Microwave Engineering
Ulm University

universität ulm

Learning on Multistatic Simulation Data for Radar-Based Automotive Gesture Recognition

Nicolai Kern, Julian Aguilar, Timo Grebner, Benedikt Meinecke, Christian Waldschmidt

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
DOI: 10.1109/TMTT.2022.3200595

Learning on Multistatic Simulation Data for Radar-based Automotive Gesture Recognition

Nicolai Kern, *Graduate Student Member, IEEE*, Julian Aguilar, Timo Grebner, *Graduate Student Member, IEEE*, Benedikt Meinecke, *Graduate Student Member, IEEE*, Christian Waldschmidt, *Fellow, IEEE*

***Abstract*—Radar-based gesture recognition can play a vital role in autonomous vehicles' interaction with vulnerable road users (VRUs). However, in automotive scenarios the same gesture produces strongly differing radar responses owed to the wide range of variations such as position, orientation, or ego-motion. Since including all kinds of modifications in a measured dataset is laborious, gesture simulations alleviate the measurement effort and increase the robustness against edge and corner cases. Hence, this paper presents a flexible geometric human target model allowing the direct introduction of a wide range of modifications, while it facilitates the handling of shadowing effects and multi-radar constellations. Using the proposed simulation model, a dataset recorded with a radar sensor network consisting of three chirp sequence (CS) radars is resimulated based on motion data simultaneously captured with a stereo video system. Completely substituting the measured by simulated data for training, a convolutional neural network (CNN) classifier still achieves 80.4 % cross-validation accuracy on a challenging gesture set, compared to 89.4 % for training on measured data. Moreover, using simulated data the classifier is shown to successfully generalize to new scenarios not observed in measurements.**

***Index Terms*—automotive radar, CNN, data augmentation, gesture recognition, human target simulation, radar sensor networks**

I. Introduction

With autonomous driving entering urban areas, autonomous vehicles (AVs) increasingly share roads with non-motorized vulnerable road user (VRU). Thus, AVs will be required to engage in bidirectional communication in order to manage complex scenarios safely [1]. To react properly, they have to sense the pedestrian intent [2], e.g. by evaluating cues such as gaze, VRU trajectories, or communication gestures [3]. Since gestures are the most explicit form of communication, they might be used to transmit messages of special importance, such as a stop signal. As a consequence, they have to be recognized with high reliability, independent of the circumstances. While camera-based algorithms can do the job of gesture recognition [4], they might be adversely affected by unfavorable environmental conditions. Radar sensors, in contrast, are known to be very robust with respect to lighting and weather, which makes them a good candidate to be part of an automotive gesture recognition system. In recent years, radar-based gesture recognition has experienced increasing attention mainly in the consumer electronics area, where it enables touchless control of devices and human-machine interaction [5]. A wide range of neural network architectures has been proposed for the task, including 2D-CNNs [6], 2D-CNNs in conjunction with recurrent models such as long short-term memorys (LSTMs) [7], or 3D-CNNs together with LSTMs [8]. Related tasks also involve machine learning algorithms for activity recognition [9], [10] and fall detection [11] based on radar data.

Despite these successes, a common issue in machine learning for radar applications remains the provision of comprehensive datasets for neural network training. Datasets need to reflect the variety with which gestures are performed in order to enable a good generalization to persons not included in the training data. While sufficient data diversity might be obtained from measurements in the field of consumer electronics, the requirement becomes difficult to fulfill in case of gesture recognition in automotive scenarios. Here, the gesture's radar response is strongly influenced by a wide range of modalities: Pedestrians can perform gestures at widely different positions with respect to the vehicle, under various orientations, and with different degrees of occlusion. In addition, the car's ego-motion has an impact on the radar response, too. Covering all variations to ensure a gesture classification algorithm can generalize to unseen modalities creates a tremendous measurement effort, such that the generation of artificial data is tempting. Data augmentation from synthesized radar data relaxes measurement requirements and exposes the classifier to modalities otherwise unseen during training. However, it has to be ensured that the artificial radar responses are good replicas of real measurements and that their quality suffices for the aim of data augmentation.

Methods for radar data synthesis have been in the focus of research for several years (cf. Table I). One approach for the generation of Doppler spectrograms that represent velocity information over time are generative adversarial networks (GANs) [12], [13]. By learning the underlying distribution of a small set of spectrograms, they are able to generate similar new data by drawing samples from it, which can be used for pre-training or direct augmentation of measurement data. However, as synthesized samples are limited to the training set distribution, they might not cover a wide range of modalities.

A second approach enabling the expansion to completely new modalities is the human target simulation. Here, a human is decomposed into a set of discrete scattering centers from which a radar response is synthesized, e.g. using a continuous wave (CW) or chirp sequence (CS) radar signal model. Approaches can be distinguished with respect to the underlying human target model, which can be based on a set of char-

All authors are with the Institute of Microwave Engineering, Ulm University, Ulm, Germany.

TABLE I
OVERVIEW OVER EXISTING APPROACHES FOR RADAR DATA SYNTHESIS

Ref.	Method	Motion Data	DA[1]	PT[2]	CD[3]
[12]	GAN	-	✓	✓	✗
[13]	GAN	-	✗	✗	✓
[14]	Simulation	MoCap	✗	✗	✗
[16], [17]	Simulation	Kinect	✗	✗	✗
[15], [21]	Simulation	Kinect	✗	✓	✗
[18], [19]	Simulation	Blender	✗	✗	✗
[20]	Simulation	Blender	✗	✗	✓
[23]	Domain Transfer	Video (Mono)	✓	✗	✓
This work	Simulation	Video (Stereo)	✗	✗	✓

[1]DA: direct augmentation, [2]PT: pre-training, [3]CD: cross-domain training

acteristic skeletal keypoints, whose positions are derived from motion capture data [14] or Kinect sensors [15], [16], [17], and models that derive scattering centers from 3D body models generated e.g. with help of computer graphics software [18], [19], [20]. Starting from these simulation approaches, some papers recently explored the potential of simulated radar data for augmentation. In [21], the authors developed a human target model based on the keypoints of the Kinect sensor to enhance activity classification from Doppler spectrograms with a convolutional neural network (CNN). By pretraining on 32,000 samples from diversified motion data and fine-tuning on a small dataset of 474 measured signatures, they realized deeper nets and higher accuracy compared to training on the measured data alone. Contrary to pretraining, the direct comparison between real and synthetic data for neural network training has been investigated as a means to assess the quality of the generated radar responses: In [20], the radar responses for seven gestures are generated from 3D Blender models, with the motion modelled only by the start and end point. After diversification of the motion data, 600 responses are simulated with the scattering centers obtained from the 3D surface mesh. By training on data from the simulation domain only, measured gestures can be distinguished with an accuracy of 84.2 %. By comparison, training on measurement data results in a classification accuracy of 94.3 % [22], however, for an extended gesture set with nine gestures. A distinctly different approach that mostly circumvents human target modeling has been published in [23]. Instead of computing radar responses from a target and an analytical radar signal model, the authors construct Doppler spectrograms from the radial velocities of the vertices of a 3D mesh, directly obtained from video footage by using a neural network. As the initial spectrograms significantly differ from measured ones, the authors train an encoder-decoder model for domain transfer on Doppler data captured synchronously with camera measurements. Generating Doppler spectrograms in this way, they obtain 81.4 % activity classification accuracy on measured data when training on simulation data only, compared to 90.2 % for training on real data.

In this paper, the challenge of creating diverse synthetic radar data for gesture recognition is tackled by developing a flexible human target model that produces high-fidelity radar data without the need for domain transfer. The pre-

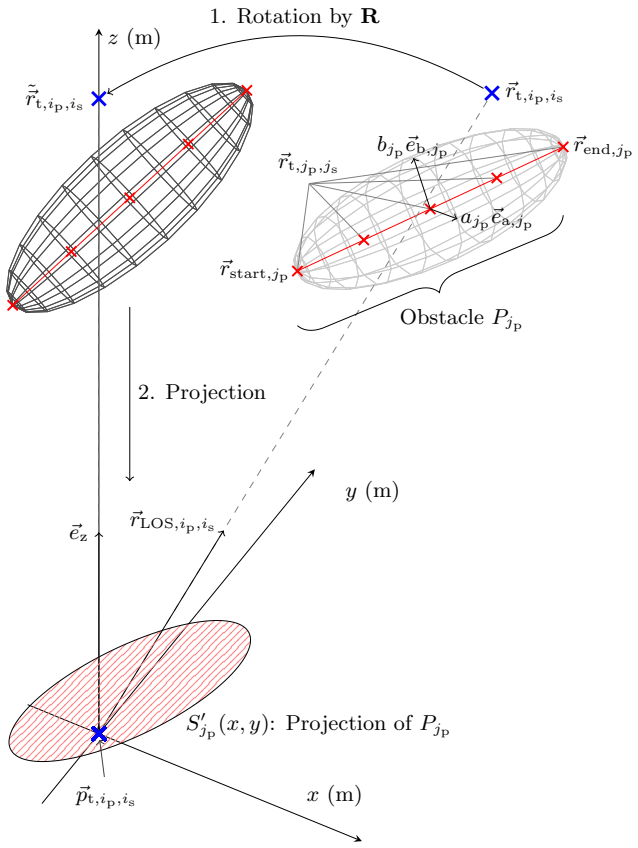

Fig. 1. Visualization of the binary shadowing computation for a point scatterer $\vec{r}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}}$ and an obstacle P_j: As the projection of the point scatterer $\vec{p}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}}$ lies inside the ellipsoid projection $S'_{j_\mathrm{p}}(x,y)$, the scatterer is occluded.

sented geometric model is particularly suited for automotive scenarios, as it enables a wide range of modifications to measurement modalities such as position or orientation at simulation level. Thus, the model facilitates the synthesis of arbitrary scenarios and radar constellations. Compared to other work, the proposed target model provides fine control over the granularity and takes into account shadowing effects due to other body parts. This is shown to be crucial when dealing with different orientations. For the assessment of the target model, the direct resimulation of a comprehensive radar dataset based on synchronized stereo video data is proposed. With the novel one-to-one resimulation approach, the simulation data quality can be assessed both by visual comparison as well as by direct replacement of measurement data during classifier training. Thereby, the potential of the presented model for the mastering of distinctly new scenarios is clearly demonstrated.

The paper is organized as follows: Section II introduces the pedestrian target model, and Section III describes the motion capture with the stereo camera system. Subsequently, Section IV presents the experimental setup used for the assessment of the simulation data in Section V and Section VI. Finally, Section VII concludes the paper.

II. ADVANCED RADAR SIMULATION MODEL

A. Pedestrian Target Model

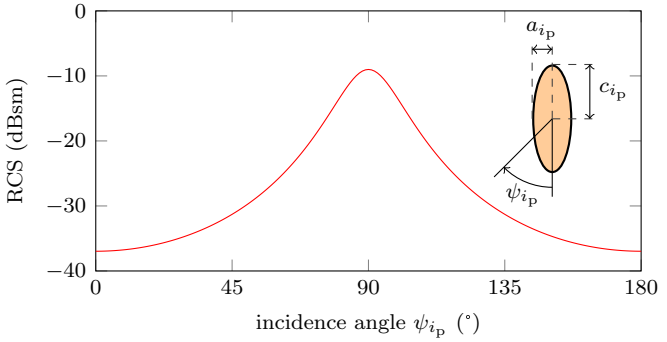

Fig. 2. Definition of the incidence angle ψ_{i_p} and illustration of the RCS value over the angle for an exemplary ellipsoid with $a_{i_\mathrm{p}} = 0.04\,\mathrm{m}$ and $c_{i_\mathrm{p}} = 0.2\,\mathrm{m}$.

1) Body Part Modelling: In order to enable the simulation of human targets based on a geometric model, the human body is decomposed into a set of body parts, from which a set of point scatterers suitable for radar simulation is derived. A simple ellipsoidal shape is assigned to each body part based on which the approximate RCS is computed. Thus, the pedestrian target model consists of N_p ellipsoids

$$P_{i_\mathrm{p}} = \{\vec{r}_{\mathrm{start},i_\mathrm{p}}, \vec{r}_{\mathrm{end},i_\mathrm{p}}, a_{i_\mathrm{p}}, b_{i_\mathrm{p}}, \vec{e}_{\mathrm{a},i_\mathrm{p}}, \vec{e}_{\mathrm{b},i_\mathrm{p}}\},\ i_\mathrm{p} \in \{1, ..., N_\mathrm{p}\}\,, \quad (1)$$

each defined by its start point $\vec{r}_{\mathrm{start},i_\mathrm{p}}$ and end point $\vec{r}_{\mathrm{end},i_\mathrm{p}}$, its minor semi-axis lengths a_{i_p} and b_{i_p}, and the unit vectors $\vec{e}_{\mathrm{a},i_\mathrm{p}}$ and $\vec{e}_{\mathrm{b},i_\mathrm{p}}$ describing the minor semi-axes. An exemplary ellipsoid with its parameters is shown in Fig. 1. The length of the body part is determined by the length of the major semi-axis $c_{i_\mathrm{p}} = |\vec{r}_{\mathrm{end},i_\mathrm{p}} - \vec{r}_{\mathrm{start},i_\mathrm{p}}|/2$. Since the ellipsoid description of body parts is only an approximation with some inevitable deviations from the actual physiognomy of real human targets, the model can be simplified by assuming $a_{i_\mathrm{p}} = b_{i_\mathrm{p}}$ for all body parts. Thus, the ellipsoids become prolate spheroids, and their RCS can be computed as [24]

$$\sigma_{i_\mathrm{p}} = \frac{\pi a_{i_\mathrm{p}}^4 c_{i_\mathrm{p}}^2}{\left(a_{i_\mathrm{p}}^2 \sin^2(\psi_{i_\mathrm{p}}) + c_{i_\mathrm{p}}^2 \cos^2(\psi_{i_\mathrm{p}})\right)^2}\,, \quad (2)$$

with ψ_{i_p} describing the incidence angle between the wave vector of the incident wave, $\vec{k}$, and the major semi-axis. The definition of the incidence angle is visualized in Fig. 2, along with the angle dependency of an exemplary ellipsoid's RCS. The positions of the ellipsoid scatterers are set to the centroids $\vec{r}_{\mathrm{c},i_\mathrm{p}}$ of the objects, i.e., $\vec{r}_{\mathrm{t},i_\mathrm{p}} = \vec{r}_{\mathrm{c},i_\mathrm{p}} = \left(\vec{r}_{\mathrm{start},i_\mathrm{p}} + \vec{r}_{\mathrm{end},i_\mathrm{p}}\right)/2$. The full target model is shown for two exemplary gestures in Fig. 3. Besides the body keypoints, the hands are represented by a refined model with additional parts between wrist and middle finger base, middle finger base and middle finger tip, and between wrist and thumb. The tips of wrist and middle finger are modeled as additional small scatterers.

2) Body Part Subdivision: While the ellipsoid model provides an approximation for the RCS of the body parts, it doesn't take into account the range and velocity spread over the body parts. As a result, the power scattered back is centered in a single range-Doppler cell. In order to approximate the spread in a geometric target model, ellipsoid targets can be

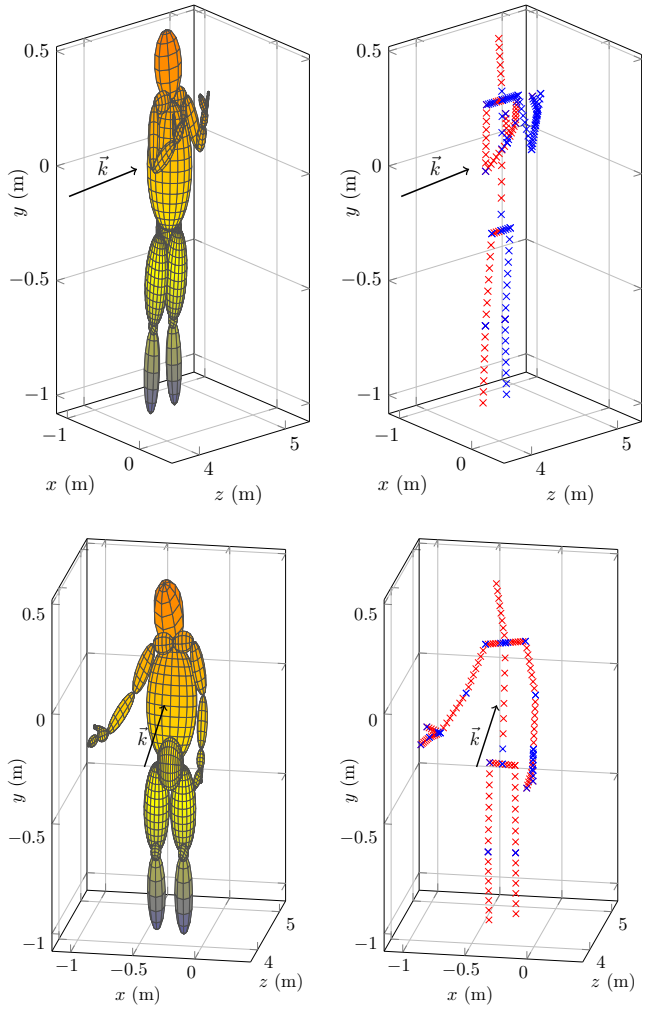

Fig. 3. Model consisting of ellipsoids (left) and point scatterers (right) for the "Come closer" and "Wave through" gesture. The colors in the right column indicate the value of the shadowing factors: red scatterers are visible to a radar sensor at $[0, 0, 0]$, while blue ones are occluded by other body parts. In addition, the wave vector $\vec{k}$ shows the direction of the incident waves.

subdivided into multiple point scatterers to reflect the actual extension as illustrated for part P_j in Fig. 1. Then, a spheroid described by $\{\vec{r}_{\mathrm{start},i_\mathrm{p}}, \vec{r}_{\mathrm{end},i_\mathrm{p}}, a_{i_\mathrm{p}}, \vec{e}_{\mathrm{a},i_\mathrm{p}}, \vec{e}_{\mathrm{b},i_\mathrm{p}}\}$ is split into $N_{\mathrm{sub},i_\mathrm{p}}$ scatterers located at

$$\vec{r}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}} = \frac{i_\mathrm{s} - 1}{N_{\mathrm{sub},i_\mathrm{p}} - 1}\vec{c}_{i_\mathrm{p}} + \vec{r}_{\mathrm{start},i_\mathrm{p}},\ i_\mathrm{s} \in \{1, ..., N_{\mathrm{sub},i_\mathrm{p}}\}\,, \quad (3)$$

with the vector $\vec{c}_{i_\mathrm{p}}$ describing the ellipsoid's major axis. The RCS value of the ellipsoid target is distributed equally over the new point scatterers. The granularity of the subdivision can differ between the parts depending on their relevance for the motion signature, or it can be skipped altogether, keeping the ellipsoidal target model. For the simulations of gestures without locomotion, the subdivision process focuses on the parts of the upper body, which exhibit the most significant motion patterns. In contrast, body parts beneath the hips are kept as ellipsoids, thereby reducing the computational complexity of the simulations. The subdivision step produces a total of $N_\mathrm{target} = \sum_{i_\mathrm{p}=1}^{N_\mathrm{p}} N_{\mathrm{sub},i_\mathrm{p}}$ scatterers. Fig. 3 visualizes

the subdivision of the ellipsoidal model for two exemplary gestures. Note that for better visualization, the number of point scatterers per body part is set higher than in the actual simulation, where up to nine point scatterers are computed for the moving parts of the upper body depending on the body part size. In contrast, the lower body is not subdivided at all.

3) Geometric Point Scatterer Shadowing: The visibility of the scattering centers is an important factor when applying transformations such as rotations or translations. Therefore, a binary geometric shadowing model is introduced for the derived point scatterers $\vec{r}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}}$. Basically, the visibility of the point targets is computed by checking whether any of the ellipsoid body parts constituting the pedestrian target model blocks the line-of-sight between the radar and the scatterer, as shown in Fig. 1. Note that the mathematical description of the model within this section assumes the radar to be located in the center of the coordinate system and aligned with its axes. The expansion to the more general case of multiple arbitrarily positioned and oriented radars is addressed in Section III-B.

For a given point scatterer at $\vec{r}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}}$, a binary shadowing factor $s^{j_\mathrm{p}}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}} \in \{0,1\}$ is computed for each body part P_{j_p} that fulfills $|\vec{r}_{\mathrm{c},j_\mathrm{p}}| < |\vec{r}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}}|$, i.e., only body parts closer to the radar than the point scatterer under investigation can be reasonable obstacles. The line-of-sight (the direction from the radar to the point scatterer) is described by the unit vector $\vec{r}_{\mathrm{LOS},i_\mathrm{p},i_\mathrm{s}} = \vec{r}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}} / |\vec{r}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}}|$. An analytical way of determining whether the line-of-sight is blocked by an ellipsoidal obstacle is to project both the point scatterer and the ellipsoid onto a plane perpendicular to the direction of propagation, i.e. with a normal vector $\vec{r}_{\mathrm{LOS},i_\mathrm{p},i_\mathrm{s}}$. The projection onto a plane is an orthographic projection, and therefore, the modelled shadowing deviates from the actual radar perspective particularly for small distances. However, since the minimum ranges for automotive gesture recognition are limited by the radar's elevation beam-width, this error is sufficiently small. After projection, the question whether the ellipsoid occludes the scatterer reduced to the question whether the projected point lies within the projection of the ellipsoid. Since the projection becomes especially simple if the direction of propagation is along one of the coordinate axes, the whole scene is rotated such that the point scatterer under investigation lies on the z-axis, i.e., $\vec{r}_{\mathrm{LOS},i_\mathrm{p},i_\mathrm{s}} = \vec{e}_\mathrm{z}$. In this case, the projection plane is the x-y-plane and the projection of the point scatterer is just the origin of this plane.

For this scene rotation (omitting point indices for the sake of clarity), a unit vector

$$\vec{v} = [v_1, v_2, v_3]^T = \frac{\vec{r}_\mathrm{LOS} \times \vec{e}_\mathrm{z}}{|\vec{r}_\mathrm{LOS} \times \vec{e}_\mathrm{z}|} \tag{4}$$

orthogonal to both the old and the new direction of propagation can be used as axis of rotation. Since both $\vec{r}_\mathrm{LOS}$ and $\vec{e}_\mathrm{z}$ are unit vectors, the rotation angle α is computed as $\alpha = \arcsin(\vec{r}_\mathrm{LOS} \cdot \vec{e}_\mathrm{z})$. From the axis of rotation and the angle, the matrix $\mathbf{R}$ that rotates $\vec{r}_\mathrm{LOS}$ onto the z-axis is obtained from Rodrigues' rotation formula as

$$\mathbf{R} = \mathbf{I} + \sin\alpha\, \mathbf{v}_\times + (1 - \cos\alpha)\, \mathbf{v}_\times^2 \,, \tag{5}$$

with the identity matrix $\mathbf{I} \in \mathbb{R}^{3\times3}$ and the matrix

$$\mathbf{v}_\times = \begin{bmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{bmatrix} . \tag{6}$$

Now, the point scatterer is rotated onto the z-axis by $\tilde{\vec{r}} = \mathbf{R}\vec{r}$, and the body part ellipsoid is rotated e.g. by applying $\mathbf{R}$ to its centroid vector and semi-axis unit vectors to obtain $\tilde{\vec{r}}$, $\tilde{\vec{e}}_\mathrm{a}$, $\tilde{\vec{e}}_\mathrm{b}$, $\tilde{\vec{e}}_\mathrm{c}$. While the projection of $\tilde{\vec{r}}$ onto the x-y-plane is just $\vec{p} = [0,0]^T$, the projection of the ellipsoid is obtained from its quadric. Computing the quadric from the semi-axis vectors and lengths, its surface can be expressed as

$$\begin{aligned} S(x,y,z) = Ax'^2 + By'^2 + Cz'^2 + 2Dx'y' \\ + 2Ex'z' + 2Fy'z' - 1 \,, \end{aligned} \tag{7}$$

with all points fulfilling the equation lying on the surface and the coordinates $x' = x - x_\mathrm{c}$, $y' = y - y_\mathrm{c}$, $z' = z - z_\mathrm{c}$ normalized w.r.t. the centroid. From (7), the projection is obtained by finding the points on the surface whose tangent planes are orthogonal to the x-y-plane, and therefore have a normal vector $\vec{n} = \nabla S(x,y,z)$ without z-component, where ∇ is the Nabla operator. Setting the z-component of the gradient to zero and eliminating z from (7), the projection of the ellipsoid can be expressed as

$$\begin{aligned} S'(x,y) = \underbrace{\left(A - \frac{E^2}{C}\right)}_{a'} x'^2 + \underbrace{\left(B - \frac{F^2}{C}\right)}_{b'} y'^2 \\ + 2\underbrace{\left(D - \frac{EF}{C}\right)}_{c'} x'y' - 1 \,. \end{aligned} \tag{8}$$

Thus, the projection of the ellipsoid is an ellipse in the x-y-plane, and the point scatterer is shadowed by the ellipsoid if its projection $\vec{p}_\mathrm{t}$ (the origin) lies inside the ellipse, i.e.,

$$s = \begin{cases} 0, & \text{if } a'x_\mathrm{c}^2 + b'y_\mathrm{c}^2 + 2c'x_\mathrm{c}y_\mathrm{c} \le 1 \\ 1, & \text{otherwise.} \end{cases} \tag{9}$$

Finally, after computing the shadowing factors for all obstacles, the point scatterer's aggregate shadowing results as

$$s_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}} = \prod_{\{k \,|\, |\vec{r}_{\mathrm{c},k}| < |\vec{r}_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}}|\}} s^k_{\mathrm{t},i_\mathrm{p},i_\mathrm{s}} \,. \tag{10}$$

The effect of the shadowing is visualized for the subdivided models in Fig. 3. As described above, the radar is located in the origin for the computation of the shadowing factor, and the viewing direction is indicated by the wave vector of the incident wave. Since in this example the model is rotated by 90° for the "Come closer" gesture, the scatterers on the body side facing away from the radar are largely occluded.

4) Geometric Ellipsoid Shadowing: For non-subdivided targets, a non-binary shadowing factor can be calculated in a similar manner. Instead of rotating the point scatterer, the centroid of the target ellipsoid under investigation is rotated towards the z-axis, together with all potential obstacles. Then, the target and all obstacles are projected onto the x-y-plane, and from the projection overlap a partial degree of shadowing can be computed, resulting in a non-binary shadowing factor.

B. Radar Signal Model

From the pedestrian target model introduced above, a list of N_{target} scatterers is generated for each frame. For each target, the corresponding radar intermediate frequency (IF) signal can be computed with the CS signal model introduced in [25]. For a target with range R and radial velocity v, the IF signal for chirp l can be expressed as

$$s_{\text{IF},l}(t) = A \exp\left(2\pi \mathrm{j}\left[\frac{2f_\mathrm{c}\left(R + vT_{\text{RRI}}l\right)}{c} + \left(\frac{2f_\mathrm{c}v}{c} + \frac{2B\cdot\left(R + vT_{\text{RRI}}l\right)}{T_\mathrm{c}c}\right)t\right]\right), \quad (11)$$

with the carrier frequency f_c, the bandwidth B, the chirp repetition time T_{RRI}, the up-chirp time T_c, and the speed of light c. The receive signal amplitude A results from the radar equation [26] with the RCS value taking into account the shadowing as $\sigma' = s\sigma$. The simulated transmit power and receiver noise figure are adjusted to reflect the average pedestrian signal-to-noise ratio observed in reference measurements.

Since the target model does not provide the radial velocity, the range information has to be interpolated to obtain the chirp-dependent ranges $R(l)$ [18]. With $R(l) = R + vT_{\text{RRI}}l$ and under the condition that the Doppler frequency is typically much smaller than the range frequency, (11) reduces to

$$s_{\text{IF},l}(t) = A \exp\left(2\pi \mathrm{j}\left[\frac{2f_\mathrm{c}R(l)}{c} + \frac{2BR(l)}{T_\mathrm{c}c}t\right]\right). \quad (12)$$

Based on (12), the IF signal for each target and each chirp can be computed, and the total radar response is obtained by superimposing all target responses.

C. Radar Signal Processing

The radar signal processing for the generation of Doppler spectrograms is identical for both measured and simulated radar responses: From the IF signals, the range-Doppler maps (RDMs) are computed by applying a 2D fast Fourier transform along the samples and chirps. Then, target lists are extracted from the RDMs using ordered statistics constant false-alarm rate (CFAR). For each CFAR detection, the azimuth angle is computed, and targets are removed in a spatial filtering step if they are more than 1.5 m apart from the human target position. While filtering is unnecessary in the processing of simulation data, it is crucial for suppressing clutter and other targets in the measured radar data, particularly in strongly reflective indoor environments such as car parks. Moreover, suppressing clutter and undesired responses in the measurements strongly reduces the need to properly model them in the simulation. After spatial filtering, Doppler spectrograms are reconstructed from the detections in the proximity of the human target. For this purpose, the detections of all measurement frames are inserted into an initially empty spectrogram of shape $N_\mathrm{v} \times N_\mathrm{f}$, with the number of velocity bins, N_v, and the number of measurement frames, N_f. The values $v_{i_\mathrm{v}}, i_\mathrm{v} \in \{1, ..., N_\mathrm{v}\}$, of the velocity bins are computed as $\vec{v} = [-N_\mathrm{v}/2\Delta v, ..., N_\mathrm{v}/2\Delta v]$, with the bin width $\Delta v = (v_{\text{max}} - v_{\text{min}})/N_\mathrm{v}$, and are therefore determined by the design parameters v_{max}, v_{min}, and N_v. When the values of the velocity bins are given, a target with velocity v is inserted at velocity bin $i_\mathrm{v} = \arg\min_{i_\mathrm{v}} |\vec{v} - v|$ and at the corresponding frame column. Finally, as the measurements can take multiple seconds, the resulting reconstructed spectrograms are subsampled into smaller snippets, each representing two seconds of observation. The subsampled Doppler spectrograms—either simulated or measured—then serve as input data to the CNN-based classifier.

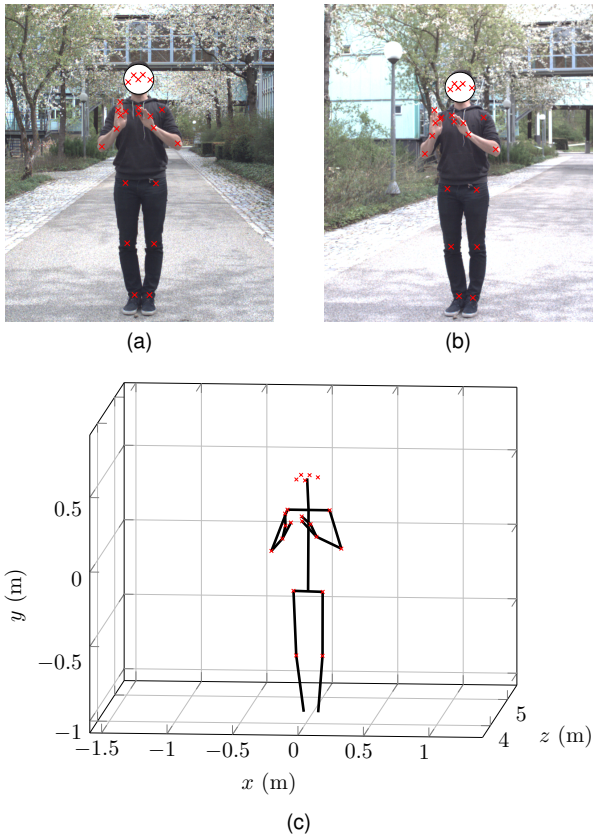

Fig. 4. Estimation of the 2D keypoints in the images of (a) the left and (b) the right camera, and (c) the computation of their 3D counterparts with the stereo calibration information.

III. Generation of Motion Ground Truth Data in the Radar Coordinate System

A multi-camera system calibrated for stereo vision is used to capture motion data for the construction of the target model. Besides providing control over the level of detail (within the limits of the camera resolution), cameras provide a relatively large flexibility when it comes to the measurement environment and range, which makes them well suited for capturing motion ground truth data simultaneously with actual radar measurements. In addition, using a camera system enables the extrinsic calibration with respect to radar sensors, which is shown to facilitate novel methods for the assessment of pedestrian simulations also for complex radar constellations.

A. Motion Capture with a Multi-Camera System

In order to generate motion data from camera measurements, first, the keypoints corresponding to the start and end points of

the targets P_{i_p} have to be found in each camera's imagery. For keypoint detection, *Detectron* [27] together with *OpenPose*'s hand keypoint detector [28] are applied. In addition, whole-body keypoints are extracted with *Openpifpaf* [29]. Generating motion data via two different ways proved to be advantageous for the compensation of the observed weaknesses of the different nets while still maintaining their benefits. For example, *Openpifpaf* produced stable keypoints but tended to miss the whole hand under less favorable lighting conditions, contrary to the more frequent but outlier-prone *OpenPose* detections. After detecting the pixel positions of a keypoint $\vec{p}_{i_\mathrm{c}}$ for each camera i_c, its 3D position $\vec{x}$ can then be found by minimizing the reprojection error back onto the image planes [30], incorporating the knowledge about the camera intrinsics and extrinsics obtained through calibration. The process of 3D motion data generation is illustrated in Fig. 4.

Since the detected keypoint data from all detectors suffer from noisy or missing detections, e.g., due to motion blur, bad light conditions, or short-time occlusions, the data have to be preprocessed before being suitable for simulations. First, the 2D keypoint positions are filtered for outliers and missing detections. Then, the resulting gaps are filled by a curve fitting spline that also smooths the 2D positions. After computing the 3D keypoint positions, the data are scanned for outliers and smoothed in the range and angle dimensions by local regression. Thus, faulty motion components are removed while still maintaining the relevant information. In addition, keypoints for which *Openpifpaf* provided information for too little frames for reasonable interpolation were removed from the simulation.

B. Radar-Centric Target Data for Arbitrary Radar Configurations

The motion data produced by the multi-camera system are described in its camera coordinate system (CCS), centered in the focal point of the reference camera $i_\mathrm{c}=0$. Using it directly for simulation creates a scenario in which a single radar is placed at the origin of the CCS. For the expansion of the simulation to multiple radars with arbitrary orientations and positions, e.g. in a radar sensor network, the pose of each radar sensor node i_n has to be defined by the 4×4 transformation matrix $\mathbf{T}_{\mathrm{LCS}_{i_\mathrm{n}},\mathrm{CCS}}$, which describes the translation and rotation between the radar local coordinate system (LCS) and the CCS. Simulations with multiple radars can then be done by transforming the target positions to the LCSs and following the simulation procedure introduced in Section II for each radar sensor individually.

C. Camera-Radar Calibration

While the radar-centric pedestrian target model allows the simulation of arbitrary radar constellations, an interesting special case is to exactly reproduce the configuration used for simultaneous radar and motion capture. In this case, positions of the radar sensors are determined by the measurement setup and the placement of the sensors relative to the multi-camera system. Since knowledge about the positioning is crucial for accurate resimulations of real multi-radar gesture measurements, an additional camera-radar calibration step is performed prior to simultaneous camera and radar measurements. For this purpose, the common field of view of the radar sensors and the camera system is crossed with a joint target consisting of a corner reflector (for the radars) attached to a checkerboard (for the camera system). Then, the 3D trajectories of the joint target are extracted from both the camera and the radar data. With the positions of the calibration target for each frame l, $\vec{t}_\mathrm{cam}[l]$ and $\vec{t}_{\mathrm{radar},i_\mathrm{n}}[l]$, the transformation matrix $\mathbf{T}$ that describes the radar positions is the one that minimizes

$$\mathbf{T}_{\mathrm{LCS}_{i_\mathrm{n}},\mathrm{CCS}} = \arg\min_{\mathbf{T}} \sum_{l} \left| \left[\vec{t}_{\mathrm{radar},i_\mathrm{n}}[l],\,1\right] - \left[\vec{t}_\mathrm{cam}[l],\,1\right]\mathbf{T} \right|, \tag{13}$$

with the positions $\vec{t}_\mathrm{cam}[l], \vec{t}_{\mathrm{radar},i_\mathrm{n}}[l] \in \mathbb{R}^{1\times3}$. The minimization problem in (13) is solved using iterative closest point (ICP) [31]. With the knowledge about the radar positions in the measurements, the generation of radar-centric target information that matches the real measurements is possible.

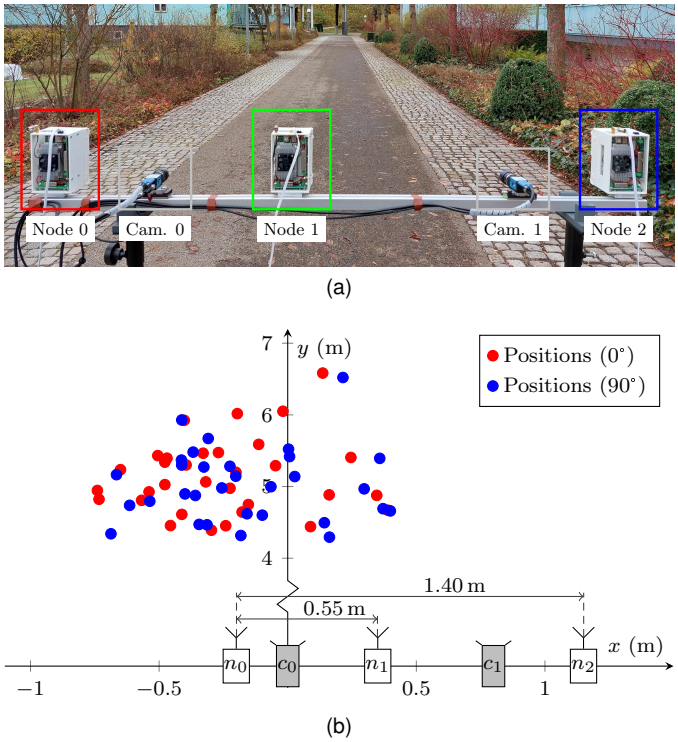

Fig. 5. Visualization of the experimental setup: (a) photography and (b) top-view sketch of the measurement system. Besides the idealized positions of the cameras (c_0, c_1) and the radar sensor nodes (n_0, n_1, n_2), the locations of the participants during the measurements are shown for both orientations.

IV. Experimental Setup

A. Multi-Radar Measurement System

For the validation of the simulation model against measurement data, synchronous radar and camera measurements are conducted with the multi-sensor measurement system illustrated in Fig. 5a. The system consists of three multiple-input multiple-output (MIMO) CS radar sensors operating at 77 GHz that are complemented by two RGB cameras calibrated for stereo vision and with the CCS located in the focal point

TABLE II
RADAR PARAMETERS USED BOTH IN MEASUREMENTS AND SIMULATIONS

Parameter	Value	Description
f_c	79.0 GHz	center frequency
B	3.36 GHz	bandwidth
T_c	33.6 µs	up-chirp time
T_RRI	138.0 µs	ramp repetition interval
N_c	128	number of chirps per sequence
N_s	336	number of samples per chirp
N_tx	3	number of transmit antennas
N_rx	4	number of receive antennas

of camera 0. All sensors are mounted on a rail, and the CS sensors are positioned approximately 0.55 m and 1.40 m apart. Fig. 5b shows a top-view sketch of the measurement setup with the ideal sensor positions in the CCS, i.e. with respect to camera 0. However, the actual placement of the sensors is subject e.g. to rotations, which fortifies the need for a preceding camera-radar calibration. The radars form an incoherent radar sensor network and are operated with a slight timing offset. Hence, target responses belonging to bistatic signal paths are mixed to intermediate frequencies outside the baseband and are subsequently suppressed by the anti-aliasing filter. As a result, each sensor processes only its monostatic responses. The separation of the radar sensors leads to slightly different viewing angles, enabling the network to exploit the RCS diversity of targets. In addition, the measurements from multiple perspectives mitigate the adverse effects of tangential motion components with their low perceived micro-Doppler, which has been shown to be highly beneficial for robust gesture recognition [32]. All radar sensors are operated with the radar parameters listed in Table II and have a range and velocity resolution of $\Delta R = 4.5\,\mathrm{cm}$ and $\Delta v = 11\,\mathrm{cm\,s^{-1}}$, respectively. The three transmit and four receive antennas form a virtual array with up to eight elements for azimuth and two elements for elevation beamforming. While the latter is only applied for the recording of camera-radar calibration trajectories, azimuth beamforming also is relevant for the spatial filtering of measurement data as described in Section II-C. Time-division multiplexing is used to separate the transmit antennas in MIMO operation.

The two cameras each have a resolution of 1280×1024 pixels, and the stereo baseline is 0.88 m. All sensors in the setup receive a common trigger signal for synchronous motion and radar data recording, with the frame rate set to 30 fps.

B. Gesture Dataset

The gesture set consists of the eight signaling gestures for traffic scenarios: "Fly" (g_0), "Come Closer" (g_1), "Slow Down" (g_2), "Wave" (g_3), "Push Away" (g_4), "Wave Through" (g_5), "Start" (g_6), and "Stop" (g_7). The characteristic poses of the gestures are shown in Fig. 6. Gestures g_0–g_5 are

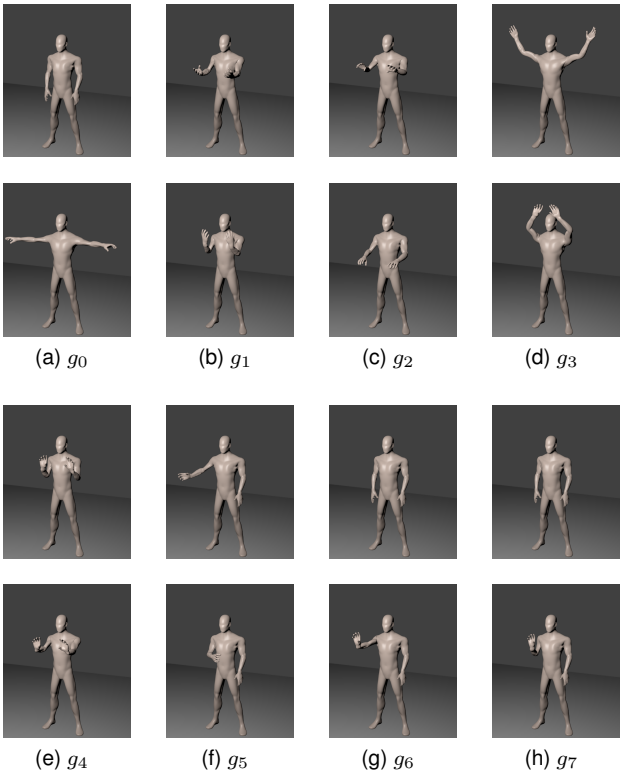

Fig. 6. Visualization of the eight gestures' start (top row) and end (bottom row) pose from [32]: (a) Fly, (b) Come closer, (c) Slow down, (d) Wave, (e) Push away, (f) Wave through, (g) Stop, and (h) Thank you.

repetitive, i.e., they are performed repeatedly without a clear start and end point. Contrary, g_6 and g_7 are non-repetitive such that executions are repeated with a pause between them. For these two gestures, subsampling is centered around the repetitions. Measurement data are collected with the multi-sensor system introduced above on a small street on the campus of Ulm University, as well as in a large indoor hall resembling a car park. Radar and camera data are available for 32 participants and under different orientations. Particularly, the orientations 0° (face pointing towards the radar, shown for g_5 in Fig. 3) and 90° (face pointing to the side, shown for g_1 in Fig. 3), which are used for the evaluation of the simulation approach, are recorded for each user. The positions of the participants in the CCS as estimated by the stereo camera system are shown for both orientations in Fig. 5b. After the radar processing previously explained in Section II-C, 3354 measured spectrograms under 0° and 3129 spectrograms under 90° are available.

C. Gesture Classifier Architecture

For the assessment of the classification performance when training on real and synthesized data, the 10-layer CNN from [32] is used. Classification is investigated for a frame rate of 15 fps, so the two seconds of observation time correspond to 30 measurement frames. The reconstructed spectrograms are resampled to obtain the required input shape of 64×64 and normalized to $[0, 1]$ before being passed to the CNN.

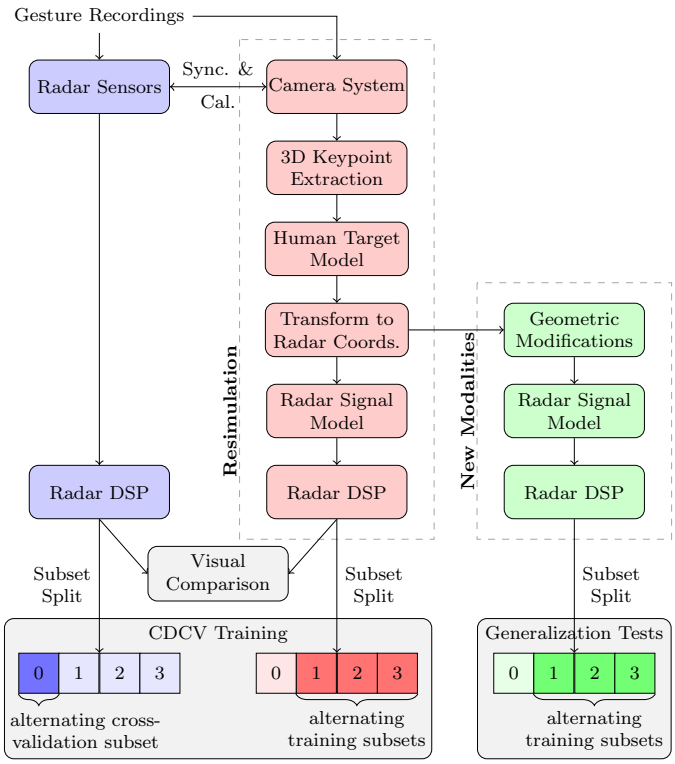

Fig. 7. Workflow for the one-to-one comparison of measured radar data and their resimulations. The resimulations are assessed by visual comparison and by cross-domain tests. The generalization to unseen modalities is tested by training on data from simulations with modified target models.

V. Assessment of Simulation Data Quality

As described above, the multi-camera system is calibrated with respect to the radar sensors and time-synchronized, which facilitates the one-to-one resimulation of a full measured dataset. This, in turn, opens up unique possibilities for the assessment of the simulation data quality. The workflow for the assessment of resimulated gesture recordings based on the motion data extracted from the multi-camera system is summarized in Fig. 7. Besides the direct visual comparison between measurements and resimulations, the artificial data are introduced in the classifier training process as a substitute for real data. The resulting cross-domain accuracy then serves as a metric for the fidelity of the simulation data. In this section, both assessment strategies are investigated for gestures recorded under 0°, as for other orientations no reliable motion data are available for direct comparison due to the occlusion of body parts.

A. Visual Comparison of Spectrograms

First, the visual one-to-one comparisons of measured spectrograms and their resimulations from the synchronized motion data are investigated. Such comparisons can give a first impression of the data quality, but they can also—particularly in the case of direct resimulations—indicate limitations and starting points for improvements of the simulation approach. Fig. 8 shows exemplary measured and resimulated spectrograms for all eight gestures. Note that the spectrograms have not been selected with regard to their quality but to represent a cross-section of the possible effects. As can be seen, under 0° the motion data for g_1, g_2, and g_4–g_7 typically are very similar to the actual motion, and the resimulated spectrograms have a high similarity with the measured ones. For gestures g_5 (“Wave Through”), g_6 (“Stop”), and g_7 (“Thank you”), the sharp velocity peaks can be attributed to the subtle motion of the fingers, clarifying the importance of both realistic motion data and a human target model with an adequate level of detail. Here, simulation e.g. based on the Kinect model from [17] with the whole hand modeled as single point scatterer results in spectrograms missing these characteristic traits. While slight deviations between the measured and simulated general motion patterns are expected due to the limited precision of the motion capture, there are also more pronounced outliers visible (exemplary ones highlighted in green in Fig. 8). These are the results of faulty keypoints, for example rarely detected hand keypoints that barely circumvent removal from simulation. While they can be better suppressed by stricter filtering of uncertain detections, it turns out that this procedure leads to worse classification results. Stricter keypoint filtering often also removes valid information in cases where the smoothing works as expected. Thus, it seems to be advantageous to accept a certain level of outliers, particularly as they follow no pattern and are thus likely to be ignored by classifiers.

Contrary to the aforementioned gestures, the resimulations of g_0 and g_3 displayed in Fig. 8 exhibit less similarity with their measured counterparts. Since under 0° the motion is predominantly tangential to the radars, the spectrograms typically show no clear motion patterns. In addition, for g_0 the hands are only visible from the side during most of the measurements (cf. Fig. 6). This increases the error susceptibility of the hand keypoint detection and makes the motion data more error-prone. To mitigate such effects, additional cameras can be used to increase the robustness of the keypoint detection against outliers and occlusions.

Besides the comparison, Fig. 8 also hints at the confusion clusters that can occur when a gesture set is not entirely designed from scratch: For example, gestures g_1 (“Come Closer”) and g_5 (“Push away”) possess highly similar motion patterns, which can deteriorate classification accuracy particularly for simulation data with insufficient level of detail. Moreover, the existence of similar gestures can penalize even small deficiencies of the target modelling.

B. Classification-based Data Assessment

The most meaningful measure for the simulation data quality is the achievable accuracy when directly using it as a substitute for real data. Therefore, the CNN is trained on both real and artificial data. All training parameters have been optimized on measurement data only and are kept constant when using synthetic data. In order to obtain a baseline accuracy, the network first is trained and tested on the measured data. For this purpose, the data are split in four subsets as shown in Fig. 7, each containing data from eight participants. Then, training is repeated four times with each of the subsets reserved as test set in one fold. The resulting 4-fold cross-validation (CV)

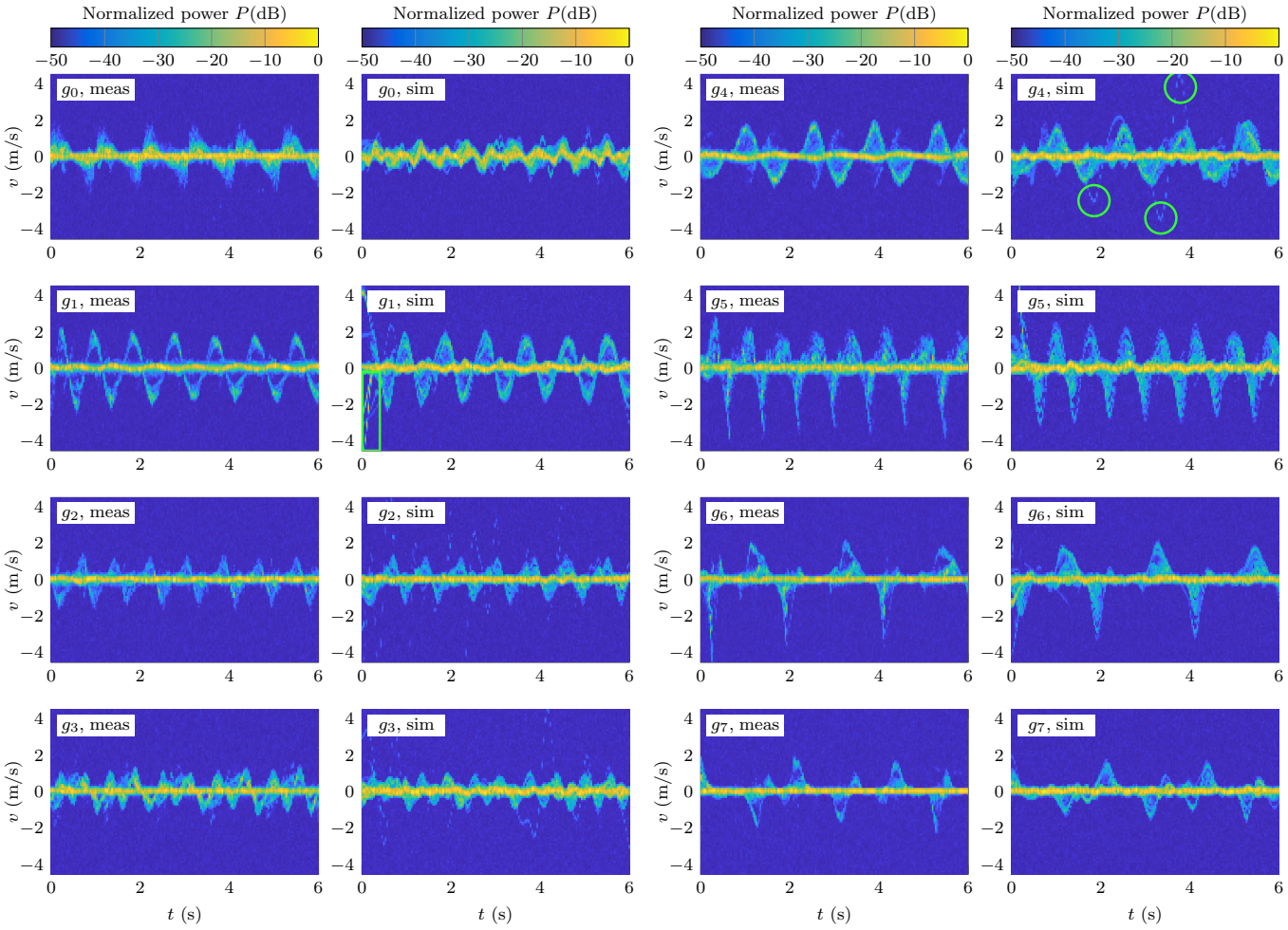

Fig. 8. Comparison of the measured and resimulated Doppler spectrograms. Note that for better comparability each spectrogram has a time span of 6 s. For classification, the spectrograms are subsampled to snippets representing 2 s. Exemplary outliers in the motion data are highlighted in green.

per-gesture accuracy is shown in Table III, and the average accuracy reaches 89.4 %. In addition, Fig. 9a shows the confusion matrix. The 4-fold CV is repeated with simulation data for training and test. For the simulation, motion data from the measurements under 0° from both *Detectron+OpenPose* and *Openpifpaf* are used, such that the simulated dataset contains twice as many samples as the measured one. Here, the classifier reaches an accuracy of 85.6 %, indicating that the samples of the different classes reflect most of the class characteristics but also that some samples are corrupted due to noisy motion data as shown above, degrading the test set accuracy compared to the previous case.

Next, the classifier is trained on simulation data only, with the test set containing measured samples. With the exact resimulations available, a straightforward test is to simply train the classifier on the whole simulation dataset and test it on the whole measured dataset. As these ideally would be identical, this is named cross-domain identity training (CDIT), and Table III shows the results of the CDIT. With perfect simulation data an accuracy very close to 100 % could be achieved. In reality, this value reaches 85.7 %, and the difference might be contributed not only to the inevitable imperfections of human target modeling but also to the partially imprecise motion capture. The latter can reduce the recognition accuracy e.g. by increasing the confusion potential between gestures g_0 and g_3 as shown in Fig. 8. Finally, the classifier is trained 4-fold on simulation data, with one subset of real data serving as test set in each fold. Hence, for cross-domain cross-validation (CDCV) the motion information underlying the simulated training data is obtained from participants not included in the respective test set, as is expected to be the case in automotive applications. Since the same 4-fold split is used as before, the training on simulation data is a direct imitation of the training on measurement data. As reported from Table III, a CDCV accuracy of 80.4 % can be reached without the classifier trained even on a single measured spectrogram. With nine percentage points, the human target model-based data synthesis achieves a similar domain transfer loss as the encoder-decoder approach presented in [23], without the need to train a neural network for domain transfer. Moreover, the gap in accuracy is slightly lower than the one obtained for the geometric model from computed animation in [20], where, however, the measurement benchmark is computed from a larger gesture set. Note that our result is achieved without

TABLE III
FOUR-FOLD CROSS-VALIDATION ACCURACY FOR GESTURES MEASURED WITH A PARTICIPANT ORIENTATION OF 0°

Metric	Fly	Come closer	Slow down	Wave	Push away	Wave through	Stop	Thank you	Avg.
CV[1], meas→meas	87.87 %	88.41 %	90.38 %	88.87 %	84.51 %	96.33 %	88.22 %	90.03 %	**89.4 %**
CV, sim→sim	85.54 %	79.80 %	89.93 %	82.60 %	82.59 %	92.41 %	87.02 %	85.30 %	**85.6 %**
CDIT[2], sim→meas	86.76 %	84.17 %	80.26 %	78.34 %	88.19 %	95.87 %	83.33 %	89.81 %	**85.7 %**
CDCV[3], sim→meas	81.57 %	83.53 %	71.46 %	77.66 %	76.17 %	92.31 %	76.72 %	83.08 %	**80.4 %**

[1]CV: cross-validation, [2]CDIT: cross-domain identity training, [3]CDCV: cross-domain cross-validation

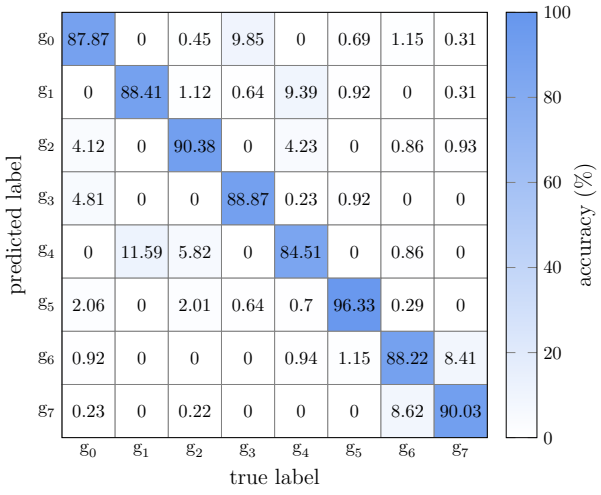

Fig. 9. Confusion matrices for 4-fold CV training on (a) measurement and (b) simulation data. In both cases, the classifier is tested on measured spectrograms.

applying any explicit diversification as proposed in [15], which might further enhance the CDCV accuracy. Besides the gesture recognition accuracy, the comparison of the confusion matrices in Fig. 9 provides further insights. Most notably, the confusion clusters after training on simulation data closely resemble the ones when training on real data. This supports the assumption that the generated and the real data have highly similar characteristics. Another interesting aspect regarding the CDCV accuracy is the fusion of data from different keypoint detectors. Using *Detectron+OpenPose* or *Openpifpaf* data alone, the accuracies are 75.1 % and 73.8 %, respectively. Thus, the combination of different keypoint detectors can partially compensate for the weaknesses of the individual models. Finally, the classifier trained on simulated data exhibits similar confusion clusters than the one trained on measured data, further validating the plausibility of the generated data. However, the confusion clusters described above appear more pronounced, so using higher-accuracy motion data might further enhance the inclusion of subtle traits.

VI. GENERALIZATION TO NEW MODALITIES BY SIMULATION DATA

A measured dataset might not sufficiently reflect all possible scenarios in challenging applications such as automotive gesture recognition. Here, synthetic data helps to fill gaps in the coverage of different modalities and edge cases, since the simulation chain can be easily adapted to include a wide range of geometric modifications (cf. Fig. 7). To put this on test, the measurement data with a participant orientation of 90° is used to represent a distinct novel scenario not covered by measured training data. Training the CNN on the measurement data recorded under this orientation, it maintains a high 4-fold CV accuracy of 84.0 %, albeit the low radial velocity for most gestures exacerbates the confusion clusters already observed in the previous section. Contrary, if only the measurement data under 0° is available, the CV accuracy with the same 4-fold split drops to 16.1 %, barely better than guessing. It must be noted that this obviously is an extreme case and not surprising, given the significantly different radar responses under the two different orientations. However, even if a large training set contains measurements under different discrete orientation angles, the generalization can still fail for certain angle ranges, not to mention other modalities.

To test the substitution in the case of missing measurement data, simulation data are generated from the same motion

TABLE IV
FOUR-FOLD CROSS-VALIDATION ACCURACY FOR GESTURES MEASURED WITH A PARTICIPANT ORIENTATION OF 90°

Metric	Fly	Come closer	Slow down	Wave	Push away	Wave through	Stop	Thank you	Avg.
CV[1], meas (90°)→meas (90°)	90.00 %	73.13 %	73.63 %	98.80 %	80.24 %	96.07 %	79.94 %	77.47 %	**84.0 %**
CV, meas (0°)→meas (90°)	0.03 %	1.09 %	46.42 %	5.67 %	0.03 %	6.48 %	56.50 %	16.46 %	**16.1 %**
CDCV[2], sim→meas	81.25 %	61.64 %	60.45 %	87.47 %	55.56 %	91.47 %	71.39 %	71.12 %	**72.8 %**
CDCV[2], sim (no sh.[3])→meas	52.37 %	50.70 %	38.29 %	91.94 %	54.35 %	91.69 %	68.17 %	63.22 %	**63.6 %**

[1]CV: cross-validation, [2]CDCV: cross-domain cross-validation, [3]: sh.: shadowing

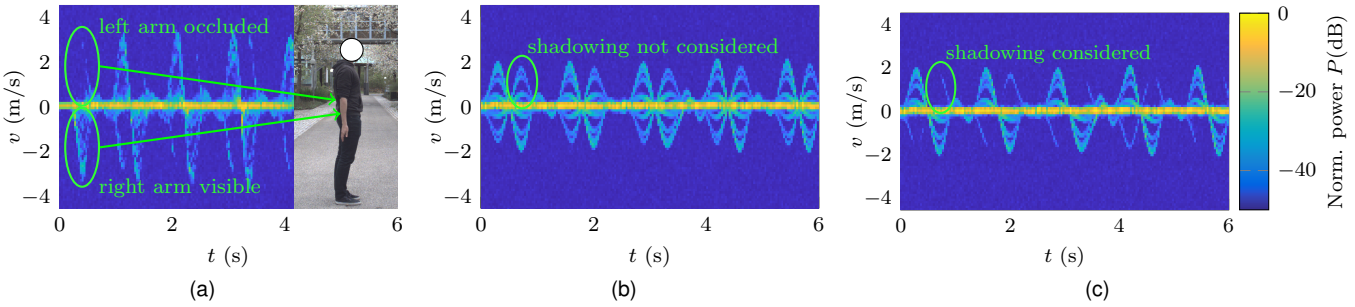

Fig. 10. Comparison of a measured (a) and two simulated (b,c) spectrograms for the "Fly" gesture with an orientation of 90°. Without shadowing (b), the arm occluded by the body is still visible. With shadowing (c), the spectrogram reflects the actual occlusion.

data as in the 0°-case but rotated around the pedestrians' center by 90°. As the orientation was only loosely instructed during the measurements, the actual values under which the radar data were recorded deviate from 90°. In order to reflect this, orientation angles of 80°, 90°, and 100° are included in the radar simulations. Training on simulation data with the same 4-fold split as used in the previous tests results in a CDCV accuracy of 72.8 %, i.e., the classifier is able to handle the previously unseen modality significantly better. With 11.2 percentage points, the difference to training on measurement data is slightly higher than in the previous section. As can be seen in Table IV, this can be attributed to the more challenging nature of the classification task under 90° due to the low Doppler frequency content. In particular, the mutual confusion leading to lower accuracies for g_1, g_2, and g_4 might be reduced by higher-resolution imagery with e.g. *Openpifpaf* hand keypoints available for more measurements.

The importance of the shadowing effect for realistic simulations is highlighted for the spectrograms of g_0 shown in Fig. 10. Note that the overall shapes of the simulated spectrograms deviate from the measured ones since it is no one-to-one resimulation: The measured spectrogram is directly recorded under 90°, whereas the motion data for the simulations are a rotated version of a recording under 0°. However, differences can be observed beyond the ones explained before. When not taking into account the shadowing effects, the simulation produces implausible data as highlighted in Fig. 10b for the motion component induced by the occluded left arm. As can be seen in Table IV, this deteriorates the CDCV accuracy not only for the classes mostly affected by shadowing (e.g. g_0) but also for other gestures that experience less significant occlusion. As a result, the generalization to the new orientation angle is more than nine percentage points worse when compared to the case where the shadowing effect is considered. This clearly demonstrates the benefit of the introduced shadowing factor for the generation of high-fidelity radar responses in automotive scenarios.

VII. CONCLUSION

In this paper, a human target model for the flexible simulation of gestures with a wide range of modalities has been presented. Based on a detailed ellipsoidal model, sub-scatterers are derived and their shadowing factors are computed. With this model, radar simulations are performed based on motion data recorded simultaneously to actual radar measurements. The simulation data are shown to have high similarity with the corresponding radar measurements, and a classifier trained on simulation data only achieves 80.4 % cross-validation accuracy, which demonstrates that the presented target model achieves better or similar cross-domain loss compared to other approaches even without diversification while maintaining the high flexibility of a geometrical model. Moreover, by testing on radar data recorded under a different viewing angle, it is shown that the simulated data can help the CNN to generalize to new modalities. By enhancing the motion ground truth data and augmenting the simulation data through diversification, the approach is expected to further narrow down the cross-domain gap.

VIII. ACKNOWLEDGMENTS

This work was supported by the Ministerium für Wissenschaft, Forschung und Kunst (MWK) Baden-Württemberg as part of the project INTUITIVER.

REFERENCES

[1] M. Colley, J. H. Belz, and E. Rukzio, "Investigating the effects of feedback communication of autonomous vehicles," in *Proc. Int. Conf. Automot. User Interfaces Interact. Veh. Appl.*, 2021, pp. 263–273.

[2] E. Ohn-Bar and M. M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 90–104, 2016.

[3] M. Milford, S. Anthony, and W. Scheirer, "Self-driving vehicles: Key technical challenges and progress off the road," *IEEE Potentials*, vol. 39, no. 1, pp. 37–45, 2020.

[4] J. Wiederer, A. Bouazizi, U. Kressel, and V. Belagiannis, "Traffic control gesture recognition for autonomous vehicles," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10 676–10 683.

[5] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, 2016. [Online]. Available: https://doi.org/10.1145/2897824.2925953

[6] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[7] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with Soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proc. Annu. Symp. User Interface Software and Technol.*, 2016, pp. 851–860.

[8] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278–3289, 2018.

[9] Y. Kim and T. Moon, "Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 8–12, 2015.

[10] 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. ACM Workshop Millimeter-wave Netw. Sens. Syst.*, 2019, pp. 51–56.

[11] B. Erol and M. G. Amin, "Radar data cube processing for human activity recognition using multisubspace learning," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 6, pp. 3617–3628, 2019.

[12] I. Alnujaim, D. Oh, and Y. Kim, "Generative adversarial networks for classification of micro-Doppler signatures of human activity," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 396–400, 2019.

[13] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Motion classification using kinematically sifted ACGAN-synthesized radar micro-Doppler signatures," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 4, pp. 3197–3213, 2020.

[14] S. S. Ram and H. Ling, "Microdoppler signature simulation of computer animated human and animal motions," in *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, 2008, pp. 1–4.

[15] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Diversified radar micro-Doppler simulations as training data for deep residual neural networks," in *Proc. IEEE Radar Conf.*, 2018, pp. 0612–0617.

[16] K. Ishak, Z. Zafar, M. Steiner, N. Appenrodt, J. Dickmann, and C. Waldschmidt, "A radar measurement setup with a ground truth system for micro-Doppler human movements," in *Proc. IEEE MTT-S Int. Conf. Microw. Intell. Mobility*, 2019, pp. 1–4.

[17] B. Erol and S. Z. Gurbuz, "A kinect-based human micro-Doppler simulator," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 30, no. 5, pp. 6–17, 2015.

[18] K. Ishak, N. Appenrodt, J. Dickmann, and C. Waldschmidt, "Human motion training data generation for radar based deep learning applications," in *Proc. IEEE MTT-S Int. Conf. Microw. Intell. Mobility*, 2018, pp. 1–4.

[19] K. Ishak, N. Appenrodt, J. Dickmann, and C. Waldschmidt, "Advanced radar micro-Doppler simulation environment for human motion applications," in *Proc. IEEE Radar Conf.*, 2019, pp. 1–6.

[20] A. Ninos, J. Hasch, M. E. P. Alvarez, and T. Zwick, "Synthetic radar dataset generator for macro-gesture recognition," *IEEE Access*, vol. 9, pp. 76 576–76 584, 2021.

[21] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "DNN transfer learning from diversified micro-Doppler for motion classification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 5, pp. 2164–2180, 2019.

[22] A. Ninos, J. Hasch, and T. Zwick, "Real-time macro gesture recognition using efficient empirical feature extraction with millimeter-wave technology," *IEEE Sensors J.*, vol. 21, no. 13, pp. 15 161–15 170, 2021.

[23] 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.*, 2021, pp. 1–10.

[24] J. Crispin and A. Maffett, "Radar cross-section estimation for simple shapes," *Proc. IEEE*, vol. 53, no. 8, pp. 833–848, 1965.

[25] V. Winkler, "Range Doppler detection for automotive FMCW radars," in *Proc. Eur. Radar Conf.* Piscataway, NJ: IEEE, 2007, pp. 166–169.

[26] M. I. Skolnik, "Introduction to radar systems," *New York*, 1980.

[27] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[28] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1145–1153.

[29] S. Kreiss, L. Bertoni, and A. Alahi, "Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association," *IEEE Trans. Intell. Transp. Syst.*, 2021.

[30] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[31] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[32] N. Kern, M. Steiner, R. Lorenzin, and C. Waldschmidt, "Robust Doppler-based gesture recognition with incoherent automotive radar sensor networks," *IEEE Sens. Lett.*, vol. 4, no. 11, pp. 1–4, 2020.



Nicolai Kern received the M.Sc. degree in electrical engineering from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2019. He is currently pursuing the Ph.D. degree with the Institute of Microwave Engineering, Ulm University, Ulm, Germany.

His research interests include machine learning applications in radar with a focus on vulnerable road users and radar sensor network signal processing.



Julian Aguilar received the B.Sc. degree in electrical engineering from Ulm University, Ulm, Germany, in 2018, where he is currently pursuing the M.Sc. degree in electrical engineering.

His main research interest is the synthesis of radar data for machine learning applications.



Timo Grebner received his Master's degree in electrical engineering from the Ulm University in 2019. He is currently pursuing the degree of Ph.D. in electrical engineering at the Institute of Microwave Engineering, Ulm University, Ulm, Germany.

His current research interests include the creation of high-resolution occupancy gridmaps by high-accuracy ego-motion estimations in a distributed and collaborative radar sensor topology.



Benedikt Meinecke received the M.Sc. degree in 2017 in electrical engineering with a focus on communication technology from Ulm University, Ulm, Germany, where he is currently working toward the Ph.D. degree in electrical engineering with the Institute of Microwave Engineering.

His research interests include system concepts for multistatic coherent radar networks and the accompanying signal processing.



Christian Waldschmidt (Fellow, IEEE) received the Dipl.-Ing. (M.S.E.E.) and the Dr.-Ing. (Ph.D.E.E.) degrees from the University Karlsruhe (TH), Karlsruhe, Germany, in 2001 and 2004, respectively. From 2001 to 2004 he was a Research Assistant at the Institut für Höchstfrequenztechnik and Elektronik (IHE), Universität Karlsruhe (TH), Germany. Since 2004 he has been with Robert Bosch GmbH, in the business units Corporate Research and Chassis Systems. He was heading different research and development teams in microwave engineering, RF-sensing, and automotive radar.

In 2013 Christian Waldschmidt returned to academia. He was appointed as the Director of the Institute of Microwave Engineering at University Ulm, Germany, as full professor. The research topics focus on radar and RF-sensing, mm-wave and submillimeter-wave engineering, antennas and antenna arrays, RF and array signal processing. He authored or coauthored over 250 scientific publications and more than 20 patents.

He is member of the executive committee board of the German MTT/AP joint chapter and member of the German Information Technology Society (ITG). He served as chair of the IEEE MTT-29 Technical Committee on Microwave Aerospace Systems and as chair of MTT-27 Technical Committee on Wireless Enabled Automotive and Vehicular Applications. He was a two-time TPC chair and general chair of the IEEE MTT International Conference on Microwaves for Intelligent Mobility. Since 2018, Christian Waldschmidt serves as associate editor for IEEE MTT Microwave Wireless Components Letters (MWCL). He is a reviewer for multiple IEEE transactions and many IEEE conferences in the field of microwaves. Christian Waldschmidt was co-recipient of 12 best paper awards since 2014.