2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

TransRSS: Transformer-based Radar Semantic Segmentation

Hao Zou[1], Zhen Xie[1*], Jiarong Ou[1], and Yutao Gao[1]

***Abstract*— Radar semantic segmentation is a challenging task in environmental understanding, due as the radar data is noisy and suffers measurement ambiguities, which could lead to poor feature learning. To better tackle such difficulties, we present a novel and high-performance Transformer-based Radar Semantic Segmentation method, named TransRSS, to effectively and efficiently feature extraction for radar segmentation. Our approach first introduces the transformer into radar semantic segmentation and deeply integrates the merits of the Convolutional Neural Network (CNN) and transformer to extract more discriminative and global-level semantic features. On the one hand, it takes advantage of the CNN with flexible receptive fields to process images thanks to the shift convolution scheme. On the other hand, it takes advantage of the transformer to model long-range dependency with the self-attention mechanism. Meanwhile, we propose a Dual Position Attention module to aggregate rich context interdependencies between the multi-view features, which achieves an implicit mechanism for adaptively feature aggregation. Extensive experiments on the CARRADA dataset and RADIal dataset demonstrate that our TransRSS surpasses the state-of-the-art (SOTA) radar segmentation methods with remarkable margins.**

I. INTRODUCTION

Radar semantic scene understanding has been receiving increasing attention from industry and academia thanks to its wide applications in various fields such as autonomous driving and smart city. Different from camera and LiDAR, frequency modulated continuous wave (FMCW) radar is a more robust sensor under severe conditions, which operates in the millimeter-wave (MMW) band (30-300GHz), leading to an excellent capability of penetrating through adverse weather and precise range detection ability. Furthermore, thanks to dense virtual antenna arrays, high definition (HD) imaging radar tackles the impact of poor angular resolution on downstream perception tasks to a certain extent [1].

Raw data of FMCW radar is analog-to-digital (ADC) signals, which are hardly understood for downstream tasks. After a series of fast Fourier transforms (FFTs) and peak detection [2], the ADC signals are able to generate radar frequency (RF) data and radar point cloud data, respectively. As a 3D tensor, radar frequency data can illustrate the range, angle, and velocity characteristics, usually referred to as Range-Azimuth-Doppler (RAD) spectrum. Since radar point cloud is similar to LiDAR point cloud, transplanting the LiDAR perception algorithms into radar can easily implement radar perception tasks. Therefore, how effectively utilizing the different kinds of data representation of the FMCW radar has become the primary challenge in radar scene understanding.

[1]The authors are with the Alibaba Group, Hangzhou, China. (Zhen Xie* is the corresponding author, email: xiezhen.xz@alibaba-inc.com)

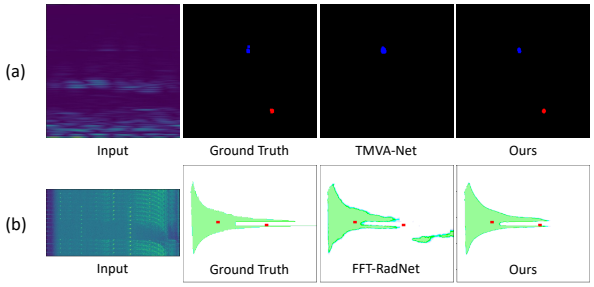

Fig. 1. Segmentation results of our TransRSS with SOTA methods [1], [13] on the CARRADA dataset (a) and RADIal dataset (b). In (a), red masks denote 'pedestrian' and bule masks denote 'car'. In (b), green masks denote 'free-driving-space' and red boxes denote 'car'.

Similar to LiDAR point cloud methods, the radar point cloud methods can be classified into two categories, i.e., the voxel-based methods [3]–[6] and the point-based methods [7]–[10]. The voxel-based methods generally conduct voxelization or project to bird's-eye-view (BEV) on the whole point cloud, so that Convolutional Neural Networks (CNN) based perception paradigms are leveraged on the compact feature representation. The point-based methods directly extract discriminative point-wise features using PointNet [11] or PointNet++ [12] from raw radar point clouds. These methods can draw on the experience of mature LiDAR perception, while radar point cloud is much sparser than LiDAR point cloud, which leads to poor feature representation learning. Another stream is RAD-based methods, like [1], [13]–[21]. Since directly processing a large volume RAD tensor will cause great computational complexity and memory consumption, the RAD-based methods usually slice or aggregate the RAD tensor along one dimension, which can be classified into single-view methods and multi-view methods. To fully explore the radar data, recent works exploit the multi-view representations of the RAD data and directly concatenate multi-view features for radar perception [13], [21], while explicitly feature aggregation could lead to the problem of feature misalignment. How to effectively aggregate multi-view features and extract more discriminative feature representation is still underexplored.

Recently, inspired by the power of transformers in natural language processing [22]–[24], the development of transformers for computer vision tasks [25]–[29] has made significant progress. These methods can easily capture global context information and naturally model long-range semantic dependencies, which could compensate for the shortcomings of CNN in feature extraction. In this work, we propose a multi-view radar segmentation method, named TransRSS, which integrates the merits of the CNN and transformer for grasping the global-level semantic features and results in more accurate segmentation, as shown in Fig. 1. It takes

advantage of the CNN to reduce local redundancy and avoid unnecessary computation and the transformer to model long-range dependency. More specifically, TransRSS comprises a pair of encoder-decoder architecture and a feature aggregation branch to extract features and fuse the feature representation from the RA and RD views, which fully exploits fine-grained and dynamic characteristics of the radar data. Inspired by shift convolution (shift-conv) [30], the encoder leverages a series of shift residual blocks (shift res-block) to efficiently extract local geometry characteristics, while a shift res-block has a larger receptive field but shares the same arithmetic complexity as a res-block [31]. The feature aggregation branch is composed of the transformer block, Atrous Spatial Pyramidal Pooling (ASPP) module [32], and Dual Position Attention (DPA) module, which further abstracts the output features of the encoder. We leverage the ASPP module and transformer block to capture the context of various receptive fields and further extract global context information, respectively. Meanwhile, for tackling the problem of feature misalignment, the proposed DPA module consists of two parallel attention sub-modules to aggregate rich contextual interdependencies between the RA and RD maps, achieving an implicit mechanism for adaptively feature aggregation. In this way, the intermediate features can be effectively fused to emerge a strong feature representation by concentrating the outputs of three modules. Finally, the decoder up-samples the aggregated features and combines them with the high-resolution feature maps from different layers of the encoder to reach reliable segmentation. Experiments on the CARRADA [33] and RADIal [1] datasets demonstrate that our TransRSS can achieve the best performance on the radar semantic segmentation task. The major contributions of our work are as follows:

1) We propose a novel radar segmentation framework TransRSS, which improves the performance of radar segmentation by taking the merits of the hybrid CNN-transformer architecture. To the best of our knowledge, TransRSS is the first work to explore the potential of the transformer in the context of radar segmentation.
2) We propose the shift res-block constructed by 1×1 shift-conv for discriminative feature extraction, which has a larger receptive field but shares the same computational complexity as a res-block.
3) We propose the Dual Position Attention (DPA) module to learn robust context information and aggregate rich contextual interdependencies between the RA and RD feature maps, which provides an implicit manner to feature aggregation adaptively.

II. Related work

A. Radar scene understanding with Point-based methods

The radar point cloud is one of the most common data formats, which has been broadly explored in object detection [3], [5], [9], [10], [34]–[37], semantic segmentation [7], [38], [39], and object tracking [6], [40]. [7] uses PointNet++ with the multi-scale grouping module and feature propagation module as a basis for radar segmentation. [9] utilizes a middle-fusion approach to fuse the radar point clouds and RGB images. The proposed RPR network utilizes radar information and image feature maps to generate accurate object proposals and distance estimations. Based on PointNet, [10] proposes a 2D object detection in radar data, which is composed of a classification network, segmentation network, and amodal 2D box estimation network. RVF-Net [34] proposes a low-level sensor fusion network for 3D detection on the nuScenes dataset [41], which fuses LiDAR, camera, and radar data. RPFA-Ne [5] proposes Pillar Features Attention instead of PointNet to extract pillar features and generate a pseudo image for proposal generation. RadarNet [3] exploits both LiDAR and radar sensors for perception, which features a voxel-based early fusion and an attention-based late fusion. Point-based methods can draw on the experience of mature LiDAR perception, but radar point cloud is much sparser than LiDAR point cloud and losses fine-grained information, which leads to poor feature representation learning.

B. Radar scene understanding with RAD-based methods

Range-Azimuth-Doppler (RAD) tensors contain rich characteristics, but are noisy and complicated. How to effectively extract RAD features is crucial for radar perception. Since the RAD tensor is cumbersome, most methods are usually considered by slicing or aggregating the tensor along one dimension in polar or Cartesian coordinates [1], [13]–[15], [17]–[20], [42]–[45]. To avoid laborious manual labeling, RSS-Net [19] correlates radar with cameras and LiDAR for achieving weakly-supervised semantic segmentation. Danet [20] proposes a Dimension Apart Module, which is lightweight and capable of extracting temporal-spatial information from the RA map. FFT-RadNet [1] leverages complex RD spectrums to construct a lightweight multi-task perception architecture. [18] proposes a cross-attention mechanism for fusing the features of RA, RD, and AD views. For exploiting the entire data, TMVA-Net [13] proposes a multi-view radar semantic segmentation in RA, RD, and AD views. The multi-view methods can exploit the entire data and tackle the challenges of the high level of noise and a large volume of RAD tensor, while how to effectively fuse the multi-view features is crucial. In comparison, our proposed TransRSS takes advantage of a hybrid CNN-transformer architecture and adaptively feature aggregation to enable discriminative feature learning and a flexible receptive field for improving the radar perception performance.

III. Method

In this paper, we propose the TransRSS, a hybrid CNN-transformer based radar semantic segmentation framework. The TransRSS comprises three branches: encoder, feature aggregation, and decoder for effectively extracting features from radar data. Since the AD view can be deduced from the RA and RD views, we only leverage the data from RA and RD views for semantic segmentation, which exploits the entire data but has a lower computational cost than using all

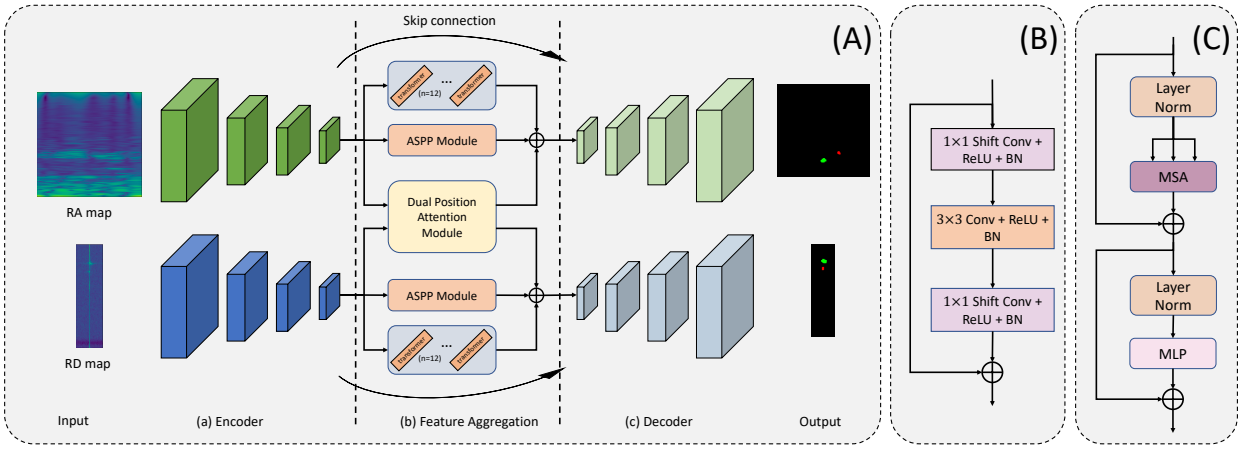

Fig. 2. Overview of the framework. (A) The pipeline of TransRSS is composed of the encoder (a), feature aggregation branch (b), and decoder (c). (B) Illustration of the shift res-block. (C) The architecture of the transformer layer.

data from three views. Moreover, unlike existing methods, our method first introduces transformer mechanisms into the radar segmentation tasks and leverages the attention schemes to extract global context information and achieve adaptively feature aggregation, respectively. The overall framework is illustrated in Fig. 2 (A).

A. Encoder-Decoder

Since the sizes of RA and RD maps are usually different, we resize them to the same size before feature abstraction. To enlarge the receptive field for more effective local feature extraction, based on shift-conv, we employ the shift-conv-shift structure to construct the shift res-block, where the kernel size is 1×1, 3×3, and 1×1, respectively. The shift-conv consists of a set of shift operations and 1×1 convolution, leading to a larger receptive field but sharing the same arithmetic complexity as a 1×1 convolution. Note that batch normalization [46] and the ReLU layer are appended after each convolution. The encoder branch consists of two shift ResNet50 constructed by the shift res-block to extract features of RA and RD maps. Each shift ResNet50 utilizes a series of shift res-block to down-sample the radar data gradually into feature maps with $1\times$, $2\times$, $4\times$, $8\times$ down-sampled sizes. There are four up-sampling blocks in the decoder to gradually decrease feature dimensions in the decoder branch. Each block is composed of two convolutions with kernel size 3×3, padding 1, and stride 2 and followed by the batch normalization [46], ReLU, and a bilinear up-sampling to recover the resolution of the feature map. For encoder-decoder architecture, the skip-connection operation passes information of the previous feature in the encoder to the decoder. A simple 1×1 convolution layer abstracts the output feature of the decoder to generate N_{cls} feature maps, where N_{cls} is the number of classes.

B. Feature Aggregation branch

In order to further extract the global contextual information and expand the receptive field, the feature aggregation branch concentrates the features from the transformer block, ASPP module, and Dual Position Attention module, which effectively extracts the global context characteristics and tackles the problem of feature misalignment between different views.

ASPP module: The ASPP module has demonstrated superiority in radar segmentation tasks in [13], which allows features to be jointly learned at different scales. Thanks to the dilation convolutions with different dilation rates, the ASPP module can enlarge the receptive field without a loss of resolution. Specifically, input features are abstracted by a 1×1 convolution and three dilation convolutions with kernel size 3×3, padding $\{1, 2, 3\}$, and dilation rate $\{1, 2, 3\}$, respectively. These features are concatenated and abstracted by a 1×1 convolution to generate the result.

Transformer block: Due to the intrinsic locality of convolution operations, CNNs generally demonstrate limitations in explicitly modeling global dependency, while transformer can capture long-range context dependencies to compensate for the shortcomings of the CNN. Specifically, given the RA or RD feature map $\mathbf{v} \in \mathbb{R}^{H \times W \times C}$, we perform the tokenization [25] to generate flattened 2D patches $\{\mathbf{v}_p^i \in \mathbb{R}^{P^2 \cdot C} | i = 1, \cdots, N\}$, where each patch size is $P \times P$ and $N = \frac{HW}{P^2}$ is the number of patches. The transformer layer utilizes linear projection to map the vectorized patches $\mathbf{v}_p$ into a latent D-dimensional embedding space. We learn specific position embeddings and add them to the patch embeddings for encoding the patch positional information as follow:

$$\mathbf{y}_0 = [\mathbf{v}_p^1\mathbf{P}; \mathbf{v}_p^2\mathbf{P}; \cdots; \mathbf{v}_p^N\mathbf{P}] + \mathbf{P}_{pos}, \tag{1}$$

where $\mathbf{P} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is the patch embedding projection and $\mathbf{P}_{pos} \in \mathbb{R}^{N \times D}$ denotes the position embedding. As shown in Fig. 2 (C), transformer block comprises L layers of Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP). The output of the l-th layer can be expressed as follows:

$$\mathbf{y}_l' = \mathrm{MSA}(\mathrm{LN}(\mathbf{y}_{l-1})) + \mathbf{y}_{l-1}, \tag{2}$$

$$\mathbf{y}_l = \mathrm{MLP}(\mathrm{LN}(\mathbf{y}_l')) + \mathbf{y}_l', \tag{3}$$

where LN is the layer normalization operator and $\mathbf{y}_l$ is the encoded image representation. The transformer block consists of 12 consecutive transformer layers to generate hidden features and reshape them to the size of input features.

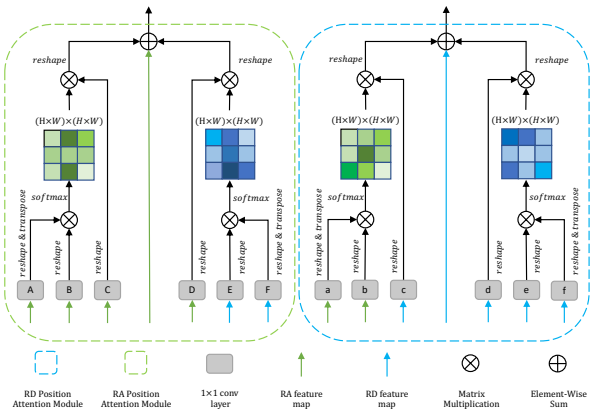

Fig. 3. The proposed Dual Position Attention (DPA) module consists of two sub-modules: RA position attention sub-module and RD position attention sub-module. It takes the RA feature map and RD feature map as inputs and computes corresponding attentional features to achieve the feature aggregation.

Dual Position Attention: The multi-view methods can make full use of the RAD data, hence how effectively and efficiently fusing the different view features is crucial. [13], [17], [47] directly concatenate intermediate feature, which struggles with the problem of feature misalignment and leads to poor feature representation. To mitigate the above mentioned issue of feature misalignment, the Dual Position Attention (DPA) module is proposed to enhance the feature representation and achieve adaptively feature aggregation. As shown in Fig. 3, the DPA module consists of the RA position attention sub-module and RD position attention sub-module, each containing self-attention and cross-attention. The RA position attention sub-module and RD position attention sub-module share the same structure. Here we take the RA position attention sub-module as an example to illustrate.

Self-attention: Given the RA feature $X \in \mathbb{R}^{C\times H\times W}$, we first feed it into three 1×1 convolution layers to generate three new feature representations X_A, X_B, and X_C, where $\{X_A, X_B, X_C\} \in \mathbb{R}^{C\times H\times W}$. Then we reshape X_A, X_B, and X_C to $\mathbb{R}^{C\times N}$, where $N = W \times H$. After that, we transpose the X_A to $\mathbb{R}^{N\times C}$ and perform a matrix multiplication between the X_A and X_B, and utilize a softmax layer to generate the self-attention map $S \in \mathbb{R}^{N\times N}$.

$$s^{ji} = \frac{exp(X_A^i \cdot X_B^j)}{\Sigma_{i=1}^{N} exp(X_A^i \cdot X_B^j)}. \tag{4}$$

After generating the self-attention map, we perform a matrix multiplication between X_C and S and reshape the result to $\mathbb{R}^{C\times H\times W}$ for generating the self-attention result.

Cross-attention: Different from self-attention, we leverage the RD feature map to calculate the cross-attention map. Given a RD feature $Y \in \mathbb{R}^{C\times H\times W}$, we feed it to two 1×1 convolution layers to generate new feature maps Y_E and Y_F, where $Y_E, Y_F \in \mathbb{R}^{C\times H\times W}$. Then we reshape Y_E and Y_F to $\mathbb{R}^{C\times N}$, where $N = W \times H$. After that, we transpose the Y_F to $\mathbb{R}^{N\times C}$ and perform a matrix multiplication between the Y_E and Y_F, and utilize a softmax layer to generate the cross-attention map $C \in \mathbb{R}^{N\times N}$.

$$c^{ji} = \frac{exp(y_F^i \cdot Y_F^j)}{\Sigma_{i=1}^{N} exp(Y_F^i \cdot Y_F^j)}. \tag{5}$$

Meanwhile, we feed a RA map X into a 1×1 convolution layer to generate a new feature map $X_D \in \mathbb{R}^{C\times H\times W}$ and reshape it to $\mathbb{R}^{C\times N}$. In addition, we perform a matrix multiplication between X_D and C and reshape the result to $\mathbb{R}^{C\times H\times W}$ to generate the final result.

Finally, we multiply the self-attention result and cross-attention result by scale parameters α and β and perform an element-wise sum operation with the RA feature to generate the final out R_{RA} as follow:

$$R_{RA} = \alpha \sum_{i=1}^{N}(s^{ji}X_C^i) + \beta \sum_{i=1}^{N}(c^{ji}X_D^i) + X^j, \tag{6}$$

where α and β are initialized to 0 and gradually learn a weight. It can be referred from Eq. 6 that the feature R_{RA} at each position is a weighted sum of the RA attention feature, RD attention feature, and original RA feature. Therefore, it not only extracts global context information but also achieves an implicit manner to adaptively feature aggregation.

C. Loss Function

Based on the fact that the labels of segmentation are unbalanced in RAD data, we employ a weighted Cross-Entropy and Soft Dice loss to learn the radar segmentation task. The weighted Cross-Entropy loss can be defined as:

$$\mathcal{L}_{wce} = -\sum_{i} \alpha_i P(y_i) log P(\widehat{y}_i), \alpha_i = 1/\sqrt{f_i}, \tag{7}$$

where f_i is the frequency of each category, $P(y_i)$ and $P(\widehat{y}_i)$ are the corresponding ground truth and predicts probability. Moreover, Soft Dice loss has proved helpful for small objects in medical image semantic segmentation [48], expressed as

$$\mathcal{L}_{dice} = \frac{1}{N}\sum_{n=1}^{N}\left[1 - \frac{2\sum_i P(y_i)P(\widehat{y}_i)}{\sum_i (P(y_i^2 + P(\widehat{y}_i^2))}\right], \tag{8}$$

where $P(y_i)$ and $P(\widehat{y}_i)$ are the same as defined in Eq. 7.

IV. Experiments

In this section, we introduce the implementation details of our method and conduct extensive experiments to quantitatively and qualitatively validate the performance of TransRSS on two radar preception datasets. We conduct detailed ablation experiments to verify the effectiveness of our proposed modules.

A. Experimental Setup

CARRADA. The CARRADA dataset provides dense semantic segmentation annotation for RA and RD views using a semi-automatic annotation method. The objects can be separated into four categories: pedestrian, cyclist, car, and background. The RA tensors have dimensions $256\times 256\times 1$. The RD tensors have dimensions $256\times 64\times 1$. Following [13], [33], the dataset splits into three parts: CARRADA-Train, CARRADA-Val, and CARRADA-Test.

RADIal. The RADIal dataset provides the RD spectrum, which contains 91 sequences of about 1-4 minutes and supplies free-driving-space segmentation and 2D box annotations for the RA view. The RD spectrum has dimensions

TABLE I

PERFORMANCE COMPARISON ON THE CARRADA-TEST. WE MARK THE BEST IN BOLD AND THE SECOND IN UNDERLINE.

View	Method	IoU(%)↑					Dice(%)↑				
		Bkg.	Ped.	Cyc.	Car	mIoU	Bkg.	Ped.	Cyc.	Car	mDice
RD	FCN [49]	99.7	47.7	18.7	52.9	54.7	99.8	24.8	16.5	26.9	66.3
	Unet [50]	99.7	51.0	33.4	37.7	55.4	99.8	67.5	50.0	54.7	68.0
	Deeplab-V3 [32]	99.7	43.2	11.2	49.2	50.8	99.9	60.3	20.2	66.0	61.6
	RSSNet [19]	99.7	0.1	4.1	25.0	32.1	99.7	0.2	7.9	40.0	36.9
	RAMP-CNN [17]	99.7	48.8	23.2	54.7	56.6	99.9	65.6	37.7	70.8	68.5
	TMVANet [13]	99.7	**52.6**	29.0	53.4	58.7	99.8	**68.9**	45.0	69.6	70.9
	Ours	99.5	49.1	**37.9**	**55.1**	**60.4**	99.8	66.2	**55.0**	**71.1**	**73.0**
RA	FCN [49]	99.8	14.8	0.0	23.3	34.5	99.9	25.8	0.0	37.8	40.9
	Unet [50]	99.8	22.4	8.8	0.0	32.8	99.9	36.6	16.1	0.0	38.2
	Deeplab-V3 [32]	99.9	3.4	5.9	21.8	32.7	99.9	6.5	11.1	35.7	38.3
	RSSNet [19]	99.5	7.3	5.6	15.8	32.1	99.8	13.7	10.5	27.4	37.8
	RAMP-CNN [17]	99.8	1.7	2.6	7.2	27.9	99.9	3.4	5.1	13.5	30.5
	TMVANet [13]	99.8	**26.0**	8.6	30.7	41.3	99.9	**41.3**	15.9	47.0	51.0
	Ours	99.6	24.9	**13.6**	**33.9**	**43.0**	99.7	37.4	**27.2**	**50.7**	**53.8**

TABLE II

FREE-DRIVING-SPACE SEGMENTATION PERFORMANCE ON THE RADIAL TEST SET.

Method	Input	mIoU(%)↑		
		Overall	Easy	Hard
PolarNet [42]	PC	60.6	61.9	57.4
FFT-RadNet [1]	RD	74.0	74.6	72.3
Ours	RD	**82.0**	**83.0**	**79.4**

$32 \times 512 \times 256$. We follow [1] and split the dataset into training, validation, and test sets and split the test set into 'hard' and 'easy' levels.

Training and Inference Details. Our method is trained in an end-to-end manner with the ADAM optimizer [51]. The exponential learning rate strategy is adopted for the learning rate decay. For the CARRADA dataset, we train the whole network with batch size 8, learning rate 1e-3 for 300 epochs on a Tesla V100 GPU. The loss function is defined as:

$$\mathcal{L}_{carrada} = \lambda_{wce}(\mathcal{L}_{wce}^{RA} + \mathcal{L}_{wce}^{RD}) + \lambda_{dice}(\mathcal{L}_{dice}^{RA} + \mathcal{L}_{dice}^{RD}), \quad (9)$$

where λ_{wce} and λ_{dice} are set as 10 and 5.

For a fair comparison with [1], we implement object detection and free-driving-space segmentation in the RADIal dataset. Similar to [1], for detection, we append a simple detection head at the end of the decoder and use a multi-task loss composed of a focal loss [52] for the classification and a smooth L1 loss for the regression:

$$\mathcal{L}_{det} = \mathcal{L}_{cls}(y_{cls}, \hat{y}_{cls}) + \mathcal{L}_{reg}(y_{reg}, \hat{y}_{reg}). \quad (10)$$

During the data preprocessing, we leverage a pre-encoder [1] to compress the RD tensor into a meaningful and compact RAD representation and split the RAD representation into RA and RD maps. Since the RADIal dataset only provides the annotations of the RA view, we only train the network using loss functions of the RA part. We train the network with batch size 8, learning rate 1e-4 for 100 epochs on a Tesla V100 GPU in the RADIal dataset. The loss function is defined as:

$$\mathcal{L}_{radial} = \lambda_{wce}\mathcal{L}_{wce}^{RA} + \lambda_{dice}\mathcal{L}_{dice}^{RA} + \lambda_{det}\mathcal{L}_{det}, \quad (11)$$

where λ_{wce}, λ_{dice}, and λ_{det} are set as 10, 5, and 100.

B. Radar semantic segmentation on the CARRADA dataset

To evaluate the performance of our method, we train it on the CARRADA-Train and CARRADA-Val and report the results on the CARRADA-Test. Following [13], we use the intersection over union (IoU) and Dice score as the evaluation metrics. Averaging these metrics over all classes yields the mean IoU (mIoU) score and mean Dice (mDice).

Comparison with state-of-the-art methods. Tab. I illustrates the performance of TransRSS on the CARRADA-Test. Our proposed method achieves the best performance for both the mIoU and mDice metric in the RA and RD semantic segmentation. For a fair comparison with [13], we fuse five frames for semantic segmentation. For the most important car class, our method outperforms previous state-of-the-art methods with remarkable margins, i.e., increasing the mIoU by 1.7%, 3.2%, and the mDice by 1.5%. 3.7% on the RD and RA views. For the performance of cyclist, our method surpasses previous methods by 8.9%, 5.0% IoU, and 10.0%, 11.3% Dice on the RD and RA views. Since the pedestrian information is not clearly reflected in the radar data, leading our method to hardly focus on the pedestrian category, the performance in the pedestrian category achieves slightly worse results. For the most important metrics of mIoU and mDice, our method achieves the best performance with 1.7%, 2.1% gain, and 2.7%, 2.8% gain on the RD and RA views, which manifests the effectiveness of TransRSS.

C. Radar Perception on the RADIal dataset

To further validate the effectiveness of our proposed TransRSS, we train our method on the newly released RADIal dataset. Following [1], we achieve free-driving-space segmentation and object detection on the RD tensor, where mIoU is leveraged for segmentation evaluation and F1 score, mean Average Precision (mAP), mean Average Recall (mAR), range accuracy, and angle accuracy are used considering an IoU threshold of 50% for detection evaluation.

Comparison with state-of-the-art methods. The performance for free-driving-space segmentation is shown in Tab. II. Our method surpasses FFT-RadNet by 8.0%, 8.4%, and 7.1% mIoU on overall, easy, and hard levels, which validates that our proposed method can effectively capture fine-grained information and model long-range dependency for improving segmentation performance. Performance for object detection is illustrated in Tab. III. We observe that our method outperforms FFT-RadNet with remarkable margins

TABLE III

OBJECT DETECTION PERFORMANCE ON THE RADIAL TEST SET.

Method	Input	Overall					Easy					Hard				
		F1(%)↑	mAP(%)↑	mAR(%)↑	R(m)↓	A(°)↓	F1(%)↑	mAP(%)↑	mAR(%)↑	R(m)↓	A(°)↓	F1 (%)↑	mAP(%)↑	mAR(%)↑	R(m)↓	A(°)↓
Pixor [53]	PC	48.42	96.46	32.32	0.17	0.25	44.66	**99.02**	28.83	0.15	0.19	54.69	93.28	38.69	0.19	0.33
Pixor [53]	RA	88.50	96.56	81.68	**0.10**	0.20	92.23	96.86	88.02	**0.09**	0.16	80.99	**95.88**	70.10	**0.12**	0.27
FFT-RadNet [1]	RD	88.91	**96.84**	82.18	0.11	0.17	94.97	98.49	91.69	0.10	0.13	76.37	92.93	64.82	0.13	0.26
Ours	RD	**94.24**	96.58	**92.01**	0.13	**0.10**	**98.22**	97.83	**98.62**	0.12	**0.10**	**83.25**	93.38	**75.11**	0.15	**0.11**

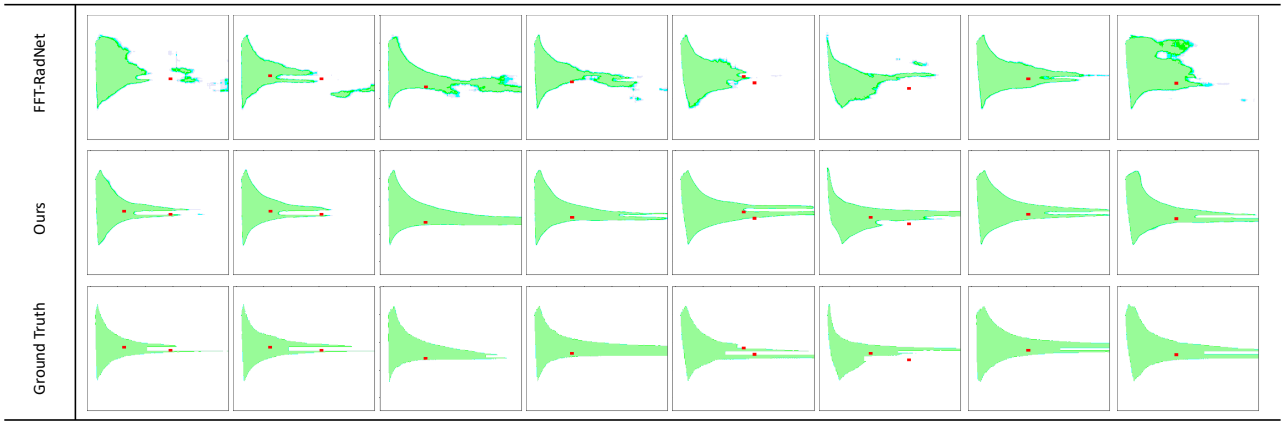

Fig. 4. Qualitative results of radar free-driving-space segmentation and detection on the RADIal test set.

TABLE IV

EFFECTS OF DIFFERENT MODULES FOR FREE-DRIVING-SPACE SEGMENTATION ON THE RADIAL VALIDATION SET.

SRB	TB	DPA	mIoU
-	-	-	78.5
√	-	-	79.6
√	√	-	81.0
√	√	√	**82.4**

TABLE V

EFFECTS OF DIFFERENT MODULES FOR OBJECT DETECTION ON THE RADIAL VALIDATION SET.

SRB	TB	DPA	F1(%)↑	mAP(%)↑	mAR(%) ↑	R(m)↓	A(°)↓
-	-	-	92.67	94.46	90.56	0.13	0.11
√	-	-	93.29	95.56	91.14	0.13	0.10
√	√	-	94.94	**97.66**	92.37	0.12	0.10
√	√	√	**96.12**	97.26	**95.00**	**0.12**	**0.09**

in F1 score, where mAR significantly increases while mAP decreases slightly. Specifically, our method achieves the best results with 5.33%, 3.25%, and 2.26% F1 score gain on overall, easy, and hard levels. All in all, our model displays a consistent performance improvement in overall main metrics. As shown in Fig. 4, we visualize the segmentation and detection results of our method and the state-of-the-art method [1]. As can be seen, our method can generate high-quality segmentation masks and detection bounding boxes. The experimental results on the RADIal dataset further validate the generalization ability and portability of our method on various datasets.

D. Ablation Studies

In this section, we conduct extensive ablation experiments to analyze individual components of our method. All models are trained on the train split and evaluated on the validation split of the RADIal dataset. It is worth noting that the baseline method includes the ASPP module and directly concatenates the intermediate features for feature aggregation. We report in Tab. IV the free-driving-space segmentation performance of the three modules: shift res-block (SRB), transformer block (TB), and Dual Position Attention (DPA) module. It can be referred from Tab. IV that each module has a positive effect on segmentation performance. The transformer block has the most segmentation gain, which proves that our transformer block can effectively extract global context information to improve the feature representation. We report in Tab. V the object detection performance of the three modules. It can be observed that the DPA module is able to improve the mAR with a remarkable margin (3.63%), while light dropping in mAP, which proves that adaptively feature aggregation can further improve the feature discrimination ability. From all the results, we notice that the SRB achieves a consistent performance improvement in segmentation and detection tasks, which validates that the SRB could effectively learn much richer contextual information and enlarger the receptive field.

V. CONCLUSION

In this paper, we propose TransRSS, the first radar semantic segmentation framework with a hybrid CNN-transformer architecture. Our method integrates the merits of the CNN and transformer to capture the global-level semantic features, and the learned discriminative features are then adaptively aggregated in an implicit manner to generate more discriminative context information for radar perception. Experimental results on the CARRADA and RADIal datasets demonstrate that our proposed method is able to effectively extract local-context feature and model long-range dependency, which significantly improves the performance of the radar segmentation task. The experiments on various datasets validate the generalization ability of our method.

REFERENCES

[1] J. Rebut, A. Ouaknine, W. Malik, and P. Pérez, "Raw high-definition radar for multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 021–17 030.

[2] M. A. Richards, *Fundamentals of radar signal processing*. McGraw-Hill Education, 2014.

[3] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, "Radarnet: Exploiting radar for robust perception of dynamic objects," in *European Conference on Computer Vision*. Springer, 2020, pp. 496–512.

[4] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1527–1536.

[5] B. Xu, X. Zhang, L. Wang, X. Hu, Z. Li, S. Pan, J. Li, and Y. Deng, "Rpfa-net: a 4d radar pillar feature attention network for 3d object detection," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 3061–3066.

[6] R. Nabati, L. Harris, and H. Qi, "Cftrack: Center-based radar and camera fusion for 3d multi-object tracking," in *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*. IEEE, 2021, pp. 243–248.

[7] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic segmentation on radar point clouds," in *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 2018, pp. 2179–2186.

[8] F. Nobis, F. Fent, J. Betz, and M. Lienkamp, "Kernel point convolution lstm networks for radar point cloud segmentation," *Applied Sciences*, vol. 11, no. 6, p. 2599, 2021.

[9] R. Nabati and H. Qi, "Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles," *arXiv preprint arXiv:2009.08428*, 2020.

[10] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer, "2d car detection in radar data with pointnets," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 61–66.

[11] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[12] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[13] A. Ouaknine, A. Newson, P. Pérez, F. Tupin, and J. Rebut, "Multi-view radar semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 671–15 680.

[14] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu, "Rodnet: Radar object detection using cross-modal supervision," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 504–513.

[15] A. Zhang, F. E. Nowruzi, and R. Laganiere, "Raddet: Range-azimuth-doppler based radar object detection for dynamic road users," in *2021 18th Conference on Robots and Vision (CRV)*. IEEE, 2021, pp. 95–102.

[16] X. Dong, P. Wang, P. Zhang, and L. Liu, "Probabilistic oriented object detection in automotive radar," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 102–103.

[17] X. Gao, G. Xing, S. Roy, and H. Liu, "Ramp-cnn: A novel neural network for enhanced automotive radar object recognition," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5119–5132, 2020.

[18] R. Kothari, A. Kariminezhad, C. Mayr, and H. Zhang, "Object detection and heading forecasting by fusing raw radar data using cross attention," *arXiv preprint arXiv:2205.08406*, 2022.

[19] P. Kaul, D. De Martini, M. Gadd, and P. Newman, "Rss-net: weakly-supervised multi-class semantic segmentation with fmcw radar," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 431–436.

[20] B. Ju, W. Yang, J. Jia, X. Ye, Q. Chen, X. Tan, H. Sun, Y. Shi, and E. Ding, "Danet: Dimension apart network for radar object detection," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, pp. 533–539.

[21] S. Azam, F. Munir, and M. Jeon, "Channel boosting feature ensemble for radar-based object detection," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 762–769.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[28] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 259–16 268.

[29] X. Pan, Z. Xia, S. Song, L. E. Li, and G. Huang, "3d object detection with pointformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7463–7472.

[30] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer, "Shift: A zero flop, zero parameter alternative to spatial convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9127–9135.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[33] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Perez, "Carrada dataset: Camera and automotive radar with range-angle-doppler annotations," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5068–5075.

[34] F. Nobis, E. Shafiei, P. Karle, J. Betz, and M. Lienkamp, "Radar voxel fusion for 3d object detection," *Applied Sciences*, vol. 11, no. 12, p. 5598, 2021.

[35] O. Schumann, J. Lombacher, M. Hahn, C. Wöhler, and J. Dickmann, "Scene understanding with automotive radar," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 2, pp. 188–203, 2019.

[36] J. Lombacher, K. Laudt, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic radar grids," in *2017 IEEE intelligent vehicles symposium (IV)*. IEEE, 2017, pp. 1170–1175.

[37] M. Meyer and G. Kuschk, "Deep learning based 3d object detection for automotive radar and camera," in *2019 16th European Radar Conference (EuRAD)*. IEEE, 2019, pp. 133–136.

[38] K. Bansal, K. Rungta, and D. Bharadia, "Radsegnet: A reliable approach to radar camera fusion," *arXiv preprint arXiv:2208.03849*, 2022.

[39] R. Prophet, A. Deligiannis, J.-C. Fuentes-Michel, I. Weber, and M. Vossiek, "Semantic segmentation on 3d occupancy grids for automotive radar," *IEEE Access*, vol. 8, pp. 197 917–197 930, 2020.

[40] N. Scheiner, F. Kraus, F. Wei, B. Phan, F. Mannan, N. Appenrodt, W. Ritter, J. Dickmann, K. Dietmayer, B. Sick *et al.*, "Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2068–2077.

[41] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[42] F. E. Nowruzi, D. Kolhatkar, P. Kapoor, E. J. Heravi, F. A. Hassanat, R. Laganiere, J. Rebut, and W. Malik, "Polarnet: Accelerated deep open space segmentation using automotive radar in polar domain," *arXiv preprint arXiv:2103.03387*, 2021.

[43] P. Li, P. Wang, K. Berntorp, and H. Liu, "Exploiting temporal relations on radar perception for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 071–17 080.

[44] K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 444–453.

[45] Y.-J. Li, J. Park, M. O'Toole, and K. Kitani, "Modality-agnostic learning for radar-lidar fusion in vehicle detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 918–927.

[46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[47] B. Major, D. Fontijne, A. Ansari, R. Teja Sukhavasi, R. Gowaikar, M. Hamilton, S. Lee, S. Grzechnik, and S. Subramanian, "Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[48] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[49] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[53] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.