Raw High-Definition Radar for Multi-Task Learning – Supplementary Material –

Julien Rebut[1] Arthur Ouaknine[1,2] Waqas Malik[3] Patrick Pérez[1]

1: Valeo.ai, Paris, France 2: Télécom Paris, Paris, France 3: Valeo North America Inc., San Mateo, CA

1. Details of the RADIal dataset

Sensor specifications. Central to the proposed RADIal dataset, our high-definition radar is composed of $N_{\text{Rx}} = 16$ receiving antennas and $N_{\text{Tx}} = 12$ transmitting antennas, leading to $N_{\text{Rx}} \cdot N_{\text{Tx}} = 192$ virtual antennas. This virtual-antenna array enables reaching a high azimuth angular resolution while estimating objects' elevation angles as well. As the radar signal is difficult to interpret by annotators and practitioners alike, a 16-layer automotive-grade laser scanner (LiDAR) and a 5 Mpix RGB camera are also provided. The camera is placed below the interior mirror behind the windshield while the radar and the LiDAR are installed in the middle of the front ventilation grid, one above the other. The three sensors have parallel horizontal lines of sight, pointing in the driving direction. Their extrinsic parameters are provided together with the dataset. RADIal also offers synchronized GPS and CAN traces which give access to the geo-referenced position of the vehicle as well as its driving information such as speed, steering wheel angle and yaw rate. The sensors' specifications are detailed in Table 1.

RADIal dataset. RADIal contains 91 sequences of 1 to 4 minutes in duration, for a total of 2 hours. These sequences are categorized in *highway*, *country-side* and *city* driving. The distribution of the sequences is indicated in Figure 1. Each sequence contains raw sensor signals recorded with their native frame rate. A Python library is provided to read and synchronize the data together. There are approximately 25,000 frames with the three sensors synchronized, out of which 8,252 are labelled with a total of 9,550 vehicles.

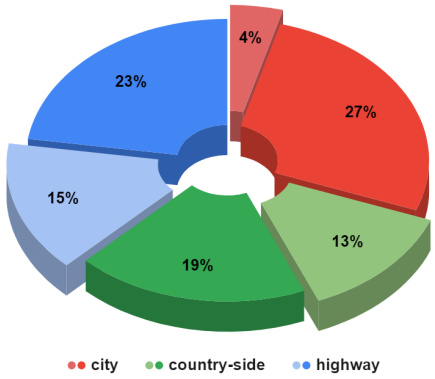

Figure 1. **Scene-type proportions in RADIal**. The dataset contains 91 sequences in total, captured on city streets, highway or country-side roads, for a total of 25k synchronized frames (dark colors), out of which 8,252 are labelled (light colors).

		HD Radar	LiDAR	Camera
FOV	Range	103 m	150 m	–
	Azimuth	180°	133°	100°
	Elevation	12°	10°	75°
resolution	Range	0.2 m	0.1 m	
	Azimuth	0.1°	0.125-0.25°	2592 px
	Elevation	1°	0.6°	1944 px
	Velocity	0.1 m·s^{-1}	–	–
Frame rate		5 fps	25 fps	30 fps
Height above ground		80 cm	42 cm	145 cm

Table 1. **Specification of the RADIal's sensor suite**. The main characteristics of the HD radar, the LiDAR and the camera are reported. Their synchronized signals are complemented by GPS and CAN information.

2. Ablation study of the MIMO pre-encoder

The role of the MIMO pre-encoder is to de-interleave the range-Doppler spectrums and to transform them into a representation that is compact and still allows, through learning, the prediction of azimuth angles along with other information on reflectors. The input of the MIMO pre-encoder is composed of the $N_{\text{Rx}} = 16$ range-Doppler spectrums *in complex numbers*, one for each Rx. The real and imaginary parts are stacked, yielding an input tensor of total size $B_{\text{R}} \times B_{\text{D}} \times 2N_{\text{Rx}}$, *i.e.*, $512 \times 256 \times 32$. The ablation study consists in evaluating the performance of FFT-RadNet's detection head while reducing the number of feature channels that the MIMO pre-encoder outputs. The maximum number of output channels is the number of virtual antennas with a complex signal (real and imaginary parts), *i.e.*,

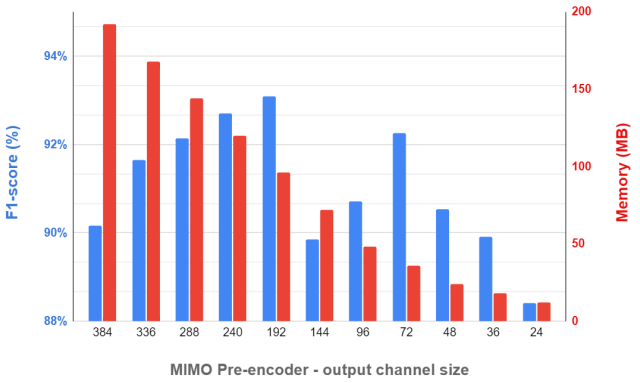

Figure 2. **MIMO pre-encoder ablation**. Influence of the number of output channels of the pre-encoder on the memory footprint and the performance of the detection head.

$N_{\text{Tx}} \cdot 2 \cdot N_{\text{Rx}} = 384$. We vary the number of output channels from a minimum of 24 to this maximum value and compute the detection performance on the validation set. The results of this ablation study are reported in Figure 2. We measure the detection performance with the F1-score, classically defined as F1-score $= \frac{\text{AP} \cdot \text{AR}}{\text{AP}+\text{AR}}$, which aggregates in a single metric both the Average Precision (AP) and the Average Recall (AR). We observe that the best performance is reached with 192 output channels, hence half of the maximum output size. This compressed output is the one that captures at best the range and azimuth information from the inputs range-Doppler spectrums toward the detection and segmentation tasks.

3. Radar versus LiDAR

RADIal dataset was designed to collect information from several sensor technologies. For safety-critical systems, such as self-driving vehicles, we believe that redundancy at various levels of the system, starting at the sensing layer, is key to guarantee safe operations. In a complete automated driving system, the combination of radars together with cameras and LiDARs will improve the overall robustness. Indeed, LiDARs provide, even at night, accurate 3D localization of objects in distance and angle, while cameras give access to a wealth of semantic and geometric information about the scene when light is sufficient. However, these two types of sensors suffer from bad weather conditions that can degrade quite significantly their performances. Radars are more robust to adverse weather conditions, provide accurate distance estimates together with the velocity of the objects, and are especially well suited to the cost and size constraints of automotive applications.

For reference, we report in Table 2 the performance on RADIal obtained by the imaging radar (with FFT-RadNet) and by the LiDAR sensor (with Pixor), respectively. The former obtains similar performances in AP and lower, though still good, performances in AR compared to the latter. This is already a remarkable result, owing to the practical advantages of the radar technology, which we reminded above. In addition, this difference of performances might be explained by the way RADIal dataset was created. The ground truth is obtained semi-automatically based on 2D detection/segmentation from the camera fused with the 3D information of LiDAR. The evaluation might thus be favorably biased toward processing LiDAR inputs.

Sensor	Model	AP(%) ↑	AR(%) ↑
LiDAR	Pixor	98.55	90.42
Radar	FFT-RadNet	96.84	82.18

Table 2. **Vehicle detection with HD radar alone and LiDAR alone**. Performance in average precision (AP) and average recall (AR) on RADIal Test split. FFT-RadNet takes range-Doppler spectrum as input, and Pixor the LiDAR point cloud.

Due to the nature of the annotation pipeline, and to the radar multi-path reflections, many sequences of complex scenes in urban or dense environments, which are present in RADIal, were not annotated. In Figure 3, we qualitatively compare vehicle detection in such complex scenes when using either the HD radar modality or the LiDAR one. We observe that the HR radar, equipped with FFT-RadNet, detects vehicles in complex situations, including beyond the first row of vehicles where neither the camera nor the LiDAR performs well.

Figure 3. **Examples of vehicle detection in complex scenes using either HD radar or LiDAR**. Comparison between Pixor trained with LiDAR point clouds ('PIXOR LiDAR' columns, green boxes) and our proposed FFT-RadNet requiring only range-Doppler as input ('FFT-RadNet', red boxes). Note that radar detections are not limited to the first row of vehicles but can see up to the second one. Also, FFT-RadNet provides vehicles' relative speed through Doppler measurements.