Anchor-free Multi-task learning for Automotive Radar using Hierarchical Attention

Anonymous WACV submission

Paper ID ****

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 a novel attention based anchor free target detection and classification using multi-task learning method for radar point clouds data. A direction field vector, as modality, is used to achieve attention inside the network. The model 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 joint feature representation and correlation to its neighborhood, leading to improvement in performance for classification. Additionally, a parameter-free target localization is proposed using Bayesian sampling conditioned on a pre-trained direction field vector. The extensive evaluation on public radar dataset shows an substantial increase in localization and classification performance compared to state-of-art (SOTA) algorithms.

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 [1, 17, 19, 47, 50]. While cameras and LiDAR have been the two major sensing modalities in the autonomous driving field, but they have their respective shortcomings. For example, object detection based on cameras is highly prone to lightning conditions, on the other hand, LiDAR is not reliable in adversarial weather situations. The several shortcomings from individual sensors can be complemented using sensor fusion [4, 18]. Sensor fusion helps by fusing information at different hierarchy-level of the network architecture (early, late, or feature) [2, 13]. However, this comes at the cost of high computational power and a large number of parameters.

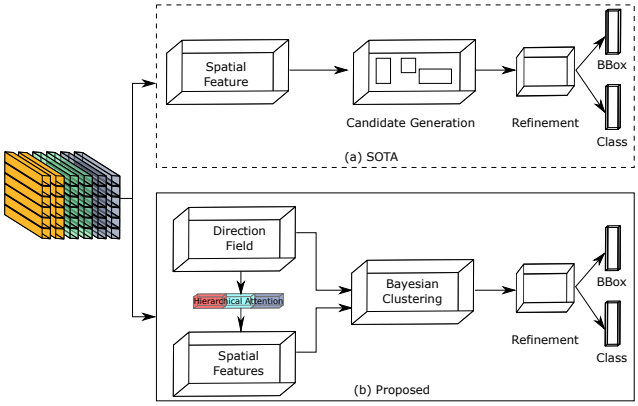

Figure 1. Comparison between the radar point cloud based state-of-the-art target detection and the proposed method. (a) The performance of SOTA relies on prior anchor box at different feature abstraction. (b) Ours proposes multi-model based target localization hierarchical attention at different feature abstraction.

Besides camera and LiDAR sensors, radar, being 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 [3, 7, 26, 42]. This is utilized to leverage multi modalities from the one sensor without introducing additional parameters, leading in an increase in both redundancy and robustness of perception in the context of autonomous driving. With recent advancement in radar systems [25], especially in processing high-resoluation data, target detection and classification are typically done via multiple sub-task specific blocks [12, 15, 27]. While target detection is based on peak detection and clustering [8, 9], target classification is typically performed via extracting target-specific parameters, e.g. Doppler spectrograms [5, 10, 34, 40, 43]. 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. Dubey et al. [11] showed that the classifi-

cation and tracking performance in a deep Bayesian metric learning framework strongly relies on the resolution of the spectrogram and in consequence on the exact radar parameters. However, the approach essentially lacks end-to-end learning.

An end-to-end solution can omit all the sub-tasks of the modular solution by mapping the high dimensional inputs from sensors directly to the desired task. Therefore, in this work, we choose to represent 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 convolution neural networks would require certain representation transformation [24, 32, 41] in order to use point clouds. The PointNet architecture [14, 30, 33] overcomes this constraint and supports point clouds as input.

In this work, we explore a different approach to better understand and interpret radar point clouds. Although the point cloud is sparse compared to LiDAR data, radar data contains strong input features in form of velocity and target reflection strength information. We begin leveraging input features by observing that the points have correlations. For example, most pedestrians' velocity components have a variance in the range of 1.69 due to their non-rigid motion but cars have a variance in the range of 0.1 in their velocity components caused by the extended rigid body. To explicitly learn and utilize this information, we learn a direction field vector for each point individually. This helps to learn correlations between points with similar direction field vectors. Thus, low and disperse correlation values indicate a false alarm, effectively rejecting the points during sampling. This helps to suppress static targets which are otherwise often classified into target bins due to multi-path reflections. Furthermore, using this information, a class agnostic Bayesian clustering can be performed over the segmented points.

To this end, we propose an anchor free target detection framework. However, the irregularity of point clouds imposes difficulties when aggregating meaningful information from a given point set, especially when such information is contained by highly unbalanced distant points. Therefore, a point cloud-based hierarchical attention using direction modalities is further proposed. To the best of the author's' knowledge, this approach is novel and has not been investigated in literature before. Our main contributions in this work are summarized as:

- Our proposition of multi-modality-based hierarchical attention for point-based target classification is compared to the traditional approach of using single modalities or dense input feature maps. This minimizes the false detection and handles the class imbalance while improving target accuracy.
- We propose a novel Bayesian localization approach using attention-based clustering. To the best of our knowledge, this is the first attempt to investigate object detection for radar in this representation.
- Development of a target localization pipeline without the need for a prior anchor (candidate) box generation or grid-based regression search algorithm. A 2D bounding box-based object detection algorithm for radar data is modified with a centeredness score to improve average precision for localization. The feasibility of the radar object detection is demonstrated with a mean intersection over union (IoU) of 0.968.

2. Related Work

While radar target detection, in general, has a long history, in the following we only discuss the related work of processing point clouds. A review of equivalent research on target localization and classification, followed by flow estimation and joint attention paradigms, is performed. Also, the literature is discussed and 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 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 [6, 29, 37, 38, 46]. Later, the proposed region undergoes a refinement based on a non-maxima suppression and the objectness score. While these methods demonstrate their success in superior performance in localizing objects and classifying them into the desired classes, the quality of the detection highly depends on the right configurations of anchor boxes. In consequence, this imposes an additional requirement, i.e. manual design optimizations of anchor box size and ratio based on the target sizes.

Anchor-Free: Recent work on 3D-BoNet [45], 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

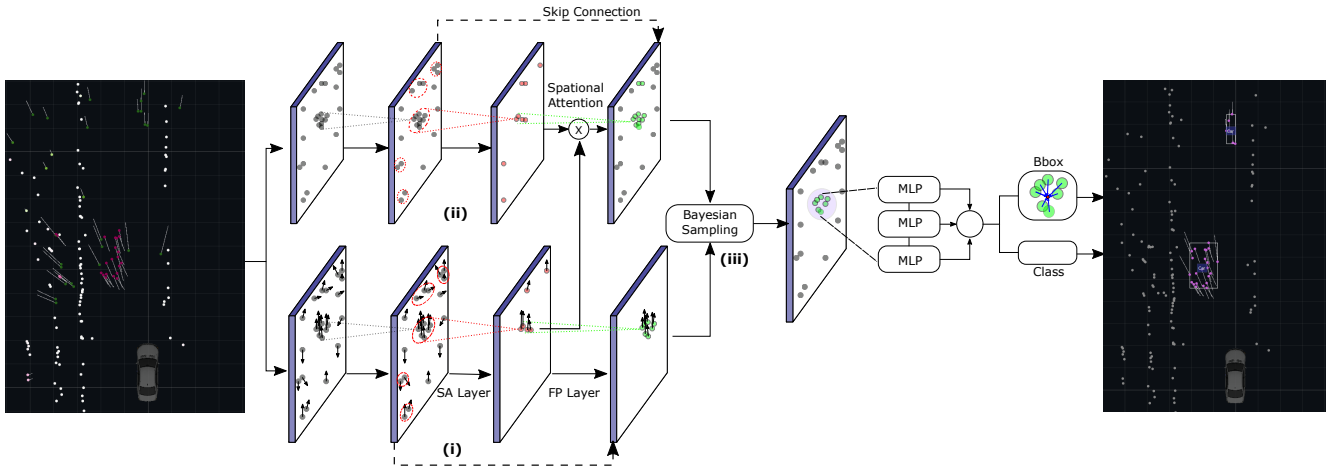

Figure 2. Overview of our methods which is composed of three modules, (i) target's direction vector estimation (dirPointNet), (ii) hierarchical spatial attention inside target segmentation network (segPointNet) using target's direction modality, and (iii) Bayesian sampling for target location using spatial location and direction vector as feature dimension.

et al. [21] introduce corner pooling, a new type of pooling layer that helps the network better localize corners. Similarly, Zhou et al. [49] suggest to detect objects by finding their extreme points.

2.2. Flow Estimator

The scene flow can help in better scene understanding and may provide cues for object segmentation and detection from its motion vectors. Although there are no direct correlations between two sampled point clouds, the authors of [23] 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 [20] 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 with the highly unbalanced data. Different methods of 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 [22,39,44,48]. The most common approaches for attention methods 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 robustness, they also add the additional requirement of multiple sensors.

3. Methodology

The given point cloud is denoted as $\mathbf{P}$ which is a point set containing n points $p_1, p_2, \ldots, p_n \in \mathbb{R}^d$ with d dimensional features. The feature vector of each point p_i consists of the global target coordinates(x_i, y_i), the azimuth angle (θ) and the signal reflection power (σ), since signal power and velocity both are key features for target identification in radar signals. 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. The performance of the network strongly depends on the richness of input features presented to the network [35]. In this case, $\mathbf{P}$ is very sparse with non-uniform distributed points for a target, making rendering reliable detection and classification very challenging.

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 a 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 used to perform a Gaussian sampling for unique cluster identification. These clusters are passed through another network for bounding box estimation. The overall framework uses the PointNet++ [33] architecture adapted for the radar scene data [36]. Targets in the radar point cloud are sparse with non-uniform feature distribution among corresponding data points. To solve this, we propose an end-to-end hierarchical, spatial, attention-based

Attention hierarchy	w/o Attention				1 Layered Attention				2 Layered Attention				3 Layered Attention			
F1-Score / Segmentation	Avg	Ped	Car	Bike	Avg	Ped	Car	Bike	Avg	Ped	Car	Bike	Avg	Ped	Car	Bike
Binary	0.92	-	-	-		-	-	-		-	-	-	0.96	-	-	-
multi-class	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

Table 1. Comparison of the effect of attention at different feature abstraction layer for both binary and multi-class segmentation.

multi-modal segmentation framework. The first network, 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 shared representations and potentially yields a faster convergence.

3.1. Direction Field Attention

Direction Field Estimation Each detection point has a radial velocity and an azimuth component (θ). Taking advantage of both values together with the sensors' yaw angle (ϕ), the direction of motion for each point is estimated as $d(\varphi) = \theta + \phi$. Both velocity and azimuth of the target are compensated with respect to its ego-motion. 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 (downsample-upsample) scheme similar to a general semantic segmentation network [31], except for changing the problem from classification to regression. The network is trained using 4D input feature tensors ($\hat{x}_{\mathrm{cc}}, \hat{y}_{\mathrm{cc}}, \hat{\theta}_{\mathrm{cc}}, \overrightarrow{\hat{v}}_{\mathrm{r}}$) to predict the motion direction vector for each traffic participant ($d(\varphi)$). 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

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

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}).

Furthermore, both input features and labels are rescaled to the range of $[0, 1)$ by applying

$$
\begin{aligned}
\hat{x}_{\mathrm{cc}} &= \frac{x_{\mathrm{cc}} - x_{i_{\mathrm{cell}}}}{s_{x_{\mathrm{cell}}}}, \\
\hat{y}_{\mathrm{cc}} &= \frac{y_{\mathrm{cc}} - y_{k_{\mathrm{cell}}}}{s_{y_{\mathrm{cell}}}}, \\
\overrightarrow{\hat{v}}_{\mathrm{r}} &= \frac{1}{v_{\max}} \vec{v}_{\mathrm{r}}, \\
\hat{\theta} &= \frac{\theta}{60^{\circ}}, \\
\hat{d(\varphi)}_i &= \frac{d(\varphi)_i}{180^{\circ}},
\end{aligned} \tag{2}
$$

with $(x_{i_{\mathrm{cell}}}, y_{k_{\mathrm{cell}}})$ representing the position of the left bottom corner of a cell. The indices (i, k), $(s_{x_{\mathrm{cell}}}, s_{y_{\mathrm{cell}}})$ 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.

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 the pre-trained dirPointNet to provide hard spatial attention inside segPointNet. As dirPointNet and segPointNet share the same number of input tensors, we preserve the flexibility of providing attention at different feature abstraction levels of the network. In contrast to dirPointNet, segPointNet uses ($\hat{x}_{\mathrm{cc}}, \hat{y}_{\mathrm{cc}}, \overrightarrow{\hat{v}}_{\mathrm{r}}, \hat{\sigma}$) as input feature tensor. Due to the difference in input features, the signal from dirPointNet is normalized prior to the spatial attention inside segPointNet. This also helps to stabilize the gradient during training.

Both networks, 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

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

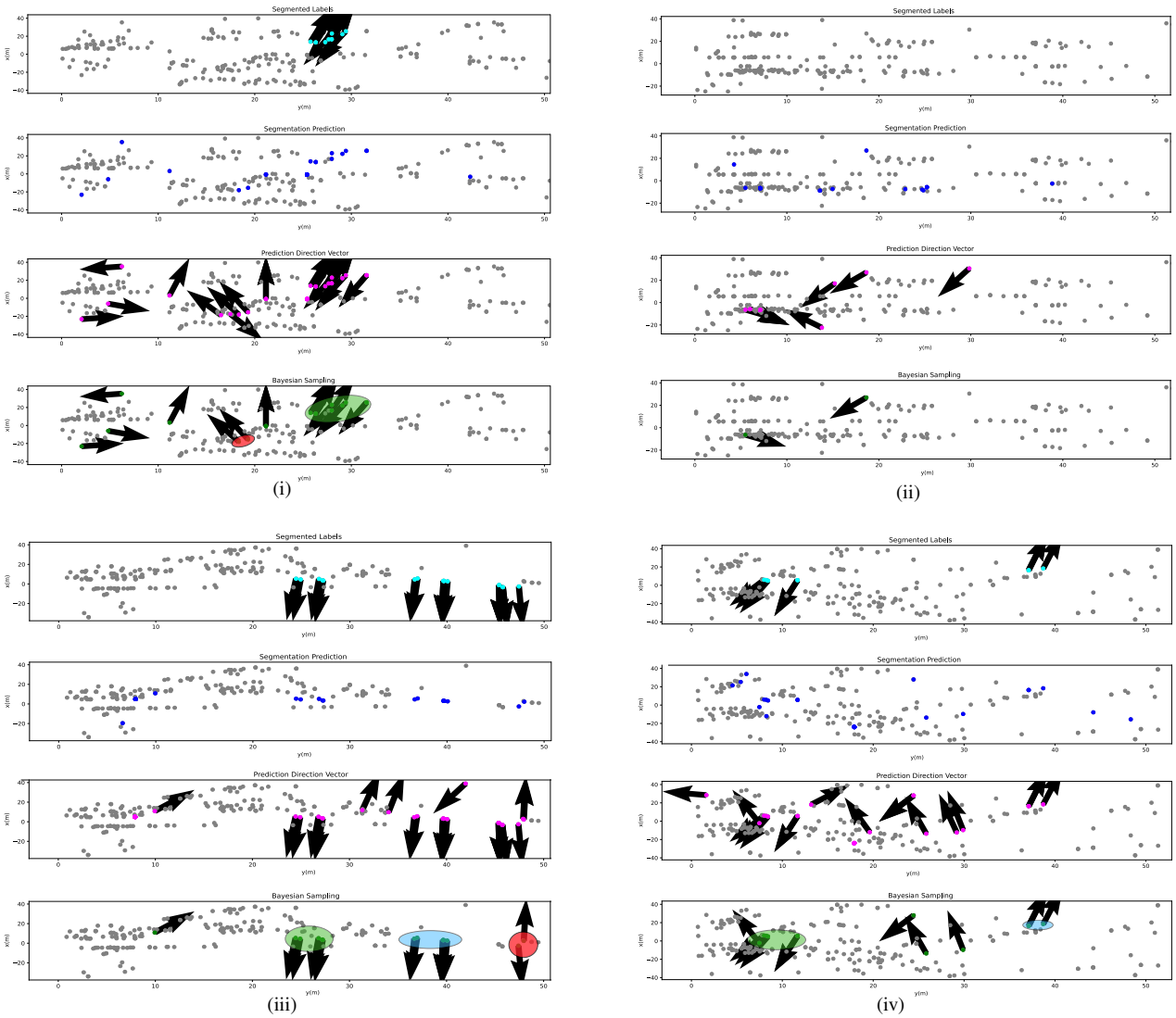

Figure 3. Visualization of the different intermediate outputs of the proposed three stage approach. The first row in each of (i-iv) represents the spatial input with the corresponding label and its associated direction field. The last three rows of (i-iv) show the predicted segmentation, estimated direction vector and Bayesian sampling for the region of interest estimation.

where, L_{cls} denotes the loss for point classification from the scene and $L_{direction}$ is for direction field estimation of classified radar targets. In Eq. (3), $y \in \{0, \ldots, K-1\}$ represents an integer class label, $\hat{\mathbf{p}} = (\hat{p}_0, \ldots, \hat{p}_{K-1}) \in [0,1]^K$ 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 attention on segPointNet by evaluating the network performance using 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 compared for binary and multicategory segmentation tasks. The binary segmentation network demonstrates a better average F1-score of 0.92 in contrast to the multi-class segmentation 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, multi-class segmentation 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 segmentation 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 [36]. This proves the advantage of

attention inside a network with different modalities which acts as an additional target feature point and improves scene recognition.

3.2. 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. In order to deal with such situations, our attention direction network can be guided not only towards the more relevant features but 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. Thereafter, the output is clustered into desired unique bins using a Gaussian mixture model operated at three scaled feature dimensions, i.e. $\hat{x}_{\mathrm{cc}}$, $\hat{y}_{\mathrm{cc}}$ and $d(\hat{\theta})_{\mathrm{cc}}$. To the best of the authors' knowledge, this approach is novel and has not been investigated in literature before.

While Table 1 demonstrates the quantitative advantage of direction vector attention, Fig. 3 shows the proposed Bayesian sampling for region extraction. To have a deeper insight on the advantages of the approach, multiple examples together with their behaviour on boundary conditions are demonstrated using 4 subplots for each example. The top row shows the input point cloud marked with the ground truth label and direction vector. The middle two rows show the output of segPointNet and dirPointNet. The last row follows the output of the Bayesian sampling layer. Fig. 3(i) show 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 high variability of direction vectors between neighboring points. As a result, the network successfully finds regions of interest. Fig. Fig. 3(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 $d(\hat{\theta})_{\mathrm{cc}}$ feature in comparison to the distributed and non-uniform spatial feature dimension ($\hat{x}_{\mathrm{cc}}$, $\hat{y}_{\mathrm{cc}}$). Furthermore, Fig. 3(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.

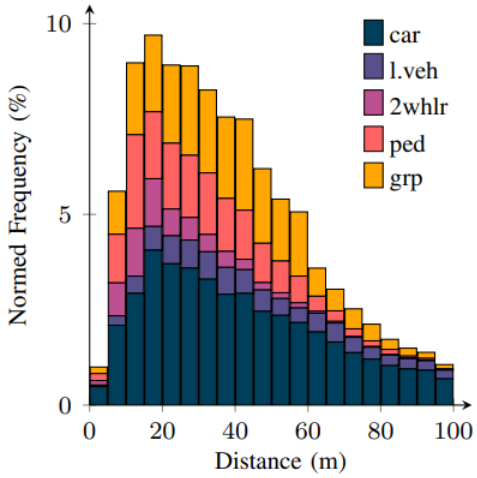

Figure 4. Distribution of objects for varying distances [36].

4. Experiments

The evaluation of the proposed framework is done on real-world data that were collected by 4 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.

car	ped.	ped. group	bike	truck	static
1.23%	0.31%	0.74%	0.11%	0.60%	97.01%

Table 2. Distribution of radar reflections among the six classes [36].

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, Fig. 4 depicts the distribution of object availability with respect to the distances to the ego-vehicle. Following object distribution over distance, we process cropped scene within a range of $80\,\mathrm{m}$ for $\hat{x}_{\mathrm{cc}}$ and an absolute range of $20\,\mathrm{m}$ for $\hat{y}_{\mathrm{cc}}$. This reduces the effect of static targets during network training.

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 cluster, with the dimension of $4 \times m$, is passed through a box regression PointNet, conceptually similar to [31]. The regression network predicts parameters of a 2D bounding box (Bbox), i.e. its center (x_c, y_c) and its size (l, w). For the box center estimation, a residual-based 2D localization is performed, similar to [28], 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 amodal 2D bounding box estimation, up to 32 points are randomly sampled from the point clusters for every radar target. The labelling 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 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 training. The, multitask loss is defined as

$$\begin{aligned} L_{multi-task} = & w_{cls} L_{cls} + (1 - w_{cls}) * L_{direction} + \\ & w_{\text{Bbox}} \left(L_{c_x,reg} + L_{c_y,reg} + \right. \\ & \left. L_{s,cls} + L_{h,reg} + L_{w,reg} \right). \end{aligned} \tag{5}$$

Both losses are weighted using the parameter w_{cls}. During the training, the weights for a target with $L_{direction}$ and L_{cls} are handled using Eq. (1) and Eq. (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.

Due to the hybrid nature of the network and its loss function, we empirically select a set of training hyperparameters, summarized in Table 4, which could further be optimized using Bayesian hyperparameter search.

Results

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, as described in Eq. (4), and the Intersection over Union (IoU) metric. The IoU compares a predicted bounding box b_{pred} with the ground truth bounding box b_{gt} and is defined as

$$\text{IoU} = \frac{|b_{gt} \cap b_{pred}|}{|b_{gt} \cup b_{pred}|}, \tag{6}$$

where $|\cdot|$ 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.

hyper parameters	point classification	direction field vector	attention classification	Bbox Estimation
Epochs	20	20	10	40
Optimizer	Adam	Adam	Adam	Adam
Batch size	8	8	8	32
Learning rate	0.0005	0.0005	0.0001	0.0001
Learning rate decay	cosine	exp	cosine	step
Weight decay	0.5	0.5		0.7
Dropout	0.5	0.5	0.5	0.25

Table 3. Training strategy and hyper-parameter settings.

The network is evaluated on the full test data set to cover the maximum number of different situations, with also the corner cases and to understand the behaviour of the networks for exemplary targets i.e. pedestrians and bikes. Additionally, the performance of the multihead box network is evaluated under three different conditions. First, only the centeredness of the input target point cluster is measured, second, both centeredness and corner points are estimated. And lastly, together with centeredness and corner points, the target point cluster is classified into the desired target class. The detailed performance is summarized in Table 4.

Segmentation \ Metric	Bbox Estimation	mIoU	Class Accuracy
binary	centeredness	**0.96**	-
multi-Class	centeredness	0.92	
multi-Class	class-agnostic Bbox	0.86	-
multi-Class	Bbox + cls	0.93	0.78

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

The network shows the maximum localization accuracy with a mean IoU of 0.96 for the binary segmented output and 0.92 for the multi-class segmented output, when only the centroid of the point cluster is estimated. The estimated centroid is used in post-processing to filter outlier input points using Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Later, all corner points are used to estimate the bounding box which is much tighter to the original point distribution. When both centroid and corner points are estimated without knowing the target class (point distribution), 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 bad localization ac-

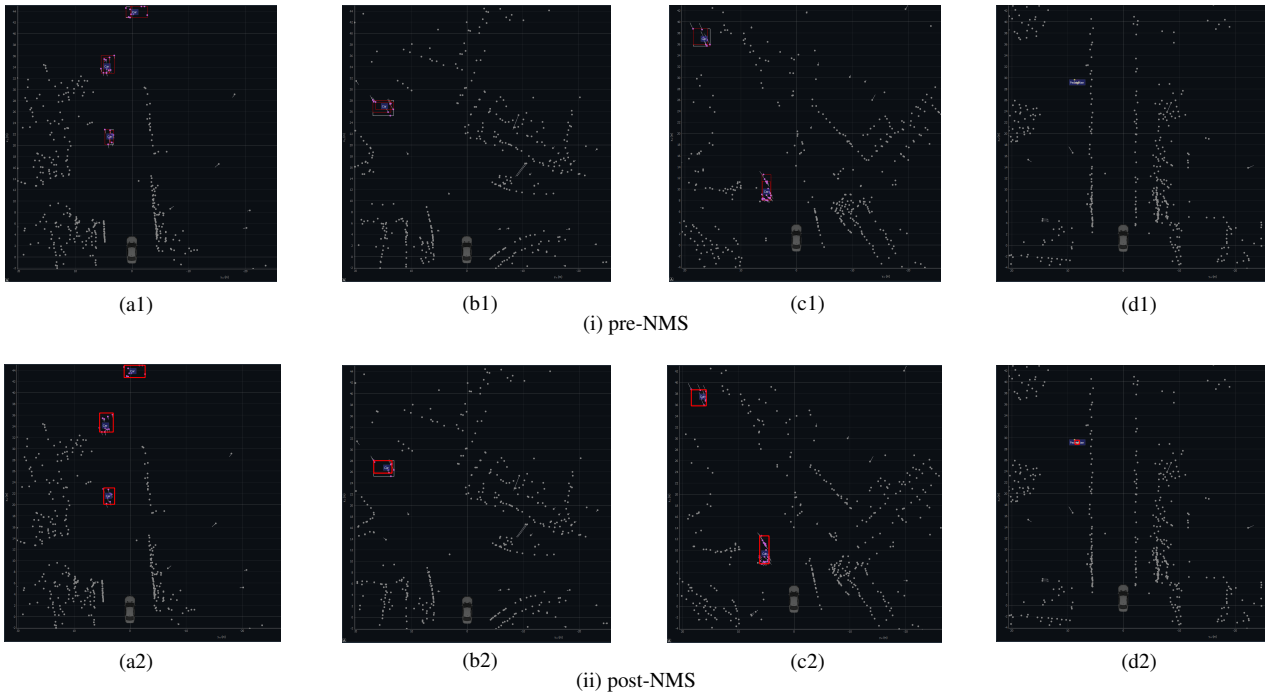

Figure 5. Combined illustration on the performance of the proposed framework. (i) The figure shows the class-aware bounding box estimation, and (ii) gives final prediction after using NMS as post-processing

curacy 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 classification accuracy for the remaining methods are omitted as the target class information for the multi-class segmentation network remains the same as its segmentation accuracy, already described in Table 1.

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 the proposed framework for successful localization of a target with just 4 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 target localization successfully, the estimated bounding box leaves out some target points treating as an outlier. As a result, network contributes to false-negative during target localization. This is caused by loss function (L_box) which do not penalize loss caused by background and foreground separately. As an alternative, in furture loss function similar to proposed in [16], can be used for better localization.

5. Conclusion

The work studies the problem of the target detection and classification on radar point clouds. Our model is built upon hierarchical attention-based anchor-free target localization and classification. To further capture and increase the target recognition accuracy, the concept of direction field vectors for each target point is introduced as a multi-modality inside the attention. Extensive experiments validate the efficacy of this approach across a wide range of examples.

Our framework generates 2D bounding boxes at the cost of two limitations. Firstly, the network uses a hard attention which limits the joint representation learning of the target's motion distribution in contrast to spatial distribution. Recent work [22,39,44,48] have demonstrated better attention mechanisms, where similar techniques are also applicable here. Secondly, the localization of our model uses Bayesian sampling-based clustering of target points along spatial and motion dimensions. Due to non-uniformity in the spatial dimension, it strongly relies mostly on a single dimension resulting in false negatives. A multi-modality and anchor-free target detection on radar point clouds is a very impor-

tant topic for future research.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Jürgen Gall, and Cyrill Stachniss. Towards 3d lidar-based semantic scene understanding of 3d point cloud sequences: The semantickitti dataset. *The International Journal of Robotics Research*, 40(8-9):959–967, 2021. 1

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Jürgen Gall, and Cyrill Stachniss. Towards 3d lidar-based semantic scene understanding of 3d point cloud sequences: The semantickitti dataset. *The International Journal of Robotics Research*, 40(8-9):959–967, 2021. 1

[3] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner. Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine*, 6(4):6–22, 2014. 1

[4] François Bremond. Scene Understanding: perception, multisensor fusion, spatio-temporal reasoning and activity recognition. July 2007. 1

[5] V.C. Chen. *The Micro-Doppler Effect in Radar*. Artech House radar library. Artech House, 2019. 1

[6] Andreas Danzer, Thomas Griebel, Martin Bach, and Klaus Dietmayer. 2d car detection in radar data with pointnets. *CoRR*, abs/1904.08414, 2019. 2

[7] J. Dickmann, N. Appenrodt, H. Bloecher, C. Brenk, T. Hackbarth, M. Hahn, J. Klappstein, M. Muntzinger, and A. Sailer. Radar contribution to highly automated driving. In *2014 11th European Radar Conference*, pages 412–415, 2014. 1

[8] A. Dubey, J. Fuchs, M. Luebke, R. Weigel, and F. Lurz. Generative adversial network based extended target detection for automotive mimo radar. In *2020 IEEE International Radar Conference (RADAR)*, pages 220–225, 2020. 1

[9] A. Dubey, J. Fuchs, M. Luebke, R. Weigel, and F. Lurz. Region based single-stage interference mitigation and target detection. In *IEEE Radar Conference 2020*, 2020. 1

[10] A Dubey, A Santra, J Fuchs, M Luebke, R Weigel, and F Lurz. Integrated classification and localization of targets using bayesian framework in automotive radars. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 6-11 June 2021, Toronto, Canada [accepted]. 1

[11] Anand Dubey, Avik Santra, Jonas Fuchs, Maximilian Lübke, Robert Weigel, and Fabian Lurz. A bayesian framework for integrated deep metric learning and tracking of vulnerable road users using automotive radars. *IEEE Access*, 9:68758–68777, 2021. 1

[12] F. Foelster and H. Rohling. Signal processing structure for automotive radar. *Frequenz*, 60:20 – 24, 2006. 1

[13] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6, 2020. 1

[14] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *CoRR*, abs/1912.12033, 2019. 2

[15] G. Hakobyan and B. Yang. High-performance automotive radar: A review of signal processing algorithms and modulation schemes. *IEEE Signal Processing Magazine*, 36(5):32–44, 2019. 1

[16] David Hall, Feras Dayoub, John Skinner, Peter Corke, Gustavo Carneiro, and Niko Sünderhauf. Probability-based detection quality (PDQ): A probabilistic approach to detection evaluation. *CoRR*, abs/1811.10800, 2018. 8

[17] Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Perspectivenet: 3d object detection from a single RGB image via perspective points. *CoRR*, abs/1912.07744, 2019. 1

[18] Zhiyu Huang, Chen Lv, Yang Xing, and Jingda Wu. Multimodal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. *CoRR*, abs/2005.09202, 2020. 1

[19] Hamid Izadinia, Qi Shan, and Steven M. Seitz. IM2CAD. *CoRR*, abs/1608.05137, 2016. 1

[20] Mingyang Jiang, Yiran Wu, and Cewu Lu. Pointsift: A sift-like network module for 3d point cloud semantic segmentation. *CoRR*, abs/1807.00652, 2018. 3

[21] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *CoRR*, abs/1808.01244, 2018. 3

[22] Xi-An Li, Lei Zhang, Li-Yan Wang, and Jian Lu. Multi-scale receptive fields graph attention network for point cloud classification. *CoRR*, abs/2009.13289, 2020. 3, 8

[23] Xingyu Liu, Charles Ruizhongtai Qi, and Leonidas J. Guibas. Learning scene flow in 3d point clouds. *CoRR*, abs/1806.01411, 2018. 3

[24] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, 2015. 2

[25] H. H. Meinel. Evolving automotive radar — from the very beginnings into the future. In *The 8th European Conference on Antennas and Propagation (EuCAP 2014)*, pages 3107–3114, 2014. 1

[26] M. Murad, I. Bilik, M. Friesen, J. Nickolaou, J. Salinger, K. Geary, and J. S. Colburn. Requirements for next generation automotive radars. In *2013 IEEE Radar Conference (RadarCon13)*, pages 1–6, 2013. 1

[27] S. Patole, M. Torlak, D. Wang, and M. Ali. Automotive radars: A review of signal processing techniques. *IEEE Signal Processing Magazine*, 34:22–35, 2017. 1

[28] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. *CoRR*, abs/1904.09664, 2019. 7

[29] Charles Ruizhongtai Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from RGB-D data. *CoRR*, abs/1711.08488, 2017. 2

[30] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 2

[31] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 7

[32] Charles Ruizhongtai Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. *CoRR*, abs/1604.03265, 2016. 2

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017. 2, 3

[34] A. Santra and S. Hazra. *Deep Learning Applications of Short Range Radars*. Artech House, 2020. 1

[35] Ole Schumann, Markus Hahn, Jürgen Dickmann, and Christian Wöhler. Semantic segmentation on radar point clouds. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 2179–2186, 2018. 3

[36] Ole Schumann, Markus Hahn, Nicolas Scheiner, Fabio Weishaupt, Julius Tilly, Jürgen Dickmann, and Christian Wöhler. RadarScenes: A Real-World Radar Point Cloud Data Set for Automotive Applications. Mar. 2021. 3, 5, 6

[37] Shaoshuai Shi, Zhe Wang, Xiaogang Wang, and Hongsheng Li. Part-a^2 net: 3d part-aware and aggregation neural network for object detection from point cloud. *CoRR*, abs/1907.03670, 2019. 2

[38] Martin Simon, Stefan Milz, Karl Amende, and Horst-Michael Gross. Complex-yolo: Real-time 3d object detection on point clouds. *CoRR*, abs/1803.06199, 2018. 2

[39] Xiang Song, Weiqin Zhan, Xiaoyu Che, H. Jiang, and Biao Yang. Scale-aware attention-based pillarsnet (sapn) based 3d object detection for point cloud. *Mathematical Problems in Engineering*, 2020:1–12, 2020. 3, 8

[40] M. Stolz, E. Schubert, F. Meinl, M. Kunert, and W. Menzel. Multi-target reflection point model of cyclists for automotive radar. In *2017 European Radar Conference (EURAD)*, pages 94–97, 2017. 1

[41] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. *CoRR*, abs/1505.00880, 2015. 2

[42] C. Waldschmidt and H. Meinel. Future trends and directions in radar concerning the application for autonomous driving. In *2014 11th European Radar Conference*, pages 416–419, 2014. 1

[43] C. Will, P. Vaishnav, A. Chakraborty, and A. Santra. Human target detection, tracking, and classification using 24-ghz FMCW radar. 19:7283–7299, 2019. 1

[44] Hang Wu and Yubin Miao. Lra-net: local region attention network for 3d point cloud completion. In *International Conference on Machine Vision*, 2021. 3, 8

[45] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *CoRR*, abs/1906.01140, 2019. 2

[46] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. *CoRR*, abs/2002.10187, 2020. 2

[47] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. *CoRR*, abs/2103.06422, 2021. 1

[48] Xin Zhao, Zhe Liu, Ruolan Hu, and Kaiqi Huang. 3d object detection using scale invariant and feature reweighting networks. *CoRR*, abs/1901.02237, 2019. 3, 8

[49] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. Bottom-up object detection by grouping extreme and center points. *CoRR*, abs/1901.08043, 2019. 3

[50] Chuhang Zou, Zhizhong Li, and Derek Hoiem. Complete 3d scene parsing from single RGBD image. *CoRR*, abs/1710.09490, 2017. 1