Machine Learning with Applications 8 (2022) 100275

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

HARadNet: Anchor-free target detection for radar point clouds using hierarchical attention and multi-task learning

Anand Dubey [a,*], Avik Santra [b], Jonas Fuchs [a], Maximilian Lübke [a], Robert Weigel [a], Fabian Lurz [c]

[a] Friedrich-Alexander-University Erlangen–Nuremberg, Cauerstrasse 9, Erlangen, 91058, Germany
[b] Infineon Technologies, Am Campeon 1-15, Neubiberg, 85579, Germany
[c] Hamburg University of Technology, Denickestrasse 22, Hamburg, 21073, Germany

ARTICLE INFO

Keywords:
Multi-task learning
Radar detection
Scene understanding

ABSTRACT

Target localization and classification from radar point clouds is a challenging task due to the inherently sparse nature of the data with highly non-uniform target distribution. This work presents HARadNet, a novel attention based anchor free target detection and classification network architecture in a multi-task learning framework for radar point clouds data. A direction field vector is used as motion modality to achieve attention inside the network. The attention operates at different hierarchy of the feature abstraction layer with each point sampled according to a conditional direction field vector, allowing the network to exploit and learn a joint feature representation and correlation to its neighborhood. This leads to a significant improvement in the performance of the classification. Additionally, a parameter-free target localization is proposed using Bayesian sampling conditioned on a pre-trained direction field vector. The extensive evaluation on a public radar dataset shows an substantial increase in localization and classification performance.

1. Introduction

A robust awareness of the environment is the indisputable key factor for realizing safe autonomous driving. In the last years, scene understanding i.e. target localization, classification, and tracking is being studied extensively, however mainly with camera or LiDAR image analysis (Behley et al., 2021; Huang et al., 2019; Izadinia et al., 2016). Besides camera and LiDAR sensors, radar as a widely-adopted sensor in traditional advanced driver assistance systems (ADAS), is very robust and reliable under different weather conditions. Radar allows instantaneous velocity estimation together with the spatial localization of measured objects (Bengler et al., 2014; Murad et al., 2013). With the recent advancement in radar systems (Meinel, 2014), especially with regard to processing high-resolution data, target detection and classification are typically done via multiple sub-task specific blocks (Hakobyan & Yang, 2019). While target detection is based on peak detection and clustering (Dubey et al., 2020a, 2020b), target classification is typically performed via extracting target-specific parameters, e.g. Doppler spectrograms (Dubey et al., 2021a, 2021b; Santra & Hazra, 2020; Stolz et al., 2017). In order to estimate a target's spectrogram, first, the target needs to be detected and separated in one of the three measurement dimensions, namely range, velocity, or angle. However, the classification essentially lacks to learn task dependency. Additionally, the success of combining all stages into one framework will heavily depend on the optimization of all sub-task specific hyperparameters.

An end-to-end solution can omit all the sub-tasks of the modular solution by mapping the high dimensional inputs from the sensors directly to the desired task. Therefore, in this work, we choose to represent the input data using point clouds, because they comprise the output with absolute values of the target's signal parameters as features, showing more information but with less complexity. However, the existing deep convolutional neural networks would require certain representation transformations (Maturana & Scherer, 2015; Qi, Su, Nießner et al., 2016; Su et al., 2015) in order to use point clouds. The PointNet architecture (Guo et al., 2019; Qi, Su, Mo et al., 2016; Qi, Yi et al., 2017) overcomes this constraint and supports point clouds as inputs.

In order to directly process point clouds, Danzer et al. (2019) extend the concept of PointNets for the detection of objects by combining a 2D object detector with a bounding box estimation. The performance of such an object detection method strongly relies on the region proposal network (RPN) which further depends on the anchor aspect ratio, the dense representation of the target/feature map and the RPN location inside the network architecture. In contrast to state-of-the-art (Danzer

* Corresponding author.
E-mail addresses: anand.dubey@fau.de (A. Dubey), avik.santra@infineon.com (A. Santra), jonas.fuchs@fau.de (J. Fuchs), maximilian.luebke@fau.de (M. Lübke), robert.weigel@fau.de (R. Weigel), fabian.lurz@tuhh.de (F. Lurz).

[1] https://github.com/ananddb90/HARadNet.git.

https://doi.org/10.1016/j.mlwa.2022.100275
Received 7 September 2021; Received in revised form 13 November 2021; Accepted 8 February 2022
Available online 14 February 2022

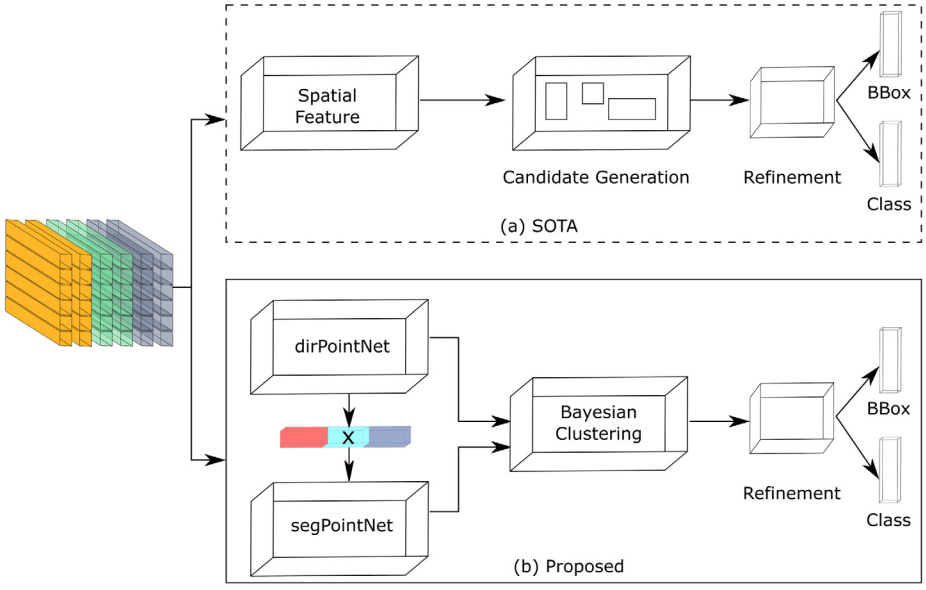

Fig. 1. Comparison between the state-of-the-art (SOTA) and our proposed framework for target detection using radar point-cloud representation. (a) The performance of the SOTA (Danzer et al., 2019) relies on prior anchor box at different feature abstraction. (b) Ours proposes multi-model based anchor-free target localization hierarchical attention at different feature abstraction.

et al., 2019), we propose, HARadNet, an anchor-free target detection and classification framework, as shown in Fig. 1. However, the irregularity of point clouds imposes difficulties when aggregating meaningful information from a given point set, especially when such information is constrained by highly unbalanced distant points. An intuitive alternative to address the problem is to make use of multi-modality (e.g. flow). Thus, we begin leveraging input features dimension by learning a direction field vector for each point individually together with the target class. This helps to learn correlations between points with similar direction field vectors. Therefore, we train the hierarchical attention based target classification using direction modalities in a multi-task framework. To the best of the author's knowledge, this approach is novel and has not been investigated in literature before. We have also released our source code for better understanding and re-usability of our work.[1] Our main contributions in this work are summarized as:

- We propose a class agnostic anchor-free localization framework for target detection on a single-frame radar point cloud.
- Further, we use the inherent properties of the radar sensor to extract the direction vector of the target of interest. This additional information is used as key feature for target segmentation and localization using hierarchical spatial attention and Bayesian clustering, respectively. The director vector not only helps to handle class-imbalance but also to solve the problem of target classification and localization by introducing uniformity over incoherence sparse spatial dimension of the radar point clouds.
- A 2D bounding box-based object detection algorithm for radar data is modified with a centeredness score to improve the average precision for the localization. The feasibility of the radar object detection is demonstrated with a mean intersection over union (IoU) of 0.968 compared to 0.73 of SOTA.

2. Related work

While radar target detection, in general, has a long history, we only discuss the related work of processing point clouds in the following. A review of the equivalent research on target localization and classification, followed by flow estimation and joint attention paradigms, is performed. Also, the discussed literature is compared with the proposed approach.

2.1. Target localization

A typical target localization uses the point cloud as its input and predicts a 2D or 3D bounding box (Bbox) for each detected target. These methods can be divided into two categories: the two-stage and single-stage methods. The former methods are commonly known as region proposal based methods, requiring an initial anchor box generation.

Anchor-based:. These methods first propose several 2D regions of interest containing objects by leveraging anchor boxes applied at different (global or local) feature levels (Danzer et al., 2019; Qi, Liu et al., 2017; Shi et al., 2019; Simon et al., 2018; Yang et al., 2020). Later, the proposed region undergoes a refinement based on a non-maxima suppression (NMS) and an objectness score. While these methods demonstrate their superior performance in localizing objects and classifying them into the desired classes, the quality of the detection highly depends on the right configurations of the anchor boxes. In consequence, this imposes an additional requirement, i.e. manual design optimizations of the anchor box size and its dimension ratio based on the target sizes. Further, the abjectness score of a target depends on a dense feature map and the location of the RPN applied over multi-scale feature maps. In recent literature, both Mao et al. (2017) and Zhang et al. (2021) proposed a multi-task learning framework where the feature map is concatenated by an additional auxiliary task prior to RPN. While both approaches demonstrate significant improvement in target detection, the performance of RPN relies strongly on auxiliary feature maps.

Anchor-free:. Recent work on 3D-BoNet (Yang et al., 2019), uses global features directly to propose the box and abjectness scores without prior anchor boxes. The predicted boxes are associated with a ground truth boxes using euclidean distance and soft IoU. Later these boxes are fused with local features to estimate the point class (mask). The performance of the approach strongly relies on hard thresholding with the resulting association problem. In contrast to treating detection as an anchor or regression problem, Law and Deng (2018) introduce corner pooling, a new type of pooling layer that helps the network to better localize corners. Similarly, Zhou et al. (2019) suggest to detect objects by finding their extreme points.

2.2. Flow estimator

The scene flow can help to better understand the scene and may provide cues for object segmentation and detection from its motion

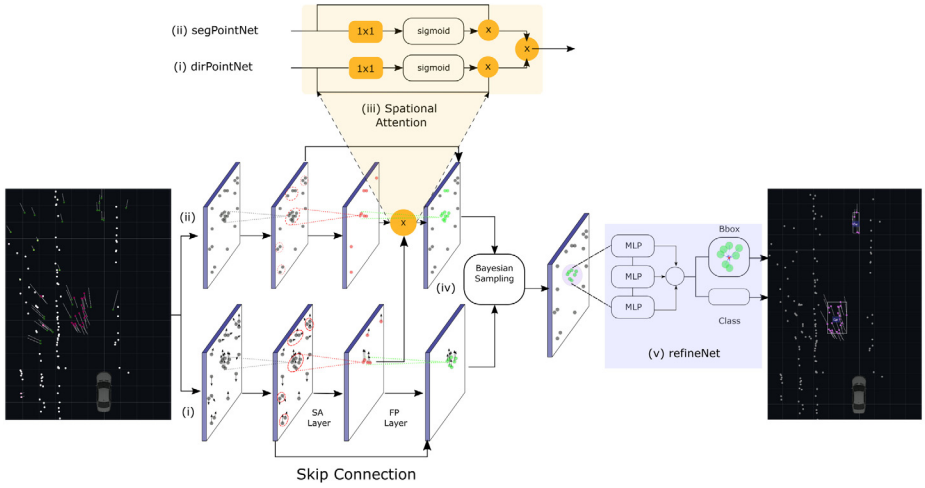

Fig. 2. Overview of our methods which is composed of five modules, (i) target's direction vector estimation (dirPointNet), (ii) hierarchical spatial attention inside target segmentation network (segPointNet) using target's direction modality, (iii) hierarchical spatial attention module, (iv) Bayesian sampling for target location using direction vector and (v) refineNet for bounding box (bbox) and target class estimation. Both dirPointNet and segPointNet follows the same architecture parameter with sampling abstraction layer (SA) and feature propagation (FP) layer.

vectors. Although there are no direct correlations between two sampled point clouds, the authors of Liu et al. (2018) propose a flow embedding layer to learn the association of points from their spatial localities and geometric similarities. Without the need for sequential data, only a single frame is used in Jiang et al. (2018) to extract unique feature representations for each point.

2.3. Attention

In contrast to LiDAR point clouds, the radar based point clouds are very sparse and non-uniform in nature. As a result, it is challenging to extract distinct target features. The complexity of the problem increases for highly unbalanced data. Different methods of the attention mechanism are proposed in the literature, which assists the network to learn important features by enhancing them at the input dimension or at the feature level (Li et al., 2020; Song et al., 2020; Wu & Miao, 2021; Zhao et al., 2019). The most common approaches of the attention method used in the literature can be group into multiview (sensors) and multi-modality (flow, depth). Here, each method can have soft, hard, Gaussian, or spatial filter based attention. While both multi-modality and multi-view based attentions bring additional knowledge inside the network and increase the robustness, they also have requirement of having multiple sensors.

3. Methodology

The given point cloud is denoted as $\mathbf{P}$ which contains n points $p_1, p_2, \ldots, p_n \in \mathbb{R}^d$ with d dimensional features. The input feature vector of each point p_i for segPointNet consists of the global target coordinate space (x_i, y_i), the azimuth angle (θ) and the signal reflection power (σ). The set of semantic labels is denoted as $\mathbf{L}$. Semantic segmentation of a point cloud is a function Ψ which assigns semantic labels to each point in the point cloud i.e. $\Psi : \mathbf{P} \longmapsto L^n$. The objective of segmentation algorithms is to find the optimal mapping from the input space to the semantic labels. However, the performance of the network strongly depends on the richness of input features presented to the network (Schumann et al., 2018).

In this work, we connect the concepts of multi-modality and attention to split the problem of target detection into three parts, as illustrated in Fig. 2. First, a one-channel direction field vector is estimated for each point. This outputs a coherent direction vector for all points belonging to an unique target. Afterwards, a direction field vector is used to provide attention inside the segmentation network to achieve better feature learning. In the end, the information from the segmented output and the direction field network is fused to perform a Gaussian sampling for unique cluster identification. These clusters are passed through another network for a bounding box estimation. The individual networks uses the PointNet++ (Qi, Yi et al., 2017) architecture adapted for the radar scene data (Schumann et al., 2021). We propose an end-to-end hierarchical, spatial, attention-based multi-modal segmentation framework, called HARadNet. The first network, named dirPointNet, learns the direction vector field for each point. This information is used inside the second network (segPointNet) to improve spatial localization. This approach increases the cross-correlation of the shared representations and potentially yields a faster convergence.

3.1. Direction field estimation

Using the inherent property of the radar sensor, the tangential velocity of the target is determined along with its spatial position in the azimuth dimension (θ) for each reflected point. Taking advantage of both values together with the sensors' yaw angle (ϕ), the direction of motion for each point is estimated as $d(\varphi) = \theta + \phi$. The velocity of the target is compensated with respect to the ego-motion, while the azimuth angle is transformed to global coordinates. In real-world scenarios, many reflections that do not belong to a moving object show a non-zero ego-motion-compensated velocity component, caused by errors in the odometry, sensor misalignment, time synchronization errors, mirror effects, or other sensor artifacts. In addition, reflections with zero velocity do not necessarily belong to a static object, since also reflections from the bottom of a rotating car wheel or body parts of a pedestrian that move perpendicular to the walking direction may show no radial velocity. As a result, multiple static targets are misinterpreted as dynamic ones. To overcome this problem, we optimize the dirPointNet network to estimate the direction of targets and suppress unwanted "noise" caused by multipath reflections.

The network follows an encode–decode scheme similar to a general semantic segmentation network, except for changing the problem from classification to regression. The network is trained using 4D input feature tensors $(\hat{x}_{\mathrm{cc}}, \hat{y}_{\mathrm{cc}}, \hat{\theta}_{\mathrm{cc}}, \vec{\hat{v}}_{\mathrm{r}})$ to predict the motion direction vector for each traffic participant $(d(\varphi))$. Furthermore, both input features and labels are rescaled to the range of $[0, 1)$ by applying

$$\hat{x}_{\mathrm{cc}} = \frac{x_{\mathrm{cc}} - x_{i_{\mathrm{cell}}}}{s_{x_{\mathrm{cell}}}}, \hat{y}_{\mathrm{cc}} = \frac{y_{\mathrm{cc}} - y_{k_{\mathrm{cell}}}}{s_{y_{\mathrm{cell}}}}, \vec{v}_{\mathrm{r}} = \frac{1}{v_{\max}} \vec{v}_{\mathrm{r}}, \hat{\theta}_{\mathrm{cc}} = \frac{\theta}{60°}, d(\hat{\varphi})_i = \frac{d(\varphi)_i}{180°} \tag{1}$$

Table 1
Comparison of the effect of attention at different feature abstraction layers for both binary (top-row) and multi-class (bottom-row) segmentation using the F1-score (see Eq. (4)).

w/o attention				1-depth attention				2-depth attention				3-depth attention			
Avg	Ped	Car	Bike	Avg	Ped	Car	Bike	Avg	Ped	Car	Bike	Avg	Ped	Car	Bike
0.92	–	–	–	0.95	–	–	–	**0.97**	–	–	–	0.96	–	–	–
0.88	0.61	0.85	0.90	0.92	0.43	0.97	0.77	0.94	0.46	0.98	0.64	**0.95**	**0.75**	**0.98**	**0.91**

with $(x_{i_{\text{cell}}}, y_{k_{\text{cell}}})$ representing the position of the left bottom corner of a cell. The indices (i, k), $(s_{x_{\text{cell}}}, s_{y_{\text{cell}}})$ resemble the cell extension in x, y-direction, while $v_{\max}$ is the maximum velocity, and $\sigma_{\max}$ the maximum signal power obtained from the whole data set. This rescaling restricts the gradient from exploding during the network training.

The resulting network optimization is still very challenging due to highly unbalanced target points and the sparsity of target features. Thus, we propose the following hybrid loss function to train the network.

$$\begin{aligned} L_{direction} &= w_{wmse} L_{wmse} + (1 - w_{wmse}) L_1, \\ L_{wmse} &= \frac{1}{n} \frac{\sum_{i=0}^{n} w_i \left(\hat{d}(\varphi)_i - d(\varphi)_i\right)^2}{\sum_{i=0}^{n} w_i}, \\ w_i &= \log\left(\hat{d}(\varphi)_i + 1\right) + 1. \end{aligned} \tag{2}$$

Here, L_1 represents absolute differences between the true value and the predicted value. The value of w_i is calculated over a number of positive samples in a batch. An empirically determined fixed value of 0.8 as used for weighted mean square error (w_{wmse}).

3.2. Direction attention

Due to sparsity, non-uniformity, and the highly imbalanced nature of target representations in radar point clouds, the actual target recognition becomes very challenging. Here we use point-wise multiplication to provide hard spatial attention inside the segPointNet. As dirPointNet and segPointNet share the same number of input tensors, we are able to preserve the flexibility of providing attention at different feature abstraction levels of the network. segPointNet uses $(\hat{x}_{\text{cc}}, \hat{y}_{\text{cc}}, \vec{\hat{v}}_{\text{r}}, \hat{\sigma})$ as input feature tensor, where $\hat{\sigma}$ is the normalized signal reflection power. We have used point-wise multiplication to provide hard attention inside network. For this purpose, prior to attention, both estimated segPointNet and dirPointNet outputs are standardized between 0 and 1. Additionally, 1×1 pointwise convolution with a residual connection to attention is used to avoid vanishing gradient. Due to the difference in input features, the estimated signal from both segPointNet and dirPointNet is standardized to range between 0 and 1 using a sigmoid function prior to the spatial attention inside segPointNet. Additionally, 1×1 pointwise convolution with a residual connection to attention is used to avoid vanishing gradient. Both network are optimized using end-to-end training. As a result, they complement each other in learning target features from different modalities. The total loss is formulated as

$$\begin{aligned} L_{attention} &= w_{cls} L_{cls} + (1 - w_{cls}) L_{direction}, \\ L_{cls} &= -\left(1 - \hat{p}_y\right)^{\gamma} \log\left(\hat{p}_y\right), \end{aligned} \tag{3}$$

where L_{cls} denotes the loss for the point classification from the scene and $L_{direction}$ is for the direction field estimation of classified radar targets. In Eq. (3), $y \in \{0, \ldots, K-1\}$ represents an integer class label, $\hat{\mathbf{p}} = \left(\hat{p}_0, \ldots, \hat{p}_{K-1}\right) \in [0, 1]^K$ is a vector representing the estimated probability distribution over the K classes and γ is a focusing parameter which specifies how much high-confidence predictions contribute to the overall loss. Table 1 gives an insight on the effect of spatial attention by dirPointNet on segPointNet by evaluating the network performance using the F1-score. The F1-score is the harmonic mean between precision P and recall R, given by

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}. \tag{4}$$

The performance of segPointNet is evaluated for both binary and multi-class segmentation tasks. While the binary segPointNet is optimized to predict only the foreground as targets of interest and the background as static-reflections or noise, multi-class segPointNet is trained to preserve the target class and background. The binary segPointNet demonstrates a better average F1-score of 0.92 in contrast to the multi-class segPointNet with an average F1-score of 0.88. This is caused by the unavailability of uniform features between points of the same class. In contrast to the case without attention, both binary and multi-class segPointNet show an improvement in the average F1-score with attention applied inside the network and validates advantage and scalability of our proposed architecture. Further, the multi-class segPointNet shows a significant improvement in the average F1-score being equal to 0.95 when attention is applied at every feature propagation layer. It is comparable to binary segPointNet with an average F1-score of 0.96. Additionally, the network shows major improvements in the recognition of pedestrians, which share the least samples of target distribution in the dataset (Schumann et al., 2021). This proves the advantage of attention inside a network with different modalities which acts as an additional target feature point and improves the scene recognition.

Table 2
Distribution of the six different target class in the dataset (Schumann et al., 2021).

Car	Ped.	Ped. group	Bike	Truck	Static
1.23%	0.31%	0.74%	0.11%	0.60%	97.01%

4. Experiments

The evaluation of the proposed framework is done on real-world data that was collected by four automotive corner radar sensors. All reflections that belong to the same physical object are grouped and annotated with a corresponding label from the following classes: car, pedestrian, pedestrian group, bike, truck and static. The distribution of the occurrences among the six classes is shown in Table 2. This gives a clear indication of the typical foreground vs background class imbalance, present in the data. Furthermore, pedestrians and bikes have the least number of training samples. Additionally, both share a lower signal reflection strength and sparse point distribution in comparison to the other target classes. As a result, the segPointNet struggles to categorize these classes correctly, as shown in Table 1. In addition, Table 2 shows the distribution of object availability with respect to the distances to the ego-vehicle. Following the object distribution over distance, we process cropped scenes within a range of 80 m for $\hat{x}_{\text{cc}}$ and an absolute range of 20 m for $\hat{y}_{\text{cc}}$. This reduces the total number of static targets during the network training.

4.1. Bayesian sampling

Compared to other vision tasks such as segmentation or categorization, localizing the object is a very complex task, mainly because the same region could also jointly belong to another target, if it is closely located or partially occluded. Additionally, due to the in-homogeneous and sparse distribution of radar points in the point cloud space, points from neighboring regions often have similar characteristics. Therefore, we intend to seek a method that has the ability to discover potential and meaningful patterns among proximity points, so that the set of points can be clustered into unique groups in a robust way.

In order to deal with such situations, our attention direction network can be guided not only towards the more relevant features but

Table 3
Comparison of the class-agnostic clustering methods for target localization with normalized features i.e. compensated spatial location in $\hat{x}_{cc}$, $\hat{y}_{cc}$ and azimuth $\hat{\theta}$), compensated target velocity ($\hat{v}_r$), reflected signal strength $\hat{\sigma}$ and motion direction vector ($\hat{d(\varphi)}$) estimated from dirPointNet.

Normalized feature dimensions						Clustering score (meanIoU [%])			
$\hat{x}_{cc}$	$\hat{y}_{cc}$	$\overline{\hat{v}}_r$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{d(\varphi)}$	BIC + GM	DBSCAN + GMM	DBSCAN + VBGM (Dirichlet process)	DBSCAN + VBGM (Dirichlet distribution)
✓	✓					0.195	0.700	0.703	0.711
✓	✓	✓				0.731	0.640	0.747	0.746
✓	✓	✓	✓			0.328	0.722	0.739	0.734
✓	✓	✓	✓	✓		0.497	0.593	0.601	0.598
✓	✓	✓	✓	✓	✓	0.640	0.652	0.661	0.652
✓	✓	✓	✓		✓	0.445	0.796	0.806	0.808
✓	✓	✓			✓	0.481	0.850	0.859	0.858
		✓			✓	**0.815**	**0.894**	**0.893**	**0.894**

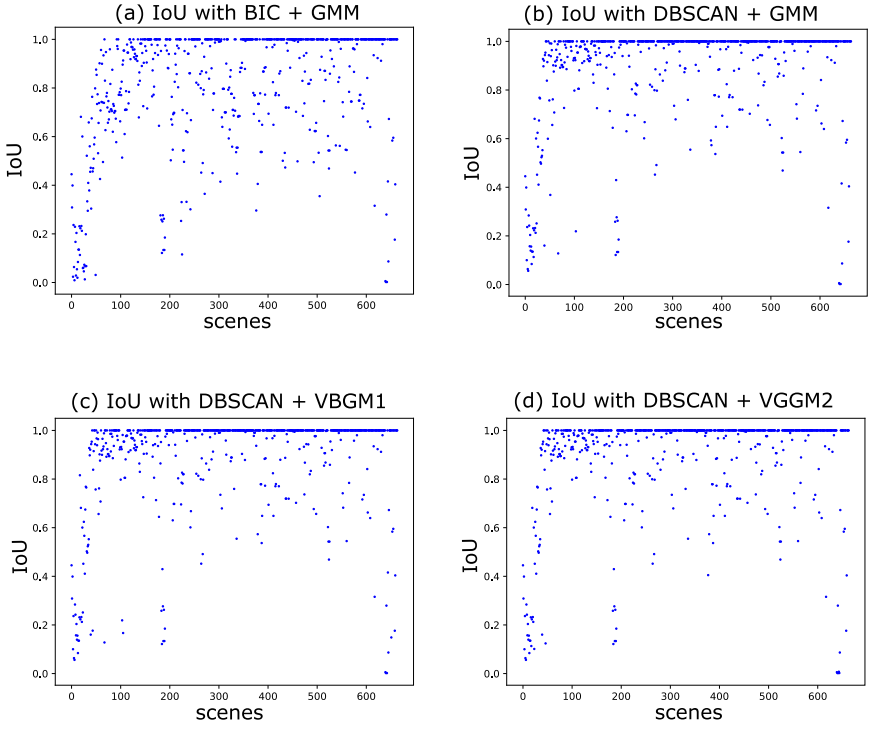

Fig. 3. Visualization of the IoU population distribution for all target of interest over the entire test sequence.

also towards the selection of unique regions, using a cluster of direction vectors as a signature distribution, in combination with spatial information. For ease of usage, we call this step Bayesian sampling, which is performed in two steps. At first, both spatial and direction information is fused and passed to different algorithms to estimate the possible number (order) of targets available in the scene. This is used by the clustering algorithm to find and localize them uniquely. In our experiment, we used the density-based spatial clustering of applications with noise (DBSCAN) and the Bayesian information criterion (BIC) for the estimation of target order. Thereafter, the output is clustered into the desired unique bins using a Gaussian mixture model (GMM) operated at different feature dimensions. To the best of the author's knowledge, this approach is novel and has not been investigated in literature before.

While Table 1 demonstrates the quantitative advantage of direction vector attention for radar point–target segmentation, Table 3 shows the advantage of the direction vector as the key feature dimension for target clustering. To compare the performance of localization, clustering algorithms are evaluated for different dimensions of features ($\hat{x}_{cc}$, $\hat{y}_{cc}$, $\overline{\hat{v}}_r$, $\hat{\sigma}$, $\hat{\theta}$, $\hat{d(\varphi)}$). Further, to evaluate the generalization of our approach, the performance of clustering is evaluated for different clustering algorithms i.e. GMM and its variant, variational Bayesian Gaussian mixture model (VBGM). The principle behind VBGM is the same as for GMM, which is expectation minimization but VBGM also adds a regularization by integrating prior distribution information. Although priors may bring initial biases, the VBGM selects a suitable number of effective clusters (targets) by avoiding the singularities which are often found in expectation–maximization solutions, and pushing weights values close to zero.

$$\text{IoU} = \frac{\left|b_{gt} \cap b_{pred}\right|}{\left|b_{gt} \cup b_{pred}\right|}, \tag{5}$$

A common evaluation metric for segmentation tasks is the mean intersection over union (mIoU). The IoU compares a predicted bounding box b_{pred} with the ground truth bounding box b_{gt} and is defined as Eq. (5). Here, $|\cdot|$ measures the total area of the underlying set. If the ground truth and the predicted bounding box are almost identical, the IoU score tends to be close to one. If the two bounding boxes do not overlap, the IoU score will be zero. While $\hat{d(\varphi)}$ helps to increase the localization accuracy, due to non-uniformity in spatial dimensions ($\hat{x}_{cc}$, $\hat{y}_{cc}$, $\hat{\sigma}$) of the radar-point clouds, the performance of clustering is strongly limited. As a result, only the estimated direction vector ($\hat{d(\varphi)}$) and its magnitude ($\overline{\hat{v}}_r$) are considered for localization. This leads to significant improvement in localization accuracy with a mean IoU of $\approx$ 90% using DBSCAN and GMM and its variant, validating the robustness of each feature combination. Fig. 3 illustrates the statistics of IoU for

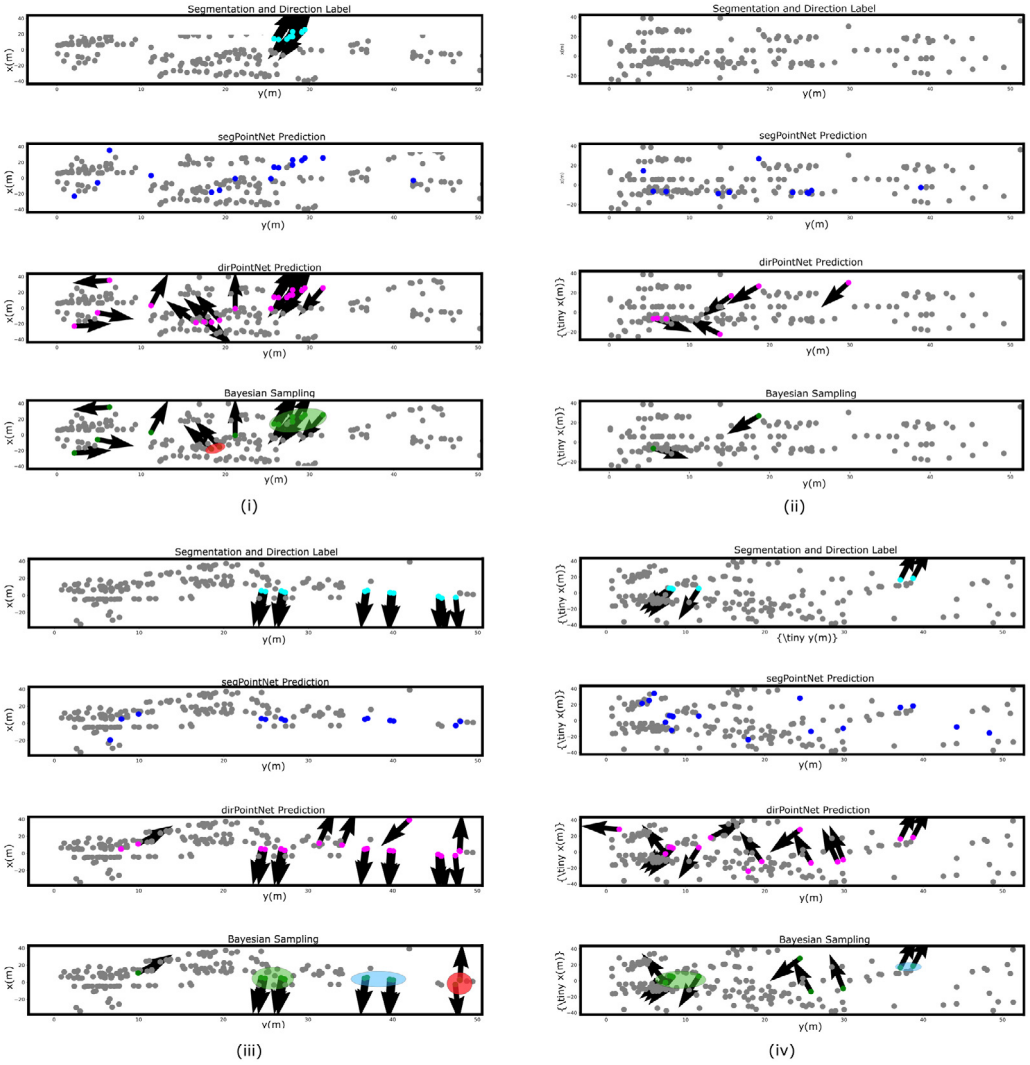

Fig. 4. Visualization of the different intermediate outputs of the proposed three stage approach. The first row in each of (i–iv) represents the spatial input with the corresponding label and its associated direction field. The last three rows of (i–iv) show the predicted segmentation, estimated direction vector and Bayesian sampling for the region of interest estimation.

all clustering methods over the entire validation scene for all targets of interest. A large percentage of all scenes is distributed around 1.0 in the vertical axis of IoU for all four subplots, indicating that many scenes in the sequence achieve a perfect match. The combination of DBSCAN and GMM gains over 50%, and finally the percentage of scenarios with an mIoU over 0.8 exceeds 82%. The best performance is achieved by the combinations of DBSCAN with VBGM using a Dirichlet distribution as priors. As a result, the rest of our experiments and evaluations are based on this clustering method.

Fig. 4 illustrates the different stages for target localization using the proposed Bayesian sampling and the effect of the estimated direction vector from the dirPointNet. To have a deeper insight on the advantages of the approach, multiple examples together with their behavior on boundary conditions are demonstrated using four random scenes from the validation set. The top row shows the input point cloud marked with the ground truth label and the direction vector. The middle two rows show the output of the segPointNet and the dirPointNet. The last row follows the output of the Bayesian sampling layer. Fig. 4(i) shows multiple false positives from both segPointNet and dirPointNet. The Bayesian sampling layer, however, suppresses false positives by fusion and later discards all the points with high variance in their feature dimension due to the high variability of the direction vectors between the neighboring points. As a result, the network successfully finds regions of interest. Fig. 4(iii) shows multiple detections for the same target and misdetections for the target due to non-coherent direction vectors. Multiple detections for the same target can be suppressed by non-maxima suppression in the post-processing stage. The examples demonstrate that the performance of Bayesian sampling strongly relies on the $d(\hat{\theta}_{cc})$ feature in comparison to the distributed and non-uniform spatial feature dimension $(\hat{x}_{cc}, \hat{y}_{cc})$. Furthermore, Fig. 4(ii) demonstrates an interesting observation and advantage of our approach where both segPointNet and dirPointNet predict a target. The Bayesian sampling layer, however, discards both points due to non-coherent direction vectors and spatial sparsity (no neighborhood). Thus, our approach also helps to suppress false positive detections.

4.2. Multi-task learning

After a joint end-to-end learning of multi-class segmentation and direction field estimation using a input dimension of $4 \times n$, the region of interest in the form of unique point clusters, with the dimension of $4 \times m$, is passed through a refineNet, conceptually similar to Qi, Su, Mo et al. (2016). The refineNet is a multi-head network with a regression and classification head. While regression head predicts parameters of a 2D bounding box (Bbox), i.e. its center (x_c, y_c) and its size (l, w) around the clustered points, the classification head is optimized to predict the

Table 4
Comparison of the localization accuracy for the class-agnostic and class-aware bounding box (Bbox) estimation.

Weighting criteria	Task weights		Classification	Bbox
	Class ($w_{s,cls}$)	Bbox (w_{Bbox})	[F1-score]	[mIoU]
Class only	1	0	0.92	0.89
Bbox only	0	1	**0.95**	0.86
Empirical weighting	0.5	2	0.78	0.93
SOTA (Danzer et al., 2019)	–	–	0.64	0.64
YOLO	2	7, 2, 5, 0.5	0.66	0.32
Weights with task uncertainty (Cipolla et al., 2018)	–	–	0.82	**0.96**

target class for the points. For the box center estimation, a residual-based 2D localization is performed, similar to Qi et al. (2019), where the network estimates the centroid over the center. To guarantee a fixed number of input points to the FC layer, the sampling process during training is considered. For the 2D bounding box estimation, up to 32 points are randomly sampled from the point clusters for every radar target. The labeling process automatically generates a bounding box from the annotated radar point targets by using the ground truth as a reference. The training is performed with a multitask loss for joint optimization of segmentation, direction field and a 2D bounding box estimation. Since the performance of the box prediction relies on the region proposal which in turn depends on the dirPointNet prediction, a trained network is used for the initialization of all weights before the training. The, multitask loss is defined as

$$\begin{aligned} L_{multi-task} &= L_{attention} + L_{refineNet} \\ L_{refineNet} &= w_{Bbox}\left(L_{c_x,reg} + L_{c_y,reg} + \right. \\ &\quad \left. L_{h,reg} + L_{w,reg}\right) + w_{s,cls} L_{s,cls}. \end{aligned} \tag{6}$$

During the training, the weight for a target with $L_{attention}$ is handled using Eqs. (2) and (3). $L_{c_x,reg}$ and $L_{c_y,reg}$ are used for the residual based center regression of the box estimation network. Furthermore, $L_{w,reg}$ and $L_{h,reg}$ are losses for width and height estimation, while $L_{s,cls}$ is for the estimation of the target class of the box with w_{Bbox} and $w_{s,cls}$ their respective task weightings. The choice of loss for box regression and target classification are smoothL1 and cross-entropy.

The network is evaluated on the full test data set to cover the maximum number of different situations, including the corner cases, in order to understand the behavior of the networks for targets like pedestrians and cyclists. Since the proposed 2D object detection method contains classification and bounding box estimation, the performance of these modules will be evaluated using the F1-score (compare Eq. (4)) and the IoU metric. Additionally, the performance of the multi-head box network is evaluated under different training conditions and compared with both region based state-of-the-art architecture and regression based YOLO. The detailed performance of all approaches are summarized in Table 4. First, the performance of the network is evaluated separately for cluster classification and class-agnostic bounding box (Bbox) estimation. This is done by enabling either $w_{s,cls}$ or w_{Bbox} exclusively. The mIoU for the class-only scenario is the same as the mIoU for the target localization since the Bbox is not fine-tuned for clustered points but achieves a classification accuracy of 0.92. On the other hand with Bbox-only, the target localization is degraded from mIoU of 0.89 to 0.86, while the classification accuracy in this case is the same as the F1-score from the multi-class segmentation task. Both centroid and corner points are estimated without knowing the target class (point distribution), thus the network considers all points as an inlier and tries to fit the bounding box to it, which results in a drift of the centroid. This leads to a bad localization accuracy of 0.86 for the multi-class segmentation network. By learning the target class distribution together with the bounding box estimation, the network shows a slightly better localization accuracy with an IoU of 0.93 and a classification accuracy of 0.78. The strong weighting for w_{Bbox} in contrast to $w_{s,cls}$ results from the much stronger average classification loss, compared to the box regression.

Additionally, the proposed framework is compared with a state-of-the-art region based object detection (Danzer et al., 2019) and a regression based detection algorithm using YOLO architecture (Simon et al., 2018). The SOTA network shows the best target classification of 0.96, at the cost of localization mIoU of 0.64. The improved target classification score is due to the reason that the performance of SOTA is evaluated over accumulated multiple frames over 500 ms, in contrast to our approach where the network is evaluated for a single frame. Additionally, the feature dimension used for localization refinement and classification of clustered point includes the original input dimensions ($\hat{x}_{cc}$, $\hat{y}_{cc}$, $\hat{\theta}$, $\hat{\sigma}$) and not the estimated direction vector ($\vec{\hat{v}}_r$). As a result, the network struggles to classify clustered point-clouds due to incoherency between spatial features. Further, we also compared our proposed framework with the YOLO architecture by optimizing it directly on our normalized input data over a single frame. The YOLO results in worse localization and classification accuracy of 0.32 and 0.66, respectively.

Although the multi-task approach aims to improve the learning efficiency by learning multiple objectives from shared representations, the performance of the multi-task network optimization is highly sensitive to weights (w_{Bbox}, $w_{s,cls}$) given to the different losses (L_{bbox}, L_{cls}). In contrast to an expensive grid-search or naive weighted sum of losses, the network is optimized using online learned weights with task (homoscedastic aleatoric) uncertainty which captures relative confidence between tasks, motivated by Bischke et al. (2019) and Cipolla et al. (2018). As a result, the Bbox and target class estimation is modified with joint-learning function $p\left(\mathbf{y}_1, \mathbf{y}_2 \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right)$ where, $\mathbf{y}_1$ and $\mathbf{y}_2$ represents bbox regression and target classification as two outputs from the multi-head network $\mathbf{f}^{\mathbf{W}}(\mathbf{x})$. This leads to the minimization objective of $-\log p\left(\mathbf{y}_1, \mathbf{y}_2 \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right)$ for the multi-output model, given as:

$$\begin{aligned} &= -\log p\left(\mathbf{y}_1, \mathbf{y}_2 = c \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right) \\ &= -\log \mathcal{N}\left(\mathbf{y}_1; \mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma_{y_1}^2\right) \cdot \mathrm{Softmax}\left(\mathbf{y}_2 = c; \mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma_{y_2}\right) \\ &= \frac{1}{2\sigma_{y_1}^2}\left\|\mathbf{y}_1 - \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right\|^2 + \log \sigma_{y_1} - \log p\left(\mathbf{y}_2 = c \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma_{y_2}\right) \\ &= \frac{1}{2\sigma_{y_1}^2}\mathcal{L}_1(\mathbf{W}) + \frac{1}{\sigma_2^2}\mathcal{L}_2(\mathbf{W}) + \log \sigma_{y_1} \\ &+ \log \frac{\sum_{c'} \exp\left(\frac{1}{\sigma_{y_2}^2} f_{c'}^{\mathbf{W}}(\mathbf{x})\right)}{\left(\sum_{c'} \exp\left(f_{c'}^{\mathbf{W}}(\mathbf{x})\right)\right)^{\frac{1}{\sigma_{y_2}^2}}} \\ &\approx \frac{1}{2\sigma_{y_1}^2}\mathcal{L}_1(\mathbf{W}) + \frac{1}{\sigma_{y_2}^2}\mathcal{L}_2(\mathbf{W}) + \log \sigma_{y_1} + \log \sigma_{y_2}, \end{aligned} \tag{7}$$

Here, $\mathcal{L}_1(\mathbf{W})$ stands for the $smoothL_1(\mathbf{y}_1, \mathbf{f}^{\mathbf{W}}(\mathbf{x}))$ for the regression loss $\mathbf{y}_1$ and $\mathcal{L}_2(\mathbf{W}) = -\log \mathrm{Softmax}\left(\mathbf{y}_2, \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right)$ for the cross-entropy loss $\mathbf{y}_2$. The network is trained to predict the log-variance $\log \sigma_y^2$ for more numerical stability and avoids gradient division when the loss is zero. The network shows the best localization accuracy, compared to our proposed framework and SOTA. This is due to the reason that the localization loss is very sensitive to both the estimated corner and the center points. As a result, the task uncertainty based approach helps the network to choose an appropriate loss weighting during the training. While this approach helps the network to improve the classification accuracy by significant amounts compared to the empirical weighting, the network still struggles to classify very sparse and spatially distributed clustered points into the desired target class without using the estimated direction vector as another feature dimension.

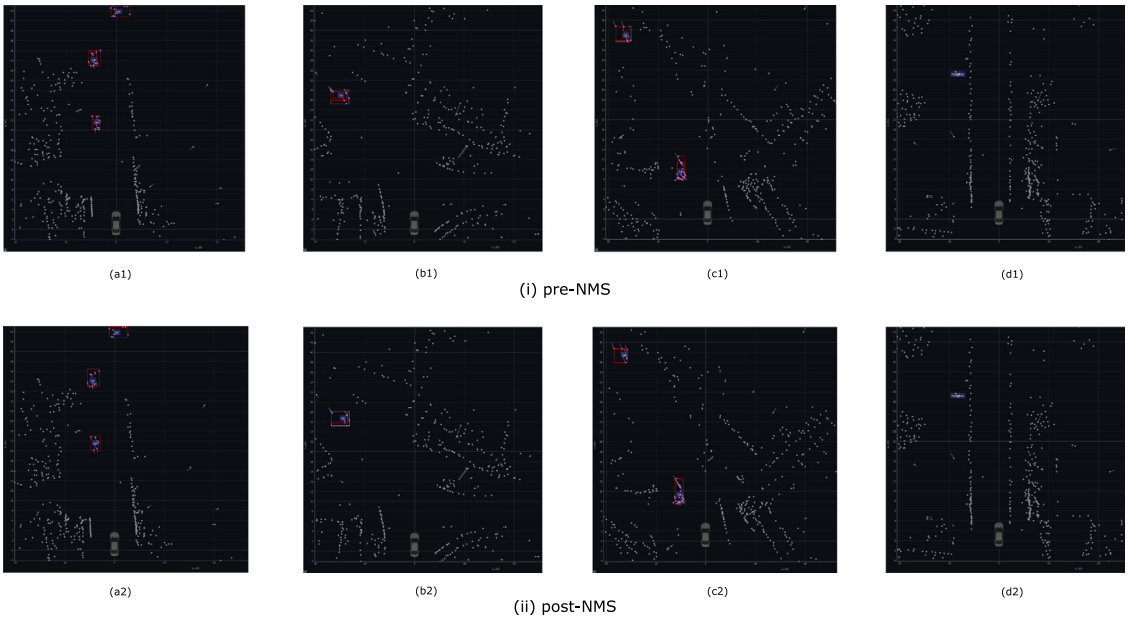

Fig. 5. Combined illustration of the performance of HARadNet. (i) shows the class-aware bounding box estimation, and (ii) illustrates the final prediction after using non-maxima suppression (NMS) as post-processing.

In addition to the quantitative evaluation, Fig. 5 illustrates few corner cases of our proposed anchor free detection framework and its localization and classification accuracy. Fig. 5(a1) shows multiple overlapping box proposals around the ground truth target due to varying target distribution. Consequently, the concept of non-maxima suppression (NMS) can be used as post-processing. Thus, all the boxes having IoU $>$ 0.5 with the ground truth are considered. Fig. 5(a2) illustrates the updated bounding box tightly coupled with the ground truth. Similarly, both Fig. 5(d1) and (d2) demonstrate the effectiveness of HARadNet for successful localization of a target with just four points. Additionally, it also preserves the target class. As a result, the need for predefined anchor boxes or grid-based regression methods can be eliminated. On the other hand, while Fig. 5(b) and (c) demonstrates the target localization, the estimated bounding box leaves out some target points treating as an outlier. As a result, the network contributes to the false-negatives during the target localization. This is caused by the loss function (L_{box}) which does not penalize loss caused by the background and the foreground separately. As an alternative, in the future, loss functions similar to the one proposed in Hall et al. (2018), can be used for better localization.

5. Conclusion

This work addresses the problem of target detection and classification on sparse non-uniform distributed radar point clouds. We proposed an anchor-free model for target localization and classification employing hierarchical spatial attention captured in the form of motion modality by using direction field vectors for each target point. This is done without using additional sensors or other dependencies such as temporal information or cross-model distillation. The entire model is trained in an end-to-end training framework using the concept of multi-task learning. This approach helps the model to converge faster for the combined tasks while sharing the learned representations. The joint motion–spatial attention mechanism for feature selection is highly essential, to select useful features from the learned representations and to improve localization performance. Furthermore, we propose a new Bayesian sampling layer which takes both spatial and motion modality to sample and cluster points with similar feature distributions. Extensive experiments validate the efficiency and robustness of this approach across a wide range of examples from the public RadarScenes dataset. For future work, we will extend our approach to Bayesian models by using the uncertainty from the Bayesian sampling layer and use probabilistic evaluation method in contrast to deterministic approaches, motivated by Hall et al. (2018).

CRediT authorship contribution statement

Anand Dubey: Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper. **Avik Santra:** Performed the analysis, Writing – original draft, Writing – review & editing. **Jonas Fuchs:** Writing – original draft, Writing – review & editing. **Maximilian Lübke:** Writing – original draft, Writing – review & editing. **Robert Weigel:** Approval of the final version. **Fabian Lurz:** Approval of the final version.

Acknowledgments

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.

References

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Gall, J., & Stachniss, C. (2021). Towards 3D lidar-based semantic scene understanding of 3D point cloud sequences: The semantickitti dataset. *International Journal of Robotics Research*, *40*(8–9), 959–967. http://dx.doi.org/10.1177/02783649211006735, arXiv:https://doi.org/10.1177/02783649211006735.

Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C., & Winner, H. (2014). Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine*, *6*(4), 6–22. http://dx.doi.org/10.1109/MITS.2014.2336271.

Bischke, B., Helber, P., Folz, J., Borth, D., & Dengel, A. (2019). Multi-task learning for segmentation of building footprints with deep neural networks. In *2019 ieee international conference on image processing (icip)* (pp. 1480–1484). http://dx.doi.org/10.1109/ICIP.2019.8803050.

Cipolla, R., Gal, Y., & Kendall, A. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 ieee/cvf conference on computer vision and pattern recognition* (pp. 7482–7491). http://dx.doi.org/10.1109/CVPR.2018.00781.

Danzer, A., Griebel, T., Bach, M., & Dietmayer, K. (2019). 2D car detection in radar data with PointNets. CoRR abs/1904.08414 arXiv:1904.08414 URL: http://arxiv.org/abs/1904.08414.

Dubey, A., Fuchs, J., Luebke, M., Weigel, R., & Lurz, F. (2020a). Generative adversial network based extended target detection for automotive MIMO radar. In *2020 ieee international radar conference (radar)* (pp. 220–225).

Dubey, A., Fuchs, J., Luebke, M., Weigel, R., & Lurz, F. (2020b). Region based single-stage interference mitigation and target detection. In *IEEE radar conference 2020*.

Dubey, A., Santra, A., Fuchs, J., Lübke, M., Weigel, R., & Lurz, F. (2021a). A Bayesian framework for integrated deep metric learning and tracking of vulnerable road users using automotive radars. *IEEE Access*, *9*, 68758–68777. http://dx.doi.org/10.1109/ACCESS.2021.3077690.

Dubey, A., Santra, A., Fuchs, J., Lübke, M., Weigel, R., & Lurz, F. (2021b). Integrated classification and localization of targets using Bayesian framework in automotive radars. In *Icassp 2021 - 2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4060–4064). http://dx.doi.org/10.1109/ICASSP39728.2021.9414131.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., & Bennamoun, M. (2019). Deep learning for 3D point clouds: A survey. CoRR abs/1912.12033 arXiv:1912.12033 URL: http://arxiv.org/abs/1912.12033.

Hakobyan, G., & Yang, B. (2019). High-performance automotive radar: A review of signal processing algorithms and modulation schemes. *IEEE Signal Processing Magazine*, *36*(5), 32–44. http://dx.doi.org/10.1109/MSP.2019.2911722.

Hall, D., Dayoub, F., Skinner, J., Corke, P., Carneiro, G., & Sünderhauf, N. (2018). Probability-based detection quality (PDQ): a probabilistic approach to detection evaluation. CoRR abs/1811.10800 arXiv:1811.10800 URL: http://arxiv.org/abs/1811.10800.

Huang, S., Chen, Y., Yuan, T., Qi, S., Zhu, Y., & Zhu, S. (2019). PerspectiveNet: 3D object detection from a single RGB image via perspective points. CoRR abs/1912.07744 arXiv:1912.07744 URL: http://arxiv.org/abs/1912.07744.

Izadinia, H., Shan, Q., & Seitz, S. M. (2016). IM2cad. CoRR abs/1608.05137 arXiv:1608.05137 URL: http://arxiv.org/abs/1608.05137.

Jiang, M., Wu, Y., & Lu, C. (2018). Pointsift: A SIFT-like network module for 3D point cloud semantic segmentation. CoRR abs/1807.00652 arXiv:1807.00652 URL: http://arxiv.org/abs/1807.00652.

Law, H., & Deng, J. (2018). CornerNet: Detecting objects as paired keypoints. CoRR abs/1808.01244 arXiv:1808.01244 URL: http://arxiv.org/abs/1808.01244.

Li, X., Zhang, L., Wang, L., & Lu, J. (2020). Multi-scale receptive fields graph attention network for point cloud classification. CoRR abs/2009.13289 arXiv:2009.13289 URL: https://arxiv.org/abs/2009.13289.

Liu, X., Qi, C. R., & Guibas, L. J. (2018). Learning scene flow in 3D point clouds. CoRR abs/1806.01411 arXiv:1806.01411 URL: http://arxiv.org/abs/1806.01411.

Mao, J., Xiao, T., Jiang, Y., & Cao, Z. (2017). What can help pedestrian detection? In *2017 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 6034–6043). http://dx.doi.org/10.1109/CVPR.2017.639.

Maturana, D., & Scherer, S. (2015). VoxNet: A 3D convolutional neural network for real-time object recognition. In *2015 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 922–928). http://dx.doi.org/10.1109/IROS.2015.7353481.

Meinel, H. H. (2014). Evolving automotive radar — From the very beginnings into the future. In *The 8th european conference on antennas and propagation (eucap 2014)* (pp. 3107–3114). http://dx.doi.org/10.1109/EuCAP.2014.6902486.

Murad, M., Bilik, I., Friesen, M., Nickolaou, J., Salinger, J., Geary, K., & Colburn, J. S. (2013). Requirements for next generation automotive radars. In *2013 ieee radar conference (radarcon13)* (pp. 1–6). http://dx.doi.org/10.1109/RADAR.2013.6586127.

Qi, C. R., Litany, O., He, K., & Guibas, L. J. (2019). Deep hough voting for 3D object detection in point clouds. CoRR arXiv:1904.09664 arXiv:1904.09664 URL: http://arxiv.org/abs/1904.09664.

Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2017). Frustum PointNets for 3D object detection from RGB-d data. CoRR arXiv:1711.08488 arXiv:1711.08488 URL: http://arxiv.org/abs/1711.08488.

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2016). PointNet: Deep learning on point sets for 3D classification and segmentation. CoRR abs/1612.00593 arXiv:1612.00593 URL: http://arxiv.org/abs/1612.00593.

Qi, C. R., Su, H., Nießner, M., Dai, A., Yan, M., & Guibas, L. J. (2016). Volumetric and multi-view CNNs for object classification on 3D data. CoRR abs/1604.03265 arXiv:1604.03265 URL: http://arxiv.org/abs/1604.03265.

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. CoRR abs/1706.02413 arXiv:1706.02413 URL: http://arxiv.org/abs/1706.02413.

Santra, A., & Hazra, S. (2020). *Deep learning applications of short range radars*. Artech House, URL: https://books.google.de/books?id=Qb-VzQEACAAJ.

Schumann, O., Hahn, M., Dickmann, J., & Wöhler, C. (2018). Semantic segmentation on radar point clouds. In *2018 21st international conference on information fusion (fusion)* (pp. 2179–2186). http://dx.doi.org/10.23919/ICIF.2018.8455344.

Schumann, O., Hahn, M., Scheiner, N., Weishaupt, F., Tilly, J., Dickmann, J., & Wöhler, C. (2021). *RadarScenes: A real-world radar point cloud data set for automotive applications*. Zenodo, http://dx.doi.org/10.5281/zenodo.4559821.

Shi, S., Wang, Z., Wang, X., & Li, H. (2019). Part-a^2 net: 3D part-aware and aggregation neural network for object detection from point cloud. CoRR abs/1907.03670 arXiv:1907.03670 URL: http://arxiv.org/abs/1907.03670.

Simon, M., Milz, S., Amende, K., & Gross, H. (2018). Complex-YOLO: Real-time 3D object detection on point clouds. CoRR abs/1803.06199 arXiv:1803.06199 URL: http://arxiv.org/abs/1803.06199.

Song, X., Zhan, W., Che, X., Jiang, H., & Yang, B. (2020). Scale-aware attention-based PillarsNet (SAPN) based 3D object detection for point cloud. *Mathematical Problems in Engineering*, *2020*, 1–12.

Stolz, M., Schubert, E., Meinl, F., Kunert, M., & Menzel, W. (2017). Multi-target reflection point model of cyclists for automotive radar. In *2017 european radar conference (eurad)* (pp. 94–97).

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. G. (2015). Multi-view convolutional neural networks for 3D shape recognition. CoRR abs/1505.00880 arXiv:1505.00880 URL: http://arxiv.org/abs/1505.00880.

Wu, H., & Miao, Y. (2021). LRA-net: local region attention network for 3D point cloud completion. In *International conference on machine vision*.

Yang, Z., Sun, Y., Liu, S., & Jia, J. (2020). 3DSsd: Point-based 3D single stage object detector. CoRR abs/2002.10187 arXiv:2002.10187 URL: https://arxiv.org/abs/2002.10187.

Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., & Trigoni, N. (2019). Learning object bounding boxes for 3D instance segmentation on point clouds. CoRR abs/1906.01140 arXiv:1906.01140 URL: http://arxiv.org/abs/1906.01140.

Zhang, X., Huo, C., Xu, N., Jiang, H., Cao, Y., Ni, L., & Pan, C. (2021). Multitask learning for ship detection from synthetic aperture radar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 8048–8062. http://dx.doi.org/10.1109/JSTARS.2021.3102989.

Zhao, X., Liu, Z., Hu, R., & Huang, K. (2019). 3D object detection using scale invariant and feature reweighting networks. CoRR abs/1901.02237 arXiv:1901.02237 URL: http://arxiv.org/abs/1901.02237.

Zhou, X., Zhuo, J., & Krähenbühl, P. (2019). Bottom-up object detection by grouping extreme and center points. CoRR abs/1901.08043 arXiv:1901.08043 URL: http://arxiv.org/abs/1901.08043.