TransRVNet: LiDAR Semantic Segmentation with Transformer

Hui-Xian Cheng, Xian-Feng Han, *Member, IEEE,* Guo-Qiang Xiao

***Abstract*—Effective and efficient 3D semantic segmentation from large-scale LiDAR point cloud is a fundamental problem in the field of autonomous driving. In this paper, we present Transformer-Range-View Network (TransRVNet), a novel and powerful projection-based CNN-Transformer architecture to infer point-wise semantics. First, a Multi Residual Channel Interaction Attention Module (MRCIAM) is introduced to capture channel-level multi-scale feature and model intra-channel, inter-channel correlations based on attention mechanism. Then, in the encoder stage, we use a well-designed Residual Context Aggregation Module (RCAM), including a residual dilated convolution structure and a context aggregation module, to fuse information from different receptive fields while reducing the impact of missing points. Finally, a Balanced Non-square-Transformer Module (BNTM) is employed as fundamental component of decoder to achieve locally feature dependencies for more discriminative feature learning by introducing the non-square shifted window strategy. Extensive qualitative and quantitative experiments conducted on challenging SemanticKITTI and SemanticPOSS benchmarks have verified the effectiveness of our proposed technique. Our TransRVNet presents superior performance over most existing state-of-the-art approaches. The source code and trained model are available at https://github.com/huixiancheng/TransRVNet.**

***Index Terms*—CNN, Transformer, Point Cloud, Range Image, Semantic Segmentation, Autonomous Driving**

I. Introduction

.

ADEQUATELY accurate, robust, reliable and real-time 3D semantic perception and understanding [1], [2] of vehicles surrounding environment are critical requirements for autonomous driving. To achieve this end, multiple types of sensors with complementary characteristics, such as cameras, Light Detection And Ranging (LiDAR) [3], Radio Detection And Ranging (RADAR) [4], are usually used in self-driving systems, among which LiDARs play an important role because they can provide distance measurements and 3D geometry with high precision. Therefore, LiDAR-based semantic scene perception, especially 3D point cloud semantic segmentation as a primary task, has been attracting increasing research attention, which aims to perform prediction of per-point class labels.

Recently, convolutional neural networks (CNNs) have brought significant performance improvement for semantic image segmentation [9], whose success is attributed to their power of learning hierarchical feature representations. However, CNNs are not suitable for 3D point clouds, since 1) standard convolution is defined over regular grids [10], while 2) 3D point cloud produced by LiDAR are inherently sparse, irregular, and unordered [11]. 3) There is lack of sufficient large-scale training datasets. Fortunately, with the availability of recent large-scale datasets (e.g. SemanticKITTI), several state-of-the-art approaches using deep neural networks have been proposed to address these problems. And according to the representation of 3D LiDAR points, these methods are commonly classified into voxel-based methods [12], [13],image-based methods [14], [15], and point-based methods [16]–[18].

For point-based approaches, although impressive progress has been achieved, high computational cost and memory consumption make this kind of methods too expensive for large outdoor scenarios in autonomous driving applications. While Voxel-based and image-based methods may be a reasonable choice that convert the raw point cloud into structured format, allowing standard 3D CNNs or 2D CNNs to be

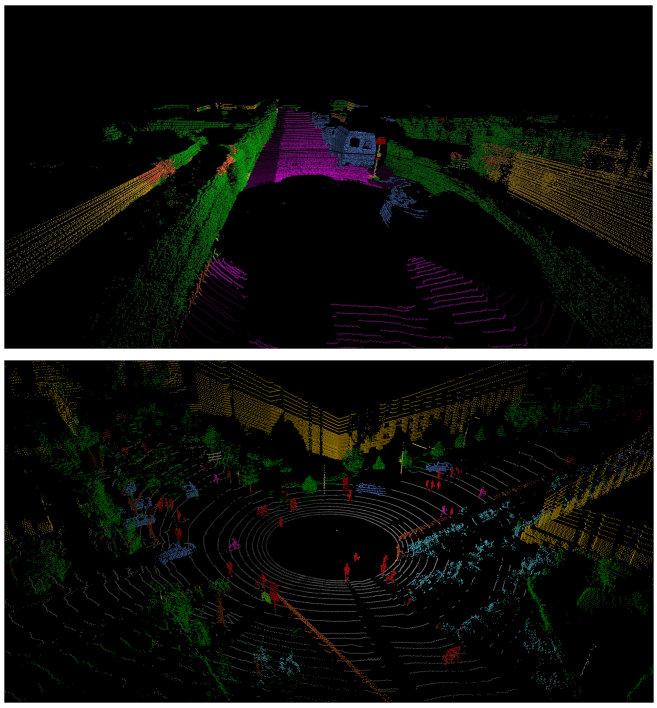

Fig. 1. **Top**: Velodyne HDL-64E [5] laser scan from SemanticKITTI dataset [6] with semantic predictions from our TransRVNet. **Bottom**: Pandora [7] laser scan from SemanticPOSS dataset [8] with semantic predictions from our TransRVNet.

Hui-Xian Cheng, Xian-Feng Han, Guo-Qiang Xiao with the College of Computer and Information Science, Southwest University, Beibei District, Chongqing, China e-mail: chenghuixian@email.swu.edu.cn, xianfeng-han@swu.edu.cn, gqxiao@swu.edu.cn Hui-Xian Cheng and Xian-Feng Han have contributed equally to this work.

Manuscript received April 19, 2005; revised August 26, 2015.

used directly [19]. However, sparsity and granular information loss challenges the application of voxel representation [20]. And image-based methods also donnot produce satisfactory performance because the projected LiDAR and RGB image live in different domains [21]. Therefore, it is necessary to design specific architecture for projection-based (i.e. LiDAR-based range image) semantic segmention, which should take into account the following problems 1) How to project the full 360 LiDAR scan into 2D surface? 2) How to learn local and global contextual information? 3) How to model long-range feature dependencies from local and global regions?

Recent advancements in Natural Language Processing (NLP) using Transformer [22] have motivated the computer vision community to investigate the adaption of Transformer models to vision tasks. ViT [23] is the first purely transformer model for image recognition. SETR [24] chooses to replace the encoder in traditional encoder-decoder network with a Transformer, and achieves competitive results on image segmentation task. However, from the other work that followed [25]–[30], we find that 1) the pure Transformer segmentation networks like SETR donnot result in very good performance. 2) The SETR is difficult to train, since it is parametric intensive and computationally expensive. 3) Additionally, the SETR fails to perform multi-scale feature extraction which is important for dealing with large scale variations of objects [21].

To tackle these issues, we contribute a novel CNN-Transformer architecture based upon an encoder-decoder skeleton, named Transformer-Range-View Network (TransRVNet). After representing the input 3D LiDAR point cloud as a 2D range image, a Multi Residual Channel Interaction Attention Module (MRCIAM) is developed to extract channel-wise multi-scale features and model context interactions within and across different information channels. Then, in our CNN-Transformer framework, a Residual Context Aggregation Module (RCAM) based on residual dilated convolution structure is used as encoding component to explore the information from different receptive fields. And, in the decoding stage, we construct a Balanced Non-square-Transformer Module (BNTM) to capture locally long-range feature dependencies using our well-designed non-square shifted window based Transformer. We perform extensive evaluation on two publicly challenging benchmarks SemanticKITTI [6] and SemanticPOSS [8]. The experimental results show that our TransRVNet is superior to several state-of-the-art models.

In summary, the key contributions of this paper are listed as follows:

- Based on attention mechanism, a Multi Residual Channel Interaction Attention Module (MRCIAM) is well designed to learn multi-scale context information from different modal channels respectively, and encodes the intra-channel as well as inter-channel interactions/feature dependencies, simultaneously.
- We propose a new Balanced Non-square-Transformer Module (BNTM) using non-square shifted window based Transformer block to better deal with size imbalanced range image, and learn locally long-range feature dependencies.
- We introduce a novel CNN-Transformer architecture, named Transformer-Range-View Network (TransRVNet) with a Residual Context Aggregation Module (RCAM) as encoder component to gain local-to-global context information from progressively enlarged receptive fields, and the proposed BNTM as decoder block. To the best of our knowledge, this is the first work focusing on the application of Transformer to semantic LiDAR point cloud segmentation.
- Comprehensive experimental evaluation on publicly available datasets SemanticKITTI and SemanticPOSS validate the superior performance of TransRVNet. And we further analyze the effect of each designed modules had on our neural network.

The rest of our paper is structured as follows. Section II gives an overview of recent related works. Section III presents the detailed description of our TransRVNet. Comprehensive experimental evaluation and analysis are performed in Section IV. Section V provides the ablation study of our proposed method. Conclusions are drawn in Section VI.

II. Related Works

A. Semantic Segmentation of 3D Point Cloud

With the remarkable advancement of deep learning technology and the availability of large-scale scene datasets, semantic 3D LiDAR point cloud segmentation for autonomous driving have achieved significant progress in recent years. The previous studies can be grouped into three categories, incuding point-based, voxel-based and image-based methods.

Point-based methods directly deal with unstructured point cloud with shared point-wise MLPs [16], or definition of point convolutional kernels [31], or graph [31]. Where pointwise MLP methods usually processes each point independently using shared Multi-Layer Perceptrons, followed by a symmetric aggregation function (e.g., max pooling) to form a global feature vector. PointNet can be considered as the pioneering work that learns point-wise feature. However, PointNet fails to fully employit the local spatial relationships in the point cloud, i.e. the contextual features among points are neglected [32] [33]. To address this problem, PointNet++ [17] develops a hierarchical framework to aggregate information from local neighboring points. But it still treats each point independently [34] [35]. Motivated by the key idea of 2D convolution, some methods concentrate on defining point convolutional kernels [34] or learn the kernel based on point positions [36] to captures local information. However, these approaches may have high complexity during learning [37]. In addition, other methods attempt to introduce graph neural network for local geometric topology aggregation. For example, DGCNN designs an Edge-Conv operator to update center point feature by aggregating features along the edges.

Theoretically and empirically, point-based approaches perform well on small point clouds. However, for LiDAR semantic segmentation, full 360° large-scale scans limit the real-time processing capability of these methods due to inefficient computational complexity and memory requirement [18].

Therefore, this kind of methods actually is less suitable for autonomous driving applications.

Voxel-based methods first discretize the point cloud into 3D volumentric representation, where each point is associated with one corresponding voxel. Then, 3D CNNs framework are used to perform voxel-wise semantic label prediction [13]. SEGCloud [12] treat the coarse voxel label prediction as a intermediate stage in point cloud semantic segmentation task. Zhu et al. [38] transformed the point cloud into a special cylindrical grid, and designed two novel CNN components. One is an asymmetric residual block to preserve the features related to cubic objects, while another is context modeling to fuse multiple low-level contexts for high-level tensor model construction. Although the voxel grid is one of ideal choice for irregular point cloud processing [39], problem of sparsity, loss of granular information and introduction of discretization artifacts limit the performance of voxel-based methods [40].

Image-based methods, here, also known as projection-based or range image-based methods, which project the 3D point cloud onto 2D image space either in multiple views [41] or in Range-View representation [42], follow by the application of 2D CNNs for semantic segmentation. The prior studies are usually built on transitional or improved encoder-decoder framework [43] [44], and Conditional Random Fields (CRF) [45] [46] is adopted as post-processing algorithm to render the final results. In subsequent works, RangeNet++ [15], 3D-MiniNet [47] utilizes a K-nearest neighbour post-processing instead of CRF to recover the consistent semantic information. SalsaNext [48] performs an uncertainty-aware semantic segmentation, and introduces a new Lovász-Softmax loss [49] to optimize the results. Triess et al. [50] proposed a Scan unfolding projection technique to solve the problems of systematic occlusions. To preserve local geometric information, ThickSeg [51] chooses to use a multi-layer image to represent the input point cloud. Then, a self-ordered 3DCNN is designed to infer the grid-wise labels. Finally, an iterative and cumulative post-processing mechanism is used to predict labels for occluded points.

B. Vision Transformers

More recently, immense progress have been made in the field of Natural Language Processing (NLP) with the advent of Transformer [22]. Inspired by its great success, researchers from computer vision community attempt to explore the application of Transformer architectures to image based tasks [52]. Detection Transformer (DETR) [53] regards the detection problem as a set prediction task, which first learns initial feature maps using a CNN backbone. Then these features are fed into transformer-based encoder-decoder structure for object detection. Vision Transformer (ViT) [23] is a pure Transformer model that achieves state-of-the-art performance on image classification. However, it is challenging for ViT to perform pixel-level dense prediction due to the following reasons. (1) ViT only outputs single-scale feature maps with low resolution. (2) ViT is computationally inefficient and parametric intensive when dealing with images, even those with smaller size. To handle these issues, many excellent works [25]–[30] have been proposed. Swin Transformer [30] constructs a hierarchical transformer model with shifted window based self-attention as fundamental component. The results on image classification, object detection and semantic segmentation are encouraging. In our paper, we adopt the core concept of Swin Transformer as the basic block of decoder part with skip connections for range images segmentation.

III. METHODOLOGY

Given a full-view 3D LiDAR scan, we aim to construct a function between input domain $\mathcal{P}$ of points and label space $\mathcal{Y}$ as $f : \mathcal{P} \longrightarrow \mathcal{Y}$ to perform efficient semantic segmentation. To this end, we first convert the point cloud segmentation into 2D range-view segmentation problem vie spherical projection. Then, the Multi Residual Channel Interaction Attention Module (MRCIAM) processes each independent modal channel for channel-wise feature learning, and captures the intra-channel as well as inter-channel interactions. In the following encoder-decoder stage, we form a Residual Context Aggregation Module as encoder block to model local and global context, while a Balanced Non-square-Transformer Module is designed as decoder component. Finally, we emploit the KNN post-processing operator to backproject the 2D range semantic to original points. Fig. 2 schematically shows the detailed architecture of the proposed TransRVNet.

A. Range-View Representation for LiDAR Point Cloud

Since (1) accurate and efficient semantic inference plays a key role in autonomous driving [47]. (2) Spherical projection representations (range view or range image) are dense, compact, and computationally efficient [54], which can include the full 360° LiDAR scan. (3) Deep learning techniques for various 2D computer vision tasks have achieved highly astounding performance.

Therefore, to obtain higher segmentation performance and take full advantage of the benefits of 2D CNNs, we choose to directly deal with 2D range image produced from LiDAR scan. Following [15], we find a mapping function $f : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ to form an one-to-one correspondence between 3D point with coordinate (x, y, z) and an pixel (u, v) in the mapped range image of size (H, W) (H and W represent the height and width of the generated range image). The spherical projection we used is formulated as,

$$u = \frac{1}{2}\left[1 - arctan\left(y, x\right)\pi^{-1}\right]W \tag{1}$$

$$v = \left[1 - \left(arcsin\left(z, d^{-1}\right) + f_u\right)\frac{1}{f_u + f_d}\right]H \tag{2}$$

where $f = f_{up} + f_{down}$ refers to the sensor's vertical field-of-view. The depth d of each point is calculated as $d = \sqrt{x^2 + y^2 + z^2}$. The final projected range image can be denoted as a tensor with the shape of $(H, W, 5)$, where each pixel contains five channels (x, y, z, d, r), r is the remission or reflection intensity information. Thought this spherical transformation, the point cloud segmentation problem is turned into image segmentation task.

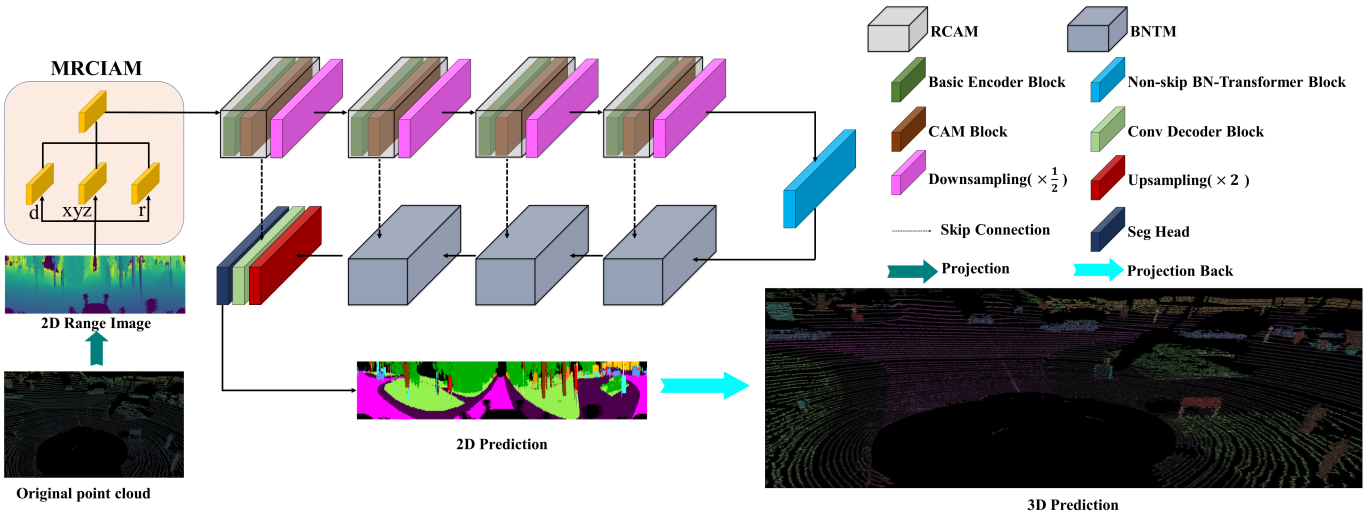

Fig. 2. The overall architecture of TransRVNet.

B. TransRVNet

The pipeline of the proposed TransRVNet is illustrated in Figure 2. In the following subsections, we will describe the main components of the designed architecture in details.

*1) **Multi Residual Channel Interaction Attention Module**:* As shown in Figure 3, it can be reported that different channels contained in projection range image actually present different modal information, and their corresponding values follow different distributions, which means different channels make different contributions to the learned features. However, most of the prior approaches [15], [44], [48] directly take the projected image as input, few works [21], [55] have noticed this phenomenon but lack more efficient feature extraction and channel interaction attention. Consequently, we decide to learn individual feature space for each independent channel with channel interaction, which is proved to be more effective and efficient.

To achieve this end, we design a Multi Residual Channel Interaction Attention Module (MRCIAM) to capture multi-scale contextual information from each channel and aggregate these features, which consists of four essential blocks, as illustrated in Figure 4. Where the first and the third ones share the same structure, including three convolutions with different kernel sizes (e.g., $3 \times 3, 5 \times 5, 7 \times 7$) and a residual connections with a 1×1 convolution to capture and fuse local-to-global context. The second block is also a residual structure but with dilated convolution. Finally, we add a Spatial-Channel Attention (SCA) module to MRCIAM for modeling the pixel-level relationships in each channel and interactions/dependencies across different channels, i.e. intra-channel and inter-channel interactions.

Overall, our MRCIAM operates on the depth/range channel, xyz channel and remission/reflectivity channel independently, and outputs local-to-global contextual features with spatial fine-grained information.

*2) **Encoder**:* The current existing Transformer-based semantic segmentation studies use either the framework of Transformer as ecnoder and CNN as decoder [24], [30], [56], [57], or the pure Transformer structure [58]–[60]. Although these models have achieved promising performance in several medical imaging datasets [61]–[63] and widely used ADE20k dataset [64], from our experiments it can be found that these two kinds of structures are less suitable for extract features from range-image-style inputs. This may be because encoding much larger patch tokens would become more difficult. In addition, it is too expensive to train these networks. Hence, we decide to use our CNN-based Residual Context Aggregation Module (RCAM) as the encoder and Transformer-based Balanced Non-square-Transformer Module (BNTM) as decoder. Specifically, the RCAM includes the following components,

Basic Encoder Block: In order to further capture more descriptive contextual features from channel-fused maps from MRCIAM, one simple to stack convolution operations with progressively enlarged kernel size to obtain much larger receptive filed. However, this strategy usually brings remarkable increase of the number of parameters. Therefore, to reduce computational cost, we design a simple but effective encoder block based on dilated convolution and residual structure (shown in Figure 5). Here, we combine two 3×3 convolutions, one 3×3 dilation convolution with dilation rate 2, as well as a residual connection with 1×1 convolution for dimension reduction. Through this design, the encoder can make full use of multi-level multi-scale multi-channel information.

Context Aggregation Module: Missing points, which we name missing noise (as illustrated in red-circled areas of GT image in Figure 3), have negative impact on performance of CNN models. This kind of noise commonly results from the limitation of sensors, mirror reflection, occlusion of the object itself, as well as spherical projection. To cope with this issue, we choose to use the Context Aggregation Module (CAM) from SqueezeSegV2 [43], and put it at the end of each Basic Encoder Block to enhance the robustness of contextual features to specific noise.

C. Decoder

The decoder is based on our Balanced Non-square-Transformer Module, which is composed of the following sub-

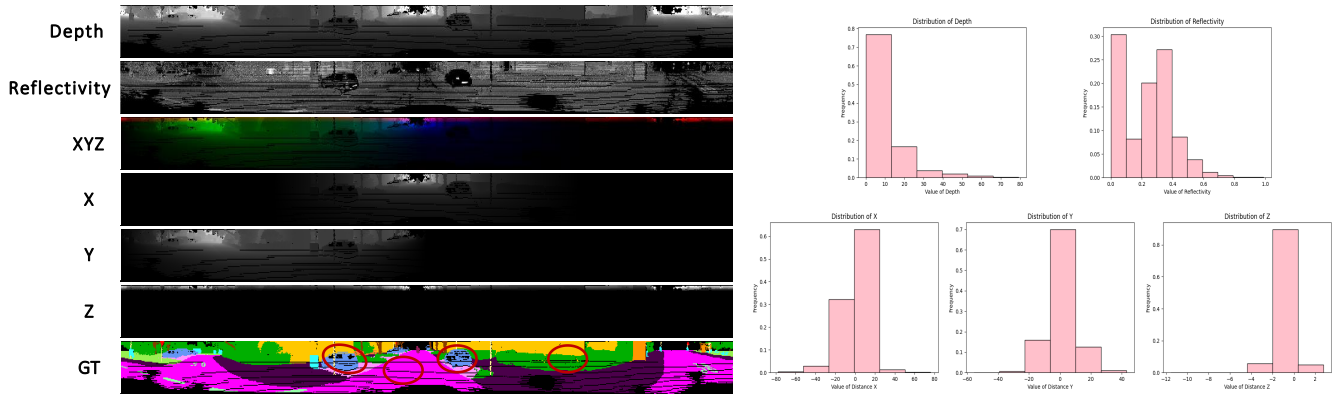

Fig. 3. Visualization of a full LiDAR scan. **Left**: Visualization of different channels of the input range image. **Right**: Quantitative analysis of the numerical distribution of the different channels. (Zoom in for more details.)

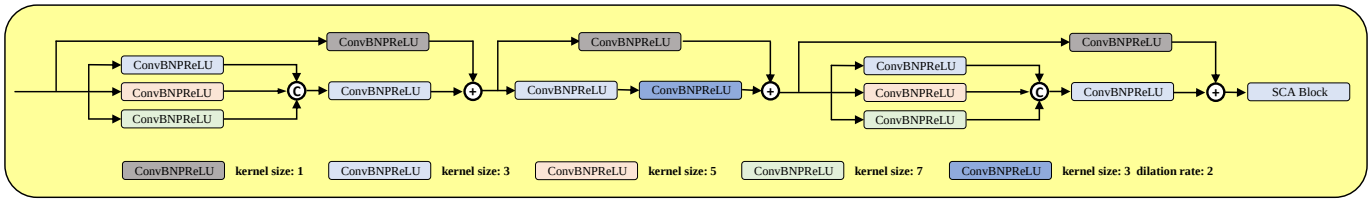

Fig. 4. The Structure of Multi Residual Channel Interaction Attention Module.

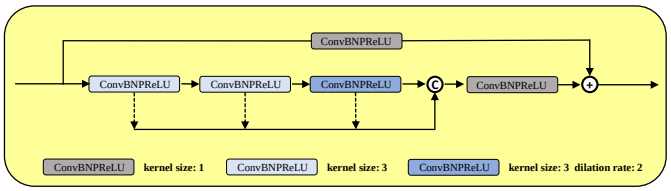

Fig. 5. The Structure of Basic Encoder Block.

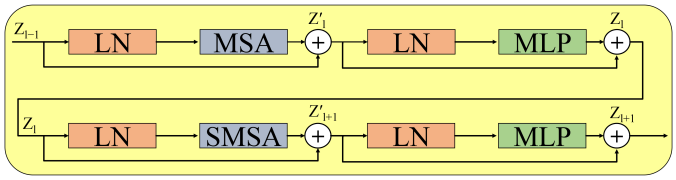

Fig. 6. The Structure of Swin Transformer Block.

blocks.

BN-Transformer Module: The standard Transformer architectures adopt the multi-head self attention (MSA) mechanism to obtain global relationships between a patch token and other tokens for image-based tasks. However, for large-resolution dense pixel-level prediction, Transformer model often has quadratic computational complexity with respect to image size [30], which limits its performance. On the other hand, the fixed size of token makes Transformer unsuitable for computer vision tasks required to deal with visual objects with varying scales. To handle these issues, Liu et al. [30] built the Swin Transformer with shifted window module MSA, MLP with GELU non-linearity, LayerNorm layer and residual connections to capture dependencies locally and reduce the computational complexity. Inspired by its tremendous success, we transfer the core idea of Swin architecture to our Transformer decoder component, as shown in Figure 6. The formulation can be summarized as,

$$\begin{aligned} Z'_l &= MSA(LN(Z_{l-1})) + Z_{l-1} \\ Z_l &= MLP(LN(Z'_l)) + Z'_l \end{aligned} \tag{3}$$

$$\begin{aligned} Z'_{l+1} &= SMSA(LN(Z_l)) + Z_l \\ Z_{l+1} &= MLP(LN(Z'_{l+1})) + Z'_{l+1} \end{aligned} \tag{4}$$

where Z'_l and Z_l represent the outputs of the (S)MSA module and the MLP module of the l_{th} block, respectively.

Here, it should be emphasized that different from original Swin Transformer, (1) we choose to design a non-square window based Transformer instead of square version to accommodate the size imbalance of range image (since range images usually have much larger aspect ratio compared to common RGB images). (2) To enable much better feature learning, each decoder module contains four stacked Transformer blocks after the upsampling layer. (3) we use Transformers with different

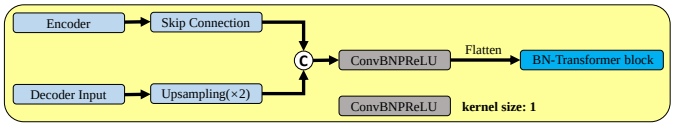

Fig. 7. The Structure of Balanced Non-square-Transformer Module.

resolutions to construct a hierarchical decoder structure for multi-scale representation. Specifically, our Balanced Non-square-Transformer Module is designed as follows (shown in Figure 7). First, we upsample the input feature maps by a factor of 2, which then are concatenated with the feature maps from the corresponding encoder together to fuse multi-scale features. Finally, we use a linear layer to adjust the channel dimension. And our designed Transformer model takes the flatten features as input to perform decoding operation. The whole decoder stage consists of three BN-Transformer Modules with different scales.

Conv Decoder Block: Although our BN-Transformer Modules have capability of achieving powerful performance, it is a little bit difficult for them to decode these feature maps with largest resolution. We, therefore, decide to put a convolution-based decoder block at the end of decoder pipeline, which includes three elements, an upsampling operation, Conv decoder, and a segmentation head.

D. Loss function

In specific to our range-view based semantic segmentation task, actually there exists three challenging problems required to be alleviated.

- The imbalanced category problem.
- The problem of optimizing the intersection-over-union (IoU).
- The ambiguous segmentation boundaries problem.

Here, three different loss functions, i.e. weighted cross-entropy loss, Lovász-Softmax loss and boundary loss, are used to supervise our model.

First, the weighted cross-entropy loss can be formulated as,

$$L_{wce} = -\sum_{c=1}^{C}(\frac{f_t}{f_c})^i y_c log(\hat{y}_c) \tag{5}$$

where y_c and $\hat{y}_c$ denotes the ground truth and prediction, respectively. To address the class imbalance, we adopt median frequency class balance strategy from SegNet [65]. And we attempt to incorporate the power operation to smooth the estimated class weights, $w_c = (\frac{f_t}{f_c})^i$, where f_c is the frequency of class c, and f_t is the median of all class frequencies.

Second, to maximize the intersection-over-union (IoU) score, we opt for the Lovász-Softmax loss function [49] represented as L_ls. Its formulation can be summarized as follows,

$$L_{ls} = \frac{1}{C}\sum_{c=1}^{C}\overline{\triangle_{J_{nc}}}\left(m\left(c\right)\right) \tag{6}$$

$$m_i(c) = \begin{cases} 1-\hat{p}_i^c & if\ c = p_i^c \\ \hat{p}_i^c & otherwise \end{cases} \tag{7}$$

where $\overline{\triangle_{J_{nc}}}$ indicates the Lovász extension of the Jaccard index,$\hat{p}_i^c \in [0,1]$, $p_i^c \in [0,1]$ are the predicted probability and ground truth for the ith pixel of class c.

The problem of boundary ambiguity between different objects/classes in the segmentation results often arises from downsampling and upsampling operations [66]–[68]. And few studies focus on finding appropriate solution to this issue for range image based segmentation [21]. Hence, motivated by [3], we introduce a boundary loss function, which can be defined as,

$$L_{bd}(\hat{y}, y) = 1 - \frac{2P^cR^c}{P^c + R^c} \tag{8}$$

where P^c and R^c denotes the precision and recall of the predicted boundary map y_{pd} with respect to the ground truth y_{gt} for class c. $\hat{y}$ and y indicate the prediction and ground truth for each class. We give the definition of boundaries as,

$$y_{gt}^b = pool(1 - y_{gt}, \theta_0) - (1 - y_{gt}), \tag{9}$$

$$y_{pd}^b = pool(1 - y_{pd}, \theta_0) - (1 - y_{pd}) \tag{10}$$

Here pool$(\cdot)$ means max-pooling operation on a sliding window of size θ_0. And in order to model vicious boundary, the θ_0 should be set as small as possible. Figure 8 shows the boundary maps of different classes with $\theta_0 = 3$.

Finally, our total loss is the weighted combination of these three loss functions, given by,

$$L = \lambda_1 L_{wce} + \lambda_2 L_{ls} + \lambda_3 L_{bd} \tag{11}$$

where λ_1, λ_2, and λ_3 are the corresponding weights to balance the weighted cross-entropy loss, Lovász-Softmax loss and boundary loss. In our experiments, we set i in L_{wce} to be 0.25 and the weights $\lambda_1 = 1.0$, $\lambda_2 = 1.5$, and $\lambda_3 = 1.0$.

E. Post-processing

The prediction label re-projection from range image to 3D point cloud possibly comes with misclassification problem. The reason is that two or more LiDAR points might be associated with the same pixel. In order to cope with this issue, we employ a k-nearest neighbor (KNN) post-processing strategy to generate a new semantic label for each 3D point, where the closest neighbors are found in the subsampled point cloud instead of the complete scene using a sliding window technique [15].

IV. EXPERIMENTS

In this section, we present quantitative and qualitative evaluations of our TransRVNet on the large-scale SemanticKITTI [6] and the newly proposed small-scale SemanticPOSS [8] benchmarks with a single NVIDIA RTX 2080Ti GPU. In the following, we first give a brief description of the datasets and evaluation metric. Then, we provide the implementation details. Finally, by comparing with other state-of-the-art approaches, we further validate the effectiveness of our proposed network with quantitative analysis and visualization.

A. Dataset

SemanticKITTI dataset is constructed by providing dense point-wise semantic annotations for all 360° scans in KITTI Odometry Benchmark [73]. This datasets contains over 43,000 scans from 21 sequences, where over 21,000 scans of sequences 00 to 10 are used for training, sequence 08 is chosen as validation set, while the remaining scans from 11 to 21 for testing.

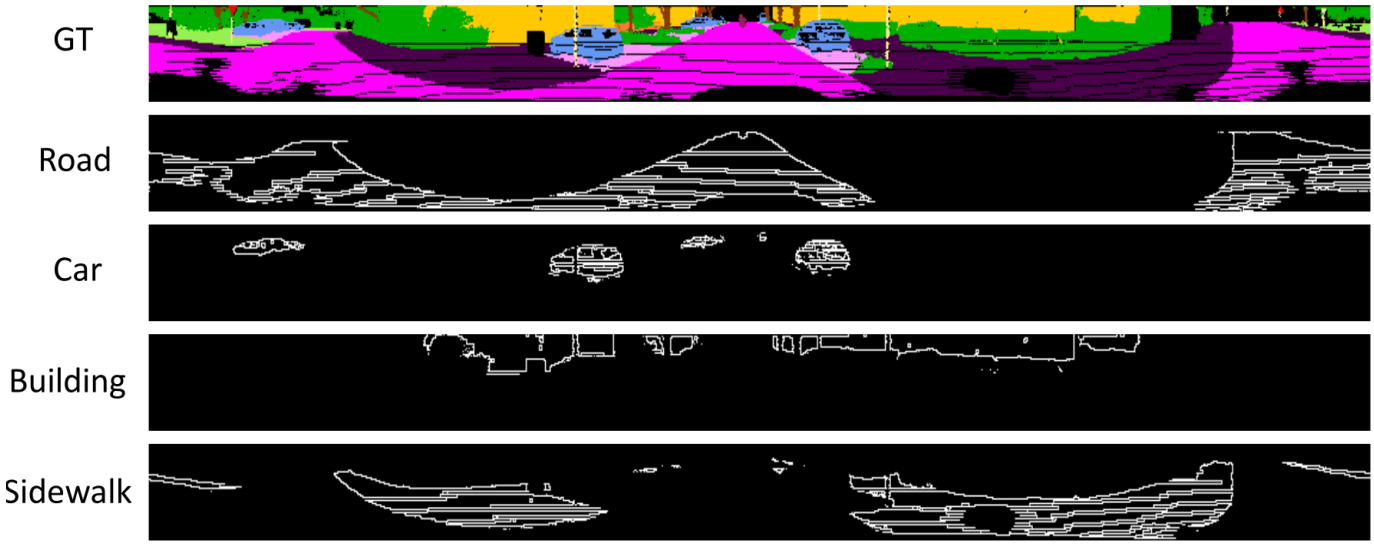

Fig. 8. Visualization of the extracted boundaries.

TABLE I
THE MEAN IOU (%) COMPARISON ON SEMANTICKITTI TEST SET. NOTE THAT AN FPS ABOVE 10 IS CONSIDERED REAL-TIME. THIS IS BECAUSE THE ACQUISITION FREQUENCY OF THE VELODYNE HDL-64E [5] LIDAR SENSOR IS 10 HZ.

Category	Method	Input Size	Car	Bicycle	Motorcycle	Truck	Other-vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other-ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic-sign	Mean IoU	FPS(Hz)
Point-based	Pointnet [16]	50K pts	46.3	1.3	0.3	0.1	0.8	0.2	0.2	0.0	61.6	15.8	35.7	1.4	41.4	12.9	31.0	4.6	17.6	2.4	3.7	14.6	2
	Pointnet++ [17]		53.7	1.9	0.2	0.9	0.2	0.9	1.0	0.0	72.0	18.7	41.8	5.6	62.3	16.9	46.5	13.8	30.0	6.0	8.9	20.1	0.1
	TangentConv [32]		86.8	1.3	12.7	11.6	10.2	17.1	20.2	0.5	82.9	15.2	61.7	9.0	82.8	44.2	75.5	42.5	55.5	30.2	22.2	35.9	0.3
	LatticeNet [35]		92.9	16.6	22.2	26.6	21.4	35.6	43.0	46.0	90.0	59.4	74.1	22.0	88.2	58.8	81.7	63.6	63.1	51.9	48.4	52.9	7
	RandLa-Net [18]		94.2	26.0	25.8	40.1	38.9	49.2	48.2	7.2	90.7	60.3	73.7	20.4	86.9	56.3	81.4	61.3	66.8	49.2	47.7	53.9	22
	MinkNet [40]		94.3	23.1	26.2	26.1	36.7	43.1	36.4	7.9	91.1	63.8	69.7	29.3	92.7	57.1	83.7	68.4	64.7	57.3	60.1	54.3	-
	BAF-LAC [69]		94.2	27.6	30.4	39.9	40.1	46.3	49.1	10.5	90.9	61.4	74.0	22.2	88.2	57.0	81.3	61.7	65.3	49.8	53.9	54.9	-
	Kpconv [34]		96.0	30.2	42.5	33.4	44.3	61.5	61.6	11.8	88.8	61.3	72.7	31.6	90.5	64.2	84.8	69.2	69.1	56.4	47.4	58.8	-
	BAAF [70]		95.4	31.8	35.5	48.7	46.7	49.5	55.7	53.0	90.9	62.2	74.4	23.6	89.8	60.8	82.7	63.4	67.9	53.7	52.0	59.9	5
Image-based	SqueezeSeg [14]	64 × 2048 px	68.8	16.0	4.1	3.3	3.6	12.9	13.1	0.9	85.4	26.9	54.3	4.5	57.4	29.0	60.0	24.3	53.7	17.5	24.5	29.5	66
	SqueezeSegV2 [43]	64 × 2048 px	81.8	18.5	17.9	13.4	14.0	20.1	25.1	3.9	88.6	45.8	67.6	17.7	73.7	41.1	71.8	35.8	60.2	20.2	36.3	39.7	50
	SqueezeSegV3 [44]	64 × 2048 px	92.5	38.7	36.5	29.6	33.0	45.6	46.2	20.1	91.7	63.4	74.8	26.4	89.0	59.4	82.0	58.7	65.4	49.6	58.9	55.9	6
	3D-MiniNet [47]	64 × 2048 px	90.5	42.3	42.1	28.5	29.4	47.8	44.1	14.5	91.6	64.2	74.5	25.4	89.4	60.8	82.8	60.8	66.7	48.0	56.6	55.8	28
	FPS-Net [55]	64 × 2048 px	91.1	48.6	37.8	37.1	30	60.5	57.8	7.5	91.1	61.9	74.6	26.0	87.4	57.4	80.9	61.2	65.0	49.9	59.2	57.1	21
	PolarNet [71]	480 × 360 × 32	93.8	40.3	30.1	22.9	28.5	43.2	40.2	5.6	90.8	61.7	74.4	21.7	90.0	61.3	84.0	65.5	67.8	51.8	57.5	54.3	16
	MPF [72]	64 × 2048 px	93.4	30.2	38.3	26.1	28.5	48.1	46.1	18.1	90.6	62.3	74.5	30.6	88.5	59.7	83.5	59.7	69.2	49.7	58.1	55.5	21
		64 × 1024 px	92.7	28.2	30.5	26.9	25.2	42.5	45.5	9.5	90.5	64.7	74.3	32.0	88.3	59.0	83.4	56.6	69.8	46.0	54.9	53.6	29
	SalsaNext [48]	64 × 2048 px	91.9	48.3	38.6	38.9	31.9	60.2	59.0	19.4	91.7	63.7	75.8	29.1	90.2	64.2	81.8	63.6	66.5	54.3	62.1	59.5	24
		64 × 1024 px	91.1	44.2	37.0	35.2	31.3	39.9	49.3	22.4	91.1	63.6	75.0	28.5	88.6	61.5	79.6	52.5	66.5	44.5	50.7	55.4	32
	RangeNet++ [15]	64 × 2048 px	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	65.0	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9	52.2	12
		64 × 1024 px	90.3	20.6	27.1	25.2	17.6	29.6	34.2	7.1	90.4	52.3	72.7	22.8	83.9	53.3	77.7	52.5	63.7	43.8	47.2	48.0	21
	TransRVNet(ours)	**64 × 1024 px**	91.9	50.6	47.0	41.8	36.7	58.3	60.5	31.7	91.1	66.9	74.9	26.7	88.9	60.6	82.5	64.4	67.4	53.7	58.0	**60.7**	26

SemanticPOSS dataset, collected by Peking University, is a small-scale much sparser benchmark with point-wise label providedd. It consists of 2,988 diverse and complicated LiDAR scenes with large quantity of dynamic instances [8]. For convenience, SemanticPOSS follows the same data format as SemanticKITTI, which is divided into 6 parts, with part 2 as testing set and the others as training set.

B. Evaluation Metric

To evaluate the performance of our TransRVNet and make a fair comparison with other state-of-the-art techniques, we use the standard mean Intersection over Union (mIoU) [74] as our evaluation metric, which can be formalized as,

$$mIOU = \frac{1}{N}\sum_{c=1}^{N}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{12}$$

where TP_c, FP_c and FN_c represents the number of true positive, false positive, and false negative point predictions for class c, respectively.

C. Implementation details

We implement our TransRVNet architecture based on PyTorch framework. All models are trained using Stochastic Gradient Descent (SGD) optimizer with momentum 0.9 and a weight decay of 0.0001. During training, random rotation, random dropping points, and random sign inverting for X and Y values with a probability of 0.5 are applied to augment the input point cloud to avoid overfitting. Specifically,

As for SemanticKITTI dataset [6], we train our model for 100 epochs with a batch size 4, and an initial learning rate of 0.005 decayed with a rate of 0.01 after each epoch. The dropout probability is 0.2. We set the height and width of

input range images to H = 64 and W = 1,024, respectively. The window size for our Transformer block is set as [4, 64]. For the KNN post-processing stage, we use window size of 7 for neighbor search, with k=7, σ=1, and a cutoff of 2 m.

For SemanticPOSS dataset [8], we set the initial learning rate to be 0.0025 that is dynamically adjusted using the Cosine Annealing scheduler with warm up. We train our TransRVNet for 50 epochs, and set the batch size as 2, the dropout probability as 0.2. The height and width of input range images are H = 64, W = 1,600. The corresponding window size of Transformer is set to [4,100]. In addition, due to the sparsity of SemanticPOSS dataset, we adjust the hyperparameter values for KNN post-processing as follows, the window size of 11, k=11, σ=1, and the cutoff of 5 m.

D. Results and Discussion

Quantitative results. The quantitative comparisons with the state-of-the-art point-wise and image-based approaches on SemanticKITTI are reported in Table I. From these experimental results, we can obtain the following observations,

- The methods taking range images with resolution 64×2048 as input usually obtain a much better mIoU compared with these using images of size 64×1024. This is because (1) larger input size means improvement of networks' learning ability to segment small-scale and fine-grained objects, such as Person, Bicycle, and Tranffic-sign in the scenes. (2) In addition, larger input also means more points back-projected to 3D point cloud. For example, 64×1024 corresponds to 65,536 points while 64×2048 is associates with 131,072 points.
- Our proposed TransRVNet achieves a superior performance (60.7%), outperforming not only all image-based methods at lower resolutions, but also point-based methods. Moreover, our method achieves a better trade-off between accuracy and computational complexity.
- The TransRVNet noticeably surpasses the counterpart image-based baseline RangeNet++ (48.0%) by a significant margin, achieving an improvement of 12.7% mIoU score, and it also improves another version of RangeNet++ (52.2%) using range image of 64×2048 as input by +8.5% mIoU value.

Table II present the mean IoU comparison with all the referred works on Sequence 02 SemanticPOSS [8] dataset. It can be clearly reported that (1) TransRVNet obtains a leading performance on both test sequences (a mIoU of 49.3%), which significantly promotes the previous state-of-the-art RangeNet++ [15] and MINet [21] by 18.4%, 6.1%, respectively. (2) All approaches have a much lower mIoU score due to the following reason that compared with SemanticKITTI, SemanticPOSS is a much smaller and more sparse dataset.

Qualitative results. The qualitative comparison between RangeNet++ and our TransRVNet are provided in Figure 9. From these results, we can see that our TransRVNet can get a much more accurate predictions, and is closer to the ground truth than RangeNet++. Figure 10 and Figure 11 shows the visualization results provided by our TransRVNet on SemanticKITTI and SemanticPOSS, respectively. These results convincingly demonstrate the effectiveness of our network.

V. Ablation Study

To better understand the proposed TransRVNet, we carry out ablation studies to discuss the contributions of important designed techniques (including network architecture design, choice of window size, role of MRCIAM and CAM module) made to segmentation performance. For all the ablation experiments, we use Sequence 08 on SemanticKITTI dataset for tesing, and the remaining sequences for training.

A. Network Architecture

In order to investigate the effectiveness of our CNN-Transformer architecture on segmentation performance, we perform the following comparisons, (1) we design two Transformer-CNN frameworks based on TransRVNet, where we replace all the Transformer decoder blocks by our Conv Decoder Block, and use the high-performance Swin-T and ViT models as encoder with the following hyper-parameter settings.

- We set the query dimension of each head in multi-head self-attention to 32, and integrate 4 expansion layers in MLP.
- For Swin-T: C = 96, layer numbers = $\{2, 2, 6, 2\}$
- For ViT: C = 64, layer numbers = $\{4, 4, 4, 4\}$

Where C denotes the number of channels in the first stage. layer numbers refers to the number of Transformer blocks in each encoder module. (2) We remove MRCIAM from the ViT-CNN model.

Table III lists the mIoU comparisons between three Transformer-CNN variants and CNN-Transformer structures. From this, it can be obviously observed that (1) Our CNN-BN-Transformer model achieves the best performance in terms of accuracy, which is much more suitable for LiDAR semantic segmentation than Transformer-CNN framework. (2) The introduction of MRCIAM leads to a significant improvement of mIoU of 5.5% over the couterpart without MRCIAM.

B. Impact of Window Size

As discussed in Section III-C, the non-square shifted window scheme is a key design element for our TransRVNet. Therefore, the window size actually have a great impact on not only the segmentation performance, but also the parameter and computation complexity. To verify its effect on segmentation result, we provide quantitative comparison of TransRVNet using different window size Table IV. As we expect, (1) the semantic segmentation performance increases with the increase of window size. This is mainly because larger window commonly corresponds to larger respective filed, which means we can aggregate much richer context information. However, larger window usually leads to higher memory cost. (2) Compared with square window, non-square window yields much better performance. For example, windows of size [4, 8], [4, 16], [4, 32] improve the mIoU over 0.2%, 0.9%, and 1.2% against the counterpart with size [4, 4]. This is largely attribute to the fact that the width of range image is far larger than its height, which we define as size imbalance problem. (3) From these results It can be concluded that non-square

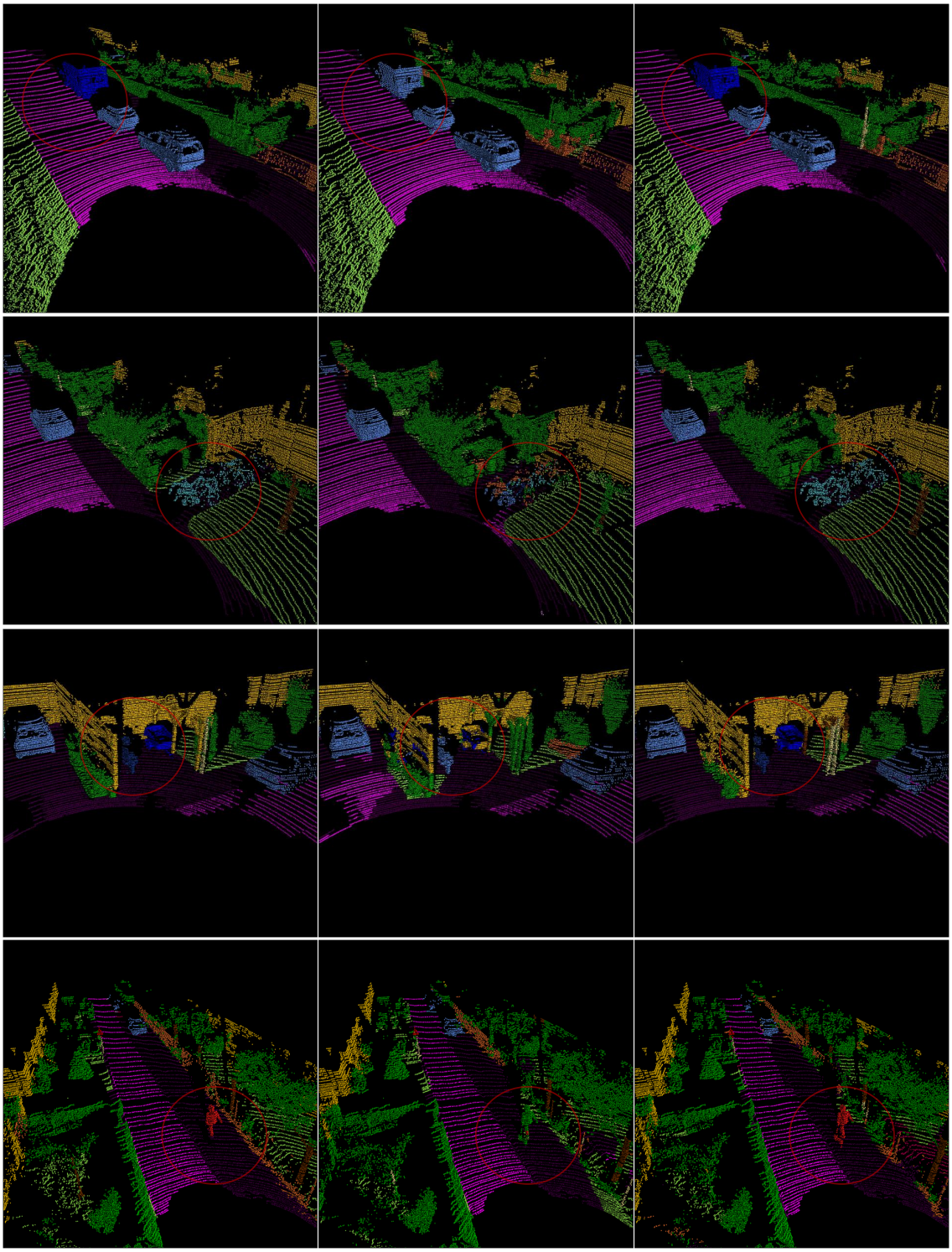

Fig. 9. Qualitative results of TransRVNet on Sequence 8 of SemanticKITTI. Color codes are: |road |car |other-vehicle |side-walk |parking |motorcycle |vegetation |terrain |trunk |building |other-structure

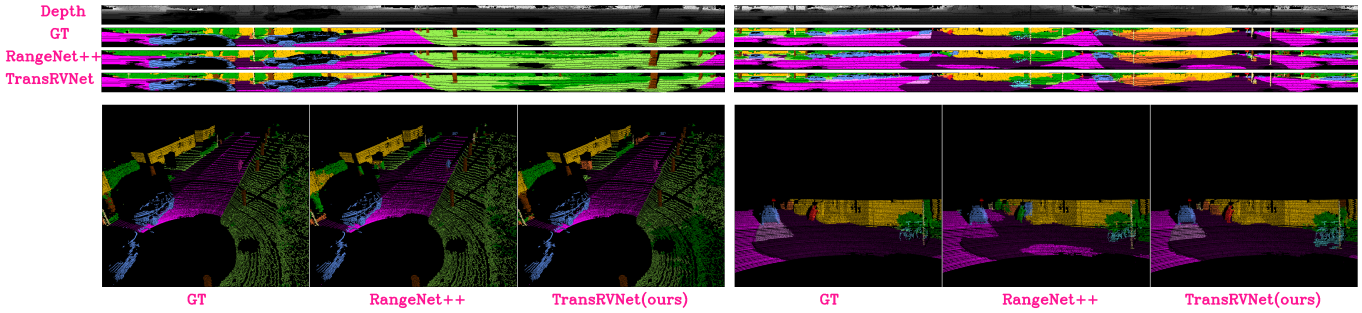

Fig. 10. Visualization example on SemanticKITTI using RangeNet++ and TransRVNet. (Zoom in for more details.)

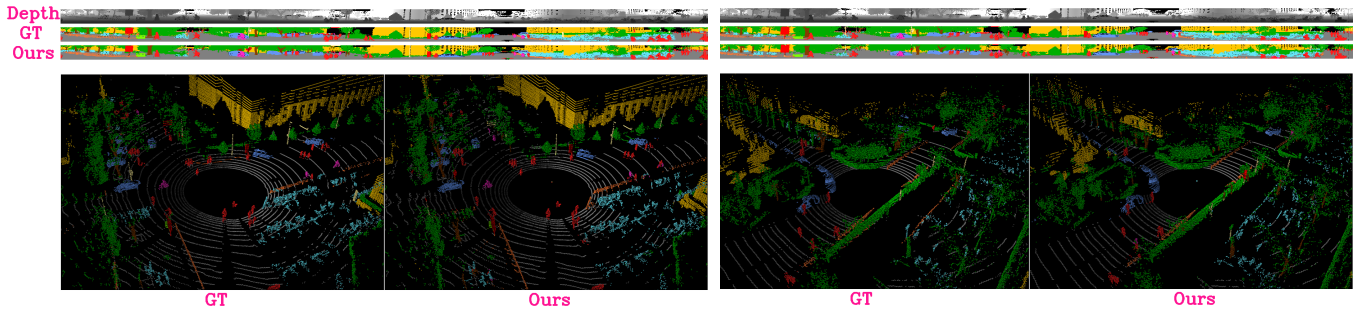

Fig. 11. Visualization examples on SemanticPOSS using TransRVNet. (Zoom in for more details.)

TABLE II
QUANTITATIVE COMPARISON WITH RECENT STATE-OF-THE-ART WORKS ON SEQUENCE 02 OF SEMANTICPOSS TEST SET USING MEAN IOU (%)

	Person	Rider	Car	Truck	Plants	Traffic-sign	Pole	Trashcan	Building	Cone/Stone	Fence	Bike	Ground	mIOU
SqueezeSeg [14]	14.2	1.0	13.2	10.4	28.0	5.1	5.7	2.3	43.6	0.2	15.6	31.0	75.0	18.9
SqueezeSeg + CRF [14]	3.8	0.6	6.7	4.0	2.5	9.1	1.3	0.4	37.1	0.2	8.4	18.5	72.1	12.9
SqueezeSegV2 [43]	48.0	9.4	48.5	11.3	50.1	6.7	6.2	14.8	60.4	5.2	22.1	36.1	71.3	30.0
SqueezeSegV2 + CRF [43]	43.9	7.1	47.9	18.4	40.9	4.8	2.8	7.4	57.5	0.6	12.0	35.3	71.3	26.9
RangeNet53 [15]	55.7	4.5	34.4	13.7	57.5	3.7	6.6	23.3	64.9	6.1	22.2	28.3	72.9	30.3
RangeNet53 + KNN [15]	57.3	4.6	35.0	14.1	58.3	3.9	6.9	24.1	66.1	6.6	23.4	28.6	73.5	30.9
MINet [21]	61.8	12.0	63.3	22.2	68.1	16.3	29.3	28.5	74.6	25.9	31.7	44.5	76.4	42.7
MINet + KNN [21]	62.4	12.1	63.8	22.3	68.6	16.7	30.1	28.9	75.1	28.6	32.2	44.9	76.3	43.2
TransRVNet	70.7	18.5	75.1	26.6	71.9	19.0	30.0	34.8	76.5	30.1	43.7	50.8	**79.5**	48.3
TransRVNet + KNN	**72.1**	**19.2**	**76.5**	**27.2**	**72.7**	**19.6**	**31.1**	**37.7**	**77.7**	**33.1**	**43.8**	**51.1**	79.0	**49.3**

window is proved to be an ideal and reasonable choice for our Transformer model.

C. Effects of the MRCIAM and RCAM module

Table V reports the exntesive ablation studies of our poposed MRCIAM and CAM modules. We add MRCIAM and RCAM to the backbone one by one to validate the effectiveness of each module. From these experimental results, we have the following observation, (1) with only MRCIAM and only RCAM block, the segmentation performance of our TransRVNet is boosted to 57.2%, 56.6%, respectively, outperforming the baseline by 2.2% and 1.6%. (2) Combining MRCIAM and RCAM together, our model achieves a significant improvement of 3.1%. To further explore the superiority of MRCIAM, we presents the quantitative results on several classes (such as Truck, Person, Parking, Building, Fence, trunk, Pole and Traffic-Sign.) in Table VI. With MRCIAM, our TransRVNet gets evident improvements for all classes in terms of class mIoU. We attribute this to the powerful ability of MRCIAM to capture multi-scale channel-wise context and model inter-channel, intra-channel interactions.

VI. CONCLUSIONS

In this paper, we transfer the encouraging performance of Transformer in NLP to image-based LiDAR semantic segmentation task by constructing a CNN-Transformer architecture, named TransRVNet. We first present a Multi Residual Channel Interaction Attention Module to aggregate contextual information and model inter-channel, intra-channel interactions. Then,

TABLE III
IMPACT OF NETWORK ARCHITECTURE ON SEMANTICKITTI VALIDATION SET

Architecture	mIoU	Params
Swin-T-CNN	46.3	34.40 M
ViT-CNN(w/o MRCIAM)	47.5	20.62 M
ViT-CNN(w MRCIAM)	53.0	20.86 M
CNN-BN-Transformer	58.1	26.52 M

TABLE IV
IMPACT OF WINDOW SIZE ON SEMANTICKITTI VALIDATION SET

Window Size	mIoU	FLOPs	GPU Memory (Batchsize=8)
[4, 4]	56.9	33.15G	7,701 MiB
[4, 8]	57.1	33.28G	7,881 MiB
[4, 16]	57.8	33.53G	8,153 MiB
[4, 32]	**58.1**	34.03G	8,757 MiB
[4, 64]	57.1	35.71G	10,513 MiB

a Residual Context Aggregation Module is used as encoder block to model multi-level multi-scale features. Finally, we design a Balanced Non-square-Transformer Module as decoder element to capture local-global dependencies. During training, we weight the weighted cross-entropy loss, Lovász-Softmax loss and boundary loss to supervise our network. Extensive experimental results on SemanticKITTI and SemanticPOSS demonstrate that our TransRVNet outperforms several state-of-the-art point-based, voxel-based and image-based models in terms of mIoU, achieving promising performance. And ablation studies further verify the effectiveness of our proposed learning modules.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (No. 62002299), and the Natural Science Foundation of Chongqing of China (No. cstc2020jcyj-msxmX0126), and the Fundamental Research Funds for the Central Universities (No. SWU120005). Hui-Xian Cheng and Xian-Feng Han are joint first authors.

TABLE V
IMPACT OF MRCIAM MODULE AND CAM MODULE ON SEMANTICKITTI VALIDATION SET

	Backbone	MRCIAM	RCAM	mIOU
Architecture	✓			55.0
	✓		✓	56.6
	✓	✓		57.2
	✓	✓	✓	58.1

TABLE VI
ANALYSIS OF MRCIAM MODULE ON SEMANTICKITTI VALIDATION SET

Architecture	Trunk	Person	Parking	Building	Fence	Truck	Pole	Traffic-Sign
Backbone	77.5	33.7	46.7	82.0	50.0	55.1	58.9	34.8
Backbone+MRCIAM	83.9	39.7	52.1	85.2	56.8	60.0	62.1	37.5
	+6.4	+6.0	+5.4	+3.2	+6.8	+4.9	+3.2	+2.7

REFERENCES

[1] J. Mei, B. Gao, D. Xu, W. Yao, X. Zhao, and H. Zhao, "Semantic segmentation of 3d lidar data in dynamic scene using semi-supervised learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 6, pp. 2496–2509, 2019.

[2] L. Ma, Y. Li, J. Li, W. Tan, Y. Yu, and M. A. Chapman, "Multi-scale point-wise convolutional neural networks for 3d object segmentation from lidar point clouds in large-scale environments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 821–836, 2019.

[3] R. Razani, R. Cheng, E. Taghavi, and L. Bingbing, "Lite-hdseg: Lidar semantic segmentation using lite harmonic dense convolutions," *arXiv preprint arXiv:2103.08852*, 2021.

[4] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[5] Velodyne HDL-64 LiDAR Sensor. Accessed: 2021-07-30. [Online]. Available: https://velodynelidar.com/products/hdl-64e/

[6] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9297–9307.

[7] Hesai Pandar40 40-Channel Mechanical LiDAR. Accessed: 2021-07-30. [Online]. Available: https://www.hesaitech.com/en/Pandar40

[8] Y. Pan, B. Gao, J. Mei, S. Geng, C. Li, and H. Zhao, "Semanticposs: A point cloud dataset with large quantity of dynamic instances," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 687–693.

[9] X. Zhang, H. Xu, H. Mo, J. Tan, C. Yang, L. Wang, and W. Ren, "Dcnas: Densely connected neural architecture search for semantic image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 956–13 967.

[10] H. Lei, N. Akhtar, and A. Mian, "Spherical kernel for efficient graph convolution on 3d point clouds," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[11] X.-F. Han, H. Laga, and M. Bennamoun, "Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1578–1604, 2021.

[12] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, "Segcloud: Semantic segmentation of 3d point clouds," in *2017 international conference on 3D vision (3DV)*. IEEE, 2017, pp. 537–547.

[13] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232.

[14] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1887–1893.

[15] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "Rangenet++: Fast and accurate lidar semantic segmentation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4213–4220.

[16] 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.

[17] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *arXiv preprint arXiv:1706.02413*, 2017.

[18] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 108–11 117.

[19] T. He, C. Shen, and A. van den Hengel, "Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 354–363.

[20] S. Fan, Q. Dong, F. Zhu, Y. Lv, P. Ye, and F.-Y. Wang, "Scf-net: Learning spatial contextual features for large-scale point cloud segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 504–14 513.

[21] S. Li, X. Chen, Y. Liu, D. Dai, C. Stachniss, and J. Gall, "Multi-scale interaction for real-time lidar data segmentation on an embedded platform," *arXiv preprint arXiv:2008.09162*, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] 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.

[24] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[25] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.

[26] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.

[27] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.

[28] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *arXiv preprint arXiv:2103.00112*, 2021.

[29] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.

[30] 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," *arXiv preprint arXiv:2103.14030*, 2021.

[31] Z. Liu, H. Hu, Y. Cao, Z. Zhang, and X. Tong, "A closer look at local aggregation operators in point cloud analysis," in *European Conference on Computer Vision*. Springer, 2020, pp. 326–342.

[32] M. Tatarchenko, J. Park, V. Koltun, and Q.-Y. Zhou, "Tangent convolutions for dense prediction in 3d," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3887–3896.

[33] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz, "Splatnet: Sparse lattice networks for point cloud processing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2530–2539.

[34] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6411–6420.

[35] R. A. Rosu, P. Schütt, J. Quenzel, and S. Behnke, "Latticenet: Fast point cloud segmentation using permutohedral lattices," *arXiv preprint arXiv:1912.05905*, 2019.

[36] Y. Liu, B. Fan, S. Xiang, and C. Pan, "Relation-shape convolutional neural network for point cloud analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8895–8904.

[37] M. Xu, R. Ding, H. Zhao, and X. Qi, "Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3173–3182.

[38] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9939–9948.

[39] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[40] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

[41] S. Xie, S. Liu, Z. Chen, and Z. Tu, "Attentional shapecontextnet for point cloud recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4606–4615.

[42] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[43] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4376–4382.

[44] C. Xu, B. Wu, Z. Wang, W. Zhan, P. Vajda, K. Keutzer, and M. Tomizuka, "Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–19.

[45] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[46] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[47] I. Alonso, L. Riazuelo, L. Montesano, and A. C. Murillo, "3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5432–5439, 2020.

[48] T. Cortinhal, G. Tzelepis, and E. E. Aksoy, "Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds," in *International Symposium on Visual Computing*. Springer, 2020, pp. 207–222.

[49] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421.

[50] L. T. Triess, D. Peter, C. B. Rist, and J. M. Zöllner, "Scan-based semantic segmentation of lidar point clouds: An experimental study," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1116–1121.

[51] Q. Gao and X. Shen, "Thickseg: Efficient semantic segmentation of large-scale 3d point clouds using multi-layer projection," *Image and Vision Computing*, vol. 108, p. 104161, 2021.

[52] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[54] E. E. Aksoy, S. Baci, and S. Cavdar, "Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 926–932.

[55] A. Xiao, X. Yang, S. Lu, D. Guan, and J. Huang, "Fps-net: A convolutional fusion network for large-scale lidar point cloud segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 176, pp. 237–249, 2021.

[56] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[57] A. Hatamizadeh, D. Yang, H. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," *arXiv preprint arXiv:2103.10504*, 2021.

[58] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo, "Segmenting transparent object in the wild with transformer," *arXiv preprint arXiv:2101.08461*, 2021.

[59] S. Wu, T. Wu, F. Lin, S. Tian, and G. Guo, "Fully transformer networks for semantic image segmentation," *arXiv preprint arXiv:2106.04108*, 2021.

[60] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[61] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[62] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu *et al.*, "A multi-organ nucleus segmentation challenge," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1380–1391, 2019.

[63] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Medical image analysis*, vol. 35, pp. 489–502, 2017.

[64] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.
[65] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2017.
[66] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8856–8865.
[67] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong, "Improving semantic segmentation via decoupled body and edge supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 435–452.
[68] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation," in *International Symposium on Neural Networks*. Springer, 2019, pp. 388–401.
[69] H. Shuai, X. Xu, and Q. Liu, "Backward attentive fusing network with local aggregation classifier for 3d point cloud semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 4973–4984, 2021.
[70] F. Zhang, J. Fang, B. Wah, and P. Torr, "Deep fusionnet for point cloud semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 644–663.
[71] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9601–9610.
[72] Y. A. Alnaggar, M. Afifi, K. Amer, and M. ElHelw, "Multi projection fusion for real-time semantic segmentation of 3d lidar point clouds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1800–1809.
[73] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.
[74] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

Hui-Xian Cheng Hui-Xian Cheng received the B.S. degree from Wuhan University Of Technology, in 2019. He is currently pursuing the M.S. degree in the College of Computer and Information Science at the Southwest University, Chongqing, China. His research interests include 3D scene understanding, point cloud semantic segmentation and multi-modal fusion.

Xian-Feng Han Xian-Feng Han received the Ph.D. degree in software engineering from Tianjin University, in 2019. He is currently a lecturer with the College of Computer and Information Science, Southwest University, Chongqing, China. His research interests include 3D point cloud processing, 3D reconstruction, and machine learning.

Guo-Qiang Xiao received the Ph.D. degree in signal and information processing from University of Electronic Science and Technology of China, Chengdu, and B.E. degree in radio technology from Chongqing University, Chongqing, China, in 1999 and 1986, respectively. Since 1986, he has been with the College of Computer and Information Science, Southwest University, Chongqing, China, where he is currently a Professor. His research interests include computer vision, pattern recognition, machine learning, and big data mining.