Weakly Supervised Semantic Segmentation for Range-Doppler Maps

Konstantinos Fatseas[#1], Marco J.G. Bekooij[*#2]

[#]Department of Computer Architectures for Embedded Systems, University of Twente, The Netherlands
[*]Department of Embedded Software and Signal Processing, NXP Semiconductors, The Netherlands
[1]k.fatseas@utwente.nl, [2]marco.bekooij@nxp.com

***Abstract* — Deep convolutional neural networks (DCNNs) have been successfully applied for object detection and semantic segmentation of radar range-Doppler (RD) maps. However, training a DCNN requires many annotated examples that are costly and difficult to create. In this work we present a method that reduces significantly the manual effort involved in the annotation of RD maps to train a DCNN for segmentation. A 40 times reduction in manual labelling effort is achieved because the annotation of each RD map includes only the class of the objects instead of drawing a polygon around the corresponding cells. The localization of the objects is performed by tracing back from the output to the input of a classification neural network. Experimental results show that our approach achieves robust localization performance in complex real-world urban scenarios as observed with a low-cost automotive radar. Furthermore, we show that our approach performs similarly to DCNNs that are trained with a publicly available dataset in which localization information is provided.**

***Keywords* — semantic segmentation, weak supervision, range-doppler maps.**

I. Introduction

Neural networks eliminate the use of heuristics and offer robust performance when processing radar data in the form of range-doppler maps. Each cell in these maps corresponds to a certain range and velocity interval. A moving object at a particular range and velocity results in a peak at the corresponding cell of the RD map. Due to the high resolution of recent radar sensors, one object often results in several detections. Therefore, more than one cell of the RD map can be related to the same object.

A key obstacle for applying neural networks in the radar domain is that they require a large amount of annotated data to train them. Adding labels such as bounding boxes or segmentation masks to RD maps, is a tedious and costly manual task. Furthermore, the maps can be hard to interpret due to the presence of clutter and ghost targets.

In this paper, we propose a method that automatically localizes objects in a radar training dataset with RD maps. Our approach requires that only the presence of objects in an RD map is manually indicated. The localization involves finding the cells that correspond to an object. Additionally, their class is also derived. These steps result to a segmented RD map.

Our approach relies on weak supervision because it localizes objects by making use of a DCNN for classification of objects as shown in Fig. 1. This network is trained using manually provided input about which type of objects are present in each RD map of a training data set. We then trace-back from the outputs of this DCNN to its inputs to identify pixels that correspond to the moving objects. The result of this tracing back are the so-called saliency maps.

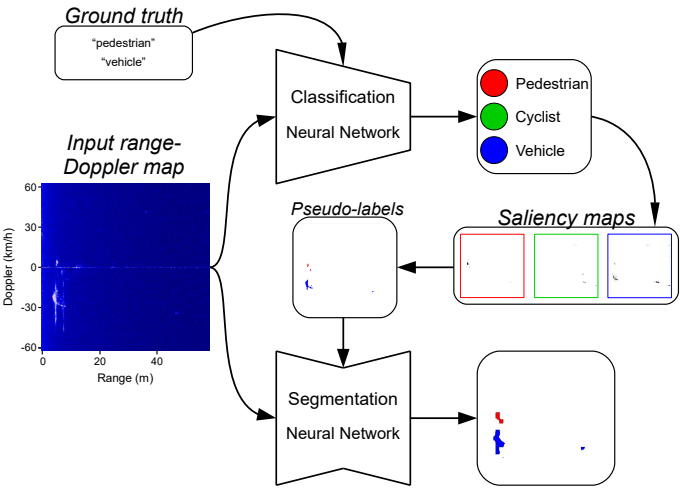

Fig. 1. Overview of the proposed neural network training procedure.

In our work we utilize the guided Grad-CAM algorithm [1] to generate a saliency map for each object class. These saliency maps are then processed and used to train a second DCNN that performs segmentation using the same training set but this time with the generated annotation. In the rest of the paper, we refer to the generated set of annotations as pseudo-labels.

The reason we create pseudo-labels instead of directly using our method for new data is the computational burden that is involved in generating saliency maps. This is because to compute the saliency maps, a forward and a backward pass of the neural network is required. In contrast, the segmentation DCNN requires only a forward pass and no further processing of its output.

By comparing our non-optimized pseudo-label generation method to a single pass of the best performing segmentation DCNN on our workstation, we found the latter to be more than X100 faster with a run-time of 3ms. Therefore, using a dedicated DCNN for segmentation is the only feasible solution for real-time automotive applications.

When compared to commonly used algorithms such as CFAR, our method has two additional beneficial properties.

One property is that it filters out unimportant detections and ghost targets. This is because the classification DCNN relies only on the important features. Thus, in the saliency map, clutter and ghost targets are suppressed as these do not contribute to the output of the classifier. Hence, this DCNN acts as a context aware filter.

The other benefit is that we use the RD maps directly as input of the DCNNs. Therefore, information in RD cells that is below the (CFAR) detection threshold is taken into account and classified, which improves the sensitivity of the radar system.

We evaluate the performance of our method using radar data obtained with a low-cost automotive radar in real-world complex urban scenarios. Additionally, we compare our results with segmentation neural networks that were trained with a publicly available dataset recorded in a controlled environment.

As performance metric we use the so-called intersection over union (IoU) metric in order to perform a direct comparison with the segmentation DCNNs trained with the Carrada dataset [2]. This metric is the percentage of correctly classified pixels over the total amount of pixels per class. Our results show that our segmentation DCNN delivers robust performance that is not affected by clutter and is comparable to fully supervised DCNNs.

II. Related Work

In this section we discuss studies related to object detection and segmentation of RD maps. We divide the related work in fully and weakly supervised neural network based methods. Additionally, we briefly investigate other cases of radar data segmentation.

A. Fully Supervised

Perez et al. [3] trained a DCNN to classify RD maps that contained a single object of the following classes: pedestrian, cyclist, and vehicle. By doing so, they were able to instantly classify targets without delay, but the targets are not localized so no further information can be derived. A method that can detect and track multiple objects in RD maps, was introduced in [4]. To detect multiple objects the authors utilized a single shot DCNN detector followed by a tracking algorithm that uses Kalman filters to track detected objects.

B. Weakly Supervised

Regarding weak supervision, authors in [5], [6] automatically produce the ground truth labels by processing data from a synchronized lidar sensor. They also train their networks under weak supervision as the labels are noisy and sometimes inaccurate.

However, a difference with our work is that we do not create the segmentation pseudo labels based on another sensor that can project its detections to a common plane with the radar. In contrast, we use abstract information about object classes to train a classification DCNN and subsequently exploit its knowledge to generate fine grained localization information.

C. Semantic Segmentation

There are several studies in which neural networks have been trained for semantic segmentation on radar data. For example, in [7] authors perform semantic segmentation on point clouds that include reflections from static and non-static objects. Prophet et al. [8] performed an evaluation of several commonly used neural networks at the task of static object point semantic segmentation. In both studies the segmentation is performed on manually labelled data that includes the azimuth direction of arrival information for each object.

More recently, the Carrada dataset [2] was made available. The authors provide range-Doppler and range-angle maps together with fine-grained labels that have been generated in a semi-automatic way. Additionally, they also trained a segmentation network as a baseline for comparison. In section IV we include results of evaluating our method with the RD maps from Carrada dataset.

To the best of our knowledge, our method is the first weakly supervised radar data segmentation method that makes use of labels generated with saliency maps.

III. Methodology

In our work we make use of the guided Grad CAM algorithm [1], which combines the gradient of the neural network and the CAM method [9] to generate saliency maps for a given input of the DCNN.

This section describes how we train a neural network to perform classification of RD maps and how we exploit its knowledge by utilizing the guided Grad CAM algorithm to generate pseudo-labels for segmentation. We also discuss the usage of the generated labels to train DCNNs to directly perform segmentation of RD maps.

A. Range-Doppler Map Multi-Label Classification

The classification neural network (Fig. 1) is trained to perform multi-label classification of range-Doppler maps. The maps may contain multiple objects that belong to different classes. Therefore, more than one neuron of the output layer can be activated simultaneously. For RD maps that do not contain any object, all neurons should remain deactivated.

Our experiments indicated that classification accuracy is improved by feeding the classifier with a series of consecutive range-Doppler maps. In doing so, the DCNN is able to exploit the temporal behaviour of each object to better predict its class. Neural networks with spatio-temporal convolutions have a significant advantage over common 2D DCNNs on our dataset. Therefore, we make use of a 3D Resnet18 [10] which we fine-tune to our training dataset. The DCNN was pre-trained for action recognition in video.

The best classification accuracy we achieved on the test dataset is 82.4%. This accuracy was reached after training the classification DCNN for 37 epochs. The learning rate was 0.01 and the batch size was 64. We utilized the Adam optimizer with its default parameters to minimize the binary cross-entropy loss.

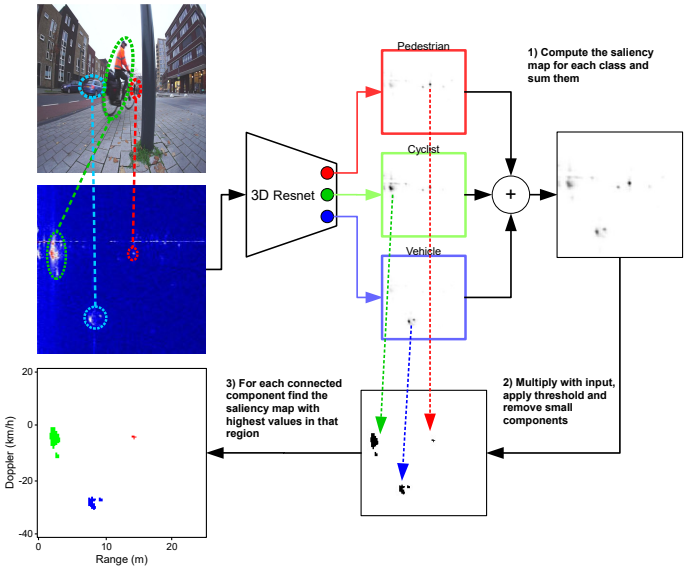

Fig. 2. Overview of the pseudo-label generation. We show only a part of the range-Doppler map for clarity. The resulting segmentation mask is depicted at the lower left corner. Note that the camera image is only used as a reference and is not used in our method.

B. Pseudo-Label Generation

The DCNN which has been trained for classification is subsequently used to generate saliency maps for each RD map of the training dataset as seen in Fig. 2. The RD maps are fed through the neural network and a saliency map for each class is computed. The maps are then summed in order to create a class agnostic binary map with all valid detections. This is done by thresholding and clustering the sum. Finally, a class is assigned to each connected component of the binary map by finding the saliency map with the largest sum of values for that region.

It can be seen in Fig. 2 that static detections and clutter are not present in the created segmentation mask. It contains only the detections that correspond to the three moving objects. In contrast, the commonly used CFAR algorithm is not able to remove ghost detections or clutter and further processing of the filtered RD map is required.

C. Semantic Segmentation

The algorithm we presented to generate pseudo-labels, is also possible to be used for computation of segmentation masks of the test data set as well, with very good results (Table 1). However, our method is significantly slower than a single forward computation of a DCNN. Therefore, we opted for utilizing the pseudo-labels to train DCNNs that can directly predict the segmentation of a RD map.

We tried two different segmentation DCNNs, a slightly modified UNet [11] and a pre-trained DeeplabV3 [12]. The only modification on the UNet was that the number of convolutional filters per layer was halved. The resulting DCNN has 7.8 million parameters while the DeeplabV3 network consists of 42 million coefficients.

Fig. 3. Images captured with the camera of the setup in four of the locations that we collected our radar and video data.

The models were trained to extract segmentation masks for each of the following classes: background, pedestrian, cyclist and vehicle. The background class represents all pixels that do not contain an object but also the static objects and clutter. We used the Adam optimizer with default parameters to minimize the sum of the IoU and the categorical cross-entropy losses. The learning rate was 0.0001 and the batch size was 16. Finally, in a same fashion as with the classification network, we allow the DCNN to observe the past by stacking the 4 most recent RD maps together with the one which is being analysed.

IV. Results and Discussion

A. Dataset

Due to the lack of a publicly available datasets with high resolution range-Doppler maps of real world scenarios, we performed measurements in urban areas with mixed traffic and pedestrians (Fig. 3). The setup used to perform our measurements consist of the TEF810X 77 GHz low-cost automotive radar transceiver and a camera with a wide-angle lens. During the recordings we acquired synchronized image and radar data with a rate of 5Hz. Our setup remained stationary throughout the recordings.

Our dataset consists of 14385 RD maps from 14 recordings with a total duration of 48 minutes. We use 12 recordings for training and the remaining 2 recordings which contain 27% of the RD maps for testing. The label for each range-Doppler map is the class of the moving objects that are present within the radar's range.

Initially, the annotation of the presence of objects from a particular class is automatically generated by detecting objects in the synchronized video with an off-the-self object detector for images. This stage was followed by manual inspection and a few corrections of the automatically created labels. We also manually annotated with segmentation masks the testing dataset such that we could evaluate the semantic segmentation performance of the DCNN that was trained using automatically generated localization information. Manually annotating the RD maps to train the classification DCNN was $\times$ 40 faster than annotating the RD maps of the test dataset with polygons, which are required to create segmentation masks.

B. Results

In this section we provide qualitative and quantitative results from the evaluation of our method. In Table 1 we present IoU score that the segmentation DCNNs achieve on our test dataset. Additionally, we also evaluate the segmentation masks that we generated for the test data.

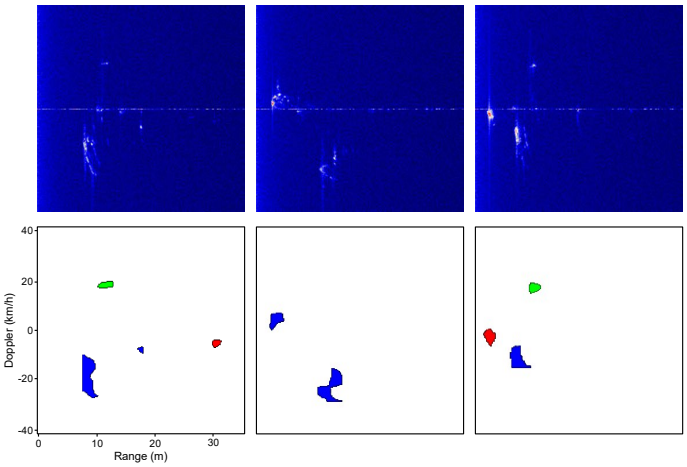

Fig. 4. Segmentation masks generated with the DeeplabV3 DCNN (bottom row) from selected RD maps (top row). Blue represents vehicles, red is for pedestrians and green for cyclists.

Interestingly, the best score for all classes is achieved by the segmentation masks that we generate with our GradCAM based method. The UNet performs worse for all classes and its performance is not improved even if we increase the number of its filters. In contrast, the DeeplabV3 is much closer to the guided GradCAM based method and the main reason for its worse performance is that it generates masks that cover an area larger than the objects. This mainly affects smaller objects such as the pedestrians. Nevertheless, it can be seen on the left column of Fig. 4 that by processing raw data the segmentation DCNN is very sensitive and can correctly detect a pedestrian which is walking 30 meters away from the radar sensor.

Besides the IoU metric, we also evaluate the DeeplabV3 in terms of object detection. We do so by detecting connected components in the manually annotated segmentation masks of the test dataset and the predicted masks from the DCNN. A detection is valid when the connected components have an IoU greater than 10%. Our results show that DeeplabV3 detects 89.1% of the moving objects while it correctly classifies 80.4% of the detections.

To evaluate our method against fully supervised DCNNs, we used our method to generate pseudo-labels for the the RD maps of the Carrada dataset. The obtained results after training DeeplabV3 with our segmentation pseudo-labels can be seen in Table 1. The IoU scores it achieves are comparable to the results of the networks that the authors trained.

V. Conclusion

In this work we present a method based on weak supervision to significantly reduce ($\times$ 40) the manual effort for annotation of RD maps to train segmentation DCNNs. The manual annotation effort is limited to indicating which object types are present in the RD maps of the training dataset because localization information is automatically derived.

We show in our evaluation that the proposed method generates pseudo-labels that allow the segmentation network to attain performance comparable to networks trained in a fully supervised manner. More specifically, with our method we created pseudo-labels for the Carrada dataset and achieved an IoU score of 44.7% by training the DeeplabV3 model. Furthermore, our results show that the performance is not compromised even in complex real-world urban scenarios that produce cluttered RD maps.

We expect that the proposed method can be generalized such that information provided by sensors, such as cameras, that measure other modalities than range and velocity, can be used to train DCNNs for segmentation of RD maps. Resulting to a fully automated data processing pipeline that can classify and localize objects in RD maps for the creation of a training dataset without any human intervention.

Table 1. Comparison of Different Segmentation DCNNs on Our and the Carrada Dataset

Dataset	Model	IoU (%)				
		Background	Pedestrian	Cyclist	Vehicle	mean IoU
Our	Resnet3D + GradCAM*	99.9	23.9	29.1	28.2	45.3
	DeeplabV3	99.9	15.6	24.3	27.4	41.8
	UNet	99.9	13.8	20.8	22.5	39.2
Carrada	FCN-8s [2]	99.7	45.2	15.5	51.3	52.9
	DeeplabV3	99.4	26.6	10.3	42.7	44.7
	FCN-16s [2]	99.6	28.9	7.2	42.1	44.5

* This is the classification model from subsection III-A, we create the segmentation masks as illustrated in Fig. 2.

References

[1] R. R. Selvaraju, M. Cogswell *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 ICCV*, 2017, pp. 618–626.

[2] A. Ouaknine, A. Newson *et al.*, "Carrada dataset: Camera and automotive radar with range-angle-doppler annotations," 2020.

[3] R. Pérez, F. Schubert *et al.*, "Single-frame vulnerable road users classification with a 77 ghz fmcw radar sensor and a convolutional neural network," in *2018 IRS*, 2018, pp. 1–10.

[4] K. Fatseas and M. J. G. Bekooij, "Neural network based multiple object tracking for automotive fmcw radar," in *2019 International Radar Conference (RADAR)*, 2019, pp. 1–5.

[5] M. Dimitrievski, I. Shopovska *et al.*, "Weakly supervised deep learning method for vulnerable road user detection in fmcw radar," in *2020 ITSC*, 2020, pp. 1–8.

[6] P. Kaul, D. de Martini *et al.*, "Rss-net: Weakly-supervised multi-class semantic segmentation with fmcw radar," in *2020 IV*, 2020, pp. 431–436.

[7] O. Schumann, M. Hahn *et al.*, "Semantic segmentation on radar point clouds," in *2018 FUSION*, 2018, pp. 2179–2186.

[8] R. Prophet, G. Li *et al.*, "Semantic segmentation on automotive radar maps," in *2019 IV*, 2019, pp. 756–763.

[9] B. Zhou, A. Khosla *et al.*, "Learning deep features for discriminative localization," in *2016 CVPR*, 2016, pp. 2921–2929.

[10] D. Tran, H. Wang *et al.*, "A closer look at spatiotemporal convolutions for action recognition," in *2018 CVPR*, 2018, pp. 6450–6459.

[11] O. Ronneberger, P. Fischer *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI 2015*, 2015, pp. 234–241.

[12] L.-C. Chen, G. Papandreou *et al.*, "Rethinking atrous convolution for semantic image segmentation," *ArXiv*, vol. abs/1706.05587, 2017.