applied sciences

MDPI

Article

Sparsity-Robust Feature Fusion for Vulnerable Road-User Detection with 4D Radar

Leon Ruddat [1], Laurenz Reichardt [1], Nikolas Ebert [1,2] and Oliver Wasenmüller [1,*]

[1] Research and Transfer Center CeMOS, Mannheim University of Applied Sciences, 68163 Mannheim, Germany; l.ruddat@hs-mannheim.de (L.R.); l.reichardt@hs-mannheim.de (L.R.); n.ebert@hs-mannheim.de (N.E.)
[2] Department of Computer Science, RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
* Correspondence: o.wasenmueller@hs-mannheim.de

Abstract: Detecting vulnerable road users is a major challenge for autonomous vehicles due to their small size. Various sensor modalities have been investigated, including mono or stereo cameras and 3D LiDAR sensors, which are limited by environmental conditions and hardware costs. Radar sensors are a low-cost and robust option, with high-resolution 4D radar sensors being suitable for advanced detection tasks. However, they involve challenges such as few and irregularly distributed measurement points and disturbing artifacts. Learning-based approaches utilizing pillar-based networks show potential in overcoming these challenges. However, the severe sparsity of radar data makes detecting small objects with only a few points difficult. We extend a pillar network with our novel Sparsity-Robust Feature Fusion (SRFF) neck, which combines high- and low-level multi-resolution features through a lightweight attention mechanism. While low-level features aid in better localization, high-level features allow for better classification. As sparse input data are propagated through a network, the increasing effective receptive field leads to feature maps of different sparsities. The combination of features with different sparsities improves the robustness of the network for classes with few points.

Keywords: 4D radar; 3D object detection; attention

Citation: Ruddat, L.; Reichardt, R.; Ebert, N.; Wasenmüller, O. Sparsity-Robust Feature Fusion for Vulnerable Road-User Detection with 4D Radar. *Appl. Sci.* **2024**, *14*, 2781. https://doi.org/10.3390/app14072781

Academic Editor: Yujin Lim and Hideyuki Takahashi

Received: 19 February 2024
Revised: 19 March 2024
Accepted: 22 March 2024
Published: 26 March 2024

1. Introduction

Detecting vulnerable road users (VRU), such as pedestrians, cyclists, etc., is a major challenge for autonomous vehicles. Vulnerable road users are defined in the ITS Directive of the Directorate-General for Mobility and Transport of the European Commission [1] as non-motorized road users, such as pedestrians and cyclists, as well as motorcyclists and persons with disabilities or reduced mobility and orientation. These road users are particularly challenging to detect due to their small size, especially when compared to other objects in a scene, frequent occlusion, and high vulnerability. For this reason, a wide variety of methods have been developed to detect these VRUs as robustly as possible, with such detection crucial for systems such as emergency braking assistants. These systems differ in terms of the sensor modality used. In scientific publications, the use of mono or stereo cameras is the most common [2], followed by 3D LiDAR sensors [3]. Both modalities show convincing results on scientific datasets under good weather conditions but are unfortunately limited in practical application. This is due to the strong dependence on environmental conditions [4]—night, rain, fog, etc.—and, in the case of LiDAR, to the substantial hardware costs.

While traditional 3D radar sensors have been an integral part of basic driver assistance systems for a long time, the recently available high-resolution 4D (or 3 + 1D) radar systems introduce new possibilities for advanced detection tasks. In 3D radars, the three dimensions are range, azimuth, and Doppler information. Azimuth and range can be interpreted as a 2D point cloud (x- and y-coordinates). Furthermore, the Doppler information can be used to

measure the speed of the detected object relative to the sensor. This limited spatial resolution is not sufficient to reliably detect VRUs. In certain situations, VRUs may be mistaken for other objects, making them undetectable by the sensor. In contrast, 4D radars provide four dimensions: range, azimuth, elevation, and Doppler information. The additional dimension compared to 3D radars is the elevation, which provides the ability to compute a 3D point cloud (x-, y-, and z-coordinates) that includes velocity, as shown in Figure 1. This new category of radar sensors allows 3D detection algorithms to be applied to these sensors. Compared to 3D radar sensors, 4D radars are much denser while maintaining robustness in challenging weather conditions. Based on their characteristics, these 4D radar sensors are well suited for VRU detection but are currently still significantly under-researched compared to competing camera or LiDAR approaches. This is partly because sensors have only recently become widely available and partly because large-scale datasets that include annotations, such as View-of-Delft [5], have only recently become accessible.

For this reason, radar sensors are often used in the field, even though the point clouds provided are significantly sparser compared to LiDAR sensors. Radars combine the advantages of relatively low costs with great robustness against weather influences.

Figure 1. The detection of vulnerable road users (VRU) from 4D radar data (colored dots) is particularly challenging due to the sparsity of radar sensor data, especially for small objects. With our SRFF, we present a method that specifically targets the detection (green boxes) of these objects.

At the same time, 4D radar sensors also present various challenges. The toughest challenge might be the limited number of 3D measurement points obtained from the sensors. In addition, these measurement points are distributed very irregularly, depending on the scene geometry. VRUs are at a natural disadvantage, as larger objects such as cars or trucks have a higher point density compared to cyclists. In addition, there are also disturbing artifacts caused by reflections and the multipath effect [6]. While classical rule-based approaches have proven to be effective at extracting geometric features such as lines [7,8], the detection of entire objects remains challenging. Consequently, classical rule-based approaches are not suitable for detection tasks. Learning-based approaches, in particular, show their full potential in this regard.

In the recent state of the art for point cloud detection, pillar-based deep learning methods have proven to be particularly effective [9]. We extend a pillar network with a multi-resolution neck to enable the detection of smaller objects, such as VRUs, with just a few points. The high-resolution features enable regions of interest (ROIs) to detect smaller

objects already in the head. In addition, we include our novel Sparsity-Robust Feature Fusion (SRFF), which combines high- and low-level features with a lightweight attention mechanism. Here, low-level features aid in better localization, while high-level features allow for better classification. The multi-level features each have different sparsities due to their effective receptive field. Our combination of features with different sparsities makes the network more robust for classes with few points. All of these characteristics enable us to accurately identify VRUs in 4D radar data. For instance, this capability allows for operation in adverse weather conditions where other sensor systems may not function correctly.

Section 2 discusses the current state of the art and highlights the most important related works. Following this, Section 3 presents our proposed network and SRFF. Section 4 presents both the qualitative and quantitative evaluations, which are then discussed in Section 5.

2. Related Works

The processing of radar data remains a challenge due to the low density and noisiness of the captured information. This severe limitation in resolution can lead to the issue that objects are only captured partially, leading to the limited classification abilities of current algorithms and incorrect bounding box sizes. For this reason, some methods include segmentation or focus only on localization [10]. To counteract this sparsity, some datasets overlap data from multiple sensors [11,12]. In addition to this inherent sparsity, conventional radar data (2D + 1D) is not able to represent the height of an object, making it difficult to separate static objects from environmental clutter and introducing errors when calibrating with respect to other sensors [13]. Additionally, ghost targets can be difficult to distinguish, with some research focusing exclusively on the detection of these radar errors [6]. While modern 4D radar sensors are able to capture height, data remain sparse. In addition to these difficulties, there are various representation methods for radar data, leading to a variety of deep learning methods.

2.1. Three-Dimensional Radar Perception

Due to the lack of height information, conventional 3D radar methods (sometimes referred to as 2D+1D) usually take a bird's-eye-view (BEV) approach. The 2D-BEV approach by Dreher et al. [14] combines a classification network with a segmentation network for semantic masking to provide input for a box regression network for detection. Meyer et al. [15] instead used a graph convolution network with a 2D-CNN, processing range–azimuth–Doppler information for object detection.

Sensor fusion is a method that has proven to counteract the sparsity of LiDAR data [16] and is a natural approach to complement radars' missing height information. Image data are dense and include height information but lack depth information, which radar can provide. LiDAR sensors provide 3D data, and while denser than radar data, these data are sparser compared to image data. Additionally, a large part of the rotating LiDAR data is not used during fusion, as the horizontal field of view of radar sensors only covers a part of the 360° data.

RODNet [17] uses an image–radar teacher network to teach a temporal range–azimuth radar network for object localization. RADDET [18] additionally includes Doppler information and uses an image–radar network to perform pseudo-labeled instance annotations. RAMP-CNN [19] utilizes three different temporal views and networks, which are fused and supervised by an image network. CenterFusion [20] uses an image network for monocular object detection, refining the results through frustum association with radar data. Zhou et al. [21] combined image and BEV networks, classical filtering, temporal radar information, and fusion in an ROI-based approach for object detection. RadarNet [22] combines LiDAR and radar in a voxel-based approach for bird's-eye-view detection. Other researchers have combined image, radar, and LiDAR modalities [23,24], even including temporal information [25], to obtain dense data.

2.2. Four-Dimensional Radar Perception

The recent advent of 4D radar and the introduction of height information have made the adaption of 3D point cloud methods, such as frameworks for 3D voxelization, to radar-only detection approaches possible.

The pillar representation divides a point cloud into an even grid. However, in contrast to voxels, each grid cell has an unlimited height, shaped like a pillar, summarizing all points inside that cell. The reduction of height allows for a 2D bird's-eye representation over the point cloud. Pillar networks apply a 2D backbone to this representation. The multi-scale backbone features are further processed by a network neck, followed by detection heads to output detection predictions. Pillar networks have proven effective, especially for accurate and robust real-time point cloud detection.

RPFA-Net [26] and VoD [5] make use of this concept and adapt the PointPillars architecture [9] in combination with attention mechanisms to create 2D pseudo-images, which are processed in a BEV perspective by a 2D-CNN. Bai et al. [27] used vector attention [28].

Four-dimensional data are comparatively new, and for this reason, there is limited research in this area. Four-dimensional radar still suffers from sparsity, and as such, there is a natural imbalance between classes (large objects such as cars have more points than small objects such as pedestrians). Attention mechanisms have been studied for a single layer but remain under-explored in aggregating multi-scale information. VRU detection has been studied with other sensor modalities [3], but it has yet to be studied with 4D radar data.

3. Method

In this section, we introduce our novel Radar Pillar Feature Fusion Network. We base our approach on PointPillars [9] for 3D object detection in LiDAR point clouds, extending the network with our novel Sparsity-Robust Feature Fusion (SRFF). The propagation of sparse data through a neural network is heavily influenced by the receptive fields of individual layers. Later layers with a larger effective receptive field have more valid information within their feature maps [29]. For this reason, SRFF uses multi-scale features to enhance the network's ability to detect particularly small objects with few points, e.g., pedestrians, in sparse radar point clouds. The complete network architecture of our Radar Pillar Feature Fusion Network is shown in Figure 2.

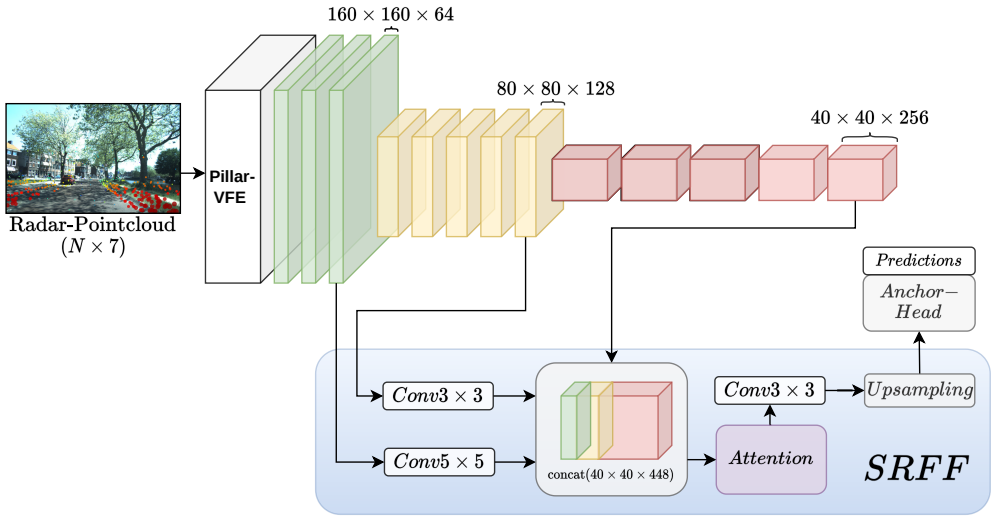

Figure 2. Representation of our Radar Pillar Feature Fusion Network. First, pseudo-images are created from the radar point cloud. Subsequently, the backbone is used to generate deep semantic features, which are, in turn, fused by the Sparsity-Robust Feature Fusion (SRFF) to ensure robust pedestrian detection by the anchor head.

3.1. Sparsity-Robust Radar Pillar Feature Fusion Network

Unlike the original PointPillars network, our Radar Pillar Feature Fusion Network uses the unprocessed and sparse radar point clouds in the form of N points $\times$ 6 as input, as shown in Figure 2. Here, 6 refers to the amount of information recorded for each point. Each point can be described by $p = [x, y, z, v_r, v_{rc}, RCS]$, where (x, y, z) are the spatial coordinates, v_r is the radial Doppler velocity, v_{rc} is the radial velocity compensated with ego-motion, and RCS is the radar cross-section. We provide comprehensive details on the data used in Section 4.1. These sparse irregular point clouds are first converted into voxels via Pillar-Voxel Feature Extraction (Pillar-VFE) [30]. Subsequently, these voxels are used to generate a two-dimensional pseudo-image of size $160 \times 160 \times 64$, which serves as input to a CNN backbone for generating deep semantic features. Note that a large number of pixels in the pseudo-image are empty, as voxels are not occupied. The backbone, depicted in Figure 2, consists of three stages, each spatially reducing the pseudo-image by a factor of 2 per stage and doubling the number of channels. We use resolutions $i \in \{1, 2, 3\}$ as $(H_i, W_i) \in M_i \times M_i$ for each stage, with $M = \{160, 80, 40\}$, respectively, and the corresponding feature depths $d_i = \{64, 128, 256\}$. Each layer in the 2D backbone is composed of a 3×3 convolution, batch normalization [31], and a non-linear activation in the form of a ReLU function. The final features of each backbone stage are fused using our novel Sparsity-Robust Feature Fusion (see Section 3.2). These fused features are processed again with a 3×3 convolutional layer and upsampled to the desired resolution. The resolution depends on the corresponding network configuration, and we test the resolutions $\{64, 128, 256\}$. Our experiments in Section 4.3 show that a higher resolution leads to better detection of small objects, which is why the final resolution is preferably set to 160×160. In the final step, we use two convolution layers as an anchor-based detection head, supporting the high resolution of input features.

We choose an anchor-based RPN head following our SRFF neck for all experiments.

3.2. Sparsity-Robust Feature Fusion

The main contribution of our Radar Pillar Feature Fusion Network is the Sparsity-Robust Feature Fusion (SRFF) neck, which is used to fuse high- and low-level features from our backbone via a lightweight attention module. Early features from a network backbone capture low-level patterns and provide localization, while deeper features represent high-level concepts and allow for better classification. Sparse data can pose a challenge for traditional convolution layers as initially, most of the information in the filter's receptive field is empty. With progressive layers, the valid information is propagated through the network and increases in density [16,29,32]. For this reason, while early low-level features can represent small structures due to their large spatial size, the sparsity of this information can be complemented by high-level features with denser information. The multi-level features each have different sparsities due to their effective receptive field. The combination of these features makes the network more robust for classes and small structures with few points.

The input of the SRFF, as shown in Figure 2, is the aggregated outputs of each backbone stage, each with a different scale. The SRFF module is inspired by the work of Shi et al. [2], which follows a similar approach in the context of 2D object detection. However, our approach is specially adapted to the processing of sparse radar data. For SRFF, several feature maps of different shapes serve as input to the module. By applying convolutions, these are standardized to have a uniform height and width and then concatenated. For the largest feature maps with a resolution of 160×160, a convolution layer of 5×5 is used, whereas for those with a resolution of 80×80, a 3×3 layer is used. The stride s of each layer is calculated as $s = \frac{k+1}{2}$, where k refers to the size of the corresponding filter kernel. The resulting feature is then passed to the attention module, which refines the fused features. The output of the attention module is also the output of the SRFF. During development, two possible lightweight attention mechanisms were used and compared in an ablation study. Specifically, we compared the Convolutional Block Attention Module (CBAM) [33]

and Channel-Spatial Attention Mechanism (CSAM) [34]. The CBAM (Figure 3a) is used to selectively emphasize or suppress the relevant channel and spatial features of a given image and is composed of two sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). The CAM sub-module aggregates the spatial information of its input x by utilizing average ($AvgPool$) and max ($MaxPool$) pooling operations on two vectors of the shape $1 \times 1 \times C$, with C corresponding to the channel depth. The inter-dependencies between channels are modeled by a multi-layer perceptron (MLP), followed by the sigmoid activation σ to produce a channel-wise attention vector. The shared MLP with one hidden layer reduces the parameter overhead with a hidden activation size of C/r, where r is the reduction ratio. The channel attention A_c can be computed as:

$$A_c(x) = \sigma(MLP(AvgPool(x)) + MLP(MaxPool(x))) \tag{1}$$

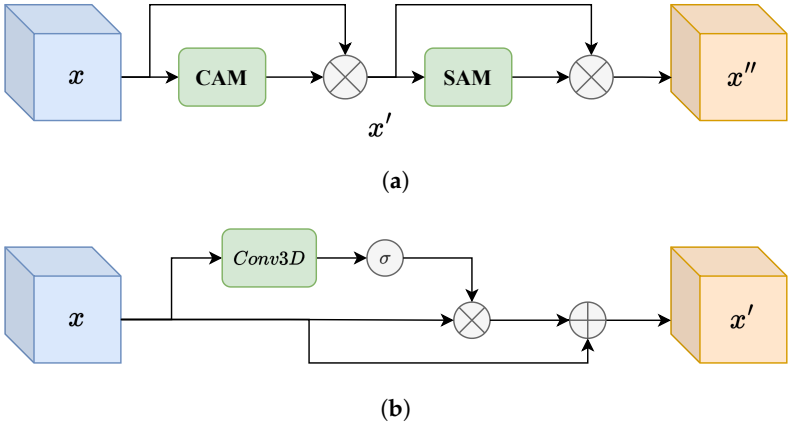

Figure 3. Overview of CBAM [33] and CSAM [34] architectures. (**a**) CBAM architecture [33]. (**b**) CSAM architecture [34].

The output of the CAM is used to modulate and reweight the original features x, resulting in x'. The SAM functions in a similar manner, compressing the channels at each spatial location by utilizing spatial average and max pooling along the channel axis and concatenating them to generate a feature descriptor. The resulting two-channel map is processed by a large kernel convolution to produce a spatial attention map A_s. This spatial attention map captures the relevance of each spatial location with respect to its large kernel neighborhood and is multiplied with the original features x' for spatial reweighting, resulting in x''. The spatial attention A_s can be computed as:

$$A_s(x) = \sigma(Conv2D([AvgPool(x'); MaxPool(x')])) \tag{2}$$

Thus, the final output x'' of the CBAM can be calculated accordingly:

$$x'' = A_s(x') \cdot x' \quad \text{with} \quad x' = A_c(x) \cdot x \tag{3}$$

The CSAM is very similar to the CBAM but differs in the way it obtains attention maps. In the CBAM, the channel and spatial attention maps are computed separately and then multiplied element-wise with the original features. In contrast, the CSAM computes feature-wise attention maps through a 3D convolution layer, followed by a sigmoid activation σ, considering both inter-channel and inter-spatial dependencies. The attention map is multiplied by the input features and a scale factor, β, before being added to the original features. The scale factor is initialized to 0, enabling the network to adaptively include attention information. This means that the CSAM provides more flexibility in determining the relative importance of the channel and spatial attention maps, whereas the CBAM assumes equal importance for both maps. The CSAM can be computed as:

$$x' = \beta\sigma(Conv3D(x) \cdot x + x) \tag{4}$$

However, the CBAM has been shown to outperform the CSAM in some benchmark datasets, suggesting that its approach of multiplying attention maps sequentially may be more effective in some cases. For this reason, in our evaluation, we examine both modules for their usability for radar data.

4. Evaluation

We conduct an evaluation and ablation study of our novel method, focusing on the detection of VRUs from 4D data. In the ablation study, we examine the impact of multi-scale information and the resolution of SRFF on detecting from sparse data. We utilize the View-of-Delft (VoD) [5] dataset, which provides calibrated and synchronized LiDAR, camera, and 4D radar data for 3D object detection of road-user classes.

4.1. Radar Data

The annotations for static and moving road users are represented by three-dimensional bounding boxes. In our work, we exclusively use 4D radar data. As mentioned in Section 3.1, the radar outputs a point cloud of shape $N \times 6$ for each scan, where N represents the number of points and 6 refers to the recorded information per point. Each point in the dataset can be described by a set of parameters, denoted as $p = [x, y, z, v_r, v_{rc}, RCS]$, where (x, y, z) represent the spatial coordinates, v_r denotes the Doppler velocity, v_{rc} corresponds to the radial velocity compensated with proper motion, and RCS refers to the radar cross-section. The VoD dataset allows for the use of radar data accumulated over three or five scans.

The VoD dataset annotates 13 classes across 8693 frames. Each object within the FOV of the camera and within 50 m of the recording vehicle is labeled with a bounding box that includes six degrees of freedom. In total, the VoD dataset contains 26,587 *pedestrian*, 10,800 *cyclist*, and 26,949 *car* labels.

4.2. Training Details

In this section, we explain the training procedure for our SRFF network and its variants. All models are implemented in the OpenPCDET framework [30]. Each model is trained for 80 epochs on the training set of the VoD dataset with a total batch size of 8. Following the methodology described in [5], we use the accumulated five scans in our experiments. The learning rate (LR) is defined using the one-cycle policy [35] with a minimum LR of 0.001 and a maximum LR of 0.01. For optimization, we use the Adam [36] algorithm. We utilize random mirroring of the point cloud along the x-axis and randomized scaling by a factor of 0.95–1.05 as data augmentation. Furthermore, only random mirroring of the point cloud on the x-axis of the recording vehicle and randomized scaling are used for data augmentation. We choose voxel sizes of $0.16, 0.16$, and 5 m for x, y, and z, respectively. Because the VoD dataset does not label beyond 50 m and the radar sensor has a limited receptive field, point clouds are clipped to 0 m–51.2 m for the x-axis, $\pm$25.6 m for the y-axis, and -3 m–2 m for the z-axis prior to voxelization. We use a maximum of $16,000$ voxels during training and $40,000$ during testing. All training runs are performed on a single Nvidia RTX 3060.

4.3. Results

The evaluation employed the Mean Average Precision (mAP) with the three-dimensional Intersection over Union (IoU), which was calculated analogously to the evaluation of the KITTI [37] dataset. IoU and mAP are two metrics commonly used in object detection tasks. IoU quantifies the overlap between the predicted and ground-truth bounding boxes to determine whether a predicted bounding box correctly localizes an object. mAP considers both the precision and recall of a model across multiple classes and aggregates IoU scores to provide an overall assessment of detection performance. The threshold for correct predictions was defined as an overlap of the predicted bounding box with the ground truth of 50% for *cars* and an overlap of 25% for *pedestrians* and *cyclists*. Two areas in the View-of-Delft evaluation split were evaluated: the entire annotated area and the driving corridor, defined as a rectangle at ground level in front of a driving train. Thus, only objects

that were at most four meters to the left or right and at most 25 m forward from the origin of the camera were analyzed.

The results of our evaluation study are depicted in Table 1. We studied the impact of our SRFF neck, comparing its performance to the baseline. At the time of writing, the PointPillars version by Palffy et al. [5] is the only comparable method following the original method that uses the accumulated five scans for training and evaluation. Additionally, we studied a *combined* version, where we fused the feature maps at three different resolutions ($40 \times 40, 80 \times 80$, and 160×160), followed by three separate detection heads.

Table 1. Evaluation of the influence of our SRFF module on the PointPillars [9] detection network, using different resolutions. We trained and evaluated on the VoD [5] dataset.

		Entire Annotated Area				**Driving Corridor**			
Class / **Metric**	**SRFF Resolution**	**Pedestrians** AP_{25}	**Cyclists** AP_{25}	**Cars** AP_{50}	**All** mAP	**Pedestrians** AP_{25}	**Cyclists** AP_{25}	**Cars** AP_{50}	**All** mAP
PointPillars [9]	-	35.8	**65.3**	40.2	47.1	45.5	**85.4**	72.2	67.7
+*SRFF-CBAM*	160×160	**36.8**	65.0	36.7	46.2	**47.2**	84.3	69.1	66.9
+*SRFF-CBAM*	80×80	29.7	60.2	33.8	41.2	39.0	73.2	68.5	60.2
+*SRFF-CBAM*	40×40	30.6	63.1	35.8	43.2	43.4	81.4	68.7	64.5
+*SRFF-CBAM*	Comb	32.9	63.3	36.1	44.1	43.2	76.9	67.9	62.7
+*SRFF-CSAM*	160×160	34.2	64.6	39.5	46.1	43.8	84.5	71.1	66.5
+*SRFF-CSAM*	80×80	31.9	60.2	34.3	42.2	39.2	81.9	68.8	63.3
+*SRFF-CSAM*	40×40	29.7	62.0	32.1	41.2	36.5	81.1	65.7	61.1
+*SRFF-CSAM*	Comb	30.3	59.4	34.3	41.4	40.9	81.1	68.4	63.4

Most notably, our results show that the resolution of the neck output had a large impact. For most comparisons, larger neck feature maps significantly improved not only the detection of VRUs but also the overall mAP. The largest increase was observed when using the CSAM, with the 160×160 version improving the mAP by 6 over the 80×80 variant. Larger neck features proved beneficial overall for the pillar network. Additionally, the choice of lightweight attention fusion had a large impact on the overall performance of smaller network neck resolutions. The CBAM outperformed the CSAM in most 40×40 and 80×80 variants. A probable explanation for this is that the initial channel reweighting in the CBAM already benefits the following spatial attention mechanism, which is not possible in the CSAM due to the combined attention calculation. Additionally, the 3D convolution kernel has a much smaller receptive field than the kernels in the CBAM, especially when reweighting the channel features. The *combined* variant of our network performed better than the small neck output feature, but worse than the 160×160 variant. This indicates that the addition of small-resolution neck features can actually be detrimental to overall network performance.

When observing VRUs, multi-scale feature fusion improved detection over the baseline. Attention fusion with the CBAM performed better than the CSAM. In the 160×160 variant, *pedestrian* detection improved by 1 mAP over the baseline, with only an additional 0.2 M parameters (ca. 1.1% of the total 18 M), whereas *cyclist* detection was the same. When considering only the driving corridor, *pedestrian* detection improved by 1.7 mAP. The results indicate that multi-resolution feature fusion is beneficial in detecting small objects.

The qualitative results in Figures 1 and 4 show that even with very few points or partial occlusion, our network reliably captured VRUs. The results were consistent in terms of visual quality for both the driving corridor and the entire scene. Notably, the height of the pedestrians was accurately determined, which can largely be attributed to the third dimension in 4D radar data. However, the images also reveal some incorrect predictions. Most notably, in the top-right image, a pedestrian was incorrectly predicted, where some points of the first car were misclassified. Another example is the missed cars in the top-left image.

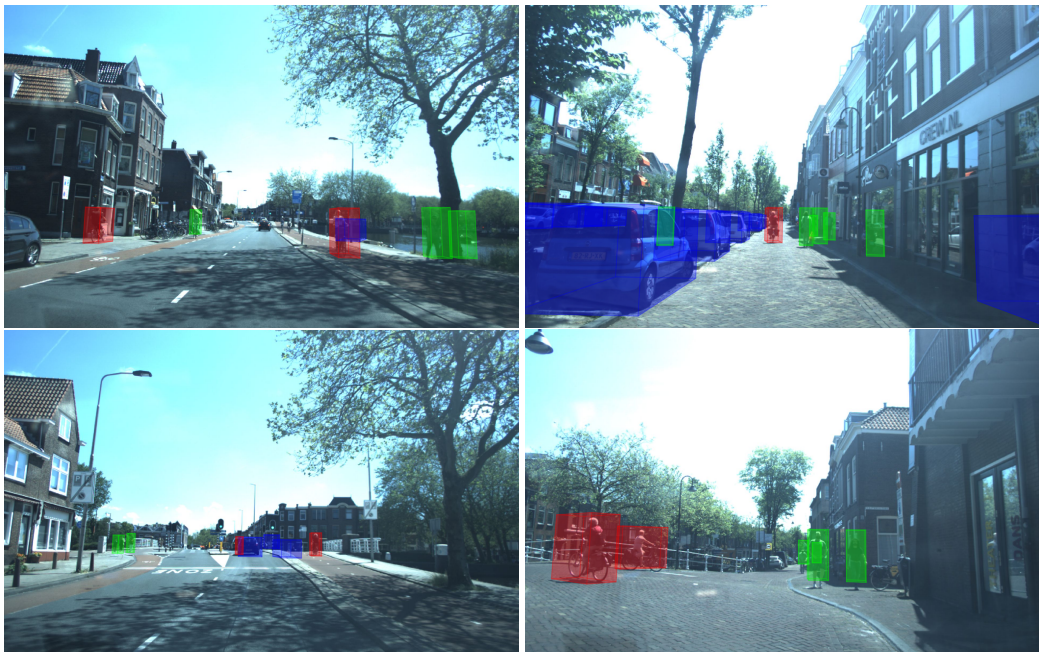

Figure 4. Qualitative evaluation of our pillar-based network with the novel SRFF module on the View-of-Delft dataset [5]. The network is able to detect smaller objects with just a few points from 4D radar data. The classes are color-coded: *cars* (blue), *pedestrians* (green), and *cyclists* (red).

5. Conclusions and Discussion

In this paper, we extended a pillar detection network with our novel Sparsity-Robust Feature Fusion (SRFF) neck to detect smaller objects with just a few points from 4D radar data. Our novel Sparsity-Robust Feature Fusion (SRFF) combines high- and low-level features from multiple resolutions to enhance both localization and classification. The low-level features aid in better localization, whereas the high-level features allow for better classification. The effective receptive field of the backbone layers significantly impacts the amount of valid data in feature maps produced from sparse data. Through this multi-resolution approach, the network is more robust for classes with few points, as it combines features with different sparsities due to their effective receptive field. We studied lightweight attention mechanisms for an intelligent fusion of these features, improving the detection performance of vulnerable road users with negligible additional overhead. Our contributions are supported by extensive experiments on the well-known View-of-Delft dataset.

At this point, we want to emphasize that our results are based on 4D radar sensors only. These have the advantage that the results are independent of weather conditions and also work reliably at night and in the rain. This cannot be proven quantitatively with the View-of-Delft dataset, but the physical properties of the sensors confirm this claim [5]. When comparing our results with detections from high-resolution images or quasi-dense LiDAR point clouds under good visibility conditions, our method exhibits worse accuracy. In bad weather conditions, however, various publications have shown that the accuracy of camera- and LiDAR-based methods drops significantly and sometimes approaches zero [38]. This property of 4D radar makes it suitable for standalone systems or in synergy with other sensors in order to compensate for their disadvantages. For maximum robustness, a combination of different sensor types (e.g., LiDAR or camera) is recommended to compensate for the disadvantages of the individual sensors. This is why we are convinced that we have made an important contribution to research in the automotive sector with a purely radar-based processing system. In the context of future work, our method could be explored in related modalities, such as LiDAR or depth sensing. Additionally, the lightweight nature of our method lends itself to subsequent research on real-time detection.

Author Contributions: Conceptualization, L.R. (Leon Ruddat), L.R. (Laurenz Reichardt), N.E. and O.W.; methodology, L.R. (Leon Ruddat), L.R. (Laurenz Reichardt), N.E. and O.W.; software, L.R. (Leon Ruddat) and L.R. (Laurenz Reichardt); validation, L.R. (Leon Ruddat) and L.R. (Laurenz Reichardt); formal analysis, L.R. (Leon Ruddat) and L.R. (Laurenz Reichardt); investigation, L.R. (Leon Ruddat); resources, L.R. (Leon Ruddat), L.R. (Laurenz Reichardt), N.E. and O.W.; data curation, L.R. (Leon Ruddat) and L.R. (Laurenz Reichardt); writing—original draft preparation, L.R. (Laurenz Reichardt), N.E. and O.W.; writing—review and editing, L.R. (Leon Ruddat), L.R. (Laurenz Reichardt), N.E. and O.W.; visualization, L.R. (Laurenz Reichardt), N.E. and O.W.; supervision, N.E. and O.W.; project administration, N.E. and O.W.; funding acquisition, O.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the Federal Ministry of Education and Research, Germany, under the project PreciRaSe (01IS23023B).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: VoD data were obtained from Plaffy et al. [5] and are available at https://github.com/tudelft-iv/view-of-delft-dataset/ (accessed on 30 October 2023) . All rights are held by Delft University of Technology, and the data can be used under the following license https://intelligent-vehicles.org/datasets/view-of-delft/view-of-delft-dataset-research-use-license/ (accessed on 30 October 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mobility and Transport. *Mobility and Transport ITS & Vulnerable Road Users*; Standard; European Commission: Brussels, Belgium, 2011.
2. Shi, Y.; Fan, Y.; Xu, S.; Gao, Y.; Gao, R. Object detection by attention-guided feature fusion network. *Symmetry* **2022**, *14*, 887. [CrossRef]
3. Fürst, M.; Wasenmüller, O.; Stricker, D. LRPD: Long range 3d pedestrian detection leveraging specific strengths of lidar and rgb. In Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020.
4. Yoshida, T.; Wasenmüller, O.; Stricker, D. Time-of-flight sensor depth enhancement for automotive exhaust gas. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1955–1959.
5. Palffy, A.; Pool, E.; Baratam, S.; Kooij, J.F.; Gavrila, D.M. Multi-class road user detection with 3+ 1D radar in the View-of-Delft dataset. *IEEE Robot. Autom. Lett.* **2022**, *7*, 4961–4968. [CrossRef]
6. Chamseddine, M.; Rambach, J.; Stricker, D.; Wasenmüller, O. Ghost target detection in 3d radar data using point cloud based deep neural network. In Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 10398–10403.
7. Liu, Z.; Xin, X.; Xu, Z.; Zhou, W.; Wang, C.; Chen, R.; He, Y. Robust and Accurate Feature Detection on Point Clouds. *Comput.-Aided Des.* **2023**, *164*, 103592. [CrossRef]
8. Xin, X.; Huang, W.; Zhong, S.; Zhang, M.; Liu, Z.; Xie, Z. Accurate and Complete Line Segment Extraction for Large-Scale Point Clouds. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *36*, 54–68. [CrossRef]
9. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
10. Zhou, Y.; Liu, L.; Zhao, H.; López-Benítez, M.; Yu, L.; Yue, Y. Towards deep radar perception for autonomous driving: Datasets, methods, and challenges. *Sensors* **2022**, *22*, 4208. [CrossRef] [PubMed]
11. Schumann, O.; Hahn, M.; Scheiner, N.; Weishaupt, F.; Tilly, J.F.; Dickmann, J.; Wöhler, C. RadarScenes: A real-world radar point cloud data set for automotive applications. In Proceedings of the IEEE International Conference on Information Fusion (FUSION), Sun City, South Africa, 1–4 November 2021; pp. 1–8.
12. Bansal, K.; Rungta, K.; Zhu, S.; Bharadia, D. Pointillism: Accurate 3D bounding box estimation with multi-radars. In Proceedings of the Conference on Embedded Networked Sensor Systems, Virtual, 16–19 November 2020; pp. 340–353.
13. Peršić, J.; Petrović, L.; Marković, I.; Petrović, I. Spatio-temporal multisensor calibration based on gaussian processes moving object tracking. *arXiv* **2019**, arXiv:1904.04187.
14. Dreher, M.; Erçelik, E.; Bänziger, T.; Knol, A. Radar-based 2D car detection using deep neural networks. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; pp. 1–8.

15. Meyer, M.; Kuschk, G.; Tomforde, S. Graph convolutional networks for 3d object detection on radar data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) **2021**, Montreal, QC, Canada, 10–17 October 2021; pp. 3060–3069.
16. Reichardt, L.; Mangat, P.; Wasenmüller, O. DVMN: Dense validity mask network for depth completion. In Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021.
17. Wang, Y.; Jiang, Z.; Gao, X.; Hwang, J.N.; Xing, G.; Liu, H. Rodnet: Radar object detection using cross-modal supervision. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 504–513.
18. Zhang, A.; Nowruzi, F.E.; Laganiere, R. RADDet: Range-Azimuth-Doppler based radar object detection for dynamic road users. *Conference on Robots and Vision (CRV)*, Burnaby, BC, Canada, 26–28 May 2021; pp. 95–102.
19. Gao, X.; Xing, G.; Roy, S.; Liu, H. Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. *IEEE Sens. J.* **2020**, *21*, 5119–5132. [CrossRef]
20. Nabati, R.; Qi, H. Centerfusion: Center-based radar and camera fusion for 3d object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 1527–1536.
21. Zhou, T.; Chen, J.; Shi, Y.; Jiang, K.; Yang, M.; Yang, D. Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection. *IEEE Trans. Intell. Veh.* **2023**, *8*, 1523–1535. [CrossRef]
22. Yang, B.; Guo, R.; Liang, M.; Casas, S.; Urtasun, R. Radarnet: Exploiting radar for robust perception of dynamic objects. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 496–512.
23. Drews, F.; Feng, D.; Faion, F.; Rosenbaum, L.; Ulrich, M.; Gläser, C. DeepFusion: A Robust and Modular 3D Object Detector for Lidars, Cameras and Radars. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 560–567.
24. Nobis, F.; Shafiei, E.; Karle, P.; Betz, J.; Lienkamp, M. Radar voxel fusion for 3D object detection. *Appl. Sci.* **2021**, *11*, 5598. [CrossRef]
25. Wang, L.; Chen, T.; Anklam, C.; Goldluecke, B. High dimensional frustum pointnet for 3d object detection from camera, lidar, and radar. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 1621–1628.
26. Xu, B.; Zhang, X.; Wang, L.; Hu, X.; Li, Z.; Pan, S.; Li, J.; Deng, Y. RPFA-Net: A 4D radar pillar feature attention network for 3D object detection. In Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 3061–3066.
27. Bai, J.; Zheng, L.; Li, S.; Tan, B.; Chen, S.; Huang, L. Radar transformer: An object classification network based on 4D MMW imaging radar. *Sensors* **2021**, *21*, 3854. [CrossRef] [PubMed]
28. Zhao, H.; Jia, J.; Koltun, V. Exploring self-attention for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10076–10085.
29. Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; Geiger, A. Sparsity invariant cnns. In Proceedings of the International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 11–20.
30. Team, O.D. OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds. Available online: https://github.com/open-mmlab/OpenPCDet (accessed on 30 October 2023).
31. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning (ICML), Lile, France, 6–11 July 2015.
32. Yan, L.; Liu, K.; Belyaev, E. Revisiting sparsity invariant convolution: A network for image guided depth completion. *IEEE Access* **2020**, *8*, 126323–126332. [CrossRef]
33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
34. Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; Shen, H. Single image super-resolution via a holistic attention network. In Proceedings of the European Conference on Computer Vision (ECCV), Heraklion, Greece, 5–11 September 2010.
35. Smith, L.N.; Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Baltimore, MD, USA, 14–18 April 2019.
36. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
37. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
38. Sheeny, M.; De Pellegrin, E.; Mukherjee, S.; Ahrabian, A.; Wang, S.; Wallace, A. RADIATE: A radar dataset for automotive perception in bad weather. In Proceedings of the International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021.