IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 26, 2021, accepted May 1, 2021, date of publication May 5, 2021, date of current version May 14, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3077690*

# A Bayesian Framework for Integrated Deep Metric Learning and Tracking of Vulnerable Road Users Using Automotive Radars

**ANAND DUBEY[1], (Member, IEEE), AVIK SANTRA[2], (Senior Member, IEEE), JONAS FUCHS[1], (Member, IEEE), MAXIMILIAN LÜBKE[1], (Member, IEEE), ROBERT WEIGEL[1], (Fellow, IEEE), AND FABIAN LURZ[3], (Member, IEEE)**

[1]Institute for Electronics Engineering, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, 91058 Erlangen, Germany
[2]Infineon Technologies AG, 85579 Neubiberg, Germany
[3]Institute of High-Frequency Technology, Hamburg University of Technology, 21073 Hamburg, Germany

Corresponding author: Anand Dubey (anand.dubey@fau.de)

This work was supported by the Electronic Components and Systems for European Leadership (ECSEL) Joint Undertaking (JU) through the Programmable Systems for Intelligence in Automobiles (PRYSTINE) Project under Grant 783190. The JU receives support from the European Union's Horizon 2020 Research and Innovation Programme and National Authorities.

**ABSTRACT** With the recent advancements in radar systems, radar sensors offer a promising and effective perception of the surrounding. This includes target detection, classification and tracking. Compared to the state-of-the-art, where the state vector of classical tracker considers only localization parameters, this paper proposes an integrated Bayesian framework by augmenting state vector with feature embedding as appearance parameter together with localization parameter. In context of automotive vulnerable road users (VRUs) such as pedestrian and cyclist, the classical tracker poses multiple challenges to preserve the identity of the tracked target during partial or complete occlusion, due to low inter-class (pedestrian-cyclist) variations and strong similarity between intra-class (pedestrian-pedestrian). Subsequently, feature embedding corresponding to target's micro-Doppler signature are learned using novel Bayesian based deep metric learning approaches. The tracker's performance is optimized due to a better separability of the targets. At the same time, the classifiers' performance is enhanced due to Bayesian formulation utilizing the temporal smoothing of the classifier's embedding vector. In this work, we demonstrate the performance of the proposed Bayesian framework using several vulnerable user targets based on a 77 GHz automotive radar.

**INDEX TERMS** Automotive radar, Bayesian framework, deep metric learning, integrated classification-tracking, unscented Kalman filter.

## I. INTRODUCTION

Both reliability and safety of autonomous vehicles require a precise perception of the operating environment, which in turn necessitates high-quality and accurate measurements of sensors [1]–[4]. Thus, reliable sensing capabilities are the key towards a successful implementation of automated or self-driving vehicles. Compared to other sensing technologies e.g., camera or Lidar, radar sensors combine a lot of advantages: they are relatively robust to bad weather conditions, work in dark environments and can sense distance and velocity of the targets simultaneously [5]. Typically, automotive radar can detect targets up to ranges of more than 200 m and provide a high range-resolution of multiple targets in its field of view. As a result, radar sensors are widely accepted and are becoming one of the major enablers for advanced driver assistance systems (ADASs) and fully automated driving [6], [7]. The typical applications for ADAS includes adaptive cruise control, forward collision avoidance (FCA), lane change assistance, parking aid, and safety of vulnerable road users (VRUs) [5], [8].

Automotive radar mmWave sensing has shifted from 24 GHz to 77 GHz due to the larger available bandwidth (76–77 GHz for long-range and 77–81 GHz for short-range applications), higher Doppler sensitivity and smaller antennas leading to small form-factors [9]. Traditional automotive radars transmit a sequence of up-chirps with low chirp times. The typical signal processing involves a chirp pulse

The associate editor coordinating the review of this manuscript and approving it for publication was Weimin Huang.

compression along the fast-time (intra-chirp time), followed by Doppler processing along the slow-time (inter-chirp time) and digital beamforming across the receive channels, generating a 3D data tensor. Following the 3D radar data tensor, automotive radar processing includes target parameter (range, Doppler and angle) estimation, followed by target tracking. However, conventional automotive radar systems face a lot of challenges, especially in complex urban environments, where the sensor needs to detect, classify and track multiple targets, e.g. VRUs like pedestrians and cyclists. While [10]–[14] propose different methods to address these challenges separately, an overview of the traditional radar signal processing is provided in [15]. Additionally, [16] provides overview on target detection and tracking and [5] examined target classification. With the recent advancement in radar systems and the processing of high-resolution data using different concepts of neural networks, target classification is typically done via extracting target specific parameters such as micro-Doppler spectrograms. Later, these spectra are fed into a classifier such as a deep convolutional neural network (DCNN) or long short-term memory (LSTM) networks for the classification of the target [9], [17]. In order to classify different targets (VRUs in our case), they need to be detected and separated first in one of the three measurement dimensions, namely range, velocity and angle [6]. In urban environments, different VRUs can be closely located and have quite similar velocity magnitudes as well, resulting in a low separability among them. E.g. where the relative velocity of vehicles on a highway can vary from 20 m/s to 80 m/s, the velocities of VRUs typically are within the range of 0 m/s to 10 m/s. Furthermore, the strength of the received signal depends on the targets surface area, visible to the radar sensor, the so-called radar cross-section (RCS). The RCS of VRU targets are up to 20 dB lower than a RCS of vehicles [9]. Thus, urban scenarios require an highly sophisticated signal processing for a reliable detection and classification of VRUs.

To increase the robustness of VRU classification and to reduce detection false-alarms, the concept of target tracking can be used. This helps to estimate the desired unknown state variables from the observed noisy measurements. The problem of target tracking, has been extensively studied in the literature [18]–[25]. These trackers are based on various motion models, like e.g. the constant velocity (CV) model, constant acceleration model, current statistic model, interacting multiple model and varied structure multiple model. The most common tracking algorithms integrating such models are the extended Kalman filter, the unscented Kalman filter (UKF), multiple hypothesis testing and particle filters. In this paper, UKF is used as a tracker to realize the overall process and measurement model through the unscented transformation that tries to approximate the distribution of a random variable which is transformed non-linearly. The state-of-the-art UKF algorithms in an automotive use-cases majorly focus on single modalities by using the target's localization information as a state vector [26], [27]. However, tracking of automotive VRUs exhibits challenges in the form of very distinct and different dynamic models, resulting in multiple switches in the associated track-ID. Additionally, an inaccurate measurement of the track association leads to divergence in the innovation squared metric [9]. In an attempt to solve the association problem, we propose an integrated Bayesian framework combining target's feature embeddings as appearance model and localization as motion model to simultaneously classify and track VRUs. As a result, the choice of feature extraction model and input data representation plays a critical role. While the reflected signal obtained from the radar sensor is processed for the estimation of the detected target's localization parameters, namely range, angle and velocity, the features corresponding to the detected target are estimated from the latent (embedding) space of a deep neural network architecture.

Conventional deep learning based approaches, trained with a cross entropy loss, requires massive amounts of data to be trained [28]. Additionally, in the context of radar sensors and micro-Doppler signatures, learned models often do not generalize well on different sensors, target orientations or inter/intra-class variations. However, for systems to work in an open set of environments, a higher distinction among the classes are required. To address these issues, deep metric learning and meta-learning have gained prominence in the literature. Deep metric learning models are optimized on certain distance metrics that aim to learn both similarity and dissimilarity among targets/classes, such that similar targets are grouped together, whereas dissimilar targets are far separated in the embedding space [29]–[35]. While techniques, such as principal component analysis or linear discriminant analysis, also project the input data into the representational space, deep metric learning utilizes a neural network to learn a rather optimized representational space using various loss functions and training approaches [36]–[42]. In [43], the authors have proposed a Siamese network for material classification of known and unknown materials using a 60 GHz radar sensor. In [44], the authors proposed a triplet loss for radar-based gesture sensing using 3D CNN for demonstrating the generalization capabilities of this approach. Further in [45], authors proposed a novel Euclidean softmax approach for learning both discriminative and separability in the feature space for human activity classification using FMCW radar sensors. Additionally, recent work [46]–[48] in computer vision domain successfully demonstrate person re-identification by leveraging the concept of metric learning. [48] propose a concept of adaptive weighted convolution which learns part-based representational learning. Whereas, to the best author's knowledge, proposed framework bring novelty of combination Bayesian features embedding inside tracker. While the latent embedding is directly being tracked by the tracker and leads to improvement in target classification, the learned variance over latent embedding help in target gating (association). As a result, framework enables an integrated full Bayesian framework. This article gives a detailed analysis on advantage of proposed framework. Additionally, similar to [48], triplet loss function is modified to have adaptive weighting over latent

embedding in contrast to euclidean distance. Furthermore, instead of triplet pairs, quadruplet pairs are mined to include both inter-class and intra-class pairs for an anchor class.

In this paper, we generate a realistic automotive radar dataset with different target maneuvering of vulnerable road users using MATLAB®'s phased-array toolbox, which we describe in Section II. In Section III, we present the traditional automotive radar signal processing involving range, Doppler, angle processing, target detection, target clustering, extraction of micro-Doppler spectra and target tracking. Section IV introduces the target appearance model and our dataset. In Section V, we present the various deep metric learning approaches including loss functions, to learn the feature embedding for the target classification problem. We also explain our extension of adding the decoder and variational auto-encoder that improves the feature representation to conventional triplet loss and quadruplet loss training. In Section VI, we introduce our integrated framework which combines the proposed metric learning embedding model to the conventional tracker to augment its state vector to track the target embedding vector along with the target location parameters to achieve a Bayesian framework. We demonstrate the superior representational learning performance of the proposed solution by comparing it with the conventional metric-learning counterparts in Section VII, and discuss the classification accuracy using k-nearest-neighbor. We further analyze the superior tracking performance of the proposed solution compared to conventional tracker, in terms of localization error and normalized innovation square metric, under exemplary scenarios in Sec. VII.

## II. SYSTEM SIMULATION

A large amount of data is needed to train and evaluate the proposed framework. In real scenarios, these can only be generated at great expense, whereby cross-influences from the environment are also always recorded and thus detailed investigations can become difficult due to reproducibility. Therefore, a simulative approach is primarily used, which is described in detail in the following. This following section introduces the simulation setups which will be the underlying basis of the further signal processing. Therefore, radar signals of different types of road users are simulated in consideration of micro-Doppler signatures. As there are several approaches of radio frequency (RF) systems and environment simulations available, a short overview, focusing on the main advantages and disadvantages of each approach is provided. This leads to describing our own environment simulation framework, regarding the used radar system model, channel model, target models, and trajectory models.

### A. STATE-OF-THE-ART RF PROPAGATION SIMULATORS

The most advanced and accurate RF wave propagation simulation is a full electromagnetic (EM) simulation, e.g. as provided by COMSOL[1] or FEKO.[2] Commonly used for static scenarios like e.g. antenna pattern or radar-cross-section simulations, the system and target characteristics can be simulated accurately. However, as it is computational expensive, it is not well suited for large and dynamic scenarios. A hybrid approach, based on a combination of a finite difference time domain solver for RF wave propagation, and computer animations for the human motion simulation, has been used in [49] to obtain micro-Doppler signatures.

Some more efficient simulators, e.g. WinProp[3] or WaveFarer[4] use ray tracing in order to obtain highly accurate approximations of propagation effects, like diffraction, reflection, scattering or multipath effects in general. They are usually based on a deterministic ray optical model, combining the Fresnel equations with the geometrical theory of diffraction (GTD) and uniform theory of diffraction (UTD) and are commonly used for propagation channel evaluations [50]–[53]. Within the simulators, scenarios and the used materials are configurable, radar-cross-sections (RCS) and antenna patterns can be imported from full EM simulations, to lower computational costs. Thus, time variant and therefore dynamic scenarios with macro movements can be simulated. Advantages are the high accuracy and flexibility, e.g. the evaluation grid size can be chosen arbitrarily. The main drawbacks, however, are the time-consuming generation of scenarios and the limited support for micro-movements, which are necessary to generate micro-Doppler signatures.

Stochastic simulators, like NYUSIM [54] or MilliCar [55], can overcome the problem of high computational costs, as they are based on statistical channel models. These models are derived from real measurements of certain scenarios. The accuracy of the results, therefore, strongly depends on the similarity of the simulated scenario and the environment of the original measurements. Results of the stochastic simulations are mainly used in the communication sector today, e.g. in designing the physical layer of new communication standards [56]. As the aim of the proposed work is to address micro-Doppler radar applications, these kinds of models will not be further evaluated here.

Low cost statistical simulations of radar target detections are provided via the Automated Driving Toolbox™[5] from MathWorks®. The considered scenario is simplified to a cubic world, where road users are approximated as cuboids, and the detections are randomly generated. It provides a simple user interface for a fast generation of scenarios, including configuration of targets' trajectories or the movement of the ego vehicle. The stochastic generation of target detections is very fast, as the RCS patterns can be set for each road user individually. A drawback is that the simulator just outputs a target list with the measurable parameters like range, velocity or angle. Additionally, only macro-movements are considered in the simulator setup. These properties make it a

[1] https://www.comsol.com/comsol-multiphysics
[2] https://www.altair.com/feko/
[3] https://www.altair.com/resource/altair-winprop-datasheet
[4] https://www.remcom.com/wavefarer-automotive-radar-software
[5] https://de.mathworks.com/help/driving/ref/radardetectiongenerator.html

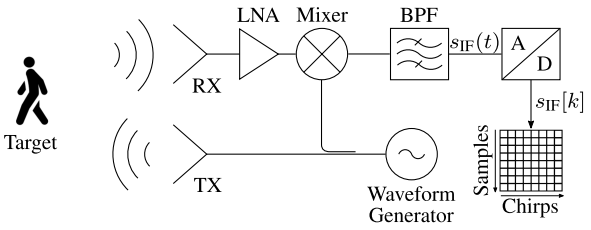


**FIGURE 1.** Simplified FMCW radar system block diagram.

common choice for use cases like occupancy grid generation in automotive radar [57].

In summary, a different approach is necessary to fulfill the needs of an automotive radar micro-Doppler simulation framework. Compared to the deterministic and stochastic simulators, with the problem of being commercial, not adaptable, or not addressing micro-Doppler scenes at all, a MATLAB® based solution seems to be the best fitting approach. Thus, we chose to further optimize the MATLAB® approach and extend it with a micro-Doppler simulation.

### B. RADAR SYSTEM SIMULATION

Automotive radar sensors typically take advantage of frequency modulated continuous wave (FMCW) signal waveforms, as they can be used to estimate both range and velocity of targets. In Fig. 1, a simplified block diagram of a typical FMCW radar system is shown. A waveform generator provides a frequency ramp, also called chirp, of duration $T_c$ with a bandwidth $B$ at a center frequency $f_c$. The signal is transmitted, gets reflected by a target at a distance of $r$, and then is collected by the receive antenna with a total delay of $\tau = 2r/c_0$. The received signal is amplified, down-mixed with the original transmit signal, as well as band-pass filtered to obtain an intermediate frequency (IF) signal. This signal is sampled with a sampling frequency $f_s$ to obtain $N_s$ samples. All target parameters of interest (i.e. range, velocity and angle) can be estimated from the sampled IF signal. In the baseband, the delay $\tau$ is converted to a frequency shift $f_{\text{range}}$ of the chirp signal. However, moving targets additionally cause a Doppler shift on the reflected signal. Utilizing a so called fast-chirp configuration, i.e. sending $N_c$ frequency ramps in a sequence, the Doppler induced frequency shift $f_D$ can be extracted. Finally, by using antenna arrays in a multiple-input multiple-output (MIMO) configuration, the angle of arrival can be obtained from the phase differences at different receive antenna positions [58]. Altogether, the frequency and phase components to be estimated, can be described as

$$f_{\text{range}} = \frac{2\,Br}{c_0 T_c}; \quad f_D = \frac{2v_r}{\lambda};$$
$$\nu(m,\phi) = \frac{2kmd\,\sin(\phi)}{\lambda}, \tag{1}$$

where the constant $c_0$ corresponds to the velocity of light in vacuum, $v_r$ is the relative velocity of a target, $\lambda$ the signal wavelength, $k$ the wavenumber, and $d$ the antenna spacing of the virtual MIMO array. The fast-time frequency $f_{\text{range}}$ gives the measured frequency shift induced by the signal delay, whereas the slow-time frequency $f_D$ describes the frequency shift induced by the relative velocity $v_r$ between the target and the radar. Finally, the spatial "frequency" $\nu(m,\phi)$ corresponds to the phase shift of the reflected signal at an azimuth angle $\phi$ observed at the antenna element $m \in [0,\ldots,N_{Rx}-1]$.

The sum of all $K_t$ incoming signal reflections is computed for each antenna element, down-mixed with the transmit signal and filtered. The sampled receive signal thus consists of a superposition of $K_t$ sinusoids with three frequency components

$$s_{IF}(l,m,n) = \sum_{i=1}^{K_t} a_i e^{j2\pi\left(\nu_i(m,\phi_i)+f_{D,i}lT_c+f_{\text{range},i}nT_s\right)} + w(l,m,n), \tag{2}$$

with $l = [0,\ldots,N_c-1]$, $m = [0,\ldots,N_{Rx}-1]$, $n = [0,\ldots,N_s-1]$, the complex receive amplitudes $a_i$, and additive white circular complex Gaussian noise $w(l,m,n)$ [59]. The three components $\nu_i, f_{D,i}$, and $f_{\text{range},i}$ correspond to spatial "frequency", relative velocity frequency and range frequency for each target, respectively. Thus, equation (2) describes the IF signal simulated for each sample point, chirp and antenna. The receive amplitude can be calculated for each target using the radar equation

$$a_i^2 = \frac{P_{Tx}G_{Tx}G_{Rx}\sigma_i\lambda^2}{(4\pi)^3\,r_i^4}, \tag{3}$$

with a total transmit power $P_{Tx}$, a transmit antenna gain $G_{Tx}$, a receive antenna gain $G_{Rx}$, and the target's RCS $\sigma_i$ and range $r_i$ [60].

In an attempt to model a realistic system setup while keeping the computations simple, a complete radar transceiver, similar to the one described in Fig 1, is simulated using the Phased Array System Toolbox™. We use a fast-chirp FMCW configuration at 77.5 GHz with a bandwidth $B$ of 1 GHz, which corresponds to a maximum unambiguous range of 50 m, typical for mid-range radar. The radar system generates $N_c = 64$ consecutive linear frequency chirps, transmitted with a peak power of $P_{TX} = 13$ dBm via one isotropic transmit antenna. During the scattering of the signals at $K_t$ targets, the respective target RCS for different road users is used. Additionally, reflections from targets visible in line of sight are considered in the simulation. This helps to have partial or complete occlusion scenario. On the receiver side, a uniform linear array (ULA) of $N_{Rx} = 8$ identical receive antennas with inter-element spacing $d$ of exactly $\lambda/2$ is utilized. Receive antennas are modeled with a gain of $G_{Rx} = 16$ dB, while phase noise is introduced within the receiver with a noise figure of NF = 4.5 dB. All relevant system parameters, used for simulations are summarized in Table 1.

### C. SIMULATION APPROACH

A detailed radar system model, based on MATLAB®'s Phased Array System Toolbox™, enables the simulation of

**TABLE 1. FMCW radar simulation parameters and resulting system properties used throughout this work.**

| Parameter | Abbr. | Value |
|---|---|---|
| Ramp center frequency | $f_c$ | 77.5 GHz |
| Bandwidth | $B$ | 1 GHz |
| Sampling frequency | $f_s$ | 1 MHz |
| Number of samples/chirp | $N_s$ | 256 |
| Unambiguous range | $r_{max}$ | 50 m |
| Range resolution | $r_{min}$ | 15 cm |
| Chirp duration | $T_c$ | 10 µs |
| Chirp repetition time | $T_{cc}$ | 150 µs |
| Number of chirps/frame | $N_c$ | 64 |
| Frame duration | $T_{Frame}$ | 9.6 ms |
| Unambiguous velocity | $v_{max}$ | 3.9 m/s |
| Velocity resolution | $v_{min}$ | 0.15 m/s |
| Number of Tx antennas | $N_{Tx}$ | 1 |
| Number of Rx antennas | $N_{Rx}$ | 8 |
| Antenna element spacing | $d$ | $\lambda/2$ |
| Transmit power | $P_{Tx}$ | 13 dBm |
| Receive antenna gain | $G_{Rx}$ | 16 dB |
| Receiver noise figure | NF | 4.5 dB |

different system and modulation parameters. In order to consider VRUs and their micro-Doppler signatures, we had to further extend the simulation framework. Therefore, the generation of raw radar data is achieved by using point target scatterer simulations in combination with realistic motion models for VRUs. In order to provide a simple and intuitive way of generating scenarios, MATLAB®'s driving scenario designer[6] is used to set up the environment, the road users and their trajectories. From this scene, radar scattering targets with an integrated motion model are created for each pedestrian and cyclist. The resulting framework yields raw radar data in time domain, which will be used throughout this work.

We divide the environment model of an automotive scenario into multiple parts: channel model, target model and trajectory model. In the following, these will be described in detail.

### 1) CHANNEL MODEL

In general, the propagation of RF signals induces a phase shift on the receive signal. Additionally, the signal is attenuated, we model this so-called path loss and the range dependent phase shift by a two-ray free-space propagation channel, assuming narrowband signals. This is a common assumption in automotive fast-chirp radar signal processing, and greatly simplifies the Doppler frequency estimation [8]. The chosen two-ray channel is essentially a very simple model of a multi-path channel with just a line-of-sight component plus a single ground reflection.

### 2) TARGET MODEL

In order to obtain an accurate receive signal including micro-Doppler components, the simulation also needs to take into account several scattering points of a human body or a bicycle. For the human motion model, an implementation of the so called Eigenwalker model is used [61]. It has some

[6]https://de.mathworks.com/help/driving/ref/drivingscenariodesigner-app.html

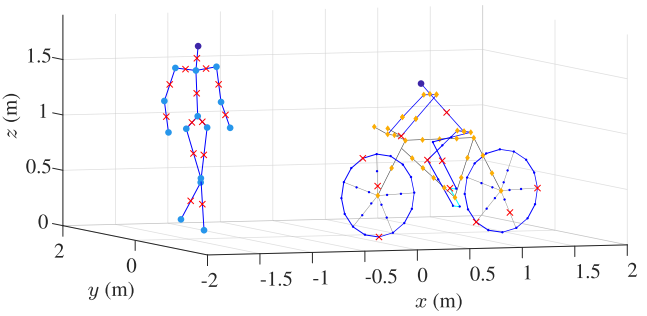


**FIGURE 2. Dynamic VRU point target models for a pedestrian and a cyclist. The pedestrian joint positions are displayed as light blue circles, while static points of the cyclist model are shown by yellow diamonds and dynamic points are highlighted as blue dots. All scattering positions are marked as red crosses.**

important benefits, such as a more realistic gait pattern in comparison to the MATLAB® default human motion model, as well as the possibility to alter the motion characteristics with respect to gender. Simulation data is generated using female, male, and the average human gait motions. Motions are calculated based on the joint positions of all extremities, whereas the middle point between two joints is considered as a radar scattering point. The full pedestrian model is shown in Fig. 2, where joints are depicted as light blue circles, whereas the scattering points are illustrated as red crosses.

To the right of the pedestrian, the cyclist model is displayed, where points with linear velocity are marked by yellow diamonds and dynamic points with rotational velocities are marked by blue circles. We use a modified version of the default MATLAB® motion model for cyclists, with overall less scattering points (marked by red crosses), to reduce the computational load and obtain less cluttered micro-Doppler spectra. The wheels are modeled with 5 wheel spokes, while just a small subset of the total available points in the model are picked as reflection points. The upper body, as well as the bike frame are kept "static", i.e. they only move linearly with the total velocity of the cyclist. In contrast, the wheels and pedals follow a circular motion, while the riders legs are fixed to the pedal and hip joint positions and thus experience an oscillating motion. Examples of relative velocities obtained from the described motion models are shown in Fig. 3, both road users move with constant velocities along a linear trajectory towards the radar sensor. The velocities in Fig. 3a are obtained from a male (solid lines) and a female (dashed lines) pedestrian model. The pedestrians are walking with 1.0 m s$^{-1}$ towards the radar sensor, noticeable by the oscillating pattern around this velocity value. Differences from the gender based gait are visible as distortions as well as different maximum velocities. However, the absolute differences are relatively small and only in the order of 0.1 m s$^{-1}$. Apart from that, Fig. 3b shows the different radial velocities of the cyclists' scattering points, with a constant velocity of $v = 3.0$ m s$^{-1}$ of the frame and body, as well as faster radial velocities for the wheel spokes, pedals and legs. Different cyclist models are obtained by using different subsets of scattering points from the total available motion points. Figure 3b shows the effect

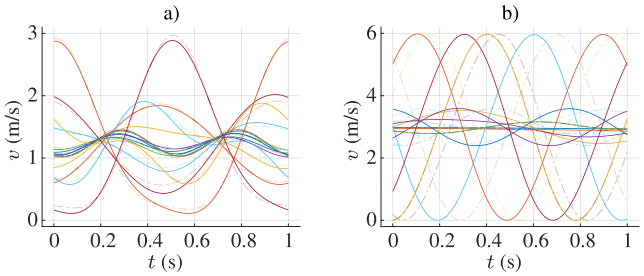


**FIGURE 3.** Relative velocities of the VRU scattering points for a female (dashed) and a male (solid) pedestrian (a) walking with 1.0 m/s and two different cyclists (b), both with a constant speed of 3 m/s.

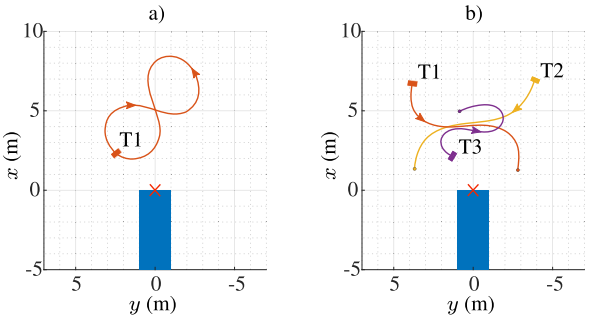


**FIGURE 4.** Simulation scenarios with one target (T1) used for training (a) and two (T1, T2) or three (T1, T2, T3) targets used for testing (b), with their respective trajectories. The radar sensor position is marked by a red cross.

on the simulated velocities for two different cyclists (solid lines vs. dashed lines) for two different subsets of scattering points. Overall the pedestrian model uses 14 scattering points, while the bicycle model uses 11 scattering points.

In the next step, the radar cross-sections for both VRU targets are derived using a simplified model. For the pedestrian model, the average RCS for the respective operating frequency is determined and then divided by the total number of scatterers to obtain the individual scattering point's RCS. For 77 GHz, an average RCS in azimuth of $\sigma_{\text{ped}} = -8.1\,\text{dB}\,\text{m}^2$ is used, in agreement with [62]. The same principle holds for the cyclist model, but the model uses a measured RCS pattern at 77 GHz from [63].

#### 3) TRAJECTORY MODEL

To have a complete automotive scenario, road users and the ego car equipped with the radar sensor need additionally to be moving along some trajectory. For this purpose, nonlinear motions for all VRUs, with constant velocities, are defined. Their trajectories are interpolated from individual waypoints as a piecewise clothoid curve, in order to obtain smooth motions. However, in reality there is an unaccountable amount of possibilities and permutations for the number and classes of targets, their individual motion, velocity and trajectory, as well as other physical and environmental parameters, e.g. radar modulations or channel characteristics. Thus, training a machine learning model on an equal distribution of possibilities from the entire space of input data is not feasible. Instead, we aim to include different features of targets and environments in our database and try to get the network to learn these features separately. Consequentially, the network should be able to infer the correct class of targets based on individual features and not on a combination of them.

For the combined target tracking and classification, we generate different scenarios of VRU targets following predefined trajectories. The scenario used to extract single target micro-Doppler signatures is shown in Fig.4a. The ego vehicle, equipped with the radar sensor, is kept stationary. The sensor position is depicted by a red cross. The target follows a trajectory with an 'eight' shape, as the actual view angle of the pedestrians and cyclists has a strong impact on micro-Doppler based classification [64]. Therefore, we want to make sure to include all the variations in our training data. To capture the whole set of variations, the 'eight' shaped trajectory is in addition rotated along the y-axis with an angle of 45°. For the testing data, we use a scenario with multiple road users and crossing trajectories, limited to a maximum of three targets. In Fig. 4b, three targets (T1-T3) with their individual trajectories, highlighted in different colors, are displayed. Each target can be either a pedestrian or a cyclist, with the aforementioned properties. All possible permutations of pedestrian and cyclist combinations are used for these two and three target scenarios.

## III. AUTOMOTIVE RADAR SIGNAL PROCESSING

The previously described simulation framework outputs raw sensor data, which needs to be further processed in order to obtain the target estimates for each data frame. This chapter gives a summary of the necessary automotive radar signal processing chain, with a special focus on the generation of micro-Doppler signatures. According to (2), the received IF signal contains all the information about the target's radial distance, relative velocity besides the spatial information in azimuth dimension. As indicated in Fig. 1, the sampled signal of a single receive channel is stored in a matrix-like format. Combining the time-domain data of multiple receive channels, a radar data cube with three dimensions, namely samples, chirps and antennas, is obtained. For a target classification and target tracking, all targets need to be successfully identified and separated in the receive signal. Thus, targets have to be resolved in either of the three available dimensions: range, velocity or angle [6]. Additionally, the extraction of a micro-Doppler signature is done based on target reflections from a specific range and, in turn, requires a target detection in the range domain.

### A. RANGE-DOPPLER-ANGLE PROCESSING

As a first step, a mean subtraction is applied along the samples and chirps, to suppress Tx-Rx leakage as well as reflections from stationary targets, also referred to as clutter. This clutter removal is also known as moving target indicator (MTI) processing [65]. In order to resolve targets, the corresponding frequencies from (1) need to be estimated from the pre-processed signal. This is accomplished by using a

**TABLE 2. Radar signal processing parameters used throughout this work.**

| Parameter | Abbr. | Value |
|---|---|---|
| CFAR threshold | PFA | 1e−5 |
| Number of FFT points (fast-time) | $\text{NFFT}_\text{r}$ | 256 |
| Number of FFT points (slow-time) | $\text{NFFT}_\text{d}$ | 256 |
| DBSCAN Epsilon neighborhood | $\epsilon$ | 4 |
| DBSCAN Minimum number of neighbors | minPoints | 6 |
| Sliding window size | $N_\text{win}$ | 200 |
| Sliding window overlap | $N_o$ | 48 |
| Number of STFT points | $N_\text{STFT}$ | 512 |

3D fast-Fourier transform (FFT) along the respective dimensions, effectively converting the samples dimension to the fast-time, the chirps dimension to the slow-time, and the antennas dimension to the azimuth angle, respectively [58], [59]. The resulting 3D spectrum is referred to as range-Doppler-angle (RDA) cube. Two targets can be resolved from any dimension in the resulting RDA cube, e.g. if they are separated by at least $r_\text{min}$ or $v_\text{min}$ in range or velocity dimension (compare the exact parameter values in Table 1).

### B. TARGET DETECTION AND CLUSTERING

Following the FFT processing, the targets' range and velocity can be extracted by searching for local maxima in the magnitude spectrum. Usually the detection is carried out in either the range-Doppler (RD) or in the range-Angle (RA) domain. The actual detection is achieved by using a constant false alarm rate (CFAR) algorithm, as it provides a better performance, compared to applying a constant threshold, under varying noise levels. It benefits from an adaptive threshold depending on each individual cell's signal-to-noise ratio (SNR). More specifically, the ordered statistics (OS)-CFAR algorithm is used in our work, as its performance is superior for closely spaced targets [65]. Applying the OS-CFAR algorithm for the peak detection, requires a clustering to group all detections of the same individual target. Therefore, a density-based spatial clustering (DBSCAN) is used [66].

### C. MICRO-DOPPLER SPECTROGRAM

Even though, reflections from VRUs contain multiple velocity components due to the micro-motions of the different extremities, most of these components can not be separated in the RD spectrum, as their range or velocities are very closely spaced. To extract the micro-Doppler components of a single target, just the FFT along the short time dimension has to be computed over the raw data $s_\text{IF}$ cube. Doing so, the range spectrum $S_r(r, t)$ is obtained for each chirp. Then, a specific target range $r_\text{det}$ is selected to obtain the signal $S_r(r_\text{det}, t)$ with an effective sampling rate equal to the chirp-to-chirp duration $T_\text{cc}$. Finally, a short-time Fourier transform (STFT) is performed, which can be described as a Fourier transform applied within a sliding window pattern of the signal, resulting in overlapping windows

$$\text{STFT}(\tau, \omega) = \int_{-\infty}^{\infty} S_r(r_\text{det}, t) w^*(t - \tau) e^{-j\omega t} dt, \qquad (4)$$

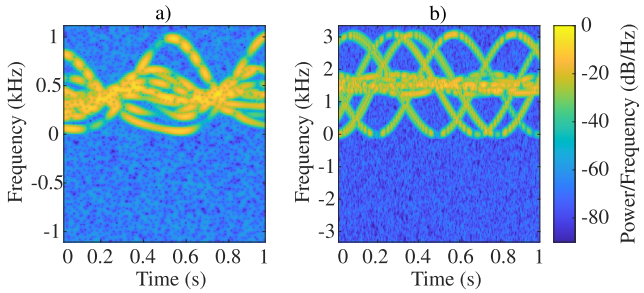


**FIGURE 5. Simulated micro-Doppler signatures of a male pedestrian (a) and a cyclist (b) from the same scenario as used for Fig. 3.**

This operation results in a Doppler frequency spectrum over time, which is usually visualized as the micro-Doppler spectrogram by taking its square magnitude [17]. Applying a smaller window function $w^*(t - \tau)$ will result in a better time resolution at the cost of frequency resolution and vice-versa [67]. Adapting this trade-off between time and frequency resolution dynamically, is still an open point in the current research. However, it could be addressed by using the concept of a wavelet transform or even learning window parameters using neural networks [68]. In this work, the STFT of the micro-Doppler signatures for both pedestrian and cyclist are estimated using a Kaiser windowing function with a sliding window size of $N_\text{win} = 200$ and an overlap of 48 samples. The number of FFT points $N_\text{STFT}$ for STFT is set to twice the window size $N_\text{win}$ which is a factor of $2^n$ i.e 512.

By combining the previously described simulation framework and this radar signal processing chain, we are now able to obtain micro-Doppler signatures for arbitrary scenarios. In order to verify the framework from chapter II, we use the same scenario like in Fig. 3, where road users are walking with a constant velocity towards the radar sensor. We extract the micro-Doppler signatures in a time window of 1 s, by using the described STFT approach. The results, obtained with an effective sampling time of 150 µs for the pedestrian scenario and 100 µs for the cyclist scenario, are shown in Fig. 5. The individual frequency components, resulting from scattering points with different relative velocities, are visible. Note the strong agreement of the overall frequency characteristics with Fig. 5, as well as the fluctuations in the total received amplitudes, resulting from the used multi-path channel.

### D. TARGET TRACKING

Conventional radar signal processing usually applies a tracking algorithm after the clustering, to filter measurements over time, and to create object detections and tracks of individual targets. False detections are usually also eliminated during the tracking. In order to avoid decreasing the measurement accuracy as well as the introduction of inherent noise, the usage of recursive filters is preferred in literature. The most widely used tracking algorithms are Kalman filters. The performance of Kalman filters relies on the state vector (i.e. the parameters

to be tracked), measurement noise, process noise and the transition from measurement to the state space.

In our case of tracking radar detections, the state vector can be described as $\mathbf{x} = \begin{bmatrix} px & py & v & \theta \end{bmatrix}^T$, where $px, py, v, \theta$ are the position coordinates along the x- and y-axis, the radial velocity and the azimuth angle, respectively. Due to low variations in the spatial dimension, azimuth information is also used as part of the state vector and improves the robustness of the target localization. We use the Root MUSIC[7] algorithm for direction of arrival (DoA) estimation. Generally, heading angle and turn-rate bring additional information for a dynamic target with non-linear motion.

However, as VRUs may have very high variations in their heading, parameters such as orientation angle, turn-rate etc are not considered in definition of the state vector. Additionally, the variance corresponding to each of the estimated localization parameters (range, Doppler, azimuth) are calculated for each detection. While SNR of each detection is used for variance over range and Doppler, variance over azimuth angle is calculated as the ratio of maximum power of the beamformer over the noise around the corresponding bin in the RD-map. Considering a radar system which is stochastic in nature and observations are either prone to noise or incomplete, a recursive Bayesian estimation algorithm becomes the popular choice. This helps to periodically predict the posterior density of the system state for every new observation.

For most general real-world (nonlinear, non-Gaussian) systems, the multi-dimensional Bayesian recursion becomes intractable and therefor, approximations have to be used. This includes methods such as Gaussian approximations (extended Kalman filters), hybrid Gaussian methods (score function EKF, Gaussian sum filters), direct and adaptive numerical integration (grid-based filters), sequential Monte Carlo methods (particle filters) and variational methods (Bayesian mixture of factor analyzers). Especially, the extended Kalman filter (EKF), a sub-optimal approximation of the recursive Bayesian framework, applied to a Gaussian random variable (GRV) of a non-linear state, is widely used. It approximates and propagates the state distribution through the first-order Taylor series linearization, which expands the non-linear state around a single-point. As a result, the EKF is not able to capture the uncertainty of the distribution, introducing large errors in the estimation of the true posterior mean and covariance, respectively. Alternatives can be unscented Kalman filters (UKF), which use deterministic sampling filters, i.e a sigma-point Kalman filter (SPKF), to approximate the GRV by a minimal set of sample points. These sample points can capture the true mean and covariance of the GRV. While Fig. 6 gives a visual overview on the prediction and update operation of the UKF to track mean $(\hat{x}_{k-1})$ and covariance $(P^x_{k-1})$ of the input state vector at a time instance $k-1$, the algorithm 1 gives a mathematical understanding on its implementation.

[7]https://www.mathworks.com/help/phased/ref/musicdoa.html

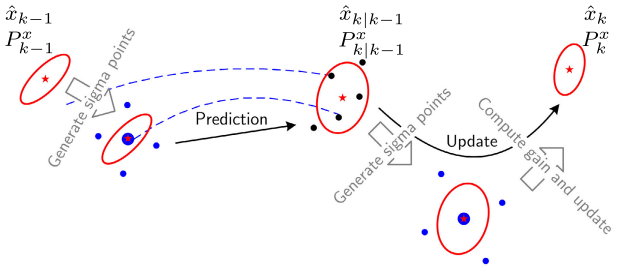


**FIGURE 6. Graphical representation of predict and update stage for an UKF where mean and variance is estimated at each stage by approximation over sigma point matrix.**

### 1) UNSCENTED TRANSFORMATION

The state-of-the-art tracking algorithms for automotive use-cases mainly focus on single modalities by having target's localization information as a state vector [26], [27]. Additional target parameters such as Doppler spectra are either ignored or computed separately. In this paper, the authors use both the localization and appearance model for the tracking of detected targets, by augmenting the state-vector target features.

To develop UKF, UT is applied at both prediction and update steps, which includes non-linear state transformation based on $f$ and $h$, respectively. As input, state vector $x_{k-1}$ of dimension $n_x$ with mean $\hat{x}^{(i)}_{k|k-1}$ and covariance $P^x_{k|k-1}$ is given. At prediction stage, sigma points $\hat{x}^{(i)}_{k-1|k-1}$ are generated which is goes under UT $(f(.))$ to estimate predicted mean $\hat{x}_{k|k-1}$ and covariance $P^x_{k|k-1}$ of state vector. Since predicted mean and variance changed, a new set of sigma point matrix is calculated due to its dependency on mean and variance. Afterwards, the new sigma point matrix is transformed into measurement space using $h(.)$ as transformation function.

A constant-velocity (CV) system is considered with the localization state vector $\mathbf{x}$. A non-linear measurement model $h(\cdot)$ accounts for the transformation of the state vector into the measurement domain. Mapping part of the localization parameters (radial range and azimuth angle) from the tracker's state vector to the measurement domain follows a non-linearity (Cartesian to Spherical), whereas mapping the radial velocity and augmented parameters (appearance embedding) corresponds to an identity mapping between state vector and measurement domain. However, the overall non-linear transformation in the process model $x^P = g(x_a)$ and the measurement model $z^P = h(x^P_a)$ can be achieved through unscented transformation, using so-called 'sigma points'. These are generated to approximate the statistical properties of the state distribution [27].

In addition, due to high similarity within input space (Doppler spectra), it would be hard for tracker to discriminate between different appearance embedding of VRUs. It is important to note that the original dimension of the micro-Doppler spectra is very large and thus, will increase computational complexity for the tracker to estimate the new state vector. As a result, the choice of feature extractor becomes very critical to bring unique appearance modalities

**Algorithm 1** Unscented Kalman Filter

**Prediction**: Generate sigma points

$$\hat{x}^{(i)}_{k-1|k-1}, i = 0, 1, \cdots, 2\,n_x$$

$$\hat{x}^{(i)}_{k|k-1} = f\left(\hat{x}^{(i)}_{k-1|k-1}\right), i = 0, 1, \cdots, 2\,n_x,$$

$$\hat{x}_{k|k-1} = \sum_{i=0}^{2\,n} W_i^{(m)} \hat{x}^{(i)}_{k|k-1},$$

$$P^x_{k|k-1} = \sum_{i=0}^{2\,n_x} W_i^{(c)} \left(\hat{x}^{(i)}_{k|k-1} - \hat{x}_{k|k-1}\right) \times \left(\hat{x}^{(i)}_{k|k-1} - \hat{x}_{k|k-1}\right)^T + Q,$$

**Update**: Generate sigma points

$$\hat{x}^{(i)}_{k|k-1}, i = 0, 1, \cdots, 2\,n_x$$

$$\hat{y}^{(i)}_{k|k-1} = h\left(\hat{x}^{(i)}_{k|k-1}\right), i = 0, 1, \cdots, 2\,n_x,$$

$$\hat{y}_{k|k-1} = \sum_{i=0}^{2\,n_x} W_i^{(m)} \hat{y}^{(i)}_{k|k-1},$$

$$P^y_{k|k-1} = \sum_{i=0}^{2\,n_x} W_i^{(c)} \left(\hat{y}^{(i)}_{k|k-1} - \hat{y}_{k|k-1}\right) \left(\hat{y}^{(i)}_{k|k-1} - \hat{y}_{k|k-1}\right)^T + R,$$

$$P^{xy}_{k|k-1} = \sum_{i=0}^{2\,n_x} W_i^{(c)} \left(\hat{x}^{(i)}_{k|k-1} - \hat{x}_{k|k-1}\right) \left(\hat{y}^{(i)}_{k|k-1} - \hat{y}_{k|k-1}\right)^T,$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + P^{xy}_{k|k-1} \left(P^y_{k|k-1}\right)^{-1} \left(y_k - \hat{y}_{k|k-1}\right),$$

$$P^x_{k|k} = P^x_{k|k-1} - P^{xy}_{k|k-1} \left(P^y_{k|k-1}\right)^{-1} \left(P^{xy}_{k|k-1}\right)^{\top},$$

into the tracker and get a better discrimination between the targets. Additionally, tracker takes variance as input noise over observed input state vector. This further imposes another challenge to find variance over extracted appearance embedding for the integration of appearance model into tracker's state vector. The details about different data processing techniques, feature extraction architecture together with making it compatible with Bayesian framework and the different optimization functions are addressed in the following Section IV and V, respectively.

## IV. TARGET APPEARANCE MODEL

In general, the appearance model of a target in computer vision consists of statistical information about the target's shape, size or motion characteristics. In order to uniquely identify the target between similar looking targets, the target's motion characteristics are considered in this paper. Therefore, different features, in our case micro-Doppler signatures of each target, are extracted to learn a unique statistical model of the target's appearance. However, the characteristics of the extracted features depend on the choice of the applied feature extractor and especially on the used data-sets. In consequence, the data preparation and the feature extraction as the two major stages are needed, which are discussed in the following.

### A. DATA PREPARATION

As the accuracy and generalization of the feature-learning algorithm depends highly on the quality of data-sets and the information variability, preprocessing becomes essential. Different techniques of preprocessing, followed by a data augmentation, resolve these issues. Nevertheless, data preprocessing is critical as it forms the basis of subsequent feature extraction methods. In general, suitable methods are selected based on the characteristics of the used datasets and the problem definition, respectively. In real radar measurement scenarios, radar cross-sections will be inherently smaller for VRUs, i.e. as opposed to vehicles, resulting in a lower signal-to-noise ratio, which in consequence leads to weaker signatures. Additionally, varying viewing angles, as discussed in Section II-C3, also distort the signatures of VRUs, resulting in deformed or missing micro-Doppler components. Therefore, the most common techniques of preprocessing, like noise removal, morphology and transformation correction, are not very well fitting for the radar domain. That's why, in this paper, the preprocessing stage is split into three consecutive steps. First, the signatures are converted from a linear to a log scale, which strengthens the weaker Doppler components of the VRUs. Afterwards, the extracted signatures are standardized to have a zero mean, right before the extracted signature magnitudes are normalized to the range of 0 to 1 in the last step.

However, in reality, the sample distribution (mean) over the training set $p\,(x_i \mid \hat{\theta})$ is not large enough in comparison to the actual distribution (population mean) $p\,(x_i \mid \theta)$. This influences the model uncertainty (epistemic) [69] and therefore, the concept of data augmentation next to data preprocessing is required. Additionally, this also addresses the bias-variance trade-off by balancing the distribution over a class. As a result, mainly two data-augmentation techniques from the literature [70] are exploited in the paper. This includes on the one hand an addition of artificial (Gaussian) noise, which can emulate superimposed clutter noise on the received echo signal. On the other hand vertical flips are used, to consider also varying Doppler frequencies caused by different directions of target's motion (towards or away) relative to the radar sensor.

A visual summary for each stage in data-processing and data-augmentation is given by Fig. 7. Here an arbitrary example of a cyclist micro-Doppler signature is considered. While Fig. 7(a) shows the Doppler spectrum of one target in linear scale, Fig. 7(b-c) corresponds to a data preprocessing step, where the linear-scale is converted to logarithmic-scale first, followed by the normalization of the spectrum. Moreover,

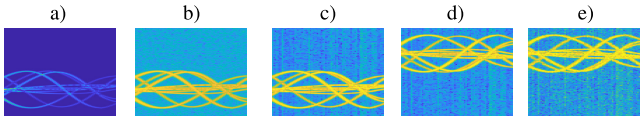


**FIGURE 7.** Visual illustration of Doppler signatures in (a) linear-scale and (b) logarithmic-scale and the normalization (c) as data preprocessing step. Additionally, (d) and (e) shows vertical flip and augmented Gaussian noise.

**TABLE 3.** Similarity indices (SSIM) for simulated micro-Doppler spectra of inter and intra classes for different VRU targets.

| | mPed | fPed | nPed | Cyc1 | Cyc2 | Cyc3 |
|---|---|---|---|---|---|---|
| mPed | 1.0 | 0.71 | 0.68 | 0.57 | 0.60 | 0.61 |
| fPed | - | 1.0 | 0.70 | 0.64 | 0.62 | 0.66 |
| nPed | - | - | 1.0 | 0.61 | 0.61 | 0.63 |
| Cyc1 | - | - | - | 1.0 | 0.52 | 0.55 |
| Cyc2 | - | - | - | - | 1.0 | 0.53 |
| Cyc3 | - | - | - | - | - | 1.0 |

Fig. 7(d-e) show vertical flips and augmented Gaussian noise as part of the data-augmentation techniques. Additionally, the signatures are resized to $64 \times 64$ and used as a gray-scale for training the network to reduce computational cost. Thus, a total of 476 samples for each class of pedestrian and 680 samples for each class of cyclist is created.

### B. FEATURE EXTRACTION

Using the processed data, the goal of feature extractor is now to find reduced sets of parameters that can be used as key features. These should contain information to define and differentiate between different available classes, being at the same time robust to environmental changes (e.g. regarding the target view points or motions). This is done by projecting the input space into a latent space dimension. With an increasing size of the input dimensions, however, the complexity of the problem statement grows exponentially. As a result, feature extractors tend to over-fit easily to the training data. In order to avoid over-fitting, different regularization methods ($L1$, $L2$) are used together with the feature learning methods. This helps by applying penalties to learned model parameters and weight coefficients.

Prior to the evaluation of the feature extraction methods, the structural similarity index measure (SSIM) over all permutations of sample classes are calculated. SSIM values closer to 1 indicate a high similarity between two images, and is in contrast to e.g. MSE more robust to noisy variations of the image [71]. Results on our dataset show a strong visual similarity among intra-class samples as well as between inter-class targets, as indicated in Table 3. Most similarity indices lie around 0.5-0.7, without noticeable differences between cyclist and pedestrian class combinations. In general this demonstrates the complexity of the problem and the importance of finding the optimum feature embeddings, which can be used for distinct appearance modeling of targets.

The history of feature extraction and selection methods in literature is substantial. The optimal choice for the feature extractor for this work was done via a systematic evaluation of

**TABLE 4.** Test accuracy on Doppler classification using different feature extractor.

| | PCA | ICA | LDA | LLE | CNN |
|---|---|---|---|---|---|
| Test Accuracy | 24.14% | 16.09% | 22.99% | 26.44% | 42.5% |

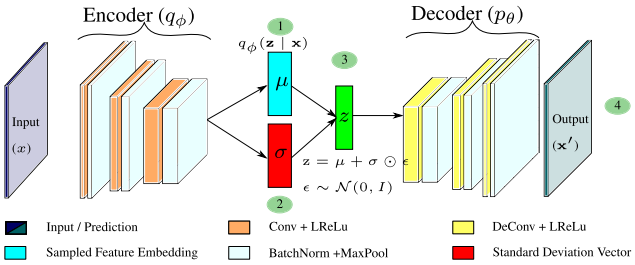


**FIGURE 8.** A summary of CNN based feature extractor illustrating different layers used for metric learning (1), Bayesian inference (1 and 2) and image reconstruction (4).

available and most commonly used methods such as principle component analysis (PCA), independent component analysis (ICA), linear discriminant analysis (LDA) and locally linear embedding (LLE). Additionally, with a huge popularity of deep learning in vision techniques in the last decades [28] and the recent interest in radar applications, a convolutional neural network (CNN) based architecture is also evaluated. In this work, CNN is trained by using the cross-entropy loss, similar to the concept of image classification. Therefore, an encoder based architecture is used during training, gradually reducing the size of layers. Table 4 shows the test accuracy of different feature extractors, comparing CNN against the PCA, ICA, LDA and LLE. The test accuracy of CNN based approaches gives a clear indication on choice of feature extractor over PCA, ICA, LDA and LLE. Additionally, it is interesting to observe that CNN performance is still less than 50% which indicates the strong correlation within pedestrian and cyclist sub-class as well as between pedestrian and cyclist class, similar to as indicated by SSIM measure in Table 3. Furthermore, the CNN-encoder based classification architectures are good to learn global learning based on spatial dimension. Whereas, it fails to provide robust local representation corresponding to input spatial dimension for unique identification of target class. This indicates on better choice of CNN architecture and different optimization function. As a result, an encoder-decoder based architecture is evaluated for the appearance modeling. Additionally, the make learned features follow Bayesian representation with mean and variance and the concept of variation inference is applied. Fig. 8 gives an overview of the structure of the network architecture used for the later experiments.

The design-space of the CNN architecture involves large number of parameters which makes it hard to find the optimum architecture for the given problem definition. Thus, the choice of the architecture design hyper-parameters is mainly inspired from [66], where the authors used a similar architecture for the target detection on sparse radar RD-maps. Both encoder and decoder have a 3-layered convolution layer

with a rectified linear unit (ReLu) as the non-linearity function. Considering the nature of the pre-processed input data, additional concepts like, batch-normalization or leaky ReLu, were not considered. While only the encoder part is integrated inside the tracker, both encoder and decoder are used during the network training.

### C. VARIATIONAL AUTO-ENCODER (VAE)

The design of VAE architecture combines concept of auto-encoder (AE) and variational inference. While the auto-encoder (AE) is used for learning the identity mapping function by reconstructing the input from its reduced representation, variational part helps to provide regularity and interpretability knowledge over latent space by learning corresponding mean and variance. This in return helps the architecture to generate new data. The architectures contain an encoder $q(.)$ network and a decoder network $f(.)$ parameterized by $\phi$ and $\theta$, respectively. The encoder network acts as a dimension-reduction by translating the higher input dimension ($x$) into the feature latent space ($\mathbf{z} = q_\phi(\mathbf{x})$). Afterwards, it is reconstructed back to the input space by the decoder ($\mathbf{x}' = p_\theta\left(q_\phi(\mathbf{x})\right)$). In consequence, the performance of the network depends on finding the optimal identity function such that $\mathbf{x}' \sim x$, which relies on an extracted feature in the latent space. Thus, network parameters $(\phi, \theta)$ are optimized using the cross-entropy loss (CE) instead of the mean square error (MSE) to avoid the vanishing gradient problem due to the non-linear *sigmoid* nature of the output [72].

Often, identity mapping functions are prone to the problem of over-fitting, specially in case of a high-dimensional input with a high redundancy (e.g. images). However, noisy augmented data helps to avoid over-fitting, as noisy data can be interpreted as a regularizer by randomly dropping (corrupting) input data, which is similar to a dropout [73]. Additionally, VAE based architecture applies a constraint (prior) over the latent space by mapping the latent to the distribution instead of the fixed latent vector. The prior on encoding vector also acts as a regularizer. To avoid intractable integrals in the process of estimating the true posterior distribution, a re-parameterization trick is used, which restricts the encoded distribution to be normal distributed. Eq. 5 shows the reparameterization, where $\odot$ is an element-wise product.

$$\begin{aligned} \mathbf{z} \sim q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right) &= \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{2(i)}\boldsymbol{I}\right), \\ \mathbf{z} &= \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}) \end{aligned} \quad ,(5)$$

As a result, the VAE architecture invokes an additional loss function, i.e. the Kullback-Leibler (KL) divergence, which brings continuity and completeness in the latent space. The total loss function is summarized as a linear combination of the CE and the KL, as depicted in Eq. 6.

$$\begin{aligned} \mathcal{L}_{\text{vae}} &= \mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{KL}} \\ &= -\mathbb{E}_{q_\phi(z|x)}(\log p_\theta(x \mid z)) \\ &\quad + \text{KL}(q_\phi(z \mid x) \| p(z)), \end{aligned} \quad (6)$$

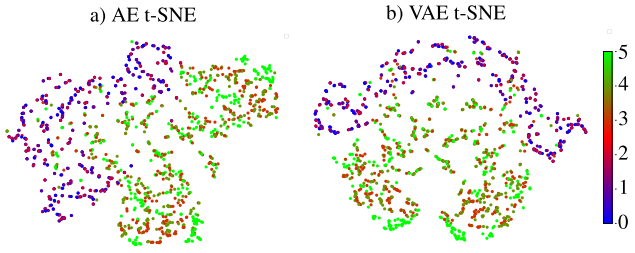


**FIGURE 9.** Illustration of the strong correlation between the sub-class of pedestrian (0, 1, 2) and cyclist (3, 4, 5) using a t-SNE plot for (a) AE and (b) VAE over feature embedding.

The network training is optimized in an iterative scheme for 10 epochs. *Adam* is used as stochastic optimizer with a learning rate of 0.0004 and keeping a default value for other hyper-parameters. After the training and to evaluate the classification accuracy, k-nearest neighbors (k-NN) algorithm [74] is used on the learned latent feature embedding. k-NN was chosen because of its non-parametric nature. However, one could also use a linear classifier such as the support vector machine (SVM). Due to constraints over the latent dimension, VAE shows an average test accuracy of 56.9% in contrast to AE having 55.25% test accuracy. To better understand the confusion between inter and intra-class, feature embedding clusters are visualized in a 2D plane using a t-distributed stochastic neighbor embedding (t-SNE)[8] tool. t-SNE helps to project the high dimensional (original) data space into the desired dimension space (2D or 3D) by projecting samples close to each other, if samples were inherently related to each-other in original space. As shown in Fig. 9(a),(b), both AE and VAE architectures fail to learn and extract distinct features for each target class. Fig. 9 shows that the both AE and VAE networks were able to cluster all sub-groups of cyclist $(3, 4, 5)$ and sub-groups of pedestrians $(0, 1, 2)$ close to each other, but unable to learn distinct features between each target sub-class like female, male and neutral labeled as 0,1 and 2 respectively. As a result and in contrast to the representation learning via CE, MSE or KL-divergence loss, a loss function is defined for learning of a similarity function, using a distance metric learning [75]. The details about different network architecture together with loss function and training environment are described in the next section.

## V. DEEP METRIC LEARNING

Deep metric learning is well studied for images, speech and language tasks and recently applied in the domain of radar sensor data. Several deep metric learning models use Siamese framework, where two or more weight shared sub-networks respectively are trained to learn distance metrics during training. The input data is projected into the embedding space via the learned model during inference. The idea is to use an anchor example, which is fed into one of the neural network

[8]https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

and learn the pair-wise or tuple-wise relationship to other examples that are fed to the other neural network during training [34]. Some challenges of training such networks are online data mining that ensures a representative dataset while training to ensure loss decreases with progressing epochs. Some of these issues are addressed through multi-class N-pair loss, constellation loss, structured loss or magnet loss in literature which not only learn relations to two or three examples per class but rather all the classes together ensuring a better representational space. In terms of distance metrics, distances such as Euclidean, cosine and Kullback-Leibler have been introduced in literature, which aim to learn the non-linear relation among data in the representational space. On the other hand, several modifications to the softmax classifier have been proposed in literature to achieve discriminative capabilities. Different techniques introduce a margin between the classes, whereas in [76] the softmax is disentangled into an individually optimizable inter- and intra-class objective. Further, Dsoftmax loss shows improved results compared to the margin based softmax approaches and center loss [77] proposes to combine the softmax objective with the Euclidean spread of the samples of a class.

In this paper, the authors combines the idea of variational inference over latent embedding in combination with metric and representational learning. This helps the network to learn both the feature embedding and the variance over it. As a result, the learnt input noise can be integrated inside a tracker, which further helps during the data association. Details about its usage and advantage are given in Section VI and VII, respectively. The green numbered blobs, shown in Fig. 10, indicate different layers, which are used to design a loss function specific for different training approaches. While blob '1' is responsible for the metric learning, blob '4' optimizes the image reconstruction error. Both '1' and '2' blobs help the network to have Bayesian inference. Further, to have a fair comparison and similarity between the different approaches for feature extraction, Fig. 10 is illustrated to summarize different architectures. The encoder and decoder network parameters are kept same throughout all the experiments, but are trained with different loss function and additional hyper-parameter optimizations. Additionally, the training environment, like the data preparation and performance metric used for the evaluation of the network, is kept identical throughout the training process. Details about the different optimization functions corresponding to the chosen network architecture are mentioned in the following subsections. Additionally, code for model definition and loss function is made available on Github respository.[9]

### A. TRIPLET NEURAL NETWORK (TNN)

The concept of siamese neural network [78] is used to design a training framework that consists of three identical sub-networks. The architecture is optimized using a triplet

[9]https://github.com/ananddb90/IntegratedBayesTracking

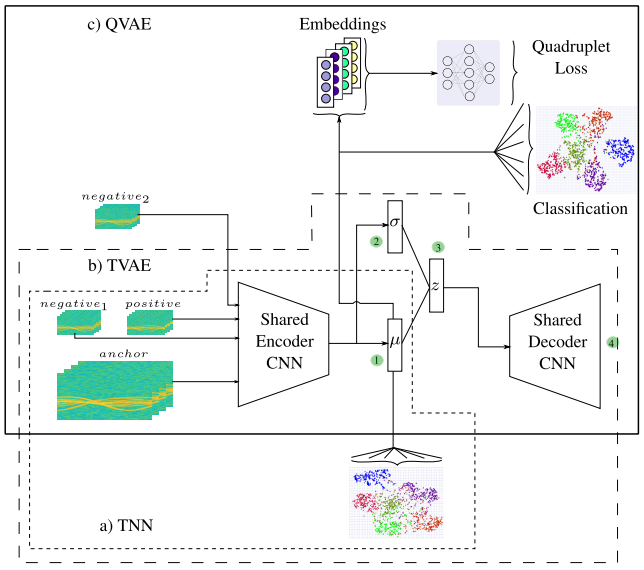


**FIGURE 10. An overview of evaluated architectures and their relation to each other. While all architecture follows k-NN classifier over mean embedding for target classification, the variational inference brings additional Bayesian knowledge over extracted mean using (b) TVAE architecture in contrast to (a) TNN. With introduction of additional negative sample from similar class-group, (c) QVAE improves distinct feature learning. This is done by learning the distance between feature embedding using MLP in contrast to usage of normal L2-Norm.**

loss [34]. The encoder network from the AE framework is considered as sub-networks, keeping the same configuration as used in the AE. Prior to the network training using the triplet loss, triplet pairs are selected. These pairs consist of anchor sample ($x_a$), i.e. any random sample, positive sample ($x_p$), which is from the same class as the anchor, and a negative samples ($x_n$), which is a sample from any different class in comparison to the anchor class. The loss function is computed over feature embedding, i.e. the latent space ($z$), as shown in Eq. 7.

$$\mathcal{L}_{\text{triplet}} = \max(\|q_\phi(x_a) - q_\phi(x_p)\|^2 - \|q_\phi(x_a) - q_\phi(x_n)\|^2 + \alpha_{\text{margin}}, 0), \quad (7)$$

The network is trained following a min-max distance learning between the triplet pairs. While the distance between the anchor and negative samples is maximized by making $d(q_\phi(x_a), q_\phi(x_p)) + \alpha_{margin}$ less than $d(q_\phi(x_a), q_\phi(x_n))$, the distance between the anchor and positive samples is minimized forcing $d(q_\phi(x_a), q_\phi(x_p))$ to 0. Here, $\alpha_{\text{margin}}$ is a hyper-parameter which defines the boundary condition between the similar and dissimilar pairs. For this process, a Euclidean distance function is considered from the available similarity metric function [75].

The choice of the input triplet pair plays an important role in learning the feature embedding. Considering the spatial complexity and similarity between training examples, the mining of input triplet pairs is done considering hard examples, i.e triplets, where the negative sample is closer to the anchor than to the positive $(d(q_\phi(x_a), q_\phi(x_n)) < d(q_\phi(x_a), q_\phi(x_p)))$ and semi-hard

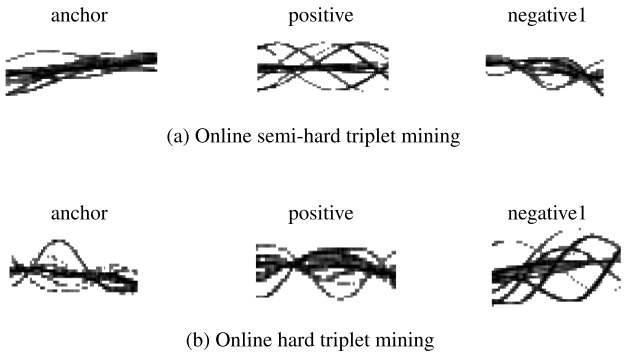


**FIGURE 11.** Sample example of (a) semi-hard and (b) hard triplet pairs of Doppler spectrum extracted during network optimization (online).

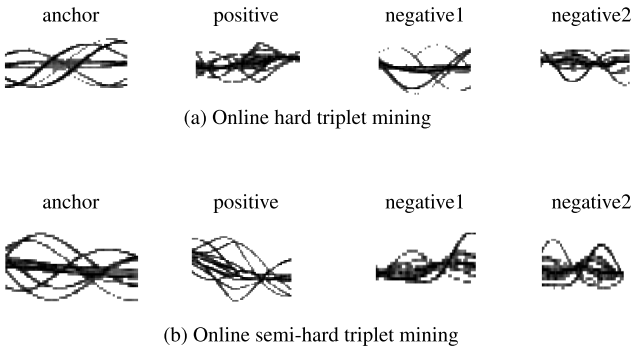


**FIGURE 12.** Sample example of (a) semi-hard and (b) hard quadruplet pairs of Doppler spectrum extracted during network optimization (online).

triplets, where $d(q_\phi(x_a), q_\phi(x_p)) < d(q_\phi(x_a), q_\phi(x_n)) < d(q_\phi(x_a), q_\phi(x_p)) + \alpha_{margin}$. The triplet mining is done in an online approach i.e. during the network training.

In Fig. 11(a-b) examples of semi-hard and hard triplet pairs selected during network training are displayed. Similar to the AE and the VAE, a k-NN classifier is used on the feature embedding to measure the classification accuracy. The network shows an improvement in feature learning by an average test accuracy of 71.2%. The t-SNE plot for TNN over feature embedding is summarized in Section VII similar to the ones illustrated in IV-C. The similar approach is followed for remaining experiments.

### B. TRIPLET-BASED VARIATIONAL AUTO-ENCODER (TVAE)

While deep metric learning is optimal for encoding the data representation and for measuring data similarity, it cannot enable probabilistic inference for the model. On the other hand, the VAE performs an approximate Bayesian inference efficiently by having continuous feature information, as discussed in subsection IV-C. In [79], the authors combined both approaches and proposed a hybrid network architecture, called TVAE, where the network is optimized by minimizing the upper-bound on the expected negative log-likelihood of the data together with the triplet loss ($L_{\text{triplet}}$). Eq. 8 gives a mathematical overview on the total loss which is a linear combination of the CE loss, KL-divergence and the triplet loss.

$$\mathcal{L}_{\text{TVAE}} = 0.7 * \mathcal{L}_{\text{reconstruction}} + 0.3 * (\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{triplet}}), \quad (8)$$

At the end of the network training, the classification accuracy is evaluated on the mean embedding vector using a k-NN classifier. The average classification accuracy of 73.85% shows the dominance of the training framework over the VAE and the triplet standalone. Simultaneously, this leads to a Bayesian inference by enabling mean and standard variance over feature embedding.

### C. QUADRUPLET VARIATIONAL AUTO-ENCODER (QVAE)

Besides the increased classification accuracy in TVAE, triplet loss, however, suffers from two major drawbacks. First, the distance metric function for an anchor is optimized with respect to the positive and negative samples. As a result, there is no discriminator part in the triplet loss function which can help to push target samples from an intra-class. This problem is avoided by including another negative sample, belonging to the same group as the first negative sample. This helps the network to have a better inter and intra-class distance by adding an extra parameter optimization to separate the negative class from each other. The resulting new loss function is termed as Quadruplet loss ($\mathcal{L}_{\text{quadruplet}}$) [80] and can be summarized by Eq. 9. It includes another hyper-parameter $\alpha_2$ which is kept to 0.5 during the training. While sample $s_i$ and $s_j$ belong to the same class and represent an anchor and positive sample, $s_k$ and $s_l$ belong to two different classes, which are also not an anchor class.

$$\begin{aligned}\mathcal{L}_{\text{quadruplet}} = &\sum_{i,j,k}^{N}\left[q\left(x_i, x_j\right)^2 - q\left(x_i, x_k\right)^2 + \alpha_1\right] \\ &+\dot{\sum}_{i,j,k,l}^{N}\left[q\left(x_i, x_j\right)^2 - q\left(x_l, x_k\right)^2 + \alpha_2\right], \\ &s_i = s_j,\ s_l \neq s_k, s_i \neq s_l, s_i \neq s_k \end{aligned} \quad (9)$$

Similar to the triplet pairs, quadruplet pairs consist of semi-hard and hard examples, which are sampled during the training of the network. Additionally, the choice of both negative samples plays an important role for network learning. As a result, SSIM score between samples were calculated for each epoch. The samples from different class than anchor class with the highest SSIM score are considered for negative samples. Fig. 12 shows a sample example of quadruplet pair with two negatives.

The distance metric is computed with a $L2$ norm (euclidean) which compares the feature embedding vectors element wise with an uniform weighting to each values. Considering the nature of the training data, i.e micro-Doppler based signatures from VRUs, small changes in Doppler frequency from the intra-class lead to an unique identification of target. At the same time, uniform weighting fails to find outliers within a small difference of the feature embedding which is usual the case for intra-class VRUs. Thus, the distance function is learned during the network training. For this purpose, a 3 layered multi-layer perceptron (MLP) based architecture is designed and optimized using the principle of siamese

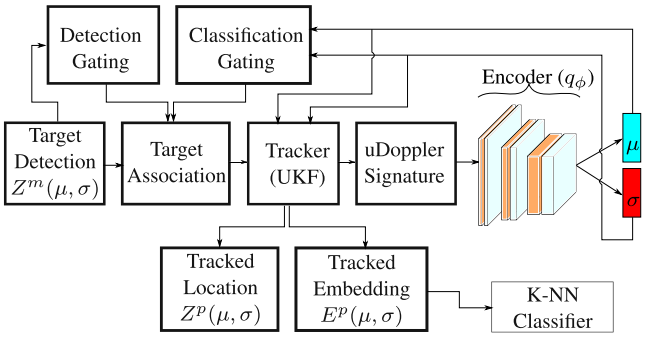


**FIGURE 13. A detailed illustration of the inference phase of the proposed framework having an integrated classification and tracking algorithm combined with the Bayesian knowledge over learned mean and variance over feature embedding.**

networks. The $q(.)$ function in Eq. 9 is the latent output from encoder. Further, the network's classification accuracy and confusion matrix are evaluated in the same way as before, using a k-NN classifier over the mean embedding. The network achieves an average accuracy of 85.04%.

It is also evident to note that the nature of clusters follows the nature of classes and sub-classes. Where corresponding data from the cyclist model follows the strong correlation between each other, it makes it hard to push their cluster apart from each other (3, 4, 5). Additionally, the Bayesian approximation at the latent space makes learned embedding continuous. This regularizes the metric loss function, relaxes the distance learning and restricts the granularity between the class clusters.

While $L2-norm$ based metric is used to calculate the distance between the class-embedding for TNN and TVAE, a pre-trained metric neural network is used for QVAE. Section VI describes the details about the proposed integrated framework for a continuous classification and tracking of the target. To show the advantage and robustness of integrating feature embedding as an appearance model, different learned embeddings (TNN, TVAE, QVAE) are used. The results are described in Section VII in detail. It is important to note that the design and choice of the feature extraction could further be improved using different architectures or optimization functions. Additionally, instead of approximating Bayesian inference from a point-estimate NN, a probabilistic NN can be used. However, this paper demonstrates advantage of the proposed integrated framework in terms of robustness of target tracking and improved accuracy of target classification over different feature extractor.

## VI. PROPOSED INTEGRATED FRAMEWORK

Fig. 13 illustrates an overview of the proposed framework for a continuous localization and classification of the detected target. This is done by augmenting the state vector of the tracker by a target's feature embedding as appearance model in combination with the target's localization as motion model, described in Section VI-A. In consequence, false alarms are suppressed by using both detection and classification gating. In addition, the framework also enables a complete Bayesian inference by using the mean and variance over detection, as described in Section III and corresponding target's features, as mentioned in Section V. As a result, the robustness of the framework is further improved by leveraging the Bayesian information associated with the input and predicted state vector and by performing data association, as discussed in subsection VI-B. The framework includes multiple processing blocks of which target detection block provides measurement data on the target's localization ($Z^m(\mu,\sigma)$) to the tracker. The encoder block ($q(\phi)$) extracts appearance embedding ($E^m(\mu,\sigma)$) and augments the tracker state vector with it for each frame. The tracker (an UKF in our case), uses these information to estimate the new position of the target and classifies the target into the defined category using a k-NN classifier. The integration of the appearance model together with the gating and data association are described below.

### A. STATE VECTOR AND FILTERING

Unlike a point estimate based encoder, the variational encoder maps the Doppler spectrogram as input to a distribution over a plausible latent embedding (i.e a feature embedding). Thus, it returns both mean (confidence $E^m_\mu$) and variance (uncertainty $E^m_\sigma$) over the feature embedding. The variance over the embedding vector is used for updating the state uncertainty corresponding to the appearance in the Kalman filter, as explained in algorithm 1. The UKF assumes a Gaussian random variable for the distribution of the state vector. Thus, the integration of the classifier output into the tracker facilitates the processing in obtaining not only the value of the current state of the classification but also the uncertainty associated with the state. Considering $\mu_i$ as the mean embedding of class $i$ with $M$ as the total dimension of the assumed embedding vector, the modified augmented state vector ($x_a$) of the tracker can be represented as Eq. 10.

$$\begin{aligned} x_a &= \begin{bmatrix} P_x & P_y & v & Az & \mu_{11} & \mu_{12} & \cdots & \mu_{1M} \end{bmatrix}^T, \\ g(x_a) &= \begin{bmatrix} p_x^P & p_y^P & v^P & Az^P & {\mu_{11}}^P & \cdots & {\mu_{1M}}^P \end{bmatrix}^T, \end{aligned} \tag{10}$$

Target's localization parameter is represented by lateral ($P_x$), longitudinal position ($P_y$), velocity ($v$) and azimuth angle ($Az$). Even though $Az$ can be estimated from $P_x$ and $Py$, $Az$ is choosen to be part of state vector [25]. This is due to fact that the tracker estimate is not a point estimate but a Gaussian distribution with mean and variance. As a result, to estimate $Az$, the formulation need to use unscented transformation (using sigma points) or Taylor expansion (as in EKF) from the distribution of both $P_x$ and $P_y$. The process model accounts for the state transition or the prediction into the next time step. The process model transformation for $x$ is dictated by the CV model and the augmented parameters are obtained by applying the non-linear process model transformation $g(\cdot)$.

### B. TARGET ASSOCIATION

The problem of data association plays a critical role in the suppression of false alarms by associating the uncertain

measurements to certain tracks. As a result, a gating operation is defined before updating the prediction for the current measurement. This involves a track creation, maintenance and deletion for single or multiple targets. The accuracy of data association relies on the choice of the distance metric which can be grouped into Bayesian or non-Bayesian based on the nature of the data. Both, measurement $(y^{(i)}_{k|k-1})$ and sigma-point transformed prediction $(x^{(i)}_{k|k-1})$ follow a Gaussian distribution, having a mean $(\hat{y}, \hat{x})$ and a covariance $(P^{y}_{k|k-1}, P^{x}_{k|k-1})$, respectively in algorithm 1. Therefore, the variance over posterior and observation is used for the data association.

Additionally, due to the nature of the state vector (distribution than point), a Mahalanobis distance as the association metric is used for the computing distance. This acts as a multivariate Euclidean norm which is described in Eq. 11. It shows that the Mahalanobis distance is a function of both the mean and covariance of the predicted state vector.

$$d = \sqrt{(\hat{x}^{(i)}_{k|k-1} - y^{(i)}_{k|k-1})^{\mathrm{T}} {P^{y_i}_{k|k-1}}^{-1} (\hat{x}^{(i)}_{k|k-1} - y^{(i)}_{k|k-1})}, \quad (11)$$

Here, $\hat{x}^{(i)}_{k|k-1}$ is the current measurement and $y_{k|k-1}$, $P^{y_i}_{k|k-1}$ are the mean and process covariance model of the predicted state vector at a particular time step. The distance $d$ is chi-square distributed with $n_z$ degrees of freedom, where $n_z$ is the dimension of the state vector which is 4 for localization and 16 for feature embedding. The measurement is associated with a particular track state only if the Mahalanobis distance is lesser than a chosen threshold. The new augmented state brings two different modalities (motion and appearance) into consideration. Thus, different thresholds for each modality are modelled which in return improves the gating operation. Overall, a threshold of 0.75 for the localization and 2.5 for the appearance model is considered. This helped to remove noisy outliers for target's localization and feature embedding (used for classification) from being associated to the states of the tracker.

## VII. RESULTS AND DISCUSSION

Considering the accuracy and the ability to learn distinct clusters for each class over the embedding space, the QVAE based training framework gives a clear indication for the choice of the feature extraction approach. Whereas, to get a better understanding on the advantages of the proposed framework over a conventional multi-target tracking (MTT) framework, all the feature extractors (TNN, TVAE, QVAE) are evaluated in an integrated framework. The extractors are thereby optimized using the metric learning. Additionally, a pre-trained feature extractor is evaluated on the micro-Doppler signatures for a particular target over a distinctly different non-linear trajectory, as described in Fig. 4(b).

Due to high variability in the nature of the training environment (architecture, data-sets, hyper-parameters), a direct comparison of the proposed framework with different methods described in the literature is not feasible. As a result, a direct bench-marking of the framework is evaluated for

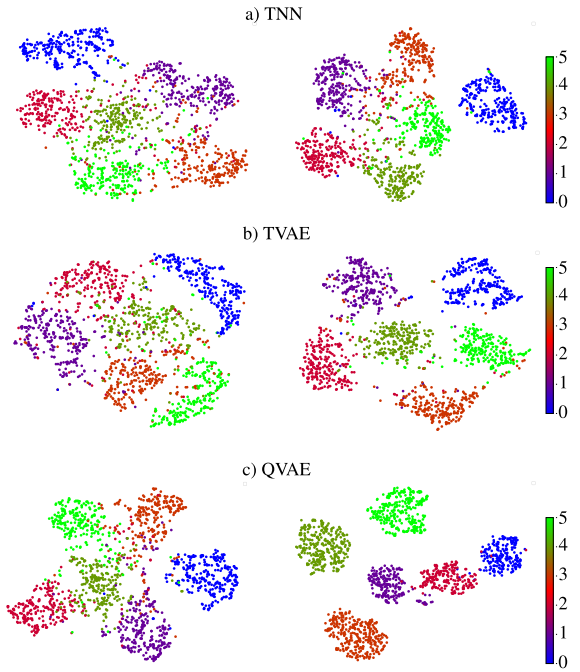


**FIGURE 14.** t-SNE plot over the feature embedding using (a) TNN, (b) TVAE and (c) QVAE based feature extractor. The left column clusters refer to latent appearance embedding from feature extractor and the right column shows an improvement in the appearance embedding for corresponding extractor when integrated inside the tracker.

inference phase using a NVIDIA Quadro P2000 GPU-based system. While the encoder part of the network requires 4.7 ms to extract the Gaussian latent feature embedding, the integrated tracker takes 3.5 ms for the estimation of new state vector. The evaluation of the integrated framework is done for three (homogeneous and heterogeneous) target classes, containing either all pedestrian or cyclist or a mixed class. During the simulation, the initial position of the targets is adjusted in such a way that each target faces partial or complete occlusion from either one of the targets, resulting in miss-detections. This helped to analyze the robustness of the integrated appearance model over the stand-alone motion model. A constant velocity model for pedestrian and cyclist classes are used through the experiments. The performance of the tracker is evaluated in three folds: the classification accuracy, localization precision and target association. Each of them is evaluated and discussed in the following paragraphs.

### A. CLASSIFICATION ACCURACY

As mentioned in Section V, the target classification accuracy is evaluated using a k-NN classifier. In addition to the feature embedding over measurement Doppler spectra, target classification is also evaluated over the tracker's estimated feature (augmented) state vector.

As the first 20 frames are used for a feature initialization buffer, the classification accuracy is evaluated from the 20th frame using the principle of a sliding window over each radar frame. For targets having a miss-detection, the Doppler embedding is taken over from the last predicted value of

**TABLE 5. Detailed quantitative analysis on the quality of the clustering and the feature embedding. It is estimated from the feature extractor and the corresponding integrated tracker framework.**

| Clustering Metric | Feature | Triplet (TNN) | TVAE | QVAE |
|---|---|---|---|---|
| Silhouette score | Embedding | 0.25 | 0.23 | 0.37 |
| | **Proposed Tracker** | **0.38** | **0.39** | **0.71** |
| Davies-Bouldin score | Embedding | 1.56 | 1.47 | 1.07 |
| | **Proposed Tracker** | **0.99** | **0.94** | **0.38** |

**TABLE 6. Detailed quantitative analysis on the accuracy of target classification using feature embedding estimated from the feature extractor and the corresponding integrated tracker framework.**

| Clustering Metric | Feature | Triplet (TNN) | TVAE | QVAE |
|---|---|---|---|---|
| Classification Accuracy | Embedding | 71.2% | 73.85% | 85.04% |
| | **Proposed Tracker** | **79.2%** | **89.01%** | **99.22%** |

the tracker. A similar approach is also applied during the localization estimation, as discussed in paragraph VII-B.

Together with the classification accuracy, both a visual and quantitative analysis is done over the measurement and estimated feature embedding. Fig. 14 gives a visual illustration of the separation between the target classes using a t-SNE over the feature embedding. While the left column of the plot shows the 2D feature clusters from a pre-trained feature extractor over a new trajectory, the right column shows estimated features from the tracker, leading to an improvement in the clustering. In consequence, also the target classification is improved. The target classes within the pedestrian group (female, male, neural) and cyclist (cyclist1, cyclist2, cyclist3) are indicated by a color coding, using blue, purple, light red and dark-red, dark-green and light-green, respectively. These target classes are numbered as 0, 1, 2, 3, 4, 5,, see the legend of Fig. 14. As mentioned before, the low variations in between the clusters, i.e between pedestrian (0,1,2) or among cyclist (3,4,5), show correlation between their appearance model. Additionally, the cyclist clusters (3,4,5) show stronger correlation in contrast to the pedestrian ones. This is due to limitations on the dynamics of its physical model and the contained reflection points out of the Matlab. Illustrating the relative improvement on distinct clustering of feature embedding by tracker in comparison to feature extractor, Fig. 14 helps to understand the generalization of the integrated tracker performance for a given feature embedding.

Additionally, a quantitative analysis over the feature clusters of Fig. 14 is evaluated. In this paper, silhouette and Davies-Bouldin coefficients are used to measure the clustering scores. The silhouette coefficient gives a similarity measure between a sample and its own cluster (cohesion) in comparison to other clusters (separation). The silhouette coefficient lies in the range of $-1$ to 1. Higher values indicate a better match of the sample to its own cluster. On the other hand, the Davies-Bouldin coefficient indicates the distance between clusters by estimating the distance of a sample between with-in and the neighboring clusters. A typical value for the Davies-Bouldin score lies in the range of 0 to 1, where lower values indicate a better clustering.

Similar to the visual understanding, Table 5 shows a quantitative improvement in target classification. Both, the Davies-Bouldin and silhouette scores presented in Table 5 are an average value over all target class for all the test samples. The silhouette scores for estimated embedding (from tracker) for Triplet is improved by $\sim$ 50% i.e. from 0.25 to 0.38. Similarly, TVAE based tracker shows an improvement by $\sim$ 70% which is from 0.23 to 0.39. Further, following similar behavior, QVAE also shows an improvement of $\sim$ 90% in silhouette score and leading to 0.71 from 0.37. In addition to it, the Davies-Bouldin coefficients for Triplet and TVAE are reduced by $\sim$ 35%. Similarly, the Davies-Bouldin coefficients for QVAE based estimated feature embedding is reduced from 1.07 to 0.38 i.e. by $\sim$ 65%. This indicates a better separability in the target features, resulting in improved classification accuracy, as shown in Table 6.

As seen from Fig. 14 and Table 5, the detailed evaluation on the confusion matrix together with the localization and the association error is done using QVAE as a feature extractor in an integrated framework. Fig. 15(a) gives a deeper insight on inter- and intra-class accuracy together with the false-alarm over the feature calculated from the QVAE. In contrast, Fig. 15(b) shows the improvement in accuracy by avoiding miss-classification with-in the target classes.

### B. LOCALIZATION ACCURACY

The tracker's state vector brings both target features for classification and target's motion for localization. In order to access and compare the localization accuracy of the proposed framework for a multi-target tracking (MTT) scenario, the similarity between the estimated and the ground truth (from simulation environment) is evaluated by using an Euclidean norm. Considering the pedestrian sub-group as a target class, Fig. 16(a) displays the localization error between the ground truth and the estimated target's position. While the localization error for male-pedestrian class gets higher between the frames 25 – 35 and 110 – 120, whereas female and neutral-pedestrian class shows higher localization error between 130 – 140 frames. This is due to the fact that measurement data for estimated target was missing due to occlusion during cross-over.

In contrast to this, the proposed framework with an integrated augmented feature embedding does not fluctuate much and therefore, helps the tracker to better associate with new measurements. This can be seen in Fig. 16(b), where the state vector (combined localization and embedding) error of the frames between 25 – 35, 110 – 120, 130 – 140 remained within a range of 0.1 – 0.2.

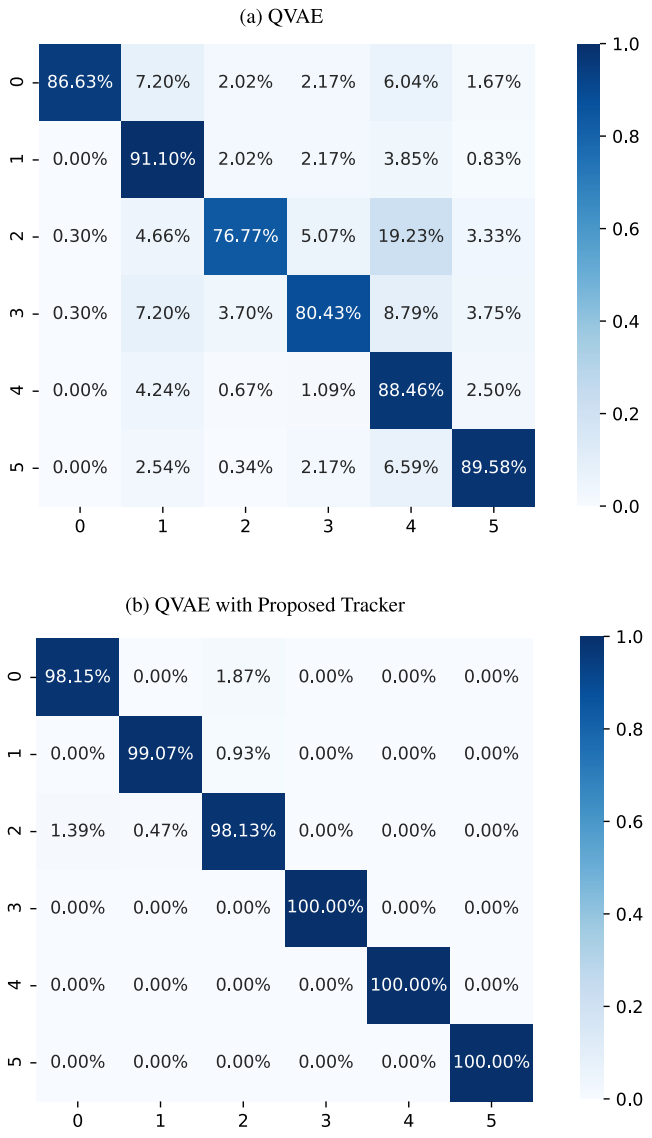


**FIGURE 15.** Confusion matrix to illustrate the improvement in the miss classification from (a) QVAE based on feature extractor in comparison to (b) proposed;QVAE integrated framework.

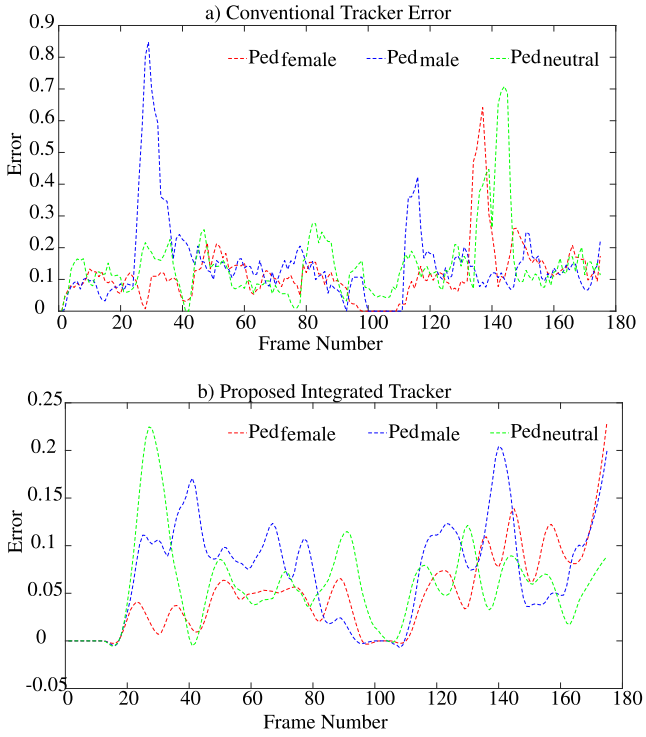


**FIGURE 16.** The error between the estimated and the ground truth state vector for (a) the conventional tracker and (b) our proposed integrated tracker.

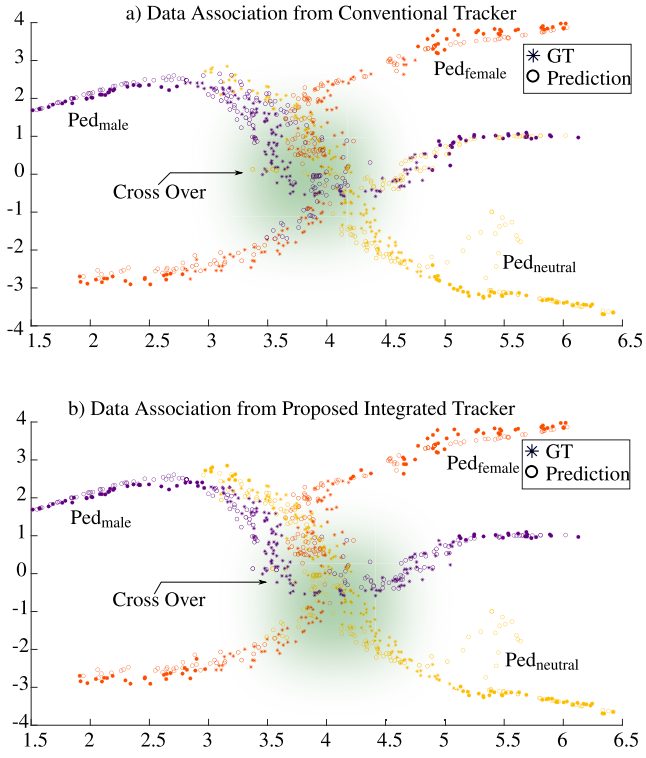


**FIGURE 17.** A comparative visual understanding on the data association from the (a) conventional tracker with the motion model and the (b) proposed integrated framework having the motion and appearance modality.

### C. ASSOCIATION

The error for tracker depends, intuitively, on the target estimation besides the cost of missed or false target associations. The advantage of an integrated appearance model is to avoid such situations, especially in case of a cross-over where a target is occluded or miss-detected. Fig. 17 illustrates a comparative analysis of the target association with and without the augmented state vector. The ground truth of the target trajectory is plotted using a circle (∘) and the estimated trajectory is shown using a star (∗). The targets female-male, male-pedestrian, and neutral-pedestrian with their initial class-ID 0, 1, 2 are highlighted in orange, purple and yellow color, respectively.

Although the tracker's motion state vector is modeled in polar coordinates, the trajectory is visualized on 2D Cartesian coordinates. The x-axis represents the lateral position $P_x$ of the target and the y-axis represents the longitudinal position $P_y$ of the target. Fig. 17(a) shows the estimated trajectory and

target association for a conventional tracker having only the motion model. As a result, during the cross-over, multiple false associations occurred between male-pedestrian (purple) - female-pedestrian (orange) and neutral-pedestrian (yellow) - male-pedestrian (purple). In contrast to this, Fig. 17(b), shows the estimated trajectory and its associated target using our proposed integrated framework which helps to suppress the false associations during the cross-over situation. Unlike in conventional tracker where only localization state vector is considered for association, the proposed framework uses 3-stage target association formulation. It considers localization, embedding and combined augmented state vector for calculation of Mahalanobis distance matrix. Later, all three matrix are combined using OR operation. In case of multiple association, distance based ranking is used to reduce false assignment. In consequence as illustrated before in fig. 17(b), the proposed integrated framework can be seen as an efficient and robust framework for a continuous target localization and classification with an improved classification accuracy compared to the conventional approaches.

### D. SUMMARY

The article proposes a novel framework to integrate both motion and appearance modalities of the target into the tracker. This is done by modifying state vector with feature embedding together with localization (x-range, y-range, velocity and angle) parameters. The target of interest considered throughout the experiment are pedestrian and cyclist as they face lot of challenges in reliable detection (due to smaller RCS) and classification (due to high correlation between their signatures). As a result, in this article author used the concept of distance metric learning applied over a latent feature vector. This helped the network distinguish and learn distinct features for each class. Moreover, the concept of variational inference is applied together with metric learning, making feature extraction fully Bayesian. The Bayesian inference from the feature extractor helped to integrate detection, classification and tracking into once framework. The continuous estimation of the features from the tracker helped to improve the classification accuracy by temporal smoothing over embedding. Additionally, during cross-over situation having partial or complete occlusion of target, framework helps to suppress false association between detection and estimation.

The entire work is done in a simulation environment to demonstrate the applicability of novel proposed framework. The framework uses micro-Doppler signatures as raw input data for the feature learning. As the estimation of Doppler spectra suffers from a time-frequency resolution trade-off, this approach gets challenging for scenarios with targets having very high varying Doppler frequency components. Those require either an adaptive sampling frequency or a wavelet transform. On the other hand, the Doppler spectra directly depend on the detection of micro-motions from the target, which then inherently depends on the pose and the view angle of the target w.r.t the radar. As a result, the diversity within Doppler spectra for a particular target class becomes very large and gets very challenging in reality. However, learning the temporal information over Doppler spectra could help to learn and model more optimized appearance model.

## REFERENCES


[1] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems: Review and future perspectives," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 4, pp. 6–22, Oct. 2014.

[2] M. Murad, I. Bilik, M. Friesen, J. Nickolaou, J. Salinger, K. Geary, and J. S. Colburn, "Requirements for next generation automotive radars," in *Proc. IEEE Radar Conf.*, Apr. 2013, pp. 1–6.

[3] J. Dickmann, N. Appenrodt, H.-L. Bloecher, C. Brenk, T. Hackbarth, M. Hahn, J. Klappstein, M. Muntzinger, and A. Sailer, "Radar contribution to highly automated driving," in *Proc. 11th Eur. Radar Conf.*, Oct. 2014, pp. 412–415.

[4] C. Waldschmidt and H. Meinel, "Future trends and directions in radar concerning the application for autonomous driving," in *Proc. 11th Eur. Radar Conf.*, Oct. 2014, pp. 416–419.

[5] B. Major, D. Fontijne, A. Ansari, R. T. Sukhavasi, R. Gowaikar, M. Hamilton, S. Lee, S. Grzechnik, and S. Subramanian, "Vehicle detection with automotive radar using deep learning on Range-Azimuth-Doppler tensors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 924–932.

[6] F. Engels, P. Heidenreich, A. M. Zoubir, F. K. Jondral, and M. Wintermantel, "Advances in automotive radar: A framework on computationally efficient high-resolution frequency estimation," *IEEE Signal Process. Mag.*, vol. 34, no. 2, pp. 36–46, Mar. 2017.

[7] F. Gustafsson, "Automotive safety systems," *IEEE Signal Process. Mag.*, vol. 26, no. 4, pp. 32–47, Jul. 2009.

[8] G. Hakobyan and B. Yang, "High-performance automotive radar: A review of signal processing algorithms and modulation schemes," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 32–44, Sep. 2019.

[9] A. Santra and S. Hazra, *Deep Learning Application Short Range Radars*. Norwood, MA, USA: Artech House, 2020.

[10] H. H. Meinel, "Evolving automotive radar—From the very beginnings into the future," in *Proc. 8th Eur. Conf. Antennas Propag.*, 2014, pp. 3107–3114.

[11] F. Fölster and H. Rohling, "Signal processing structure for automotive radar," *Frequenz*, vol. 60, nos. 1–2, pp. 20–24, Jan. 2006.

[12] J. Hasch, E. Topak, R. Schnabel, T. Zwick, R. Weigel, and C. Waldschmidt, "Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 3, pp. 845–860, Mar. 2012.

[13] S. M. Patole, M. Torlak, D. Wang, and M. Ali, "Automotive radars: A review of signal processing techniques," *IEEE Signal Process. Mag.*, vol. 34, no. 2, pp. 22–35, Mar. 2017.

[14] M. Heuer, A. Al-Hamadi, A. Rain, and M.-M. Meinecke, "Detection and tracking approach using an automotive radar to increase active pedestrian safety," in *Proc. IEEE Intell. Vehicles Symp. Proc.*, Jun. 2014, pp. 890–893.

[15] A. M. Richards, *Principles of Modern Radar Basic Principles*. London, U.K.: Institution of Engineering and Technology, 2010.

[16] L. Hammarstrand, L. Svensson, F. Sandblom, and J. Sorstedt, "Extended object tracking using a radar resolution model," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 3, pp. 2371–2386, Jul. 2012.

[17] V. C. Chen, *The Micro-Doppler Effect in Radar*. Norwood, MA, USA: Artech House, 2019.

[18] Z. Duan, X. R. Li, C. Han, and H. Zhu, "Sequential unscented Kalman filter for radar target tracking with range rate measurements," in *Proc. 7th Int. Conf. Inf. Fusion*, 2005, pp. 1–8.

[19] A. Lin and H. Ling, "Three-dimensional tracking of humans using very low-complexity radar," *Electron. Lett.*, vol. 42, no. 18, pp. 1062–1063, Aug. 2006.

[20] S. Chang, M. Wolf, and J. W. Burdick, "An MHT algorithm for UWB radar-based multiple human target tracking," in *Proc. IEEE Int. Conf. Ultra-Wideband*, Sep. 2009, pp. 459–463.

[21] Y. Kim and H. Ling, "Through-wall human tracking with multiple Doppler sensors using an artificial neural network," *IEEE Trans. Antennas Propag.*, vol. 57, no. 7, pp. 2116–2122, Jul. 2009.

[22] E. Cortina, D. Otero, and C. E. D'Attellis, "Maneuvering target tracking using extended Kalman filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 27, no. 1, pp. 155–158, 1991.
[23] S. Chang, R. Sharan, M. Wolf, N. Mitsumoto, and J. W. Burdick, "UWB radar-based human target tracking," in *Proc. IEEE Radar Conf.*, Oct. 2009, pp. 1–6.
[24] S. Chang, M. Wolf, and J. W. Burdick, "Human detection and tracking via ultra-wideband (UWB) radar," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 452–457.
[25] W. Sun, W. Huang, Y. Ji, Y. Dai, P. Ren, P. Zhou, and X. Hao, "A vessel azimuth and course joint re-estimation method for compact HFSWR," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1041–1051, Feb. 2020.
[26] C. Will, P. Vaishnav, A. Chakraborty, and A. Santra, "Human target detection, tracking, and classification using 24-GHz FMCW radar," *IEEE Sensors J.*, vol. 19, no. 17, pp. 7283–7299, Sep. 2019.
[27] P. Vaishnav and A. Santra, "Continuous human activity classification with unscented Kalman filter tracking using FMCW radar," *IEEE Sensors Lett.*, vol. 4, no. 5, pp. 1–4, May 2020.
[28] Z. Li, W. Yang, S. Peng, and F. Liu, "A survey of convolutional neural networks: Analysis, applications, and prospects," *CoRR*, vol. abs/2004.02806, 2020. [Online]. Available: https://arxiv.org/abs/2004.02806
[29] F. Wang and J. Sun, "Survey on distance metric learning and dimensionality reduction in data mining," *Data Mining Knowl. Discovery*, vol. 29, no. 2, pp. 534–564, Mar. 2015.
[30] J. Lu, J. Hu, and J. Zhou, "Deep metric learning for visual understanding: An overview of recent advances," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 76–84, Nov. 2017.
[31] E. Xing, M. Jordan, S. J. Russell, and A. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems*, vol. 15, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA, USA: MIT Press, 2003, pp. 521–528.
[32] S. Sun and Q. Chen, "Hierarchical distance metric learning for large margin nearest neighbor classification," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 7, pp. 1073–1087, Nov. 2011.
[33] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," 2015, *arXiv:1506.07310*. [Online]. Available: http://arxiv.org/abs/1506.07310
[34] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015. [Online]. Available: http://arxiv.org/abs/1503.03832
[35] J. Hu, J. Lu, and Y.-P. Tan, "Deep metric learning for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2056–2068, Nov. 2016.
[36] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: http://arxiv.org/abs/1610.02984
[37] M. Chen, Y. Ge, X. Feng, C. Xu, and D. Yang, "Person re-identification by pose invariant deep metric learning with improved triplet loss," *IEEE Access*, vol. 6, pp. 68089–68095, 2018.
[38] X. Yang, P. Zhou, and M. Wang, "Person reidentification via structural deep metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 2987–2998, Oct. 2019.
[39] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737
[40] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2612–2620.
[41] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016, pp. 1857–1865.
[42] W. Ge, W. Huang, D. Dong, and M. Scott, "Deep metric learning with hierarchical triplet loss," *CoRR*, vol. abs/1810.06951, 2018. [Online]. Available: http://arxiv.org/abs/1810.06951
[43] J. Weis and A. Santra, "One-shot learning for robust material classification using millimeter-wave radar system," *IEEE Sensors Lett.*, vol. 2, no. 4, pp. 1–4, Dec. 2018.
[44] S. Hazra and A. Santra, "Short-range radar-based gesture recognition system using 3D CNN with triplet loss," *IEEE Access*, vol. 7, pp. 125623–125633, 2019.
[45] T. Stadelmayer, M. Stadelmayer, A. Santra, R. Weigel, and F. Lurz, "Human activity classification using mm-wave fmcw radar by improved representation learning," in *Proc. 4th ACM Workshop Millimeter-Wave Netw. Sens. Syst.*, New York, NY, USA, 2020, pp. 1–6.
[46] D. Yuan, X. Chang, P.-Y. Huang, Q. Liu, and Z. He, "Self-supervised deep correlation tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 976–985, 2021.
[47] D. Yuan, W. Kang, and Z. He, "Robust visual tracking with correlation filters and metric learning," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105697.
[48] X. Shu, D. Yuan, Q. Liu, and J. Liu, "Adaptive weight part-based convolutional network for person re-identification," *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 23617–23632, Jun. 2020.
[49] S. Vishwakarma, A. Rafiq, and S. S. Ram, "Micro-Doppler signatures of dynamic humans from around the corner radar," in *Proc. IEEE Int. Radar Conf. (RADAR)*, Apr. 2020, pp. 169–174.
[50] B. Colo, A. Fouda, and A. S. Ibrahim, "Ray tracing simulations in millimeter-wave vehicular communications," in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 1–4.
[51] R. Hoppe, G. Wolfle, P. Futter, and J. Soler, "Wave propagation models for 5G radio coverage and channel analysis," in *Proc. 6th Asia–Pacific Conf. Antennas Propag. (APCAP)*, Oct. 2017, pp. 1–3.
[52] Z. Zhang, J. Ryu, S. Subramanian, and A. Sampath, "Coverage and channel characteristics of millimeter wave band using ray tracing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 1380–1385.
[53] M. Lubke, H. Hamoud, J. Fuchs, A. Dubey, R. Weigel, and F. Lurz, "Channel characterization at 77 GHz for vehicular communication," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Dec. 2020, pp. 1–4.
[54] M. K. Samimi and T. S. Rappaport, "3-D millimeter-wave statistical channel model for 5G wireless system design," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 7, pp. 2207–2225, Jul. 2016.
[55] M. Drago, T. Zugno, M. Polese, M. Giordani, and M. Zorzi, "MilliCar: An ns-3 Module for mmWave NR V2X Networks," in *Proc. Workshop*, Jun. 2020, pp. 9–16.
[56] *5G;Study on channel model for frequencies from 0.5 to 100 GHz*, document 38.901 V 16.1.0, 3rd Generation Partnership Project (3GPP), Dec. 2019.
[57] R. Prophet, A. Deligiannis, J.-C. Fuentes-Michel, I. Weber, and M. Vossiek, "Semantic segmentation on 3D occupancy grids for automotive radar," *IEEE Access*, vol. 8, pp. 197917–197930, 2020.
[58] S. Sun, A. P. Petropulu, and H. V. Poor, "MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 98–117, Jul. 2020.
[59] I. Bilik, O. Longman, S. Villeval, and J. Tabrikian, "The rise of radar for autonomous vehicles: Signal processing solutions and future research directions," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 20–31, Sep. 2019.
[60] J. A. Scheer, "The radar range equation," in *Principles of Modern Radar: Basic Principles*. London, U.K.: Institution of Engineering and Technology, 2010, pp. 59–86.
[61] N. F. Troje, "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns," *J. Vis.*, vol. 2, no. 5, p. 2, Sep. 2002.
[62] N. Yamada, Y. Tanaka, and K. Nishikawa, "Radar cross section for pedestrian in 76 GHz band," in *Proc. Eur. Microw. Conf.*, 2005, p. 4.
[63] D. Belgiovane and C.-C. Chen, "Bicycles and human riders backscattering at 77 GHz for automotive radar," in *Proc. 10th Eur. Conf. Antennas Propag. (EuCAP)*, Apr. 2016, pp. 1–5.
[64] R. Du, Y. Fan, and J. Wang, "Pedestrian and bicyclist identification through micro Doppler signature with different approaching aspect angles," *IEEE Sensors J.*, vol. 18, no. 9, pp. 3827–3835, May 2018.
[65] A. M. Richards, *Fundamentals Radar Signal Processing*. New York, NY, USA: McGraw-Hill, 2014. [Online]. Available: https://www.ebook.de/de/product/19460046/mark_a_richards_fundamentals_of_radar_signal_processing_second_edition.html
[66] A. Dubey, J. Fuchs, M. Lubke, R. Weigel, and F. Lurz, "Generative adversial network based extended target detection for automotive MIMO radar," in *Proc. IEEE Int. Radar Conf. (RADAR)*, Apr. 2020, pp. 220–225.
[67] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: Phenomenon, model, and simulation study," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 42, no. 1, pp. 2–21, Jan. 2006.
[68] T. Stadelmayer and A. Santra. (2020). *Parametric Convolutional Neural Network for Radar-Based Human Activity Classification Using Raw ADC Data*. [Online]. Available: https://doi.org/10.36227/techrxiv.12896108.v1
[69] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *CoRR*, abs/1703.04977, pp. 1–7, Oct. 2017.
[70] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[71] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[72] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *Schedae Informaticae*, vol. 1, pp. 1–9, Oct. 2017.
[73] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[74] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.
[75] Y. Wang and Q. Yao, "Few-shot learning: A survey," *CoRR*, abs/1904.05046, p. 5, Oct. 2019.
[76] L. He, Z. Wang, Y. Li, and S. Wang, "Softmax dissection: Towards understanding intra- and inter-class objective for embedding learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 10957–10964.
[77] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9911, Oct. 2016, pp. 499–515.
[78] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "Learning a distance function with a Siamese network to localize anomalies in videos," *CoRR*, vol. abs/2001.09189, 2020. [Online]. Available: https://arxiv.org/abs/2001.09189
[79] H. Ishfaq, A. Hoogi, and D. Rubin, "TVAE: Triplet-based variational autoencoder using metric learning," 2018, *arXiv:1802.04403*. [Online]. Available: https://arxiv.org/abs/1802.04403
[80] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," *CoRR*, vol. abs/1704.01719, 2017. [Online]. Available: http://arxiv.org/abs/1704.01719

**ANAND DUBEY** (Member, IEEE) received the B.Tech. degree from JIITU, Noida, India, in 2012, and the M.Sc. degree in electronics and communication and automotive software engineering from TU Chemnitz, Chemnitz, Germany, in 2018. In September 2018, he joined the Institute for Electronics Engineering as a Research Assistant. His main research interests include digital radar signal processing algorithms and applied Bayesian machine learning for radar signal processing. Prior to his research, he has worked as a Software Developer in an automotive industry for more than 3.5 years. He is a member of the IEEE Microwave Theory and Techniques Society (IEEE MTT-S) and the IEEE Signal Processing Society (IEEE SPS).

**AVIK SANTRA** (Senior Member, IEEE) received the M.S. degree *(cum laude)* in signal processing from the Indian Institute of Science, Bengaluru, in 2010. He is currently working as Senior Staff Algorithm Engineer with Infineon, Neubiberg, developing signal processing and machine learning algorithms for industrial and consumer radars and depth sensors. Earlier in his career, he has worked as a System Engineer for LTE/4G modem with Broadcom Communications; and a Research Engineer at Airbus, developing cognitive radars. He is author of the book titled *Deep Learning Applications of Short-Range Radars* (ArTech House). He has filed more than 40 patents and published 30 research articles related to various topics of radar waveform design, radar signal processing, and radar machine/deep learning topics. He is a Reviewer of various IEEE and Elsevier journals. He was a recipient of several outstanding reviewer awards.

**JONAS FUCHS** (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Friedrich-Alexander University Erlangen-Nuremberg (FAU), Erlangen, Germany, in 2016 and 2018, respectively. In 2018, he joined the Institute for Electronics Engineering as a Research Assistant. His main research interests include radar signal processing algorithms, direction-of-arrival estimation, and applied machine learning for radar signal processing. He is a member of the IEEE Microwave Theory and Techniques Society (IEEE MTT-S) and the IEEE Signal Processing Society (IEEE SPS).

**MAXIMILIAN LÜBKE** (Member, IEEE) received the B.S. degree in medical engineering and the M.S. degree in electrical engineering from Friedrich-Alexander University Erlangen-Nuremberg (FAU), Erlangen, Germany, in 2017 and 2018, respectively. In 2019, he joined the Institute for Electronics Engineering, FAU, as a Research Assistant. His current research interests include joint radar and communication system and circuit design with respect to automotive applications. He was a recipient of the Best Paper Award of the ACM International Conference on Nanoscale Computing and Communication, in 2019, and the EAI International Conference on Bio-Inspired Information and Communications Technologies, in 2020.

**ROBERT WEIGEL** (Fellow, IEEE) was born in Ebermannstadt, Germany, in 1956. He received the Dr.Ing. and Dr.Ing.habil. degrees in electrical engineering and computer science from the Technical University of Munich, Munich, Germany, in 1989 and 1992, respectively. From 1994 to 1995, he was a Guest Professor of SAW technology with the Vienna University of Technology, Vienna, Austria. He was a Research Engineer, a Senior Research Engineer, and a Professor of RF circuits and systems with the Technical University of Munich, until 1996. From 1996 to 2002, he was the Director of the Institute for Communications and Information Engineering, University of Linz, Linz, Austria, where he co-founded the Company DICE, in 1999, which split into an Infineon Technologies (DICE) and an Intel (DMCE) company, which are devoted to the design of RFICs and MMICs. In 2000, he was appointed as a Professor of RF engineering with Tongji University, Shanghai, China. Since 2002, he has been the Head of the Institute for Electronics Engineering, University of Erlangen–Nuremberg, Erlangen, Germany, where he co-founded the companies eesy-id and eesy-ic, in 2009 and 2012, respectively. He has authored or coauthored more than 900 articles. He has been engaged in research and development of microwave theory and techniques, electronic circuits and systems, and communication and sensing systems. He served in many roles for the IEEE MTT-S and UFFC-S. He has been the Founding Chair of the Austrian COM/MTT Joint Chapter, a Region 8 MTT-S Coordinator, a Distinguished Microwave Lecturer, an IEEE MTT-S AdCom Member, and the 2014 MTT-S President.

**FABIAN LURZ** (Member, IEEE) received the B.Sc. and M.Sc. degrees in information and communication technology and the Dr.Ing. degree from Friedrich-Alexander University Erlangen-Nuremberg (FAU), Erlangen, Germany, in 2010, 2013, and 2019, respectively. In 2013, he joined the Institute for Electronics Engineering, FAU, as a Research Assistant. From 2017 to 2020, he was the Group Leader of the Circuits, Systems and Hardware Test Group. Since June 2020, he has been a Senior Engineer and the Research Group Leader of the Institute of High-frequency Technology, Hamburg University of Technology, Hamburg, Germany. His research interests include microwave circuits and systems, especially for low-cost and low-power metrology applications. He was a recipient of the First Prize of the High Sensitivity Radar Student Design Competition of the IEEE International Microwave Symposium, in 2014, 2017, and 2018, and the IEEE Microwave Theory and Techniques Society Graduate Fellowship Award, in 2016.