Telecom-Paris

IMA206

MULTI-VIEW RADAR SEMANTIC SEGMENTATION FOR SCENE UNDERSTANDING IN AUTONOMOUS DRIVING

Realized By:

Eya GHAMGUI

Siwar MHADHBI

Saifeddine BARKIA

Taher ROMDHANE

2020 - 2021

Plan

1. Introduction

With the rapid growth of autonomous vehicles, more and more requirements for environmental perception are being required. Cameras, lidars and radars are now used combined in this field to ensure the safety of the passengers on the roads. The redundancy of the information coming from those three types of sensors is crucial and important to secure this task. Recently, because of its inexpensive cost, adaptability in diverse climates (they have the distinct advantage of being unaffected by rain, snow, or fog.), and motion detection capability, radars are one of the most often utilized sensors. The radar can give a variety of data kinds to meet the needs of different levels of autonomous driving. As a common and necessary perceptive sensor on automated vehicles, it enables long measuring distance range, low cost, dynamic target detection capacity, and environmental adaptability, which enhances the overall stability, security, and reliability of the vehicle. Long Range Radar (LRR) can detect targets within the range of 250 m. This is extremely important for safe automobile driving. It can also use the Doppler effect to assess the relative velocity of targets (resolution up to 0.1 m/s), which is critical for motion prediction and driving decisions. It is an indispensable sensor for intelligent vehicles because of these features and its low cost, and it has already been deployed to production automobiles, particularly for advanced driving-assistance systems (ADAS).

In this report, we are going first to introduce the context of the proposed work: Radar images the Range Angle and Range Doppler measures and then introduce the dataset used in this work. Secondly, we are going to shed light on the related works in the field of radars. After that, we will explore the idea of TMVA-Net. Finally, we will walk you through our proposed approach and compare both the quantitative and qualitative results.

2. Context Radar images: RA and RD measures

The CARRADA dataset is a set of synchronized camera and radar recordings with range-angle-Doppler annotations. The acquisition is performed using a FMCW radar and a camera on a stationary car. The Frequency-Modulated Continuous Wave (FMCW) radar system is the most common automotive radar system to detect the range and velocity of targets through stretch processing. Moreover, recent automotive radar systems are taking advantage of multiple-input-multiple-output (MIMO) antenna arrays to provide the azimuth information of targets. Depending on the MIMO antenna configuration, it is also possible to exploit the elevation information of the targets.

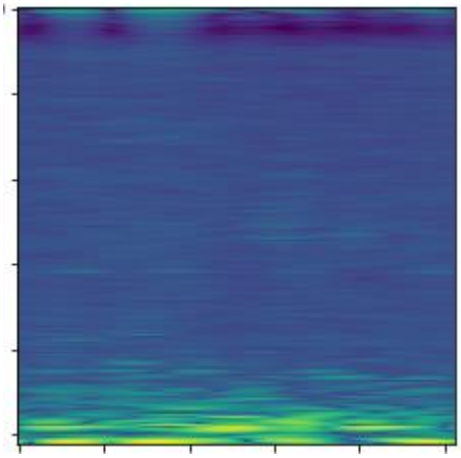
The applications of these radars are important as they are often used for the classification of targets and/or their activities. Most of the real-world targets are not rigid bodies. Motions or vibrations induced by different parts of the targets produce additional Doppler shifts, which is known as Doppler effects and can be used to identify target features. For instance, motions induced by human body parts produce a Doppler signature which can be used to identify human activities.

The radar recording gives us information about object signatures. The detected images are synchronized to have the same frame rate in the dataset. In our case, they are in both range-angle and range-Doppler representations for each sequence. While discovering these frames we found that they are presenting cars, pedestrians, and cyclists. This dataset contains range-angle and range-Doppler raw radar representations, and they are annotated with sparse points, bounding boxes and dense masks to localize and categorize the object signatures.

Camera Image: 000119

Range - Angle representation:

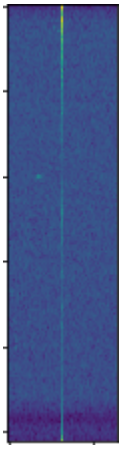
The range-angle representation is a radar scene in polar coordinates. Each instance detected in the frame is characterized by a feature point. They are projected onto the range angle representation by converting the Cartesian coordinates to polar coordinates.



Range-Angle 000119

Range - Doppler representation:

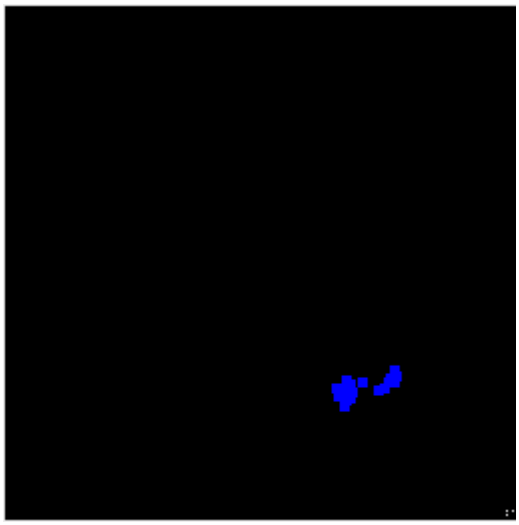
The previous points are projected also onto the range-Doppler representation using the radial velocity and the distance is computed with the real-world coordinate.



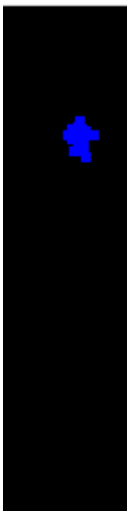
Range-Doppler 000119

Annotations:

The semi-automatic algorithm presented in the paper "***CARRADA Dataset: Camera and Automotive Radar with Range-Angle-Doppler Annotations***" [1] generates precise annotations on raw radar data. Sparse points are the range-Doppler and range-angle representations of the clusters. The bounding box is defined as a rectangle parameterized by {(xmin, ymin), (xmax, ymax)}, where xmin is the minimum coordinate of the set, xmax is the maximum, and similarly for the y-coordinates. Finally, the dense mask annotation is obtained by dilating the sparse annotated set with a circular structuring element.



Range-Angle-Mask 000119



Range-Doppler-Mask 000119

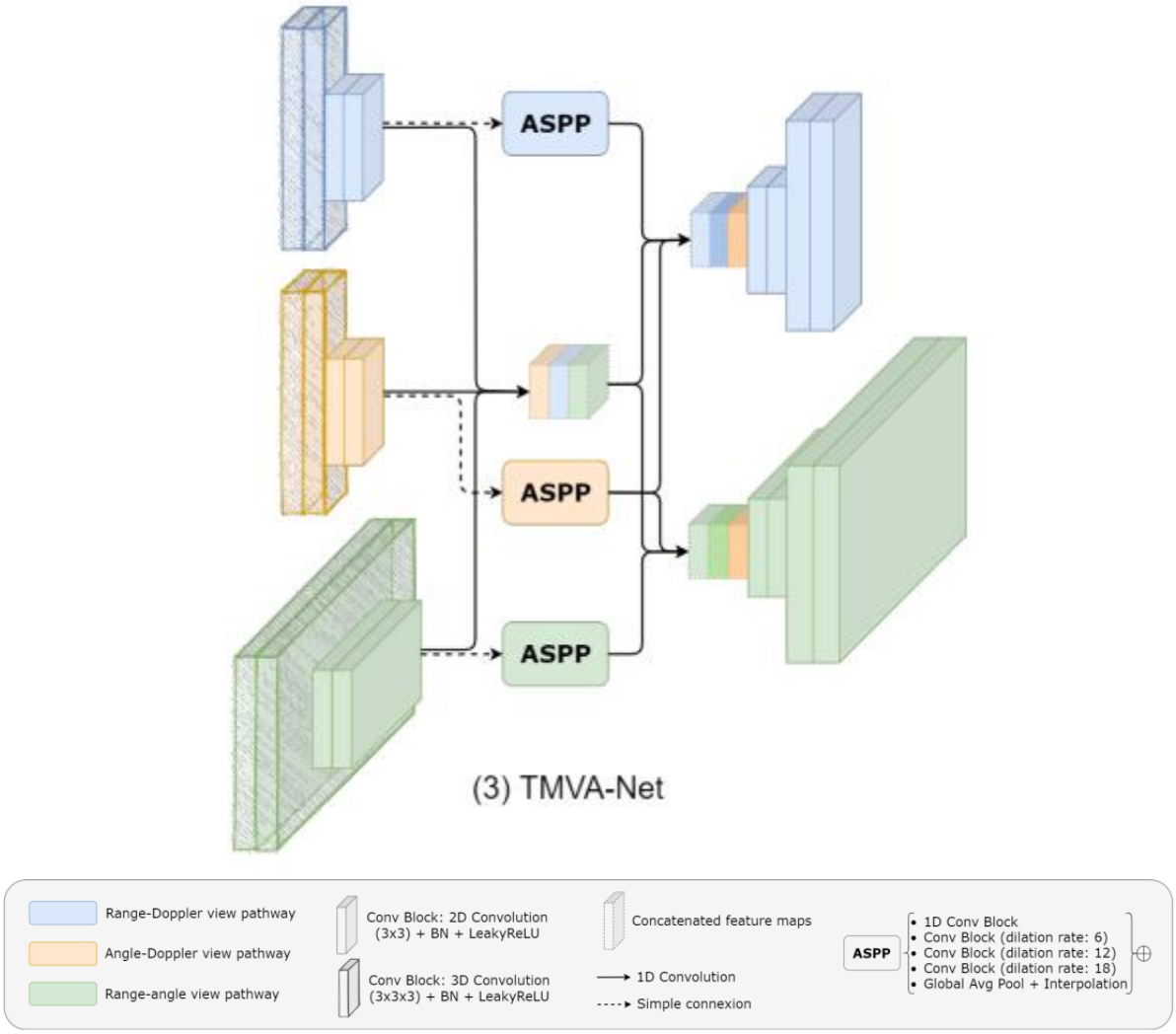
3. TMVA model

Many models have been proposed in the literature to treat this problem. However, the idea of the TMVA model is unique in a way that we will explain further.

For radar semantic segmentation, the time dimension is quite useful. It aids in estimating the form of an object's signature despite high noise levels, as well as distinguishing objects with identical velocities that are close to each other. The TMVA-Net architecture expands the other models by explicitly utilizing the temporal component.

For this architecture, each encoder branch replaces the 2D convolutions in the first block with 3D convolutions, allowing it to learn the spatio-temporal properties with a small increase in the number of parameters. Then for each output we will apply the ASPP block; ASPP: Prior to convolution, Atrous Spatial Pyramid Pooling (ASSP) is a semantic segmentation module that resamples a given feature layer at numerous rates. This entails probing the original image with various filters with complementing effective fields of view,

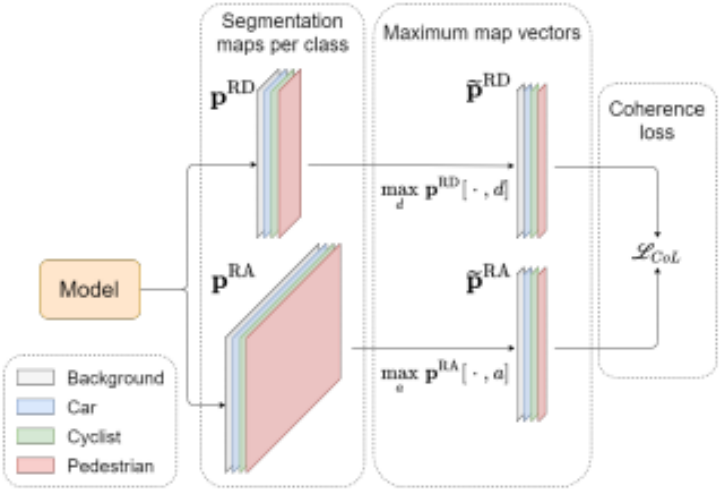
collecting both objects and important image context at different scales. Rather than resampling features, multiple parallel atrous convolutional layers with various sampling rates are used to achieve the mapping.



The feature maps generated from each encoding backbone are stacked into a shared latent space. From there, two decoders segment respectively the RD and RA views. They take as input the stacked features from the processed latent space and the multi-scale feature maps from the dedicated ASPP modules.

In fact, the goal of multi-view radar semantic segmentation is to segment various views of the aggregated RAD tensor simultaneously. The things we want to identify are visible in distinct radar views, therefore it is obvious that there needs to be some coherence across the segmented views. A pedestrian should not be represented in one view while a cyclist should be represented in another. In addition to the cross-entropy and the soft dice

loss, to maintain consistency between the model's predictions, the authors used a coherence loss (CoL). We will detail this loss in the incoming sections.



After that, the idea is to train the model on three combined losses. The Cross Entropy loss, which is specialized in pixel-wise classification. The sDIce, for good shape segmentation and finally the Coherent loss for make both the prediction of RD and RA view coherent.

4. Proposed approaches

4.1. Losses

The following section details the loss functions applied to each segmented view (Range Doppler (RD) view and Range Angle (RA) view) to train TMVA-Net architecture. We propose a different expression for the total loss and a different expression for the introduced ***Coherence loss*** in [2] with the intention to further enforce consistency over the two views of the scene and ameliorate the predictions.

Three different loss functions are proposed in [2]: Weighted Cross Entropy, Soft Dice, and Coherence. The different strengths of these losses are combined in the following final loss to train multi-view architectures:

$$\mathcal{L} = \lambda_{\text{wCE}}(\mathcal{L}_{\text{wCE}}^{\text{RD}}+\mathcal{L}_{\text{wCE}}^{\text{RA}})+\lambda_{\text{SDice}}(\mathcal{L}_{\text{SDice}}^{\text{RD}}+\mathcal{L}_{\text{SDice}}^{\text{RA}})+\lambda_{\text{CoL}}\mathcal{L}_{\text{CoL}}$$

Where:

$$\lambda_{wCE},\ \lambda_{SDice}\ and\ \lambda_{CoL}\ are\ set\ empirically.$$

The Coherence Loss

The objects we wish to detect are observed in the different radar views, thus it is clear that a certain coherence must be maintained between the segmented views, which is the goal of the coherence loss (CoL). It encourages the network to predict high probability values at the same distance for both views, the problem of being in the same class is treated by the Weighted Cross Entropy (wCE) loss and the Soft Dice (SDice) loss. In most cases, RA view misclassifies the detected object, this error is penalized by wCE and SDice losses.

The coherence loss is defined as the mean squared error (MSE) between the maximum range probability vectors.

$$\mathcal{L}_{\text{CoL}}(\mathbf{p}^{\text{RD}}, \mathbf{p}^{\text{RA}}) = \left\| \tilde{\mathbf{p}}^{\text{RD}} - \tilde{\mathbf{p}}^{\text{RA}} \right\|_{\text{F}}^{2}$$

Where:

- $||.||_F$ denotes the Frobenius norm.
- P^{RD} and P^{RA} are the prediction vectors.
- $\tilde{P}^{RD}$ and $\tilde{P}^{RA}$ are obtained by applying a max(.) operator on P^{RD} and P^{RA} along the Doppler or the angle axis.

We propose a new expression of the total loss and a new expression of the Coherence loss.

4.1.1. New expression for the Total Loss

4.1.1.1 Motivation

The Range Angle measures are less reliable than the Range Doppler measures. They are less precise in representing the range distance compared to the Range Doppler measures. This can be easily visualized in the ground truth of the CARRADA dataset where Range Angle maps represent very restrictive range distance of the detected object whereas Range Doppler maps represent larger and more extended signatures of the detected object. In practice, we notice that misclassification cases are mostly by the Range Angle view. Thus, our aim is to reduce the impact of the low precision of RA view by associating weights with respect to each view. As we aim to penalize more the loss associated with the RD view, we assign a lower weight to RA.

The weighting factors ($\boldsymbol{\lambda}^{RD}$ and $\boldsymbol{\lambda}^{RA}$) are set practically and chosen as follows.

4.1.1.2 Expression

$$L = \lambda_{wCE}(\lambda^{RD}\, L^{RD}_{wCE} + \lambda^{RA}\, L^{RA}_{wCE}) + \lambda_{SDice}(\lambda^{RD}\, L^{RD}_{SDice} + \lambda^{RA}\, L^{RA}_{SDice}) + \lambda_{CoL}\, L_{CoL}$$

$Where:$

$$\begin{cases} \lambda^{RD} &= 0.7 \\ \lambda^{RA} &= 0.3 \end{cases}$$

A reformulation of the total loss is given by:

$$L = \lambda^{RD}_{wCE}\, L^{RD}_{wCE} + \lambda^{RA}_{wCE}\, L^{RA}_{wCE} + \lambda^{RD}_{SDice}\, L^{RD}_{SDice} + \lambda^{RA}_{SDice}\, L^{RA}_{SDice} + \lambda_{CoL}\, L_{CoL}$$

$Where:$

$$\begin{cases} \lambda^{RD}_{wCE} = \lambda_{wCE} \times \lambda^{RD} \\ \lambda^{RA}_{wCE} = \lambda_{wCE} \times \lambda^{RA} \\ \lambda^{RD}_{SDice} = \lambda_{SDice} \times \lambda^{RD} \\ \lambda^{RA}_{SDice} = \lambda_{SDice} \times \lambda^{RA} \end{cases}$$

4.1.2 New expression for the Coherence Loss

4.1.2.1 Motivation

The coherence term is supposed to be a restrictive term in the optimization of the network. In fact, the goal of the Coherence loss is to look for the matching that can exist between RA and RD. Hence, it puts constraints on the predictions of both views. It was revealed that this might increase probabilities associated to the range distance of the object and better localize the positions of the prediction in the accurate class which ameliorates the performances in RA. However, it deteriorates the performances in RD since the signatures in the RA vector are very restricted and with low resolution which influences the signature in RD.

Therefore, we aim to solve this problem by minimizing the impact of the constraint of the coherence loss. It might be a reasonable proposition to introduce the ground truth in the coherent loss to consider a specific zone of the RD and RA probability vectors. This by filtering the probabilities using the ground truth vectors. However, the RA ground truth vector y^{RA} is very restrictive and limits the information on the signature. Therefore, to match areas in both vectors, we introduce a common ground truth vector y^{R}. We propose to set the common ground truth vector y^{R} equal to the RD ground truth vector y^{RD}. because the resolution and the annotations are better regarding RD information.

4.1.2.2. Expression

$$L_{CoL}(p^{RD}, p^{RA}) = ||\tilde{p}^{RD} \odot y^R - \tilde{p}^{RA} \odot y^R||$$

$Where:$

$$y^R = y^{RD}$$

4.2. Experiment results

4.2.1 Quantitative results

During our project, we have trained 4 models. The first model is the model implemented in the paper TMVA-Net. The second model is trained on a modified global loss where we added more weight to the RD than RA view. The third model is trained using the added weights on the global loss and with the introduction of the ground truth on the coherent loss. Finally, the fourth model is trained on a modified coherent loss without adding supplement weights to any of the views.

Range-Doppler				
	First Model: Initial Model	Second model	Third Model	Fourth Model
Precision	0,651	0,675	0,656	0,665
recall	0,768	0,769	0,763	0,792
miou	0,579	0,594	0,578	0,596
dice	0,699	0,715	0,699	0,718

Range-Angle				
	First Model: Initial Model	Second model	Third Model	Fourth Model
Precision	0,469	0,469	0,452	0,456
recall	0,532	0,566	0,522	0,526
miou	0,401	0,409	0,389	0,393
dice	0,494	0,506	0,478	0,483

Interpretations

From the Range - Doppler results, we notice that the second and the fourth model have better results than the initial model. These approaches are ameliorating the performance of the model in the quantitative way. In addition, the fourth model outperforms

the second model. It gives 59.6% of the miou and 71.8% of dice. The third model does not add any improvement to the model.

From the Range-Angle results, we remark that the second model has the best performance for all the metrics with 40.9% for the miou and 50.6% of the dice which is higher than 49.4% of the initial model.

➔ We can deduce that the second model is globally the model that gives better results. By adding weights to the model, we give privilege to the data with relevant information, and we diminish the impact of the other data.

4.2.2 Qualitative results

Now we are going to explore the result from a qualitative point of view. We will see if the qualitative results are coherent with the quantitative ones. To do that, we will consider 2 frames from 2 different sequences and compare the results.

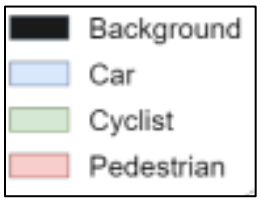
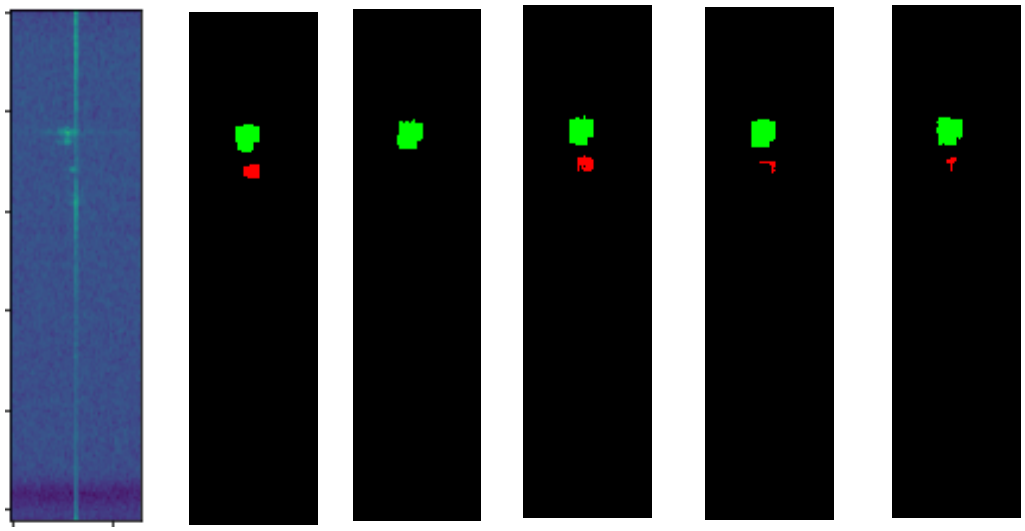
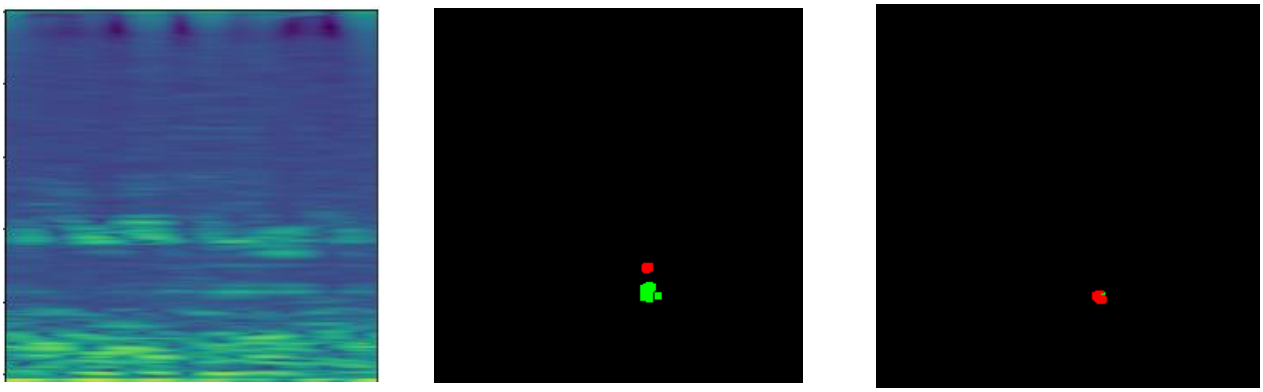


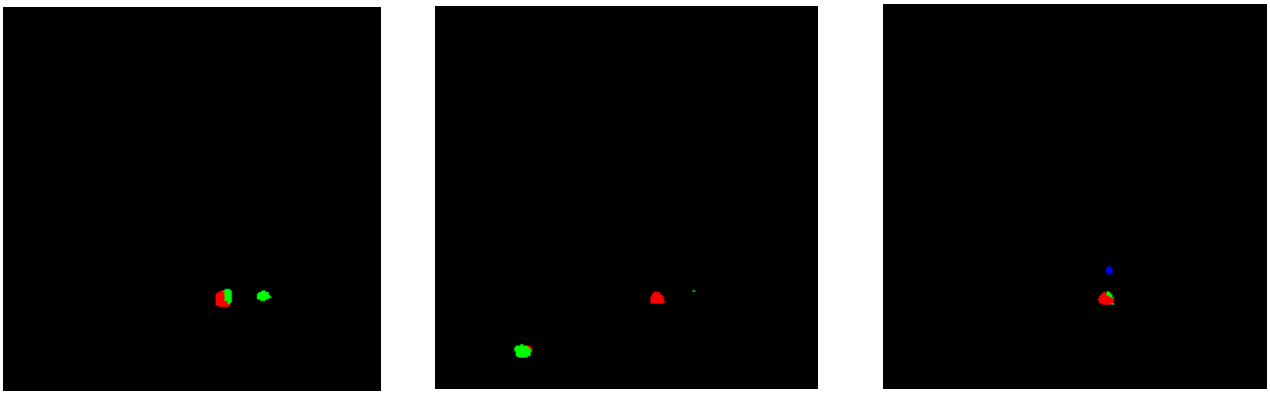
Image camera



(a) RA raw (b) RA GT (c) RA model 1 (d) RA model 2 (e) RA model 3 (f) RA model 4



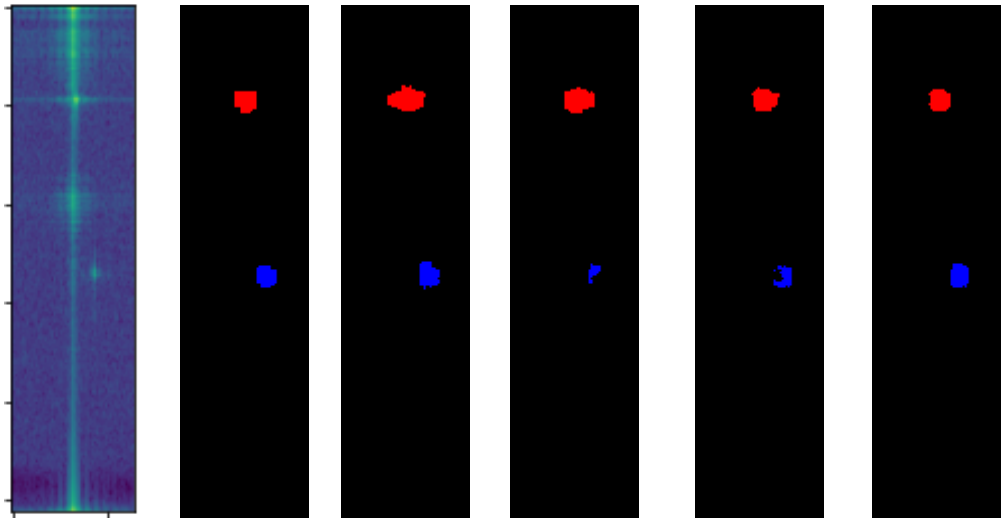
RA raw (a) RA original (b) RA model 1 (c)



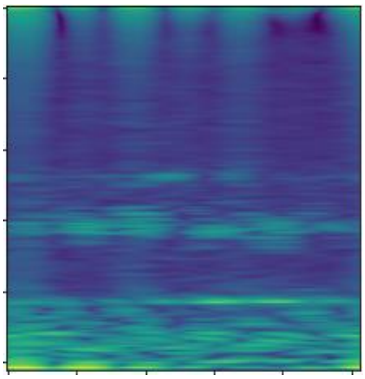
RA model 2 (d) RA model 3 (e) RA model 4 (f)

In this image, we can see that the predictions between RA/RD views of our models are coherent except for the 4th model where he predicted the presence of a far car somewhere on the Range angle view. This may have happened because of the introduction of the ground truth in the coherent loss which impacted the other losses. Also, we can see that the models that we have trained using our approach did better than the TMVA-model. Model 2 to 4 have correctly predicted the presence of both the pedestrian and the cyclist on the Range doppler view. In general, the RA resolution is not perfect for all 4 models, and it is a general problem with radar images. As for the location of the classes, the model 2 gave the best results. It has the highest dice and intersection over union with the ground truth mask for both the RA and RD views. The new models gave better results than the first model and there is a coherence between all the views except the fourth model.

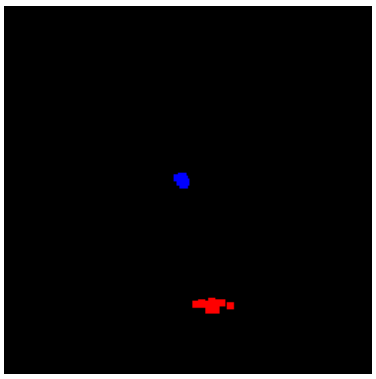
Image camera 2



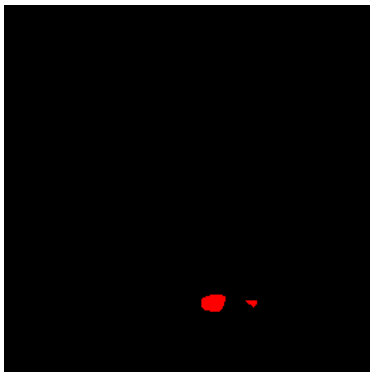
(a) RA raw (b) RA GT (c) RA model 1 (d) RA model 2 (e) RA model 3 (f) RA model 4



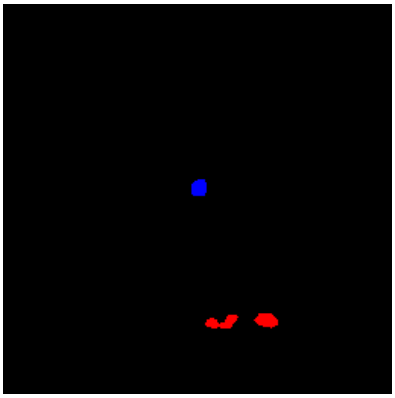
RA raw



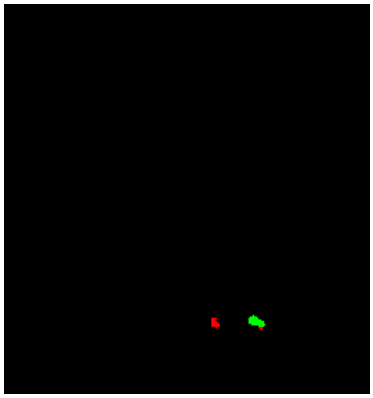
RA original



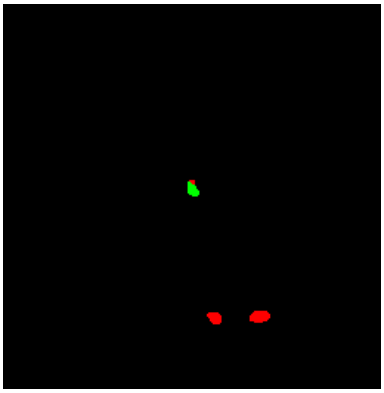
RA model 1



RA model 2



RA model 3



RA model 4

For the RA view, all the models have successfully predicted the presence of the car and the pedestrian in the image. In this example, the TMVA-net architecture, and model 4 gave the best results which are the closest to the RA ground truth mask. However, we can see that for model 3 and 4, there is no coherence in the prediction of the image. Both models have bad Range angle resolution, they have misclassified the car to a cyclist, and this is maybe because the car appears very small in the picture. If We combine both prediction on both views, we can see that model 2 and model 1 have done globally good job predicting the presence of the true present classes. Model 2 has better predictions on the RA view. Model one has better results on the RD view.

5. Conclusion

After understanding the architecture of the model, its code and its function, we have studied the loss function and especially the coherence loss. The aim of this project is to ameliorate the quantitative and qualitative results of the TMVA-Net model. Many ideas are suggested. The starting point was from discussing the relevance of each type of data (the RA raw or the RD raw). That is why, two approaches were posed. The first one is to add weights to the element of the loss and the second one is to take into consideration the RD masks. After applying the convenient transformation, we trained the model. We noticed that globally, model 2 which is the loss with weights gives better qualitative and quantitative results.

Add to that, many other ideas are proposed, and their experiment may be useful to ameliorate the results. The first one is to try the parameter tuning by trying different values of weights. The second approach is to add a Mathematical morphology to the results of the predicted matrices multiplied with masks. In addition, we can use the fact that the data information depends on the time.

This project is very interesting, and it may open other roots of research to contribute in the autonomous vehicle domain. Any improvement will participate in offering the opportunity for safe, efficient, accessible, and affordable transportation. They promise not only a novel system of mobility, but also a novel approach to the urban lifestyle.

Bibliography

[1] A. Ouaknine, A. Newson, P. Pérez, F. Tupin, and J. Rebut. Multi-View Radar Semantic Segmentation. *ArXiv*, 2021.

[2] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez. CARRADA dataset: camera and automotive radar with range-angle-Doppler annotations. In ICPR, 2020.