivPose3D is suitable for shared human-robot workspaces, where privacy sensitivity is crucial. It enables robust human tracking for tasks involving human-robot collaboration, such as Speed and Separation Monitoring, and human-robot interaction through gestures while maintaining privacy [21], [22].

Our work makes the following contributions to human pose estimation and gesture recognition using RaDAR- and Radio-Frequency-based sensors:

- Using a gesture command recognition module in parallel with the pose estimation module for human-to-robot communication, addressing the larger error in the body regions with low reflective power such as the human's arm.
- Open-sourcing a dataset containing ten participants performing walking in the workspace, left-hand wave, and right-hand wave for training the mmPrivPose3D model.

The remainder of this paper is organized as follows. Section II provides an overview of existing approaches to human pose estimation and gesture recognition using RaDAR- and Radio-Frequency-based sensors. Section III presents mmPrivPose3D parallel architecture. The system was validated, and the experimental results are presented in Section IV.

II. Related work

Our study focuses on parallel pose estimation and gesture recognition. In these topics, well-known methods such as OpenPose [23] or MediaPipe [24] do not perform well under varying lighting and weather conditions [25]. Our approach is based on RaDAR- and radio-frequency-based methods aiming to address these challenges. In this section, we summarize the relevant approaches using these sensing technologies.

A. Antenna-based models

A Radio Frequency (RF) antenna array transmits radio frequency signals at a power level one-thousandth that of a Wi-Fi signal. Skeleton-tracking models that use the signals transmitted and received by these antennas to estimate human poses are known as antenna-based models.

In this domain, RF-pose3D was introduced [18], featuring three models for feature extraction and a region proposal network to detect individuals' bounding boxes using a sliding window for multi-person pose estimation. The architecture performs 4D convolution and is, decomposed into 3D convolutions across two planes and the time axis. An open-pose model using a 12-camera system provides ground truth, tracking 14 keypoints over a 3-second sliding window.

However, the model has limitations, including struggles with complex hand motions due to weaker RF reflections from small body parts, such as the hands. It also requires specific hardware, including a T-shaped antenna and a wide 1.78 GHz bandwidth RF signal, which limits its practical use [18]. To address these issues, the following section introduces a Wi-Fi-based model.

B. WiFi-based models

One study explored the use of Wi-Fi signals for 3D skeleton reconstructions [26]. The setup for this Wi-Fi-based model includes one transmitter and multiple receivers to collect the Wi-Fi signal data. This method utilizes Channel State Information filtered from the noise of the Wi-Fi signals. The received data are then fed as input into a deep neural network architecture, which includes four layers of CNN and three layers of Long Short-Term Memory (LSTM) networks, a type of RNN known for capturing temporal dependencies in consecutive data samples. The model outputs features representing human joint rotations and its accuracy is validated using 3D keypoints from an externally installed Vicon motion capture system. However, a significant limitation of this model is its inability to estimate the keypoints of dynamic humans, and the assumption that activities do not involve location changes.

[1] https://perma.cc/S5L4-4DPH
[2] https://perma.cc/MU8A-LVEP

C. RaDAR-based models

1) For pose estimation: Expanding upon WiFi and antenna-based methodologies, a study was conducted using FMCW RaDAR [27]. The mmpose model, a prominent example of a real-time network, distinguishes itself by identifying 15 human skeleton keypoints from RaDAR reflections. It processes RaDAR point-clouds on both XY and XZ planes using a forked 2D CNN. For the ground-truth calibration, this model employed Mathworks' skeletal tracking algorithm via the MATLAB API of Kinect. However, this model has several limitations. In particular, projecting a 3D point-cloud onto two separate images, also used in another work called MARS [28], disturbs the 3-D spatial correlation among points [25]. Furthermore, it depended on two AWR1642 boost[3] FMCW RaDARs for data capture. Additionally, the AWR line is suitable for automotive applications but not for industrial applications.

In the RPM approach, two identical AWR2243 RaDAR sensors[4] are used, with one positioned horizontally and the other vertically. The two RF signals are concatenated and fed into a Feature Fusion Network (FFN) for feature extraction, followed by a Spatio-Temporal Attention Module to recover the remaining body part keypoints, as the reflected signal is only from a subset of limbs [29]. However, this model uses two 76 to 81 GHz RaDAR sensors that are not certified for industrial applications. In addition, the pose estimation loss function has not taken the RaDAR data outliers into account and it suffers from performance degradation when processing data in a new scene as the dataset only contains walking as human activity [29].

Finally, mmpose-FK was recently introduced as a model that addresses the need to maintain the 3D spatial coherence by utilizing a voxel data-processing method with a single 60GHz IWR6843ISK-ODS RaDAR sensor and one Azure Kinect sensor to generate ground truth joint positions. In addition, the model has a forward kinematic layer that enhances the stability of keypoint positions [25]. However, this model neglected the effect of point-cloud outliers in the Mean Squared Error (MSE) as the position loss function and was validated based on a simple walking movement along a straight line rather than the entire sensor's field of view.

It is worth noting that the aforementioned approaches demonstrated the highest keypoint Mean Absolute Error (MAE) in the wrist and arm regions owing to their low RF reflective power [19]. This results in missing data points in the RaDAR point-cloud and therefore negatively influences the accurate detection of complex hand gestures.

2) Gesture command recognition: In addition to pose estimation, sensor signals can be used for gesture recognition, for example, as a command to the robot [21]. Regarding mmWave RaDAR, a recent approach investigated a Spiking Neural Network (SNN) to perform gesture recognition on a 60 GHZ mmWave RaDAR over a long-range [30]. In this approach, the RaDAR range-Doppler matrix was converted into a 16×16 image, which served as the input for the liquid state machine model. This model sends spike signals as input to a logistic regression for classification. Despite its superior accuracy compared to traditional classifiers, this model has a delay of 0.5 to 1 second in inference, which limits its real-time capability.

In a different study using the IWR6843AOPEVM sensor, it was observed that converting feature spectrum maps to images required a large high amount of computational resources. The study proposed the extraction of six feature types per gesture from the radar's 2D range-Doppler heatmap, and feeding them to an Artificial Neural Network (ANN) as a multi-dimensional parameter set [31]. However, this method takes 0.8 seconds to estimate a gesture, which is too slow for human-robot collaboration. For example, a robot such as the UR10e can move 1 meter in that time, posing a danger to humans.

III. The mmPrivePose3D architecture

There is increasing interest in the use of RaDAR sensors for human pose estimation and gesture recognition. However, the models developed thus far for pose estimation use bulky hardware setups and RaDAR sensors that are unauthorized for industrial usage[5] and often suffer from larger errors in wrists and hands [19]. This makes the detection of hand gestures difficult. In this work, we propose a system called mmPrivPose3D that aims to address both challenges by integrating pose estimation and gesture command recognition modules through a parallel architecture to ensure an acceptable real-life data inference time for human-robot collaboration applications.

The mmPrivPose3D system consists of four building blocks as illustrated in Fig. 1 including data collection, data processing, model selection, and integration.

A. Data collection

In this phase, we simultaneously gathered the 3D point-clouds of humans from an IWR6843AOPEVM RaDAR with 10 Hz data frequency and an RGBD image from an Intel RealSense L515 camera[6] with a 30 frame rate per second (fps) as the ground truth in real-time, see Fig. 1A. This camera was used as the ground truth to enable accurate comparisons with previous work, which also used an RGBD camera as the ground truth [25]. Additionally, these RGBD cameras have a similar level of precision to the Azure Kinect [32]. An offset between the sensors (3.9 cm in both the horizontal and vertical directions) and a 90-degree rotation around the local x-axis were applied to the camera data for spatial calibration and alignment with the RaDAR's coordinate system. In addition, the radar is mounted 1.15 meters above the ground at the base of the collaborative robot (cobot), similar to other sensors, such as SICK laser scanners. Because the cobot operates at a higher level, it does not obstruct the sensor, thereby ensuring proper functionality and full human body coverage in a 120-degree field of view of the sensor in azimuth and elevation. To account for potential misalignment or drift, the experimental setup was

[3] https://perma.cc/6ZRF-P5CK
[4] https://perma.cc/D3YU-9MSL
[5] https://perma.cc/YX33-RVJG
[6] https://perma.cc/96K7-GDC3

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3578094

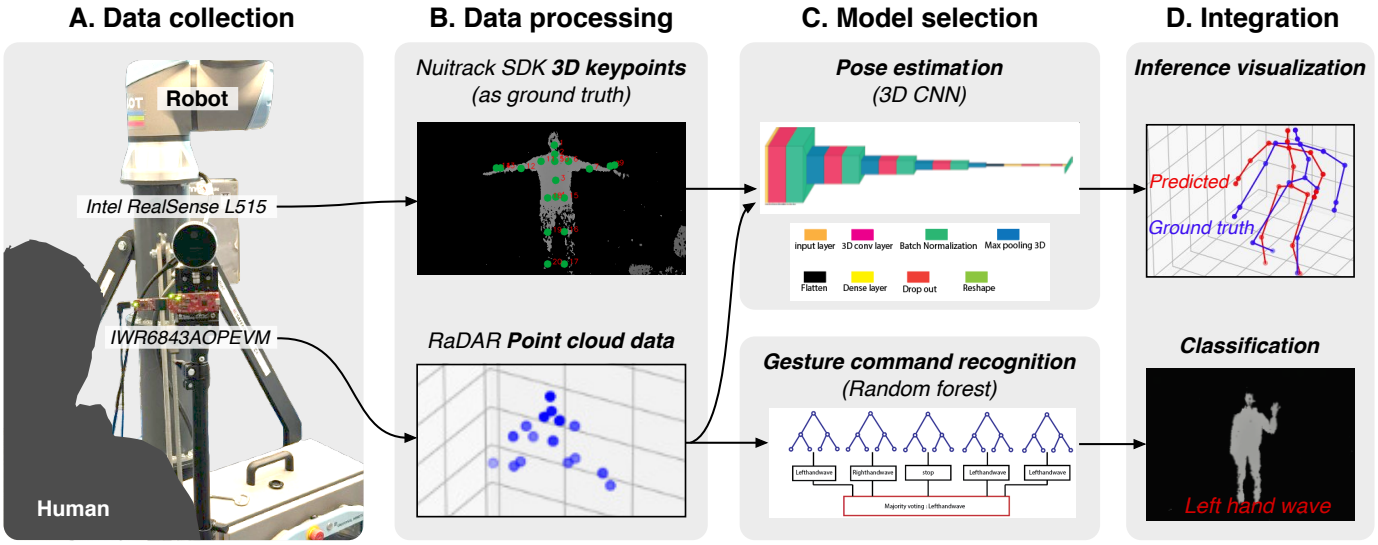


Fig. 1. The mmPrivePose3D system consists of four building blocks namely Data collection, Data processing, a novel model performing pose estimation and gesture command recognition in parallel based on FMCW RaDAR data, and Integration.

regularly inspected to verify that both alignment and mounting distance remained consistent. Periodic recalibration checks were performed using known reference points to maintain accurate sensor alignment and address any shifts or drift [20].

The point-cloud, derived from RaDAR data, results from a signal processing algorithm running on the Digital Signal Processor (DSP) of the sensor's evaluation module. This algorithm performs three Fast Fourier Transforms (FFTs) on the Intermediate Frequency (IF) signal, which is the mix of the transmission and reflected chirps (signals with linear frequency over time), after Analog-to-Digital Converter (ADC) sampling. The RaDAR's signal is at a frequency range of 60 GHz to 64 GHz and transmits new packet data from its DSP every 100 milliseconds.

In mmPrivPose3D, rather than converting RaDAR data into images or directly processing the RaDAR User Datagram Protocol packets in the host computer before feeding them into the model, the data processing codes of the IF signals involving the range FFT, Doppler FFT, and Angle of Arrival (AoA) calculations are directly flashed into the RaDAR DSP[7]. This allows the RaDAR to process the data internally and provides a 3D point-cloud per human, which is the result of clustering reflective points with non-zero doppler using DBSCAN [33], as its output. This approach significantly reduces the overall latency of the system by leveraging on-chip processing instead of working with RAW RaDAR data.

Data collection was conducted in a public environment at the Brubotics Lab, Vrije Universiteit Brussels. We recruited 9 male participants and one female participant, aged between 24 and 37 years, with heights ranging from 1.7 to 1.8 meters, and diverse body sizes. It should be noted that the sensors used in this work are independent of demographic variables such as gender and age, as they only capture the mechanical aspects of movement. The sensors are also not sensitive to height. While this range may not cover all possible heights globally, it is a reasonable height range in Europe. This reflects our practical and logistical conditions rather than any intentional exclusion. RaDAR is resilient to environmental variations and therefore supports the reliability of this dataset across different environmental conditions [20]. The data collection was ethically approved by the Ethics Committee for Human Sciences of the Vrije Universiteit Brussel on 26th April 2024 (ECHW_511). This dataset is part of our published mmPrivPose3D dataset [20]. It features more participants than existing RaDAR-based human pose estimation datasets such as MARS [28] (with only four participants) and HuPR [34] (with six participants), and it utilizes a 60 GHz RaDAR authorized for industrial use rather than a 77 GHz RaDAR which is not authorized. Additionally, it permits free walking across the field of view, unlike datasets such as mmpose-NLP [35] or mmpose-FK [25], which primarily include samples of walking back and forth along a straight line and arm-swing activities. Informed consent for data collection and processing was obtained from all participants involved in the study.

B. Data processing

This section details how the training data is obtained, by combining the frames captured by the RGBD cameras with the frames sensed by the RaDAR. Data from both sensors are transferred to a Linux computer via UART and USB-3 interface. This computer is equipped with an Intel Core i7-13700H CPU and an NVIDIA A2000 Ada GPU. Our multi-threaded Python program implements data acquisition of both sensors concurrently. Both threads are synchronized, capturing data from both sensors with matching timestamps and saving them in a CSV file to form the training dataset.

The camera thread processes the RGBD image of the L515 camera to estimate the human body's 3D keypoints. For this purpose, Nuitrack SDK[8] was used which offers 19 keypoints for human-pose estimation. To address the missing

[7] https://bit.ly/3y9WZXx

[8] https://perma.cc/6BMG-HDAC

regions in the image caused by invalid depth data, a hole-filling filter was applied to the depth image prior to keypoint estimation. These keypoints were used as the ground truth for the mmPrivePose3D pose estimation section. It is important to note that the classical version of this SDK is unable to distinguish between a person facing forward or backward during keypoint identification [36].

Fig. 1B illustrates the human keypoints detected by Nuitrack SDK, derived from the camera's depth image captured alongside the RaDAR point-cloud. The algorithm running on RaDAR DSP encapsulates the processed data into a packet. Our custom algorithm, which operates on a Linux-based local computer, decodes this packet in real-time in the RaDAR thread. This decoding process is crucial for extracting the reflection points from the RaDAR, which are essential for constructing the point-cloud, which is a finite set of 3D coordinates $\{(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)\}$, for each detected human. This data, along with the estimated keypoints from the L515 camera, is passed with the same timestamp into the model.

C. Model selection

1) Pose estimation: For this section of mmPrivePose3D, a 3D CNN algorithm is used. As shown in Fig. 1C-top, the model consists of four 3D convolution layers, four pooling layers, and two fully connected layers. This setup was selected to achieve a balance between inference processing time that has to happen in real-time and prediction accuracy, which is improved by adding more convolution layers. In addition, we use an LSTM layer to consider the temporal dependencies and lack of enough information from human motion in a single RaDAR frame while maintaining real-time inference per frame. This layer aggregates information over 8 consecutive frames. The input 3D point-cloud of RaDAR is voxelized with fixed cube sizes of 32 cm^3 within a fixed grid space of 0.84 m^3 to maintain the spatial relation between the points [37]. Following each pooling layer, the model incorporates a batch normalization layer to enhance the training efficiency.

We chose the Huber loss function [38] for its resilience against outliers, which is a crucial attribute given the sparse nature and outlier susceptibility of RaDAR point-clouds due to environmental reflections. The Huber loss function uniquely combines the elements of MSE and MAE. When the difference between the predicted value and the ground truth is small, the Huber loss behaves like MSE. Conversely, in instances where this error is significantly large owing to the presence of outliers in the data, it takes on the characteristics of MAE. The Huber loss function is employed independently for each 3D coordinate system, as follows:

$$H(Y, Y_G, \delta_i) = \begin{cases} \frac{(Y-Y_G)^2}{2} & \text{if } |Y - Y_G| \leq \delta_i, \\ \delta_i |Y - Y_G| - \frac{1}{2}\delta_i^2 & \text{otherwise,} \end{cases} \quad (1)$$

where δ_i is a hyper-parameter that must be tuned based on the training results for each of the 3 axes, Y_G is the coordinate of the ground truth keypoints in each of the 3 directions, and Y represents these coordinates predicted by the model. In this project, the above loss function was fine-tuned by assigning specific weights to each coordinate ϕ_i, thereby prioritizing the error minimization in all directions. For the x- and y-axes, a higher weight is applied because of the RaDAR's heightened 3D point sparsity in these directions, as opposed to the z-axis. Additionally, keypoints linked to human arms are given increased importance, and separate weights (w_k) are assigned to them for enhanced precision. This is because of the lower reflective capabilities of the arms compared to other body parts, which significantly impacts the 3D point-cloud data derived from the RaDAR [19]. The overall loss calculation involves aggregating the weighted Huber losses for all keypoints and coordinates, as follows:

$$\mathrm{L}(Y, Y_G) = \sum_{k=1}^{19} w_k \times \sum_{i \in \{x,y,z\}} \phi_i \times H\left(Y, Y_G, \delta_i\right) \quad (2)$$

where H represents the Huber loss function applied to each coordinate of the 3D keypoints, w_k is the extra weight applied if keypoint k belongs to a human arm. This model predicts 19 3D keypoints of the human body using the estimated keypoints from the RGBD camera image as the ground truth.

2) Gesture command recognition: A parallel-running random forest classifier was implemented to identify hand gestures to address the limited details captured by the pose estimation model for the human arms and hands. The classifier was trained using two hand gestures, see Fig. 1C-bottom. These two gestures were selected from mmPrivPose3D dataset as: right- and left-hand waves [20]. The dataset can be further expanded to include other gestures that are being performed in human-robot interactions. These gestures were performed across three locations within the field-of-view of the sensor. Consequently, the classifier enhances the capabilities of mmPrivPose3D by providing more detailed information about a human's hand position in parallel which can also be used for tasks such as stopping and starting the robot or triggering another human-robot interaction process.

D. Integration

This block integrates and threads the outputs from the models operating in parallel, delivering 19 keypoints that form the human skeleton, accompanied by a label indicating whether one of the trained gestures is being performed, see Fig. 1D. A visualization environment was developed to visualize and evaluate the output of the parallel models, displaying both the ground truth data (depth image and Nuitrack skeleton) and the model predictions for accuracy assessment.

IV. Results

A. Pose estimation

The dataset for training the 3D CNN model was gathered by allowing each participant to move freely within an area covered by the 70-degree horizontal field of view of both sensors between 2.0 and 4.0 m distance. The motions included walking and freely waving both hands. Participants avoided side poses to prevent body occlusion and faced the sensors throughout. The choice of a minimum distance of two meters

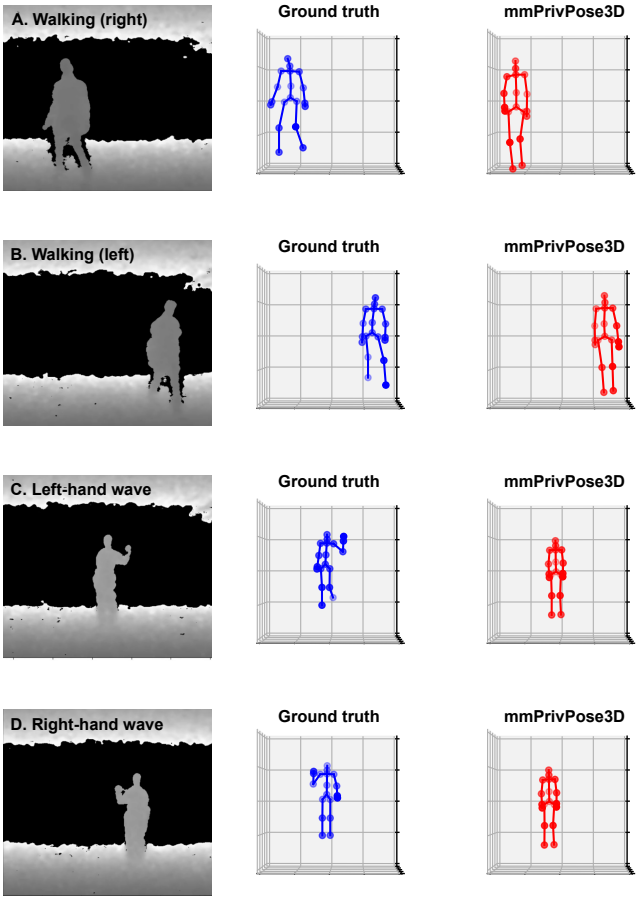


Fig. 2. Results of the pose estimation module of mmPrivPose3D in selected movements compared to the ground truth. Left: L515 ground truth depth image. Center: Ground truth depth map processed by Nuitrack SDK to keypoints. Right: mmPrivPose3D pose estimation. Hand-waving is *not* detected which is addressed through the parallel gesture command recognition module.

was to ensure that human safety was maintained while human-robot interaction as it is above the reach of long collaborative robots such as UR10e (1.3 m) [20]. From 24005 samples of 15 participants in our dataset [20], the sample of ten participants consisting of RaDAR point-clouds and their corresponding 3D keypoints were gathered. 80% of this data was used for training and the rest were utilized as the test set. Owing to the hardware-neutral nature of the input point-cloud data structure, this dataset can also be used to train models running on other mmwave RaDAR sensors. In addition, the dataset covers a wide range of motion as participants were asked to perform random movements such as running or waving while gathering the dataset. The model was trained using a k-fold cross-validation approach to reduce the bias associated with random sampling of the data. Out of the 10 subjects, 9 were used to train the model while the 10^{th} was used for evaluation as an unseen subject. The results are shown in Fig. 2 for selected movements including normal walking and hand waves.

The model was evaluated on the test set to determine the absolute Mean Per Joint Position Error (MPJPE) which measures the L2 distance between the ground truth and the

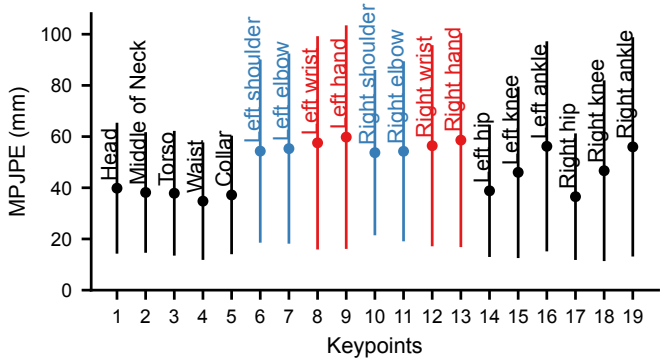


Fig. 3. Mean Per Joint Position Errors (MPJPE) and standard deviations of 19 3D keypoints. The largest errors occurred at keypoints 8, 9, 12, and 13, corresponding to the wrists and hands – highlighted in red. The other arm keypoints are highlighted in blue.

TABLE I
COMPARISON OF MPJPE VALUES OF ALL KEYPOINTS AND SPECIFIC BODY KEYPOINTS AMONG DIFFERENT DEVELOPED MODELS (UNIT: MM).

Methods	Neck	Hip	Wrist	Knee	All
mmPrivPose3D	38.1	**38.8**	**57.5**	**46.6**	**48.3**
RFPose3D [18]	73.7	107.6	159.4	149.5	134.1
RPM [29]	49.0	51.5	65.8	60.5	59.2
RPM 2.0 [40]	**37.0**	47.1	69.4	58.3	57.5

predicted joints in the world coordinates [29]. Fig. 3 shows the MPJPE for each joint. It is noted that the largest errors are associated with keypoints 8, 9, 12 and 13, which are located on the wrists and hands of the left and right arms. This observation is reasonable, considering that this body part has a lower reflective capability [18]. To reduce this error, integration of RaDAR reflectors on the hands was considered. However, workers are generally averse to wearing additional gear since it can interfere with their operations [39], and hence was not considered an option.

We compared the performance of our pose estimation model with previously developed models, that is, RFPose3D [18], RPM [29], and RPM 2.0 [40]. Table I shows the MPJPE values of all the joints among the different models[9]. The mmPrivPose3D achieved the lowest MPJPE across all predicted keypoints. In addition, unlike the other models, our model requires only a single industry-certified 60GHz mmWave RaDAR for operation, which ensures a more compact form factor of the overall setup. Nonetheless, similar to the other models such as RPM 2.0 [40], the wrist and hand keypoints exhibit higher error rates (max. 10 cm). In addition to the low reflective power of these body parts [18], the small size and fast movement of these parts may have contributed to this error rate. Therefore, the pose estimation is insufficient to accurately recognize hand gesture commands, as shown in Fig. 2C-D. This further supports our decision to integrate the gesture command recognition module in parallel with the pose estimation module.

[9]The mmpose-FK model [25] was not included in the evaluation as it was validated based on a simple walking movement along a straight line, rather than the entire sensor field of view as other works.

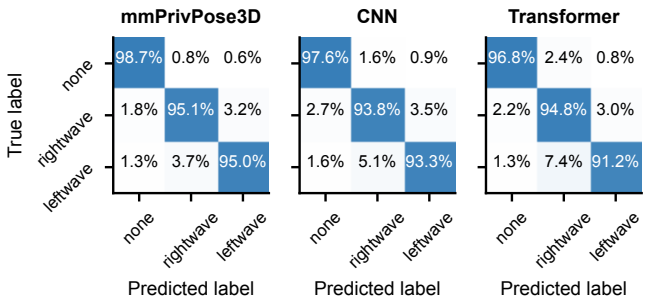


Fig. 4. Confusion matrices for evaluation of the gesture command recognition module, indicating that the mmPrivPose3D model obtained the overall accuracy of 96.2%, compared to 94.8% of CNN and 94.3% of Transformer.

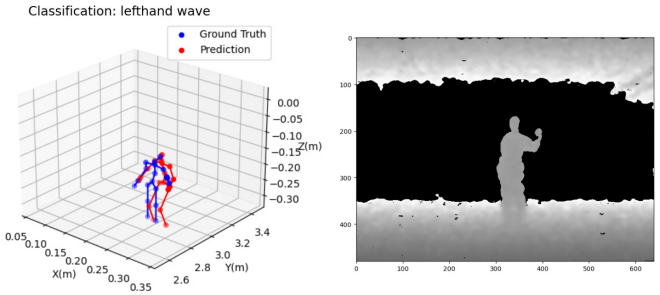


Fig. 5. Visualization of the integrated pose estimation and gesture command recognition results with respect to the ground-truth data from the camera.

B. Gesture command recognition

The dataset used to train the gesture command recognition module of mmPrivPose3D was compiled by having each participant perform two gestures (right-hand and left-hand waves) at different locations within the RaDAR's field of view. Free walking for pose estimation and left/right-hand waves for gesture command recognition were chosen based on their popularity in human-robot collaboration applications [41] and relevant datasets ([27], [16], [34], [42], [29]). Other gestures can be trained using the same method. The dataset was recorded separately from the pose estimation data, with each participant given a 3-minute time window to perform one of the two gestures in various locations. Subsequently, an additional 3 minutes was allotted for the second gesture. Of a total of 86483 samples in mmPrivPose3D dataset [20], 80% of the data from ten participants were utilized for model training with 5 folds, while the remaining 20% were reserved for validation as the 6-th fold through k-fold cross-validation. The mean cross-validation score was 96.3%. The failure cases can be attributed to the variation in the speed of performing these two gestures by different participants.

We compared the performance of mmPrivPose3D with a set-transformer model [43] and a CNN model, see Fig. 4. Our model achieved an inference accuracy of 96.2%, higher than the other two models. The CNN model, consisting of three layers of convolution, pooling, and fully connected, achieved 94.9% accuracy. The set-transformer, employing an encoder-decoder framework and multi-head self-attention, achieved 94.3% accuracy.

C. Inference

After training the modules for pose estimation and gesture command recognition, we ran them concurrently in two parallel threads for real-time processing of the RaDAR data with gesture command recognition being performed continuously using a sliding window approach with a window size of 4 frames to have a balance between the inference time and accuracy. Furthermore, we introduced a third thread to align the RaDAR data with the ground-truth data of the camera, which operates at a different frequency. The outputs from these models were integrated into our visualization environment along with the ground truth data of the camera, see Fig. 5. To assess the real-time effectiveness of our models, we measured their inference times: the pose estimation module recorded a 20 ms inference time, and the gesture command recognition module recorded 12 ms. Given the RaDAR data frequency of 10 Hz, the cumulative inference time of these modules (31 fps) affirms their real-time operational capability as it can provide the inference result in the timing window between two consecutive RaDAR detection frames.

V. Conclusion

In this study, we introduce mmPrivPose3D, a human pose estimation and gesture command recognition system with a parallel architecture using an FMCW RaDAR for privacy compliance. The model uses two parallel modules. The pose-estimation module is capable of estimating 19 keypoints. This model demonstrated the lowest MPJPE of 48.3 mm for all keypoints compared to prior techniques, although it exhibited higher errors in the wrist and arm keypoints. To compensate for this drawback, the mmPrivPose3D system incorporates a parallel gesture command recognition module to classify human gestures more accurately. This model achieved a 96.2% accuracy rate in recognizing two representative gestures, that is, right-hand and left-hand waves, and the dataset can be expanded to other gestures. Future developments include expanding the mmPrivPose3D dataset [20] to encompass a broader range of gestures, using a Spiking Neural Network (SNN) as an alternative for the parallel gesture command recognition module to improve the accuracy of gesture recognition, reduce the inference time, and exploring 140 GHz RaDAR which has higher angular and range resolution owing to a higher frequency and bandwidth [44]. In addition, the technique of forward kinematics used in mmpose-FK [25] can be utilized on top of the pose estimation module to provide more accurate real-time information regarding a human's 3D position and performed gestures.

Declaration

Conflict of interest. The work presented in this manuscript is related to a patent request ID 2024/248 (imec).

Data availability. The dataset is available at https://doi.org/10.17632/pmdr5rgn8c.1.

Ethical Statement. This research was conducted in full accordance with the Declaration of Helsinki and approved by the

Ethics Committee for Human Sciences of the Vrije Universiteit Brussel on 26th April 2024 (ECHW_511). Informed consent for data collection and processing was obtained from all individual participants involved in the study.

Author's Contribution. All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Nima Roshandel. The first draft of the manuscript was written by Nima Roshandel, Constantin Scholz and Hoang-Long Cao. All authors commented on previous versions of the manuscript. All authors have read and approved the final manuscript. The funding was acquired by Constantin Scholz, Bram Vanderborght and Jan Genoe.

REFERENCES

[1] R. Raffik, R. R. Sathya, V. Vaishali, S. Balavedhaa, *et al.*, "Industry 5.0: Enhancing human-robot collaboration through collaborative robots–a review," in *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pp. 1–6, IEEE, 2023.

[2] M. H. Zafar, E. F. Langås, and F. Sanfilippo, "Exploring the synergies between collaborative robotics, digital twins, augmentation, and industry 5.0 for smart manufacturing: A state-of-the-art review," *Robotics and Computer-Integrated Manufacturing*, vol. 89, p. 102769, 2024.

[3] C. Scholz, H.-L. Cao, I. El Makrini, and B. Vanderborght, "Antropo: An open-source platform to increase the anthropomorphism of the franka emika collaborative robot arm," *Plos one*, vol. 18, no. 10, p. e0292078, 2023.

[4] M. Doyle-Kent and P. Kopacek, "Collaborative robotics making a difference in the global pandemic," in *Digitizing Production Systems: Selected Papers from ISPR2021, October 07-09, 2021 Online, Turkey*, pp. 161–169, Springer, 2022.

[5] C. Scholz, H.-L. Cao, I. El Makrini, S. Niehaus, M. Kaufmann, D. Cheyns, N. Roshandel, A. Burkiewicz, M. Shhaitly, E. Imrith, *et al.*, "Improving robot-to-human communication using flexible display technology as a robotic-skin-interface: a co-design study," *International Journal of Intelligent Robotics and Applications*, pp. 1–18, 2024.

[6] E. Matheson, R. Minto, E. G. Zampieri, M. Faccio, and G. Rosati, "Human–robot collaboration in manufacturing applications: A review," *Robotics*, vol. 8, no. 4, p. 100, 2019.

[7] P. Chemweno, L. Pintelon, and W. Decre, "Orienting safety assurance with outcomes of hazard analysis and risk assessment: A review of the iso 15066 standard for collaborative robot systems," *Safety Science*, vol. 129, p. 104832, 2020.

[8] D. Pascual-Hernández, N. O. de Frutos, I. Mora-Jiménez, and J. M. Canas-Plaza, "Efficient 3d human pose estimation from rgbd sensors," *Displays*, vol. 74, p. 102225, 2022.

[9] M. Slembrouck, H. Luong, J. Gerlo, K. Schütte, D. Van Cauwelaert, D. De Clercq, B. Vanwanseele, P. Veelaert, and W. Philips, "Multiview 3d markerless human pose estimation from openpose skeletons," in *Advanced Concepts for Intelligent Vision Systems: 20th International Conference, ACIVS 2020, Auckland, New Zealand, February 10–14, 2020, Proceedings 20*, pp. 166–178, Springer, 2020.

[10] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8877–8886, 2023.

[11] W. Ma, K. Wang, J. Li, S. X. Yang, J. Li, L. Song, and Q. Li, "Infrared and visible image fusion technology and application: A review," *Sensors*, vol. 23, no. 2, p. 599, 2023.

[12] C. Scholz, H.-L. Cao, E. Imrith, N. Roshandel, H. Firouzipouyaei, A. Burkiewicz, M. Amighi, S. Menet, D. W. Sisavath, A. Paolillo, X. Rottenberg, P. Gerets, D. Cheyns, M. Dahlem, I. Ocket, J. Genoe, K. Philips, B. Stoffelen, J. Van den Bosch, S. Latre, and B. Vanderborght, "Sensor-enabled safety systems for human–robot collaboration: A review," *IEEE Sensors Journal*, vol. 25, no. 1, pp. 65–88, 2025.

[13] J.-T. Huang, C.-L. Lu, P.-K. Chang, C.-I. Huang, C.-C. Hsu, Z. L. Ewe, P.-J. Huang, and H.-C. Wang, "Cross-modal contrastive learning of representations for navigation using lightweight, low-cost millimeter wave radar for adverse environmental conditions," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3333–3340, 2021.

[14] C. Zhu, Z. Zhao, Z. Shan, L. Yang, S. Ji, Z. Yang, and Z. Zhang, "Robust target detection of intelligent integrated optical camera and mmwave radar system," *Digital Signal Processing*, vol. 145, p. 104336, 2024.

[15] B. Ubezio, C. Schöffmann, L. Wohlhart, S. Mülbacher-Karrer, H. Zangl, and M. Hofbaur, "Radar based target tracking and classification for efficient robot speed control in fenceless environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 799–806, IEEE, 2021.

[16] A. Sengupta and S. Cao, "mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[17] M. Strobel, S. Schoenfeldt, and J. Daugalas, "Gesture recognition for fmcw radar on the edge," in *2024 IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNeT)*, pp. 45–48, 2024.

[18] 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 *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pp. 267–281, 2018.

[19] S. H. Javadi, A. Bourdoux, N. Deligiannis, and H. Sahli, "Human pose estimation based on isar and deep learning," *IEEE Sensors Journal*, 2024.

[20] N. Roshandel, C. Scholz, H.-L. Cao, M. Amighi, H. Firouzipouyaei, A. Burkiewicz, S. Menet, F. Ballen-Moreno, D. W. Sisavath, E. Imrith, *et al.*, "mmprivpose3d: A dataset for pose estimation and gesture command recognition in humanrobot collaboration using frequency modulated continuous wave 60hhz radar," *Data in Brief*, p. 111316, 2025.

[21] I. El Makrini, S. A. Elprama, J. Van den Bergh, B. Vanderborght, A.-J. Knevels, C. I. Jewell, F. Stals, G. De Coppel, I. Ravyse, J. Potargent, *et al.*, "Working with walt: How a cobot was developed and inserted on an auto assembly line," *IEEE Robotics & Automation Magazine*, vol. 25, no. 2, pp. 51–58, 2018.

[22] H.-L. Cao, C. Scholz, J. De Winter, I. E. Makrini, and B. Vanderborght, "Investigating the role of multi-modal social cues in human-robot collaboration in industrial settings," *International Journal of Social Robotics*, vol. 15, no. 7, pp. 1169–1179, 2023.

[23] D. Osokin, "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose," *arXiv preprint arXiv:1811.12004*, 2018.

[24] J.-W. Kim, J.-Y. Choi, E.-J. Ha, and J.-H. Choi, "Human pose estimation using mediapipe pose and optimization method based on a humanoid model," *Applied sciences*, vol. 13, no. 4, p. 2700, 2023.

[25] S. Hu, S. Cao, N. Toosizadeh, J. Barton, M. G. Hector, and M. J. Fain, "mmpose-fk: A forward kinematics approach to dynamic skeletal pose estimation using mmwave radars," *IEEE Sensors Journal*, 2024.

[26] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3d human pose construction using wifi," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pp. 1–14, 2020.

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

[28] S. An and U. Y. Ogras, "Mars: mmwave-based assistive rehabilitation system for smart healthcare," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 20, no. 5s, pp. 1–22, 2021.

[29] C. Xie, D. Zhang, Z. Wu, C. Yu, Y. Hu, and Y. Chen, "Rpm: Rf-based pose machines," *IEEE Transactions on Multimedia*, 2023.

[30] H. Cappelle, A. G. Daronkolaei, J. Tsang, B. Debaillie, and I. Ocket, "Radar-based human-robot interfaces," in *Artificial Intelligence for Digitising Industry–Applications*, pp. 221–237, River Publishers, 2022.

[31] Q. Li, L. Liu, S. Hao, and G. Wan, "Dynamic gesture recognition method based on millimeter-wave radar," in *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pp. 63–67, IEEE, 2022.

[32] A. Lopez Paredes, Q. Song, and M. H. Conde, "Performance evaluation of state-of-the-art high-resolution time-of-flight cameras," *IEEE Sensors Journal*, vol. 23, no. 12, pp. 13711–13727, 2023.

[33] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, pp. 232–238, 2014.

[34] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5715–5724, 2023.

[35] A. Sengupta and S. Cao, "mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars,"

IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 8418–8429, 2023.

[36] R. Li, W. Si, M. Weinmann, and R. Klein, "Constraint-based optimized human skeleton extraction from single-depth camera," *Sensors*, vol. 19, no. 11, p. 2604, 2019.

[37] Y. Xu, X. Tong, and U. Stilla, "Voxel-based representation of 3d point clouds: Methods, applications, and its potential use in the construction industry," *Automation in Construction*, vol. 126, p. 103675, 2021.

[38] W. Ding, Z. Cao, J. Zhang, R. Chen, X. Guo, and G. Wang, "Radar-based 3d human skeleton estimation by kinematic constrained learning," *IEEE Sensors Journal*, vol. 21, no. 20, pp. 23174–23184, 2021.

[39] M. Javdan, M. Ghasemaghaei, and M. Abouzahra, "Psychological barriers of using wearable devices by seniors: A mixed-methods study," *Computers in Human Behavior*, vol. 141, p. 107615, 2023.

[40] C. Xie, D. Zhang, Z. Wu, C. Yu, Y. Hu, and Y. Chen, "Rpm 2.0: Rf-based pose machines for multi-person 3d pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[41] R. C. Hsu, P.-C. Su, J.-L. Hsu, and C.-Y. Wang, "Real-time interaction system of human-robot with hand gestures," in *2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pp. 396–398, 2020.

[42] C. Shi, L. Lu, J. Liu, Y. Wang, Y. Chen, and J. Yu, "mpose: Environment- and subject-agnostic 3d skeleton posture reconstruction leveraging a single mmwave device," *Smart Health*, vol. 23, p. 100228, 2022.

[43] G. Li, L. Guo, R. Zhang, J. Qian, and S. Gao, "Transgait: Multimodal-based gait recognition with set transformer," *Applied Intelligence*, vol. 53, no. 2, pp. 1535–1547, 2023.

[44] A. Visweswaran, K. Vaesen, M. Glassee, A. Kankuppe, S. Sinha, C. Desset, T. Gielen, A. Bourdoux, and P. Wambacq, "A 28-nm-cmos based 145-ghz fmcw radar: System, circuits, and characterization," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 7, pp. 1975–1993, 2021.