TransRadar: Adaptive-Directional Transformer for Real-Time Multi-View Radar Semantic Segmentation

Yahia Dalbah   Jean Lahoud   Hisham Cholakkal

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)

Abstract

Scene understanding plays an essential role in enabling autonomous driving and maintaining high standards of performance and safety. To address this task, cameras and laser scanners (LiDARs) have been the most commonly used sensors, with radars being less popular. Despite that, radars remain low-cost, information-dense, and fast-sensing techniques that are resistant to adverse weather conditions. While multiple works have been previously presented for radar-based scene semantic segmentation, the nature of the radar data still poses a challenge due to the inherent noise and sparsity, as well as the disproportionate foreground and background. In this work, we propose a novel approach to the semantic segmentation of radar scenes using a multi-input fusion of radar data through a novel architecture and loss functions that are tailored to tackle the drawbacks of radar perception. Our novel architecture includes an efficient attention block that adaptively captures important feature information. Our method, TransRadar, outperforms state-of-the-art methods on the CAR-RADA [26] and RADIal [28] datasets while having smaller model sizes. https://github.com/YahiDar/TransRadar

1. Introduction

Automotive systems rely on radar sensing for most of the tasks that require deterministic distance measurements, such as collision avoidance, blind spot detection, and adaptive cruise control. The prevalence of radar sensors in these tasks has been attributed to their relatively low cost, low processing time, and ability to measure the velocity of objects. On the other hand, LiDAR sensors have risen in popularity as the main automotive perception tool for autonomous driving due to their relatively higher resolution and ability to generate detailed point-cloud data. This popularity is noticeable in recent literature, where LiDAR sensors are dominantly used in object detection and semantic segmentation tasks. However, LiDAR sensors suffer from few drawbacks originating from the shorter wavelength of their signals. LiDAR sensors are highly prone to errors,

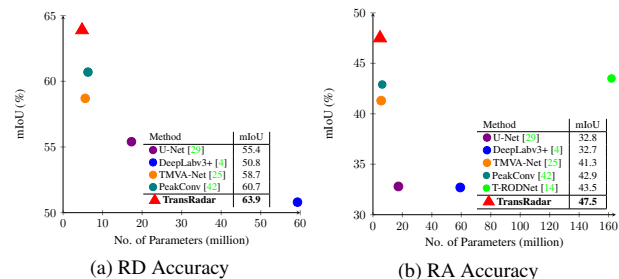(a) RD Accuracy   (b) RA Accuracy

Figure 1. mIoU scores vs No. of Parameters (millions) of state-of-the-art models in semantic segmentation on the CARRADA dataset. Our method, TransRadar, outperforms previous state-of-the-art methods in the semantic segmentation task with an mIoU of 63.9% for RD maps and 47.5% for RA maps.

weather fluctuations, and occlusion with raindrops and/or dust [7]. Moreover, LiDAR signals' higher frequencies result in a rapid attenuation of their strength with respect to distance traveled, which results in a maximum range of operation of 100 to 200m.

Unlike LiDARs, frequency-modulated continuous wave radars operate in the millimeter wave band in which signals do not get significantly attenuated when faced with occlusions, allowing operation ranges of up to 3,000m. Radars function in adverse weather conditions more robustly than other commonly used sensing methods like cameras and LiDARs. Radar signals are also rich in information as they contain Doppler information that includes the velocity of the objects. These radar features have motivated its usage not only in deterministic instrumentation but also for computer vision tasks [33, 39]. The radar signals can be processed to be used in an image-like pipeline in the form of Range-Angle (RA), Range-Doppler (RD), and Angle-Doppler (AD) maps. These maps are sliced views of the total 3D Range-Angle-Doppler (RAD) cube, and obtaining any two combinations allows for the calculation of the third.

The task of semantic segmentation using raw/processed radar data has been a growing task in the radar perception community and has shown promising development in recent years [8, 14, 22, 23, 26, 27, 30, 33, 39]. Nonetheless, segmenting radar images still poses a challenge due to the noisy and sparse nature of the data, as well as the high imbal-

ance between the foreground and background. Also, despite the information-rich nature of radar data and the ability to obtain multiple views from a single sensing instance, most works do not utilize these benefits and tend to limit their approaches to Convolutional Neural Network (CNN) models on a single view, resulting in models that do not adequately capture global information from these maps. To circumvent that, we propose a novel attention-based approach for semantic segmentation using radar data signals in radar learning. Our technique extends the definition of attention models to apply attention to adaptively sampled variations of our input feature maps, tackling the sparse nature of radar data. The adaptability nature of our attention block allows it to attend to multiple views of the Range-Angle-Doppler (RAD) cube in an efficient way. We also combine our model with a loss function tailored to sparse and highly imbalanced data of our task. We propose a combination of class-agnostic, multi-class, and multi-view consistency losses.

Contribution: In this work, we propose an automotive radar sensing method that outperforms previous state-of-the-art works and sets new top scores in the reported metrics (Figure 1). Our main contributions are:

- We introduce a novel adaptive-directional attention block that efficiently captures information from a sparse receptive field and simultaneously tackles the multi-input multi-output nature of our task.

- We propose a novel loss function for the radar semantic segmentation task tailored to address the inherent main drawbacks of radar data. These drawbacks include the noisy and sparse nature of radar signals and the disproportional level of background/foreground objects.

- Our proposed approach results in state-of-the-art performance in radar semantic segmentation of two recent datasets for radar perception, CARRADA [26] and RADIal [28], and achieves state-of-the-art results in the object detection task of the RADIal dataset.

2. Related Work

Low-cost frequency modulated continuous wave radars have been historically used in multiple applications involving machine learning and pattern recognition such as human activity and hand gesture recognition [41, 43, 44]. In the context of automotive driving and autonomous vehicles, LiDAR sensors are more popular with a common data output in the form of a point cloud. While multiple works have explored point-cloud fusion of radars and LiDARs [1, 7], radar signals processing usually yields different physical representation than the LiDAR.

The low resolution and high sparsity of radar data make the point-cloud format and associated architectures unsuitable. While some datasets provide point-cloud radar

data [2, 30], recent approaches to radar processing use the full/split processed RAD tensors in the shape of 3D/2D image-like data. Common radar datasets provide either a single view of the data (either RA or RD) [28, 33], the original raw and unprocessed radar signals [28], or the full RAD tensors [25, 39]. RAD tensors provide cohesive information of the radar data; however, it is often undesirable to use 3D data due to the increased complexity of models when associated with the density of radar data, especially when taking multiple frames from the temporal domain. In this work, we focus our efforts on getting an automated radar perception model through sliced radar RAD tensors and comparing our method to similar works.

With the recent emergence of radar datasets [26, 28], few methods have been proposed for semantic segmentation and object detection. While common methods for image semantic segmentation can be employed, such as UNet [29]and DeepLabv3+ [4], these methods are not tailored to the noisy and sparse nature of radar images. We highlight the most recent and relevant works that process radar data. TMVA-Net [25] is a multi-view method that is composed of an encoding block, a latent-space processing, and a decoding block. It fully consists of convolutional layers and presents a strong baseline for predictions in RD and RA maps on the CARRADA dataset. RAMP-CNN [9] is a CNN-based model that was mainly designed for processing 3D RAD tensors but was re-purposed for this dataset. T-RODNet [14] is a recent model utilizing Swin Transformers [20] but does not produce RD predictions and operates only on RA inputs. While T-RODNet shows improved RA scores, we focus on simultaneous prediction of the RD and RA semantic segmentation maps. PeakConv [42] applies the convolution operation with a receptive field consisting of the peaks of the signal. While this approach achieves improved segmentation performance compared to TMVA-Net, it also increases the number of parameters.

Sparse variants of attention have been proposed in the literature. ReLA [35] replaces the softmax activation with ReLu to achieve sparsity in attention and uses layer normalization to improve translation tasks. The sparsity can range from switching off attention to applying attention to all the input. On the other hand, our method learns the offsets to which the attention is applied and targets consistent efficiency for the radar segmentation task. Other sparse attention methods, such as NPA [36] and SCAN [40] address point clouds, which are sparse in nature. Our method aims at learning to select important locations in the radar map dense grid.

3. Baseline

TMVA-Net starts by encoding the RA, RD, and AD input maps to reduce the input size to one-fourth of its original resolution. Each output is then passed into an Atrous Spatial
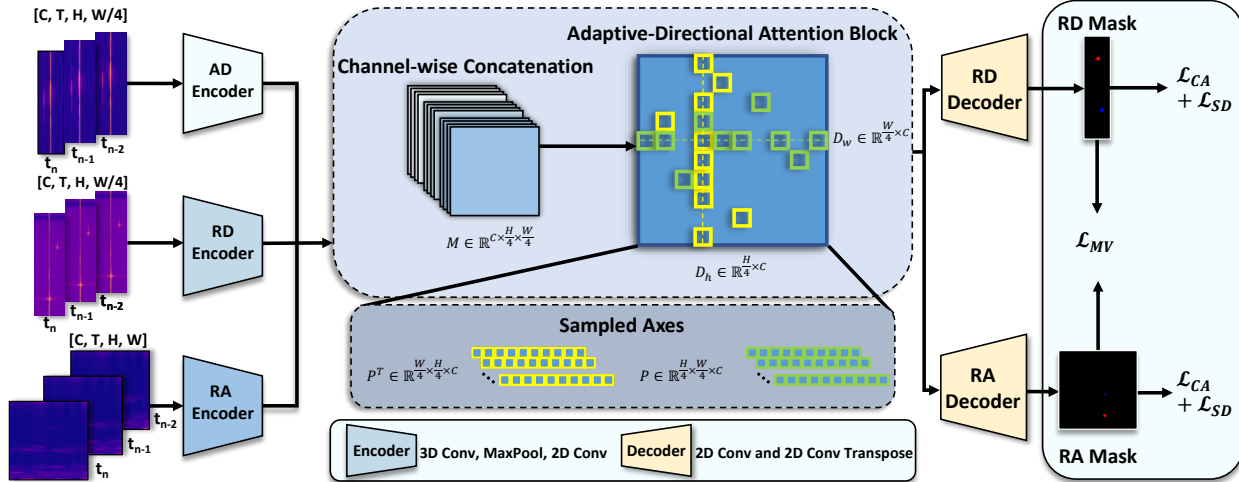
Figure 2. Overview of our proposed method for radar semantic segmentation. The model starts by encoding multiple frames of the Angle-Doppler (AD), Range-Doppler (RD), and Range-Angle (RA) maps. The encoded features are concatenated into a single block of feature maps that is then passed into our adaptive-directional attention block. The adaptive-directional attention blocks sample rows and columns following Eqs. 1 & 3 and apply self attention following Eq. 2 after each sampling instance. The outputs are then split into two decoders generating RD and RA masks that are passed into our three loss functions described in Section 4.4.

Pyramid Pooling (ASPP) block [3], and is also concatenated into a single feature maps holder. Both the ASPP output and the concatenation are then passed into a two-branches (RA and RD) decoding space that produces prediction maps. TMVA-Net uses a combination of three loss functions: a weighted Cross-Entropy loss, where the weights correspond to the frequency of classes in the dataset, a weighted Soft Dice loss, and a coherence loss. The coherence loss is a mean-square error of the RD and RA outputs to ensure coherence of predictions from different views.

3.1. Limitations

The mentioned models yield state-of-the-art results in radar semantic segmentation on the CARRADA dataset. Nonetheless, these models have limitations pertaining to the nature of the implementation and the task. First, the models are limited to convolution layers that learn local spatial information of the multi-input data. While increasing the number of feature maps at every layer would slightly improve the accuracy of these models, it imposes a large computation burden. This impedes the model from further improving without increasing the number of parameters with the majority of parameters being employed in the convolutional layers. The second limitation is the ability of these models to learn and retain information from other maps. T-RODNet processes RA maps only, while TMVA-Net concatenates all feature maps in the bottleneck along with the ASPP outputs. For the rest of the model, all combined feature maps are treated as a single set of feature maps coming from one source that gets split into two prediction heads.

Another important aspect to be considered in these methods is the number of parameters. TMVA-Net pro-

duces multi-view results with $50\times$ less parameters than T-RODNet. Lastly, all reported models were trained using the combination of losses which are not optimally designed for the task of radar semantic segmentation. Therefore, we propose an alternative approach in Section 4.4.

4. The Proposed Method

4.1. Motivation

Our proposed method is designed to address the limitations we observed in the state-of-the-art models discussed previously. We aim to create a compact model that improves upon previous methods by addressing the issues observed in model learning through a proposed novel architecture and loss functions. Our method overcomes the hurdle of introducing attention in deep learning models by minimizing the number of tokens to keep the model fast and small. We also take into consideration the sparse nature of the radar data while implementing our method. We propose a loss function tailored specifically for the task of radar learning by taking the acquisition structure into consideration. We extend our approach to addressing the issue of class imbalance in a more refined way compared to weighted cross-entropy, and we tackle the poor localization ability of the proposed models in our loss functions. Lastly, we propose a new multi-view range matching loss that addresses the drawbacks of fused multi-view inputs.

4.2. Overall Architecture

We propose a lightweight attention-based neural network architecture, shown in Figure 2, which addresses the limitations of the previous works. The model starts by us-

355

ing a similar encoding module as the one used in TMVA-Net [25], with $x_i \in \mathbb{R}^{1 \times T \times H \times W}$ where x_i is an RA, RD, or AD feature map, T is the number of past frames taken from the range $[t_0 - T, t_0]$, and H and W are the height and width of the radar frequency map, respectively. The feature maps generated from the encoders are expressed as $x_{en} \in \mathbb{R}^{C \times H_d \times H_d}$, where x_{en} is an encoded feature map, C is the number of feature maps, and H_d and W_d are the downsampled heights and widths, respectively. The produced feature maps are then channel-wise concatenated into a single latent space that constitutes the input to our adaptive-directional attention block. In convolution-based competing methods, we noticed that reducing the feature maps below 128 channels in the latent bottleneck greatly reduces the mIoU, so we adopt an attention-based approach that achieves similar scores with smaller feature maps.

Contrary to other attention-based approaches in radar perception [14], we do not need to use convolutional layers or heavy positional embeddings. Instead, we shed light on the way the dataset is constructed, where the multi-view input has implicit information that can be shared across axes and channels. Figure 2 illustrates the operation mechanism of our adaptive-directional attention block after the concatenation of the inputs' encoding.

4.3. Adaptive-Directional Attention

In our model architecture, we propose a novel adaptive-directional attention block that composes the backbone of our model. Similar concepts of sampling straight-vector axes were previously proposed in the literature [11, 12, 32]. However, our adaptive-directional attention tackles the sparse nature of radar data by utilizing attention that can extend further than single-column/row attention. In this way, it ensures a comprehensive outlook of the information space while being computationally efficient. For a 2D input image of shape $C \times H_d \times W_d$, we obtain two attention variations, one of the shape $H_d \times W_d \times C$ and another of the shape $W_d \times H_d \times C$. For example, for a width W_d, we have W_d sampled vectors of size $H_d \times C$. The rationale behind incorporating the channels in our sampling traces back to the rich information provided by the radar data's feature maps. We sample our axes by employing vertical and horizontal iteration limits of sizes k_h and k_w, respectively. We also define the horizontal and vertical shifts, Δh and Δw, that constitute the offset limits of sampling. Lastly, we define learnable parameters θ_h and θ_w that perform a modulating operation to limit the effect of noise seen in data, allowing the model to learn to suppress insignificant regions. Using these definitions, we then write the sampling operation that occurs before the attention on the columns as:

$$x_{i,j} = \sum_{k=1}^{w} (\theta_h)_k \cdot X_{H,C}^{(i,j+\Delta h_k)} \tag{1}$$

where $x_{i,j}$ is the value of the column with indices i, j belonging to the axes as $i \in [0, H]$ and $j \in [0, C]$. Parameter w refers to the horizontal iterations limit (i.e. how many pixels we iterate over), belonging to the previously defined parameter k_w. $(\theta_h)_w$ is the corresponding modulation weight for the associated shift, and Δh_w covers how far we sample from the axis center (i.e. the starting column).

After the sampling operation, we obtain W_d vectors of size $H_d \times C$. The query, key, and values (**q**, **k**, **v**) are then obtained through multi-layer perceptron layers, where the multi-headed self-attention (MSA) is then calculated as:

$$SA(q, k, v) = Softmax(\frac{qk^T}{\sqrt{d_k}})v$$
$$MSA = [SA_1; SA_2; ...; SA_s] \tag{2}$$

for s heads obtained from the input, following the formulation in vision transformers [6]. We note that we first sample by columns (i.e. produce W_d vectors of size $H_d \times C$) and apply MSA, then sample by rows (i.e. produce H_d vectors of size $W_d \times C$) and apply the second MSA. The formulation for the MSA applied to the rows is similar to that of the columns, with the following row sampling:

$$x_{i,j} = \sum_{k=1}^{h} (\theta_w)_k \cdot X_{W,C}^{(i+\Delta w_k, j)} \tag{3}$$

Unlike convolution-based transformers or other types of attention modules, the nature of our adaptive-directional attention allows us to alleviate the need for convolutional channel mixing or expansions. The adaptive sampling reduces the model complexity significantly by incorporating a convolution-like operation before applying attention.

4.4. Proposed Loss Function

Model learning in both semantic segmentation and object detection can prove difficult due to the large ratio of background to foreground pixels. This disparity was historically studied in multiple works that addressed the issue either through employing multi-stage detectors [18, 31] in object detection, or targeting the way models learn through innovative loss functions that handle class imbalance in semantic segmentation [19, 38]. Radar-based datasets have a larger proportion of background pixels when compared to actual objects (foreground). This discrepancy is notably present in the datasets we operate on, where the background class consists of more than 99% of the total dataset pixels [26, 28]. In addition to the class imbalance between background and foreground pixels, the annotated objects are relatively small in pixel size. Lastly, RD, RA, and AD maps' noisy nature is a learning hurdle for the models. To tackle these issues, we propose an Object Centric-Focal loss (OC) and a Class-Agnostic Object Localization Loss (CL). We

add both of them in a single term, the Class-Agnostic Object Loss (CA), and propose a new multi-view range matching loss (MV) that suits our multi-output architecture.

4.4.1 Class-Agnostic Object loss

Object Centric-Focal Loss: The main highlight of this loss is the weighing of the binary cross-entropy between the background and foreground of the predictions, with higher weight being given to the foreground. This is defined as:

$$\mathcal{L}_{OC} = (1 - y_{pred})(\delta\mathcal{L}_{BCE_{FG}} + (1 - \delta)\mathcal{L}_{BCE_{BG}}) \quad (4)$$

where δ is a weighing factor (set to 0.6) and $\mathcal{L}_{BCE}$ is the binary cross entropy, calculated with the two classes 'background' and 'foreground'. While our semantic segmentation objective includes multi-class labels, we aim to use this loss to penalize the model on hard background prediction, keeping it only to a binary background/foreground calculation. While other loss functions [19] propose a power factor on the $(1 - y_{pred})$ term, we instead remove it and use one-hot prediction masks. Both operations come in favor of having a balanced approach between ground truth probabilities and loss value, and heavily penalizing misclassification between the background and foreground.

Class-Agnostic Object Localization Loss: To illustrate the rationale of proposing this localization loss, we show RA and RD input maps with their output predictions, along with the corresponding RGB image in Figure 3. Any other object signature seen in the RA input image can be attributed to speckle noise, Doppler-induced noise, or any other sort of undesired noise that is unaccounted for. Due to this noisy nature of radar data, producing a significantly larger amount of false positives was a noticeable pattern across tested models. We also noticed similar behavior in the opposite way, where the model learns the noise as part of the background and confuses objects with similar signatures as the noise for being part of the background, resulting in many false negatives. Therefore, we propose an intersection-based loss that penalizes the model on false background/foreground predictions. This builds on the previous object-centric loss by creating an IoU-based loss that penalizes mislocalization of objects, defined as:

$$\mathcal{L}_{CL} = 1 - \frac{TP}{TP + FN + FP}, \quad (5)$$

where TP refers to true positives, FN to false negatives, and FP to false positives. Similar to $\mathcal{L}_{OC}$, we extend our implementation to focus on the one-hot predictions instead of the probability maps, which imposes a larger penalty for making a false background prediction. Adding $\mathcal{L}_{OC}$ and $\mathcal{L}_{CL}$ terms yields our class-agnostic object loss: $\mathcal{L}_{CA} = \mathcal{L}_{OC} + \mathcal{L}_{CL}$.
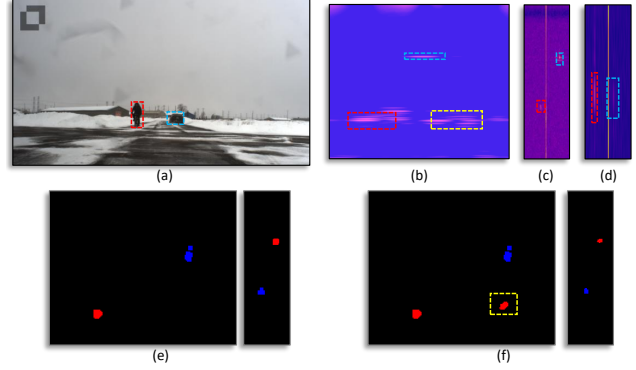

Figure 3. Radar (b) RA, (c) RD, and (d) AD maps with (a) synchronized RGB image. Red and blue annotation boxes correspond to the person and car, respectively, shown in the RGB image. We highlight a sample random noise appearing on the RA map with a yellow box. (e) shows the ground truth mask for the RA and RD maps (left to right) of this scene, and (f) shows a false segmentation with the noise seen as an object. The noise shown in the RA map does not appear as frequently in RD maps. The contrast of the maps was edited for illustration purposes.

4.4.2 Multi-Class Segmentation Loss

To include the multi-class nature of our dataset and localization of different class predictions, we use a similar Soft Dice loss (SD) term to the one used in [25], described as:

$$\mathcal{L}_{SD} = \frac{1}{K} \sum_{k=1}^{K} [1 - \frac{2\sum \mathbf{y}\mathbf{p}}{\sum \mathbf{y}^2 + \mathbf{p}^2}] \quad (6)$$

where $\mathbf{y}$ and $\mathbf{p}$ refer to the ground truth and probability map output of the model. Unlike the previous terms, we do not use a one-hot binary map prediction and instead use the original continuous probability map. We also do not limit $\mathcal{L}_{SD}$ to background/foreground classes since we use it for multi-class predictions.

4.4.3 Range Consistency Loss

In addition to the class-agnostic object loss and multi-class segmentation loss, we define a Multi-View range matching loss (MV) as:

$$\mathcal{L}_{MV} = \begin{cases} \frac{1}{2}(RD_m - RA_m)^2 & |RD_m - RA_m| < 1 \\ |RD_m - RA_m| - \frac{1}{2} & otherwise \end{cases}$$
$$\quad (7)$$

where RD_m and RA_m are the max-pooled RA and RD probability maps, leaving only the R direction. The analytical term of this loss is a special case of the Huber loss [13] and was proven to be more robust than mean-square error when dealing with outliers.

Overall Loss: Our total loss is then defined as the weighted

sum of all proposed losses with weights α_1, α_2, and α_3 as:

$$\mathcal{L}_{total} = \alpha_1 \mathcal{L}_{CA} + \alpha_2 \mathcal{L}_{SD} + \alpha_3 \mathcal{L}_{MV} \qquad (8)$$

5. Experiments

5.1. Datasets

To test the effectiveness of our proposed approach, we use the CARRADA [26] dataset as the main multi-class radar semantic segmentation dataset. We also test our proposed method on RADIal [28] dataset and compare to previous state-of-the-art methods in radar semantic segmentation and object detection.

CARRADA: The CARRADA [26] dataset consists of synchronized camera-radar recordings of various driving scenarios containing 12,666 frames. The annotations of the data were done semi-automatically and provided for the RD and RA views [26]. The dataset contains four object categories: pedestrian, cyclist, car, and background. The input are the RA, RD, and AD maps decomposed from the 3D RAD tensor. RA maps have a size of $1 \times 256 \times 256$ while RD and AD have a different resolution of $1 \times 256 \times 64$. We use the 2D decomposition of the RAD tensor to reduce the model complexity, which is an important factor in radar perception in automotive driving.

RADIal: The RADIal [28] dataset is a new high-resolution dataset consisting of 8,252 labeled frames. RADIal varies from CARRADA in that it does not provide a multi-view input and depends only on RD input. The outputs are also produced and compared to projected annotated RGB images, unlike the CARRADA dataset that compares annotation directly in the RD/RA planes. RADIal also provides a high-definition input, where the input size is $32 \times 512 \times 256$. RADIal provides annotations for two classes only: free-driving-space and vehicle annotations (i.e. free or occupied).

5.2. Evaluation Metrics

We follow the same evaluation metrics used in previous works, which are the common intersection over union (IoU), the Dice score (F1 score), and the mean of each across the classes. The mIoU is also used to evaluate the semantic segmentation task on the RADIal dataset. The combination of the mIoU and the Dice score creates a fair and comprehensive assessment of the results. For the object detection task in RADIal, we use the same metrics as [28] with Average Precision (AP), Average Recall (AR), and regression errors.

5.3. Implementation Details

We implement and train TransRadar using the PyTorch library on a single NVIDIA A100 GPU. All reported models on the CARRADA dataset were trained with a batch size of 6 and using 5 past frames. We use Adam optimizer [17], initial learning rate of 1×10^{-4}, and an exponential scheduler (step = 10). For our final TransRadar model, we use $8\times$ cascaded blocks of our adaptive-directional attention block. For the testing, we use a batch size of 1 and a similar number of past frames.

For the RADIal dataset training, we replace FFTRadNet [28] backbone with our proposed model. We employ a single-view encoding/decoding paradigm similar to the one shown in Figure 2. We use the same segmentation and detection heads from the FFTRadNet model, and the same optimizer and scheduling as CARRADA dataset training.

5.4. State-of-the-art Comparisons

Semantic Segmentation on the CARRADA: Table 1 shows the quantitative comparisons of the proposed approach with existing state-of-the-art frameworks for radar semantic segmentation. The results listed in the table show that TransRadar outperforms state-of-the-art methods in both the mIoU and mDice metrics. A large part of this is attributed to the introduction of the CA loss, which will be discussed in detail in the ablation studies in Section 5.5. Our model achieves new state-of-the-art performance with an RD mIoU score of 63.9%, which outperforms the closest baseline by 3.2%, and has a mDice score of 75.6%. For the RA map predictions, our method yields a mIoU of 47.5%, outperforming the state-of-the-art score by 4.0%, with a mDice of 59.3%. We also point out that our model significantly outperforms other models in the Cyclist class, where we note a large gap of 12.0% between our model and the second-best model in the RA map, and 13.1% in the RD map. This can be attributed to the consistency with RD as well as the ability to predict harder examples. Across the board, our model sets new state-of-the-art scores except for the car class IoU and Dice in the RA maps, where T-RODNet has a slightly higher score.

Figure 4 shows two qualitative results on a hard scene and a normal scene from the test split of CARRADA. The first scene shows a good segmentation with instances of mislocalization in all tested methods, with TransRadar and UNet giving the best prediction results. We then present a well-segmented RD and RA predictions in the second scene relative to the mask from our method when compared to other models. We also notice a coherent translation of the RD to RA views in the range dimension in both scenes.

Semantic Segmentation on RADIal: We further look at the semantic segmentation results of the RADIal dataset shown in Table 2. Our method outperforms all previously reported models in the semantic segmentation task with a mIoU of 81.1% and less than half the model size of the most recently reported state-of-the-art method, C-M DNN [15]. These results showcase the ability of our proposed method,

View	Method	Params (M)	IoU (%)					Dice (%)				
			Bkg.	Ped.	Cycl.	Car	mIoU	Bkg.	Ped.	Cycl.	Car	mDice
RD	FCN-8s [21]	134.3	99.7	47.7	18.7	52.9	54.7	99.8	24.8	16.5	26.9	66.3
	U-Net [29]	17.3	99.7	51.1	33.4	37.7	55.4	99.8	67.5	50.0	54.7	68.0
	DeepLabv3+ [4]	59.3	99.7	43.2	11.2	49.2	50.8	99.9	60.3	20.2	66.0	61.6
	RSS-Net [16]	10.1	99.3	0.1	4.1	25.0	32.1	99.7	0.2	7.9	40.0	36.9
	RAMP-CNN [9]	106.4	99.7	48.8	23.2	54.7	56.6	99.9	65.6	37.7	70.8	68.5
	MVNet [25]	2.4	98.0	0.0	3.8	14.1	29.0	99.0	0.0	7.3	24.8	32.8
	TMVA-Net [25]	5.6	99.7	52.6	29.0	53.4	58.7	99.8	68.9	45.0	69.6	70.9
	PeakConv [42]	6.3	-	-	-	-	60.7	-	-		-	72.5
	TransRadar	4.8	**99.9**	**57.7**	**36.1**	**61.9**	**63.9**	**99.9**	**73.2**	**53.1**	**76.5**	**75.6**
RA	FCN-8s [21]	134.3	99.8	14.8	0.0	23.3	34.5	99.9	25.8	0.0	37.8	40.9
	U-Net [29]	17.3	99.8	22.4	8.8	0.0	32.8	99.9	25.8	0.0	37.8	40.9
	DeepLabv3+ [4]	59.3	99.9	3.4	5.9	21.8	32.7	99.9	6.5	11.1	35.7	38.3
	RSS-Net [16]	10.1	99.5	7.3	5.6	15.8	32.1	99.8	13.7	10.5	27.4	37.8
	RAMP-CNN [9]	106.4	99.8	1.7	2.6	7.2	27.9	99.9	3.4	5.1	13.5	30.5
	MVNet [25]	2.4	98.8	0.1	1.1	6.2	26.8	99.0	0.0	7.3	24.8	28.5
	TMVA-Net [25]	5.6	99.8	26.0	8.6	30.7	41.3	99.9	41.3	15.9	47.0	51.0
	T-RODNet [14]	162.0	99.9	25.4	9.5	**39.4**	43.5	99.9	40.5	17.4	**56.6**	53.6
	PeakConv [42]	6.3	-	-	-	-	42.9	-	-		-	53.3
	TransRadar	4.8	**99.9**	**30.3**	**21.5**	38.2	**47.5**	**99.9**	**46.6**	**35.3**	55.3	**59.3**

Table 1. Semantic segmentation performance on the test split of the CARRADA dataset, shown for the RD (Range-Doppler) and RA (Range-Angle) views. Columns from left to right are the view (RD/RA), the name of the model, the number of parameters in millions, the intersection-over-union (IoU) score of the four different classes with their mean, and the Dice score for the same classes.
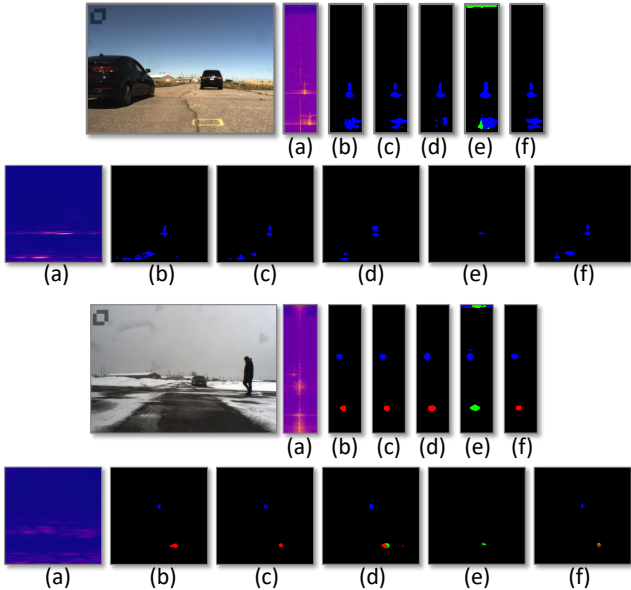

Figure 4. Qualitative results on two test scenes from the CAR-RADA test split showing the RGB camera view with results of semantic segmentation from different methods. For every image, (top) depicts the RD and (bottom) depicts RA. (a) RD/RA inputs, (b) ground-truth, (c) TransRadar, (d) TMVA-Net [25], (e) MVNet [25], and (f) UNet [29]. All RD outputs were rotated for visual coherency. Different colors correspond to different classes. Blue: Car, Green: Cyclist, Red: Pedestrian. Black: background.

which is tailored to radar data, to tackle various datasets.

Object detection on RADIal: Object detection results on the RADIal dataset are shown in Table 3. Our method outperforms all previously reported models in this task as well

with significantly higher AR and lower angular prediction error. Despite our method not being designed for the task of object detection, the model still sets a new record for this task. All taken into account, our model sets a new standard for state-of-the-art predictions in these two datasets.

Backbone	# Params. (m)	% mIoU
PolarNet [24]	-	60.6
FFTRadNet [28]	3.8	74.0
C-M DNN [15]	7.7	80.4
TransRadar	**3.4**	**81.1**

Table 2. Semantic segmentation results on the RADIal dataset [28]. Our method outperforms most recent state-of-the-art methods in both metrics. The best scores per column are in bold. '-' is an unreported value with no replicable results.

Backbone	% AP ↑	% AR ↑	R(m) ↓	A(°)↓
Pixor [37]	96.6	81.7	0.10	0.20
FFTRadNet [28]	96.8	82.2	**0.11**	0.17
C-M DNN [15]	96.9	83.5	-	-
TransRadar	**97.3**	**98.4**	**0.11**	**0.10**

Table 3. Object detection results on the RADIal dataset. Our method yields an increase in the average recall and a significant decrease in the angle regression error. The best scores per column are in bold. '-' is an unreported value with no replicable results.

5.5. Discussion & Ablation Study

Different Backbone Architectures: To evaluate the effect of using our loss function, we compare different

Architecture	Param. (M)	mIoU$_{RD}$(%)	mIoU$_{RA}$(%)
UNETR [10]	165.0	52.5	34.2
CSWin [5]	83.0	25.0	21.9
ViT [6]	238.9	28.5	36.9
UNet [29]	184.4	53.1	38.4
TMVA-Net [25]	5.6	60.7	43.1
No Adaptive	**4.0**	61.9	42.3
TransRadar	4.8	**63.9**	**47.5**

Table 4. Different backbones using our proposed loss configuration. The best scores are in bold. 'No Adaptive' refers to the implementation of our method where no offset or modulating to the axis sampling is introduced (i.e. straight-line rows and columns).

other backbones using the same configuration on the CARRADA dataset. Tested backbones include available state-of-the-art methods and other transformer architectures such as ViT [6], UNETR [10], ConViT [34], and CSWin Transformer [5]. This allows us to evaluate both the loss function with other state-of-the-art models and our adaptive-directional attention with other attention-based techniques. Table 4 lists the quantitative comparison between them. Other than TMVA-Net, models were implemented with the same encoding and decoding as our adaptive-directional attention block. We notice that our loss improves TMVA-Net's performance significantly in both RD and RA mIoU scores. TransRadar still outperforms all other attention models and shows that the sparse nature of the adaptive-directional attention yields the best results in radar perception. To evaluate the effect of the adaptive sampling, we implement our model by applying attention to unshifted and unmodulated axes. Adding adaptive-directional sampling yields an increase of 1.40% in the RD mIoU and a 4.04% increase in the RA mIoU, while using less parameters than previous state-of-the-art methods.

Model	mIoU$_{RD}$	mIoU$_{RA}$
Sampling only	62.9	47.4
Attention without sampling	63.0	43.3
Attention with normal sampling	64.1	45.7
TransRadar	63.9	47.5

Table 5. Ablation experiment for the adaptive-directional attention head. We report segmentation performance on CARRADA dataset in terms of mIoU for the RA and RD maps.

Ablation for the adaptive-directional attention: We also perform ablation experiments on the adaptive-directional attention head. We show the semantic segmentation performance on the test split of the CARRADA dataset in Table 5. Noticeably, attention contributes to the increments in RD map performance, while the directional sampling contributes to RA's mIoU.

Evaluation of Loss Functions: We further test the effect of the loss functions on the learning of our method, where we

Loss					RD	RA
$\mathcal{L}_{OC}$	$\mathcal{L}_{CL}$	$\mathcal{L}_{SD}$	$\mathcal{L}_{CoL}$	$\mathcal{L}_{MV}$	mIoU	mIoU
✓	✓			✓	3.7	7.5
✓	✓	✓			61.9	37.5
		✓		✓	61.2	45.9
✓		✓		✓	62.3	42.2
✓	✓	✓	✓		62.9	47.4
✓	✓	✓		✓	**63.9**	**47.5**

Table 6. Comparison of performance of loss functions. $\mathcal{L}_{OC}$ is the object centric-focal loss, $\mathcal{L}_{CL}$ is the class-agnostic object localization loss, $\mathcal{L}_{CA}$ is the sum of previous terms, $\mathcal{L}_{SD}$ is the soft Dice loss, $\mathcal{L}_{MV}$ is the multi-view range matching loss, and CoL is the coherence loss used in [25]. The best scores are in bold.

test our model under different combinations of the functions defined in Section 4.4. Removing $\mathcal{L}_{SD}$ yields poor prediction scores, which showcases its necessity in this task. Using our model without RA-RD coherence yields a poor RA score, while using a coherence loss boosts RA's score by at least 3.5%. We also report the effects of $\mathcal{L}_{OC}$ and $\mathcal{L}_{CL}$, separately, or both combined ($\mathcal{L}_{CA}$). Removing $\mathcal{L}_{OC}$ from the $\mathcal{L}_{CA}$ term reduces RD score heavily while removing $\mathcal{L}_{CL}$ from $\mathcal{L}_{CA}$ reduces RA score. Localization is a harder task in RA maps than it is in RD due to its larger resolution which results in a more pronounced effect from $\mathcal{L}_{CL}$. Lastly, we compare the effect of introducing our $\mathcal{L}_{MV}$ loss instead of the baseline coherence loss. Following our discussion in Section 4.4, $\mathcal{L}_{MV}$ remedies the problem of RA reducing RD's accuracy, where we notice an increase in the accuracy of RA without compromising RD scores.

6. Conclusion

We introduce a novel attention-based architecture for the task of semantic segmentation on radar frequency images, named TransRadar. Our method uses an adaptive-directional attention block and a novel loss function tailored to the needs of radar perception. Our model achieves state-of-the-art performance on two semantic segmentation radar frequency datasets, CARRADA [26] and RA-DIal [28], using a smaller model size. Our proposed method also achieves improved performance for the task of object detection in radar images.

Paths of future works include implementing approaches that fuse radar input with RGB images to produce more robust predictions. The ability to fuse both data sources is promising in creating a new standard for automotive driving.

References

[1] Kshitiz Bansal, Keshav Rungta, and Dinesh Bharadia. Rad-segnet: A reliable approach to radar camera fusion. *arXiv preprint arXiv:2208.03849*, 2022. 2

[2] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar robotcar dataset: A

radar extension to the oxford robotcar dataset. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Paris, 2020. 2

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2, 7

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 8

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4, 8

[7] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2021. 1, 2

[8] Xiangyu Gao, Guanbin Xing, Sumit Roy, and Hui Liu. Experiments with mmwave automotive radar test-bed. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1–6, 2019. 1

[9] Xiangyu Gao, Guanbin Xing, Sumit Roy, and Hui Liu. Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. *IEEE Sensors Journal*, 21(4):5119–5132, 2021. 2, 7

[10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 8

[11] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019. 4

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019. 4

[13] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. 5

[14] Tiezhen Jiang, Long Zhuang, Qi An, Jianhua Wang, Kai Xiao, and Anqi Wang. T-rodnet: Transformer for vehicular millimeter-wave radar object detection. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. 1, 2, 4, 7

[15] Yi Jin, Anastasios Deligiannis, Juan-Carlos Fuentes-Michel, and Martin Vossiek. Cross-modal supervision-based multi-task learning with automotive radar raw data. *IEEE Transactions on Intelligent Vehicles*, pages 1–15, 2023. 6, 7

[16] Prannay Kaul, Daniele de Martini, Matthew Gadd, and Paul Newman. Rss-net: Weakly-supervised multi-class semantic segmentation with fmcw radar. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 431–436, 2020. 7

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 4, 5

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 7

[22] Michael Meyer and Georg Kuschk. Automotive radar dataset for deep learning based 3d object detection. In *2019 16th European Radar Conference (EuRAD)*, pages 129–132, 2019. 1

[23] Farzan Erlik Nowruzi, Dhanvin Kolhatkar, Prince Kapoor, Fahed Al Hassanat, Elnaz Jahani Heravi, Robert Laganiere, Julien Rebut, and Waqas Malik. Deep open space segmentation using automotive radar. In *2020 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, pages 1–4, 2020. 1

[24] Farzan Erlik Nowruzi, Dhanvin Kolhatkar, Prince Kapoor, Elnaz Jahani Heravi, Fahed Al Hassanat, Robert Laganière, Julien Rebut, and Waqas Malik. Polarnet: Accelerated deep open space segmentation using automotive radar in polar domain. *CoRR*, abs/2103.03387, 2021. 7

[25] Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. Multi-view radar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15671–15680, 2021. 1, 2, 4, 5, 7, 8

[26] Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Pérez. Carrada dataset: Camera and automotive radar with range- angle- doppler annotations. In

2020 25th International Conference on Pattern Recognition (ICPR), pages 5068–5075, 2021. 1, 2, 4, 6, 8

[27] Andras Palffy, Jiaao Dong, Julian FP Kooij, and Dariu M Gavrila. Cnn based road user detection using the 3d radar cube. *IEEE Robotics and Automation Letters*, 5(2):1263–1270, 2020. 1

[28] Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17021–17030, 2022. 1, 2, 4, 6, 7, 8

[29] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]). 1, 7, 8

[30] Ole Schumann, Markus Hahn, Nicolas Scheiner, Fabio Weishaupt, Julius F Tilly, Jürgen Dickmann, and Christian Wöhler. Radarscenes: A real-world radar point cloud data set for automotive applications. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2021. 1, 2

[31] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016. 4

[32] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pages 108–126. Springer, 2020. 4

[33] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: Radar object detection using cross-modal supervision. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 504–513, 2021. 1, 2

[34] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 8

[35] Shuangjie Xu, Rui Wan, Maosheng Ye, Xiaoyi Zou, and Tongyi Cao. Sparse cross-scale attention network for efficient lidar panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2920–2928, 2022. 2

[36] Ruixiang Xue, Jianqiang Wang, and Zhan Ma. Efficient lidar point cloud geometry compression through neighborhood point attention. *arXiv preprint arXiv:2208.12573*, 2022. 2

[37] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. 7

[38] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022. 4

[39] Ao Zhang, Farzan Erlik Nowruzi, and Robert Laganiere. Raddet: Range-azimuth-doppler based radar object detection for dynamic road users. In *2021 18th Conference on Robots and Vision (CRV)*, pages 95–102, 2021. 1, 2

[40] Biao Zhang, Ivan Titov, and Rico Sennrich. Sparse attention with linear units. *arXiv preprint arXiv:2104.07012*, 2021. 2

[41] Guoqiang Zhang, Haopeng Li, and Fabian Wenger. Object detection and 3d estimation via an fmcw radar using a fully convolutional network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4487–4491. IEEE, 2020. 2

[42] Liwen Zhang, Xinyan Zhang, Youcheng Zhang, Yufei Guo, Yuanpei Chen, Xuhui Huang, and Zhe Ma. Peakconv: Learning peak receptive field for radar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17577–17586, 2023. 1, 2, 7

[43] Zhenyuan Zhang, Zengshan Tian, Ying Zhang, Mu Zhou, and Bang Wang. u-deephand: Fmcw radar-based unsupervised hand gesture feature learning using deep convolutional auto-encoder network. *IEEE Sensors Journal*, 19(16):6811–6821, 2019. 2

[44] Zhenyuan Zhang, Zengshan Tian, and Mu Zhou. Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor. *IEEE Sensors Journal*, 18(8):3278–3289, 2018. 2