2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

RegionNet: Region-feature-enhanced 3D Scene Understanding Network with Dual Spatial-aware Discriminative Loss

Guanghui Zhang[1,2], Dongchen Zhu[1], Xiaoqing Ye[3], Wenjun Shi[1,2], Minghong Chen[1,2], Jiamao Li[1,2], and Xiaolin Zhang[1,2,4]

Abstract— **Neural networks have recently achieved impressive success in semantic and instance segmentation on 2D images. However, their capabilities have not been fully explored to address semantic instance segmentation on unstructured 3D point cloud data. Digging into the regional feature representation to boost point cloud comprehension, we propose a region-feature-enhanced structure consisting of adaptive regional feature complementary (ARFC) module and affinity-based regional relational reasoning (AR^3) module. The ARFC module aims to complement low-level features of sparse regions adaptively. The AR^3 module emphasizes on mining the potential reasoning relationships between high-level features based on affinity. Both the ARFC and AR^3 modules are plug-and-play. Besides, a novel dual spatial-aware discriminative loss is proposed to improve the discrimination of instance embedding. Our proposal-free point cloud instance segmentation network (RegionNet) equipped with the region-feature-enhanced structure and dual spatial-aware discriminative loss achieves state-of-the-art performance on S3DIS dataset and ScanNet-v2 dataset.**

I. INTRODUCTION

3D scene understanding plays a critical role in many robotics applications, such as outdoor autonomous navigation and indoor service robots. However, 3D point cloud data is sparse, non-uniform and unordered. It remains some challenges on the point cloud semantic instance segmentation, which is an important task of scene understanding.

Point cloud semantic instance segmentation includes semantic segmentation and instance segmentation tasks. Semantic segmentation labels each point with an object category which it belongs to, while instance segmentation distinguishes each object. Effective feature extraction acts a vital part in the two segmentation tasks. Nowadays feature extraction methods can be roughly categorized into voxel-based approaches [1], [2], which bring high computational and memory costs, and point-based approaches, which are flexible and efficient. PointNet [3] is the pioneer work of point-based methods, which learns the point-level feature embedding on unordered point clouds directly and exploit max pooling to aggregate the point features. As thus, PointNet fails to capture local structures represented by neighboring points. Many methods [4], [5], [6], [7], [8], [9], [10] have been proposed to solve this problem. PointNet++ [4] is the most representative among them. It hierarchically processes a set of points sampled (centroids) to capture local structures. However, [4] has two limitations that restrict its performance. For one thing, ball query adopted in [4] neglects the density non-uniformity of point clouds, which is not conducive to the presentation of low-level features. The multi-scale or multi-resolution strategies slightly alleviate the problem at high computational cost. For another, regional relational reasoning plays a key role in accurate 3D scene understanding for humans. Relational reasoning aims at explaining the interactions between local regions. For instance, the legs of tables are symmetrical, and tables are usually near chairs. Due to inadequate mining of relational reasoning in [4], it is easy to cause part or instance confusion leading to low segmentation accuracy in segmentation tasks.

Aiming at the first problem, we propose a lightweight adaptive regional feature complementary module to complement low-level features in sparse regions by combining the max-relative features of their kNN regions. With regard to the second problem, inspired by relational networks [11], [12], [13], we propose an affinity-based regional relational reasoning module to mine potential reasoning relationships between high-level regional features.

In addition, instance segmentation is more challenging than semantic segmentation. Proposal-free methods [14], [15], [16], [17], [18] generate point-level feature embedding and then apply a cluster algorithm to group points to segment 3D instances. They can avoid the expensive non-maximum suppression to prune dense object proposals of proposal-based methods [19], [20]. Thus, we design a proposal-free network for 3D semantic instance segmentation. Generating discriminative instance embedding is a key to boosting instance segmentation performance. Intuitively, adjacent instances in space are grouped into one instance due to similar feature embedding more easily than nonadjacent instances. The idea is important, but ignored by existing methods. Accordingly, we design a dual spatial-aware discriminative loss to learn more discriminative instance embedding. To sum up, our main contributions are as follows:

- A novel region-feature-enhanced structure including ARFC and AR^3 modules is proposed. The structure boosts low and high level regional features by ARFC

*This work was supported in part by the National Key R&D Plan of the Ministry of Science and Technology under Grant No. 2017YFC0805502, in part by the National Natural Science Foundation of China under Grant 61671014, in part by the Shanghai Municipal Science and Technology Major Project (ZHANGJIANG LAB) under Grant 2018SHZDZX01, in part by Shanghai Sailing Program 19YF1456000, and in part by Youth Innovation Promotion Association, Chinese Academy of Sciences(2018270).

[1] Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China jmli@mail.sim.ac.cn

[2] University of Chinese Academy of Sciences, Beijing 100049, China.

[3] Baidu, Shanghai 201210, China.

[4] ShanghaiTech University, Shanghai 201210, China.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

and AR^3 modules, respectively. The former complements the low-level features of shallow sparse regions and the later reasons about the relationships between high-level regional features based on affinity. The two modules are plug-and-play, which can be directly embedded into other point-based architectures.

- The dual spatial awareness for discriminative loss is presented, which aims to narrow the intra-instance gap and enlarge the inter-instance margin by leveraging intra and inter location space knowledge.
- A proposal-free point cloud instance segmentation framework (RegionNet) equipped with region-feature-enhanced structure and dual spatial-aware discriminative loss is designed. The RegionNet achieves state-of-the-art performance on S3DIS and ScanNet-v2 datasets.

II. Related Work

A. Deep Learning on Point Clouds

The recent availability of indoor scene datasets has sparked research in point clouds by deep learning. One of popular research methods is voxel-based methods [1], [2], which convert point clouds into regular volumetric occupancy grids and perform voxel-level predictions via 3D convolutional neural networks. However, voxel-based methods have the disadvantages of high complexity and storage redundancy, which make them difficult to address large-scale 3D scenes. Some methods [3], [4], [5], [6], [7], [8], [9], [21] have been designed to process point clouds directly. PointNet [3] and PointNet++ [4] are widely used in point cloud feature extraction. Recurrent Neural Networks (RNNs) and Graph Convolutional Networks (GCNs) are two extended pipelines following the spirt of PointNet. RNNs-based methods [5], [6] and GCNs-based methods [7], [8], [22] focus on exploring long-range spatial dependencies and construction of graph. The most above-mentioned methods determine the local regions by ball query or kNN searching ignoring density non-uniformity of point clouds. [4] designed multi-scale and multi-resolution strategies to alleviate the problem at high computational cost. Our lightweight ARFC module exploits max-relative features of kNN regions to complement low-level features to solve density non-uniformity of point clouds.

B. Relational Reasoning

Relational modules have been designed to solve some problems [23], [12], [24], [11]. Santoro *et al.* [23] present a relational network (RN) for a visual question answering task. Zhou *et al.* [12] propose a temporal relational network to explain the interactions between frames of videos. Inspired by the success of these methods, relational reasoning on 3D data has only started to be tackled in the literature [13], [25], [26]. More recently, [25] proposes clustering the local region features and point coordinates for learning the relationships between regions. Duan *et al.* [13] design a structural relational reasoning module for 3D semantic segmentation, which is the most relevant work to ours. However, simple concatenation makes half of the feature channels remain unchanged, which limits the representation of the relationship in [13]. Our AR^3 module learn the relationships based on affinity.

C. Instance Segmentation

Works on 2D instance segmentation can be roughly classified into proposal-based and proposal-free methods. Proposal-based methods [27], [28], [29] exploit region proposals to locate the object and then obtain the corresponding mask to classify whether the patch contains an object. Proposal-free methods [30], [31], [32], [33] are usually composed of segmentation branch and clustering-purpose branch. The pixel-wise mask prediction is obtained by the segmentation branch, and pixels belonging to a certain instance are clustered by clustering-purpose branch. Similarly, proposal-free methods [14], [15], [16], [17], [18] on 3D point clouds generate point-level feature embedding and then cluster them to segment instances. SGPN [14] learns to group point features through a similarity matrix and uses double-hinge loss to supervise the similarity matrix. [15], [16], [34] jointly optimize semantic and instance segmentation and use discriminative loss [32] to learn point-level instance embedding. Note that, [17], [18], [34] are voxel-based methods, whereas [14], [15], [16] are point-based methods, which are same as ours. Another recent pipeline is single-stage 3D-BoNet [35] regressing 3D bounding box and predicting a point-level mask for each instance.

III. Our Approach

In this section, we first describe the whole network (Section III-A). Then we introduce the proposed ARFC module complementing low-level regional features in sparse regions (Section III-B) and the AR^3 module mining the reasoning relationships between high-level regional features (Section III-C). Finally, the proposed dual spatial-aware discriminative loss supervising the instance embedding learning is introduced (Section III-D).

A. Network Architecture

An overview of our approach is illustrated in Fig. 1. In the encoding stage, we use region-feature-enhanced structure (*i.e.* ARFC and AR^3 modules) to reinforce regional features after SA (set abstraction) [4] operation at each layer, whereas [15] only adopts stacked SA. Significantly, the proposed ARFC module is utilized to complement low-level features. With the network getting deeper and the receptive field getting larger, the reasoning relationships between high-level regional features are learned through our AR^3 module. In the decoding stage, one of the decoders is for point-level semantic predictions, while the other one is for instance embedding learning [15]. The proposed dual spatial-aware discriminative loss aiming to narrow the intra-instance gap and enlarge the inter-instance margin by combining location space knowledge is employed to learn instance embedding. Compared with naive discriminative loss in baseline [15], our core contribution lies in dual spatial awareness. Finally, mean-shift clustering and BlockMerging algorithm [14] are utilized to obtain instance labels. The mode of semantic of the points within the same instance as semantic label.

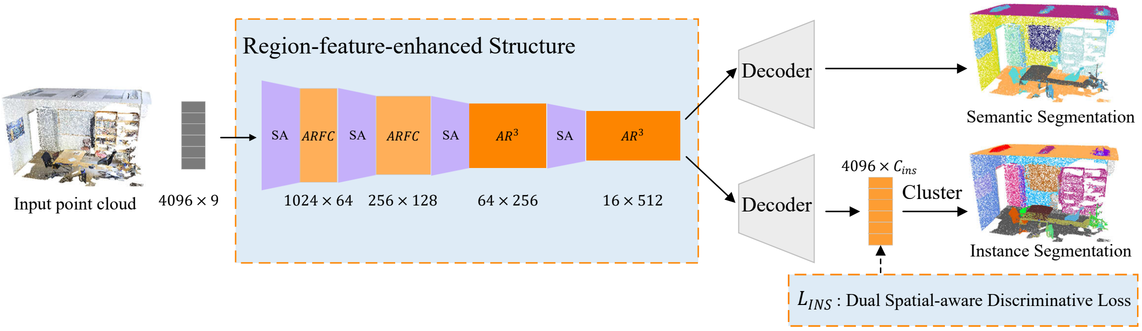


Fig. 1. An overview of the proposed RegionNet. The inputs are point clouds with xyz coordinates and RGB attributes (9 means xyz, RGB, and normalized xyz). The output includes two parts: semantic labels and instance labels. The proposed region-feature-enhanced structure includes ARFC and AR^3 modules, that is, embedding the ARFC and AR^3 modules into vanilla PointNet++ architecture (without multi-scale grouping). The ARFC module aims at adaptively complementing low-level features in shallow sparse regions (Fig. 2), and the AR^3 module aims at reasoning about relationships between high-level regional features (Fig. 3). Besides, the dual spatial-aware discriminative loss is proposed to supervise instance embedding learning (Fig. 4). SA denotes set abstraction [4]. The down-sampling in SA samples the input point clouds into 1024, 256, 64, and 16 points sequentially.

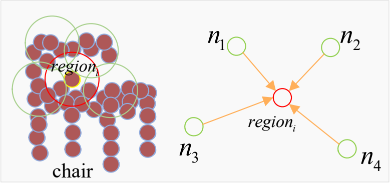


Fig. 2. ARFC mechanism. Given input feature of sparse region $region_i$, complement the features by exploiting max-relative features of kNN regions.

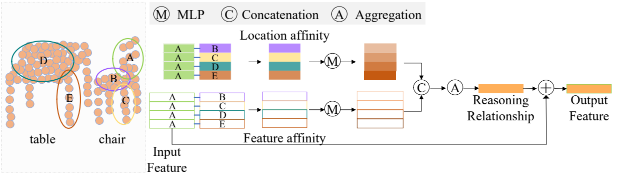


Fig. 3. AR^3 module. Taking local region A as an example, affinities between A and remaining regions are calculated in location and feature embeddings. Then potential reasoning relationship is achieved by MLP, concatenation, and aggregation progressively. Local regions in our network are obtained by SA, which do not require any extra supervision signals.

B. Adaptive Regional Feature Complementary Module

Due to the density non-uniformity of the point clouds, there are few points in $region_i$ (marked in red in Fig. 2) obtained by ball query. Thus, we propose an ARFC module to adaptively complement the low-level features of such sparse regions to alleviate the problem of density non-uniformity by exploiting the max-relative features of its kNN regions (*i.e.* dissimilar missing features), as shown in Fig. 2.

Given a region set $\mathbb{R} = \{r_1, r_2, ..., r_N\}$. We denote F_{r_i} as input feature of region r_i. F'_{r_i} is output complementarity feature. We formulate the ARFC module as

$$F'_{r_i} = f_c(F_{r_i}, F_{n_j}),\ n_j \in N_k(r_i) \tag{1}$$

where $N_k(r_i)$ denotes the kNN regions of r_i. f_c denotes the complementarity process taking F_{r_i}, F_{n_j} as inputs and F'_{r_i} as output, which will be elaborated below.

Inspired by [22], we use dynamic max-relative GCN to hierarchically absorb the relative features of neighboring regions. To begin with, we exploit a max aggregator to aggregate the feature difference, which is used for extracting max relative features of neighborhood regions. Then, we concatenate the feature $F_{r_i^l}$ and the max-relative features of kNN regions $F_{N_k}(r_i^l)$, followed by a multilayer perceptron (MLP). The $l+1$ layer regional feature $F_{r_i^{l+1}}$ is acquired after adding the $F_{r_i^l}$ (Eq. 2).

$$\begin{aligned} F_{N_k(r_i^l)} &= max(F_{n_j^l} - F_{r_i^l}),\ n_j^l \in N_k(r_i^l) \\ F_{r_i^{l+1}} &= F_{r_i^l} + MLP\Big(\big[F_{r_i^l}, F_{N_k}(r_i^l)\big]\Big) \end{aligned} \tag{2}$$

where l denotes the layer, $l = 0, 1, ..., l_{max}$. $[,]$ denotes concatenate operation. Note that, we draw lesson from the idea of dynamic graph construction in [8] to dynamically select useful information of different levels. So for the 1st layer, i.e., $l = 0$, we search the kNN regions of r_i according to the location coordinates (xyz) of the centroids, when $l \geq 1$, according to the feature obtained from the last layer.

Finally, we fuse the features from different layers to realize the complementary feature F'_{r_i} (Eq. 3).

$$F'_{r_i} = MLP\big([F_{r_i^1}, F_{r_i^2}, ..., F_{r_i^{l_{max}}}]\big) \tag{3}$$

Remarkably, our ARFC module has the same input and output dimensions, which can be directly embedded into point-based architectures to complement low-level features. In our network, we embed the module (in Fig. 1) into the first two layers of the region-feature-enhanced structure.

C. Affinity-based Regional Relational Reasoning Module

Local regions in a scene are potentially relevant. For example, the legs of tables are symmetrical, and tables are usually near chairs. With the exploitation of reasoning relationships, the models are able to understand 3D objects more comprehensively. From the perspective of bionics, we generally first observe shapes or outlines of sub-regions to judge the simple affinity rather than concatenate them like [13] when reasoning about the relationships between sub-regions, which has been indirectly proved effective in [21]. Then further relationships between them are mined through progressive reasoning. Therefore, we propose the AR^3 module, as shown in Fig. 3.

Given a regional feature set $\mathbb{F} = \{F_1, F_2, ..., F_N\}$. The AR^3 module can be formulated as

$$F_i' = F_i + f_r(F_i, F_j),\ F_j \ \forall \ \mathbb{F} \tag{4}$$

where f_r denotes the reasoning process taking F_i, F_j as inputs and reasoning relationship as output (see Fig. 3), which will be described in detail in the following.

Firstly, we calculate the affinity between local regions. The two-norm and one-norm distances are simple ways to calculate the affinity. However, this symmetric strategy is problematic since the impact of region A on B is the same as that of B on A [21]. In fact, A can deduce B, while B may not be able to deduce A. Consequently, we model the embedding difference as affinity, which is simple but effective. As shown in Fig. 3, the AR^3 module first calculates the affinity of location embedding as well as feature embedding. Location embeddings refer to xyz coordinates of centroids (i.e, sampled points) here. We argue that the affinity of location embedding represents real distance and structure information between two regions, which also implies the receptive filed size of the sub-region. Reasoning in feature embedding can quarry potential repetitive regional patterns.

Secondly, further reasoning processes on location and feature affinities are accomplished by respective MLP. Then we concatenate and aggregate them effectively to realize reasoning relationship. The aggregation operation is implemented by Eq. 5.

$$h\big([x, y]\big) = \mathcal{F}_C^C\Big(\sum_{i=1}^{N-1} \mathcal{F}_{3+C}^C\big([x, y]\big)\Big) \tag{5}$$

where $\mathcal{F}$ denotes the 1×1 convolution. The subscript represents the number of input channels, and the superscript represents the number of output channels.

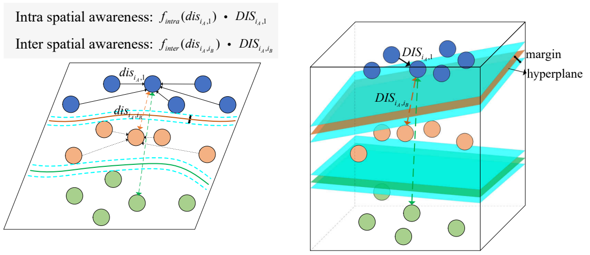
Eventually, we add reasoning relationship to input feature, and the enhanced high-level regional feature F_i' is achieved.

The AR^3 module is also a plug-and-play module without extra supervision signals that can be easily integrated into point-based architectures to enhance feature representation of large regions. We plug the AR^3 module (in Fig. 1) into the last two layers of region-feature-enhanced structure to extract reasoning relationships in multi-scale local features.

To be clear, although both AR^3 module and ARFC module are designed to enhance regional feature, there are at least three obvious differences between them. Firstly, the AR^3 module acts in the high-level features with large receptive field, while the ARFC module acts in the low-level features of shallow sparse regions. Secondly, the AR^3 module exploits MLPs to reason about the relationships between a sub-region and all other sub-regions, while ARFC module adopts a max aggregator to hierarchically aggregate max-relative features of kNN regions to complement. Thirdly, location information is used in quite a different way. AR^3 module learns location relationships, while ARFC module finds neighboring regions.

D. Dual Spatial-aware Discriminative Loss

As shown in Fig. 4, a good segmentation can be achieved by the hyperplane with large margin. Thus we hope the margin between the two instances in feature space to be large and the gap of feature embeddings of points belonging to the same instance to be small. Based on discriminative loss [15], we propose the dual spatial awareness for improvement, dubbed dual spatial-aware discriminative loss including intra spatial-aware and inter spatial-aware discriminative losses.


Intra spatial awareness: $f_{intra}(dis_{i_A,1}) \cdot DIS_{i_A,1}$
Inter spatial awareness: $f_{inter}(dis_{i_A,i_B}) \cdot DIS_{i_A,i_B}$
$dis_{i_A,1}$
dis_{i_A,i_B}
$DIS_{i_A,1}$
DIS_{i_A,i_B}
margin
hyperplane

(a) Location space

(b) Feature space

Fig. 4. Dual spatial-aware discriminative loss. Dots of different colors denote different instances. dis denotes distance in location space, DIS denotes distance in feature space. f_{intra} and f_{inter} refer to $w_{i,j}$ and w_{i_A,i_B} in Section III-D, respectively.

Within an object instance, feature embeddings of points closer to the center are more likely to be similar to the center feature embedding, while feature embeddings of points farther from the center (near the edge) are more likely to be different. The points farther from the center should be given a greater suction force, namely intra spatial awareness. The intra spatial-aware loss is defined as

$$L_{intra} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{N_i}\sum_{j=1}^{N_i} w_{i,j}\left[\|f_j - \mu_{f,i}\|_1 - \delta_v\right]_+^2 \tag{6}$$

where I is the number of ground-truth instances, N_i is the number of points in instance i and $[x]_+ = max(0; x)$ means the hinge. $\|\cdot\|$ denotes ℓ_1 distance. $\mu_{f,i}$ is the mean feature embedding of instance i. f_j is a feature embedding of a point. δ_v is a margin. The intra spatial awareness is implemented by $w_{i,j}$, which is inspired by [17]. However, their loss weakens intra loss item, and experimentally we found the origin loss does not provide improvement in our network. We avoid weakening by scale transform and adopt informative points.

Specifically, given an instance i with M points, we first use Principal Component Analysis (PCA) to obtain informative points. The representative center location coordinate $\mu_{l,i}$ is obtained by further averaging informative points. We define the intra spatial awareness based on Laplacian kernel as

$$w_{i,j} = \frac{2}{1 + e^{-\|l_j - \mu_{l,i}\|_1}}, \quad w_{i,j} \in (1, 2) \tag{7}$$

where l_j is a location coordinate of a point.

Between different object instances, feature embeddings of the two centers closer to each other are more likely to be similar, while feature embeddings of the two centers farther apart are more likely to be different. The closer the centers of two instances, the greater the repulsion, namely inter spatial awareness. The inter spatial-aware loss can be expressed as

$$L_{inter} = \frac{1}{I(I-1)}\sum_{i_A=1}^{I}\sum_{i_B=1}^{I} w_{i_A,i_B}\left[2\delta_d - \|\mu_{f,i_A} - \mu_{f,i_B}\|_1\right]_+^2 \tag{8}$$

where $i_A \neq i_B$. δ_d is a margin. Similar to intra spatial awareness, the inter spatial awareness is constructed as

$$w_{i_A,i_B} = 1 + \frac{2}{1 + e^{\|s_{l,i_A} - s_{l,i_B}\|_1}},\ w_{i_A,i_B} \in (1,2) \qquad (9)$$

Instead of averaging informative points obtained by PCA in intra spatial awareness term, we calculate the distance between informative point set s_{l,i_A} and informative point set s_{l,i_B} directly to take shape of objects into account. Besides, we scale the weight range to (1, 2), which can preserve the role of the discriminative loss [15].

Additionally, like [15], regularization term is incorporated into instance segmentation loss to keep the embedding values bounded, which is defined as $L_{reg} = \frac{1}{I}\sum_{i=1}^{I}\|\mu_{f,i}\|_1$. So

$$L_{INS} = L_{intra} + L_{inter} + \alpha \cdot L_{reg} \qquad (10)$$

where α is a hyperparameter. As for semantic prediction, the cross-entropy loss L_{SEM} is chosen in this paper. Above all, our total train loss is $L = L_{SEM} + L_{INS}$.

IV. Experiment

A. Datasets and Implementation Details

Datasets We evaluate our approach on 3D instance segmentation on the two datasets: Stanford 3D Indoor Semantics Dataset [36] (S3DIS) and ScanNet-v2 dataset [37]. The S3DIS dataset contains 3D scans from Matterport Scanners in 6 areas including 272 rooms. Each point in the scene point cloud has an instance label and one of the semantic labels from 13 categories. The ScanNet-v2 dataset is obtained by fusing multiple scans from different views. It contains 1513 scanned and reconstructed indoor scenes, and each point has an instance label and one of the semantic labels from 40 categories. We employ 1201 scenes as the training set and the rest 312 scenes as the test set.

Implementation Details Our framework was implemented based on TensorFlow using Adam optimizer on single NVIDIA Titan-X with 12GB memory. We train our network for 50 epochs, with batch size 12, base learning rate 0.001, which is divided by 2 every 300k iterations. Same as [15], δ_v and δ_d are set to 0.5, 1.5, respectively. α is set to 0.001. C_{ins} is set to 5. l_{max} and k are set to 1, 16 in ARFC module respectively. Following [15], [35], we split the rooms into 1m^2 overlapped blocks containing 4096 points.

Evaluation Metrics Following [3], [15], we report our experiments testing on Area 5 and 6-fold cross validation results on S3DIS dataset. For evaluation of semantic segmentation, overall accuracy (oAcc), mean accuracy (mAcc) and mean intersection over union (mIoU) across all the categories are calculated and reported. For evaluation of instance segmentation, (weighted) coverage (Cov, WCov) [38], [15], as well as classical metrics mean precision (mPrec) and mean recall (mRec) with IoU threshold 0.5 are also calculated and reported. Besides, as for Scannet-v2 dataset, we use general average precision (AP) as the instance evaluation metric with different IoU thresholds, such as 0.5, 0.25.

TABLE I
INSTANCE AND SEMANTIC SEGMENTATION RESULTS ON S3DIS.

Method	Instance metric				Semantic metric		
	mCov	mWCov	mPrec	mRec	oAcc	mAcc	mIoU
Test on Area 5							
SGPN [14]	32.7	35.5	36.0	28.7	-	-	-
JSIS3D [16]	32.6	35.6	39.7	29.1	82.8	49.8	42.0
3D-BoNet [35]	41.5	44.6	50.1	39.0	86.9	59.2	51.8
ASIS [15]	47.1	50.1	51.8	44.5	86.5	61.3	53.0
Ours	**50.2**	**53.4**	**58.8**	**47.5**	**87.9**	**64.0**	**55.6**
Test on 6-fold Cross Validation							
SGPN [14]	37.9	40.8	38.2	31.2	-	-	-
JSIS3D [16]	37.3	41.0	49.5	33.4	79.9	59.8	48.5
3D-BoNet [35]	48.4	52.4	65.6	47.6	86.3	69.3	59.4
ASIS [15]	53.1	56.9	62.0	49.0	86.2	71.7	59.6
Ours	**54.6**	**58.6**	**66.0**	**52.1**	**87.9**	**72.9**	**62.6**

TABLE II
PER CLASS RESULTS ON S3DIS DATASET AREA 5. MOST METHODS DO NOT PERFORM WELL ON "BEAM", WHICH HAS FEW POINTS (0.029%).

	mean	celing	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
Instance metric WCov														
JSIS3D [16]	35.6	82.5	86.0	52.5	0.0	3.6	51.3	14.8	42.4	54.0	5.3	36.1	5.7	28.4
3D-BoNet [35]	45.7	89.4	87.8	71.8	0.0	10.8	6.3	10.3	46.3	68.2	10.5	45.3	47.2	43.3
ASIS [15]	50.1	88.9	**89.8**	70.4	0.0	10.9	61.7	6.0	50.4	70.5	35.8	**56.4**	**64.7**	45.5
Ours	**53.4**	**91.3**	88.3	**75.1**	0.0	**12.9**	**70.6**	**19.1**	**52.1**	**77.5**	**37.0**	53.5	63.4	**52.6**
Semantic metric IoU														
JSIS3D [16]	42.0	89.3	96.3	70.3	0.0	5.9	47.9	14.8	57.3	61.8	9.1	46.7	7.8	38.5
3D-BoNet [35]	58.9	91.5	**98.4**	76.7	0.0	**16.7**	51.8	28.0	68.5	74.8	19.2	57.5	42.7	48.0
ASIS [15]	53.0	91.7	96.8	74.4	0.0	8.2	48.6	16.0	72.4	80.4	38.2	**59.0**	**53.1**	50.3
Ours	**55.6**	**93.2**	97.6	**78.8**	0.0	15.0	**54.6**	**29.2**	**74.7**	**80.6**	**38.4**	56.9	50.7	**53.2**

B. Evaluation on S3DIS Dataset

We compared our method with baseline ASIS [15], which is based on vanilla PointNet++ architecture, and other state-of-the-art methods, including SGPN [14], JSIS3D [16], 3D-BoNet [35] on S3DIS dataset. Table I reports instance segmentation results. Our method achieves 58.6 mWCov, which outperforms ASIS[1] by 1.7 when evaluating by 6-fold cross validation. In terms of mPrec, a larger 4.0 gain is yielded. Besides, the recall rate mRec also achieves significative 3.1 improvement. When testing on Area 5, the improvements are consistent across the four evaluation metrics. We show some instance segmentation results in Fig. 5. With the help of the proposed region-feature-enhanced structure, our method performs better than other methods in those regions having sparse points or potential reasoning relationship. Besides, nearby different instances (e.g., door and wall, chair and chair) are distinguished well as the proposed dual spatial-aware discriminative loss helps network enlarge the margin in feature space. As shown in Table II, performance gains on door, chair are in line with our observations.

Table I also illustrates the semantic segmentation results. Our method outperforms other state-of-the-art methods. As a whole, performances of more than half of object categories achieve the highest score, as depicted in Table II. Some examples are visualized in Fig. 6. Our method helps semantic segment completely and correctly. Taking the sofa (line 3 in

[1] We reproduced the results of ASIS using the code at github, published by the authors. Note, limited to resources, we set the batchsize to 12, and train it for 50 epochs, which is same as our method.

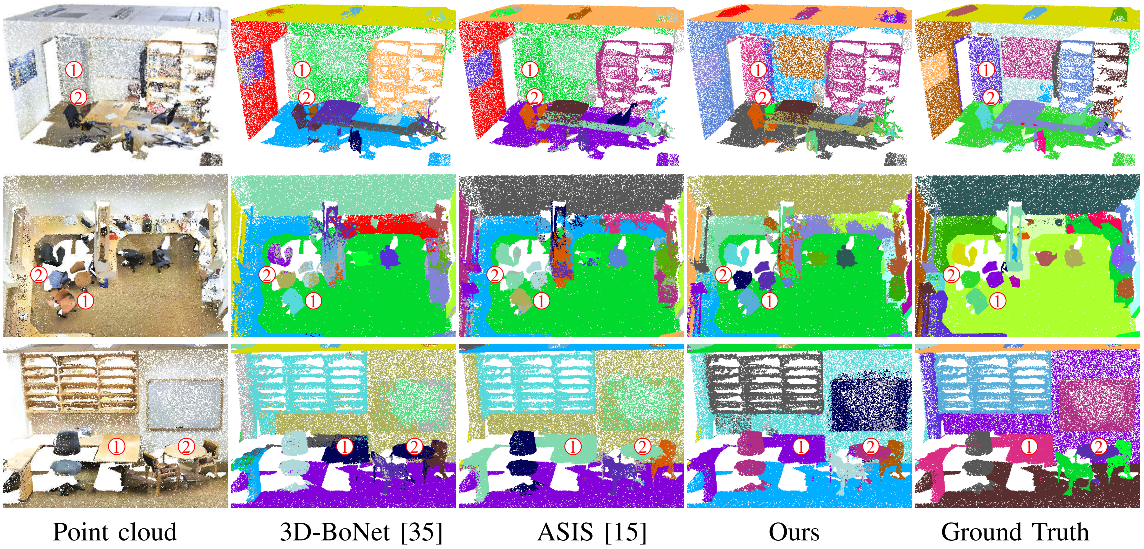


Fig. 5. The visualization comparison of our method and [35], [15] for instance segmentation on S3DIS. Different colors represent different instances.

TABLE III
INSTANCE SEGMENTATION RESULTS ON SCANNET-V2 DATASET. POINT-BASED METHODS INCLUDING SGPN [14], GSPN [20], ASIS [15] ARE REPORTED. 3D-BONET [35] ADOPTING VOXEL-BASED BACKBONE WHEN EVALUATING ON THE DATASET IS EXCLUDED.

	SGPN [14]	GSPN [20]	ASIS [15]	Ours
$AP_{0.5}$	35.1	37.8	44.5	**46.1**
$AP_{0.25}$	-	53.4	64.0	**69.1**

Fig. 6) as an example, the points on its surfaces are sparse due to illumination. Thanks to ARFC module complementing low-level regional feature and AR^3 module reasoning about relationships between high-level regional features, the better segmentation result is realized.

C. Evaluation on ScanNet-v2 Dataset

Further, we conduct experiments on Scannet-v2 dataset. ScanNet [37] presented a voxel-based coarse prediction framework, most methods based on which are voxel-based methods. Instead, our method is point-based. So we make comparisons with SGPN [14] and GSPN [20]. Table III reports the experiment results. Our method outperforms [20] by 8.3 at $AP_{0.5}$. Significantly, our approach is superior to [14] by a large margin at $AP_{0.5}$. Besides, we also retrain and test baseline ASIS on the dateset, our methods still obtain better results. We submit our results to ScanNet-v2 benchmark, and achieve state-of-the-art performance in point-based methods with $AP_{0.5}$ 24.8 and $AP_{0.25}$ 47.4, while SGPN [14] 14.3 and 39.0. This further proves our RegionNet is effective and adaptable to various indoor datasets.

D. Ablation Study

To evaluate the effectiveness of each component of our framework, we conduct ablation experiments on Area 5 of S3DIS dataset. The base network is ASIS based on vanilla PointNet++ (without multi-scale grouping) and discriminative loss. Table IV reports the ablation experiments of main components including ARFC module, AR^3 module, and dual spatial-aware discriminative loss (DSDLoss). Specifically, our method equipped with any one of the components achieves better performance than base. In terms of mPrec, they all have a more than 3 improvement. Further, we test the model equipped with ARFC and AR^3, namely regional-feature-enhanced structure, which is superior to the model equipped with only ARFC or AR^3. Finally, the full framework achieves optimal performance across four instance evaluation metrics and three semantic evaluation metrics. This shows that our designed regional-feature-enhanced structure and dual spatial-aware discriminative loss are effective.

In order to further prove the effectiveness of our ARFC and AR^3 modules, we visualize a close-up example in Fig. 7. Comparing (b) and (c) (boxed in solid line), the back of the chair with sparse points is segmented completely through adding ARFC module. Then comparing (c) and (d) (boxed in dotted line), the AR^3 module improves legs of the table.

TABLE IV
ABLATION STUDY THE REGION-FEATURE-ENHANCED STRUCTURE(CONSISTING OF ARFC, AR^3), AND DUAL SPATIAL-AWARE DISCRIMINATIVE LOSS (DSDLOSS) ON S3DIS DATASET AREA 5.

	Module		Loss	Instance metric				Semantic metric		
	ARFC	AR^3	DSDLoss	mCov	mWCov	mPrec	mRec	oAcc	mAcc	mIoU
Base				47.1	50.1	51.8	44.5	86.5	61.3	53.0
Ours	✓			48.1	51.2	55.0	44.9	87.7	62.1	54.7
		✓		48.6	51.6	54.9	45.4	87.3	62.0	54.1
			✓	47.9	51.0	55.8	45.6	87.2	63.7	55.1
	✓	✓		50.0	53.0	56.0	45.5	87.7	62.7	55.2
	✓	✓	✓	**50.2**	**53.4**	**58.8**	**47.5**	**87.9**	**64.0**	**55.6**

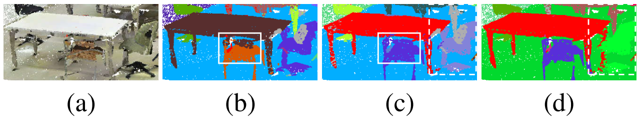


Fig. 7. A close-up part confusion example of region-feature-enhanced structure. (a)RGB; (b)Base; (c)Base+ARFC; (d)Base+ARFC+AR^3.

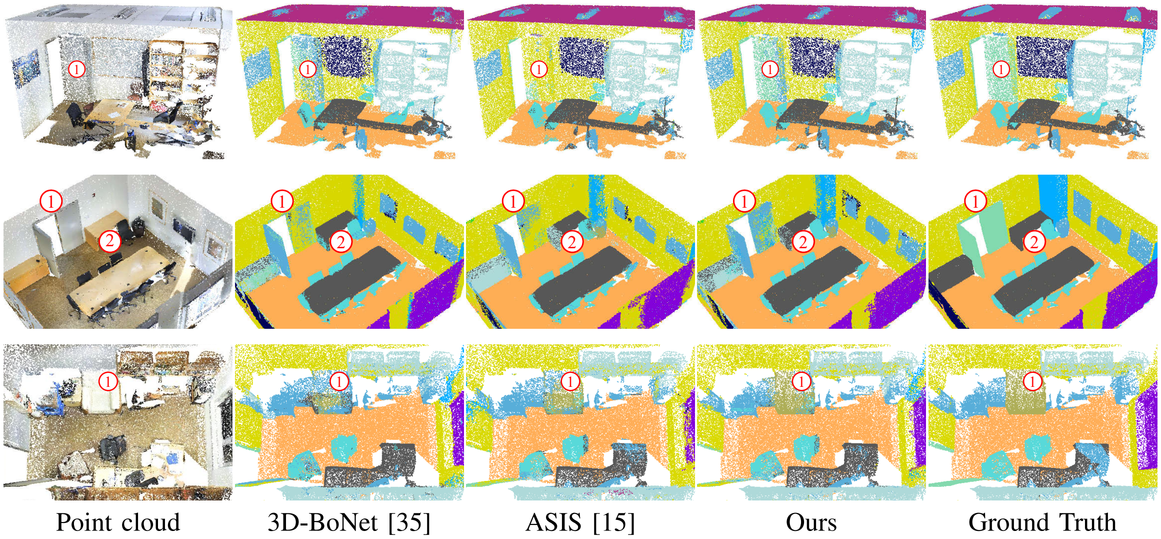


Fig. 6. The visualization comparison of our method and state-of-the-art methods for semantic segmentation on S3DIS dataset.

TABLE V

Ablation study for ARFC module.

	mWCov	mIoU	inference speed(ms)
Base	50.2	53.0	**151**
Base-MSG [4]	50.1	53.7	178
Base-ARFC(ours)	**51.2**	**54.7**	155

TABLE VI

Ablation study for Dual Spatial-aware Discriminative loss.

	DSDLoss		Instance metric	
	Intra	Inter	mPrec	mRec
Base			51.8	44.5
	✓		53.7	43.4
		✓	54.6	44.1
	✓	✓	**55.8**	**45.6**

TABLE VII

Generality analysis for our ARFC and AR^3 modules.

	mWCov	mIoU
3D-BoNet [35]	44.6	51.8
3D-BoNet-ARFC	48.6	54.9
3D-BoNet-AR^3	46.3	52.5
3D-BoNet-ARFC-AR^3	**49.1**	**55.1**

They confirm that the ARFC module complements the low-level features in sparse regions, and the AR^3 module has the role of reasoning on high-level regional features to alleviate part confusion. Besides, we also compare our ARFC module with multi-scale grouping (MSG), which is a density adaptive strategy in PointNet++. As shown in Table V, our ARFC module exploiting max-relative features of neighboring regions achieves higher performance than MSG strategy with less inference time at the same input size ($4k$ points).

Additionally, we also prove the role of intra spatial awareness and inter spatial awareness in our DSDLoss. Table VI reports the results. It gets limited improvement compared with base when only intra spatial awareness is exploited, as shown in Fig. 8 (b) and (c) (marked in red). Because it just intends to converge the points in an instance without emphasizing the distinction between instances. When the inter spatial awareness is added, our precision is improved significantly, as shown in Fig. 8 (c) and (d) (marked in blue).

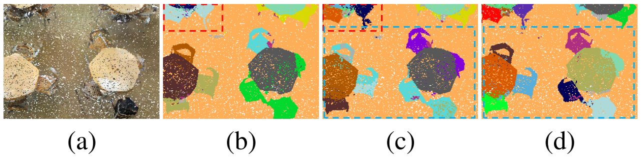


Fig. 8. A close-up instance confusion example of dual spatial-aware discriminative loss. (a) RGB; (b) Base; (c) Base+Intra; (d) Base+Intra+Inter.

E. Generality and Run-time Analysis

To verify the generality of the proposed plug-and-play ARFC and AR^3 modules, we embed them in other point-based networks, such as point-based 3D-BoNet [35]. As shown in Table VII, our ARFC and AR^3 modules significantly boost the performance of [35], which proves they have good generality. As for network computational complexity, the ARFC module exploits non-parameter max-pooling operation, and the AR^3 module is embedded into deep layers with least sampled points, which need little computational cost. Actually, the inference speed of the baseline equipped with the ARFC and AR^3 modules is 159ms, whereas the baseline ASIS is 151ms at the same input size ($4k$ points). This shows that the ARFC and AR^3 modules are lightweight.

V. Conclusion

We put forward a proposal-free region-feature-enhanced network termed RegionNet, focusing on reinforcing regional feature representation and improving discriminative point feature embedding. Our RegionNet successfully improves the performance of 3D semantic instance segmentation on indoor datasets, which demonstrates the effectiveness of incorporating low and high-level feature relationships between regions and dual spatial-aware knowledge. For future work, we plan to learn the relationships between concrete instances or parts instead of abstract regions generated SA.

ACKNOWLEDGMENT

Project supported by Shanghai Municipal Science and Technology Major Project (Grant No. 2018SHZDZX01, ZHANGJIANG LAB).

REFERENCES

[1] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 922–928.

[2] G. Riegler, A. Osman Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3577–3586.

[3] 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.

[4] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in neural information processing systems*, 2017, pp. 5099–5108.

[5] Q. Huang, W. Wang, and U. Neumann, "Recurrent slice networks for 3d segmentation of point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2626–2635.

[6] X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, "3d recurrent neural networks with context fusion for point cloud semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 403–417.

[7] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4558–4567.

[8] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," vol. 38, no. 5. ACM New York, NY, USA, 2019, pp. 1–12.

[9] W. Wu, Z. Qi, and L. Fuxin, "Pointconv: Deep convolutional networks on 3d point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9621–9630.

[10] L. Pan, P. Wang, and C.-M. Chew, "Pointatrousnet: Point atrous convolution for point cloud analysis," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4035–4041, 2019.

[11] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3588–3597.

[12] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–818.

[13] Y. Duan, Y. Zheng, J. Lu, J. Zhou, and Q. Tian, "Structural relational reasoning of point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 949–958.

[14] W. Wang, R. Yu, Q. Huang, and U. Neumann, "Sgpn: Similarity group proposal network for 3d point cloud instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2569–2578.

[15] X. Wang, S. Liu, X. Shen, C. Shen, and J. Jia, "Associatively segmenting instances and semantics in point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4096–4105.

[16] Q.-H. Pham, T. Nguyen, B.-S. Hua, G. Roig, and S.-K. Yeung, "Jsis3d: Joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8827–8836.

[17] Z. Liang, M. Yang, and C. Wang, "3d graph embedding learning with a structure-aware loss function for point cloud semantic instance segmentation," *arXiv preprint arXiv:1902.05247*, 2019.

[18] C. Liu and Y. Furukawa, "Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation," *arXiv preprint arXiv:1902.04478*, 2019.

[19] J. Hou, A. Dai, and M. Nießner, "3d-sis: 3d semantic instance segmentation of rgb-d scans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4421–4430.

[20] L. Yi, W. Zhao, H. Wang, M. Sung, and L. J. Guibas, "Gspn: Generative shape proposal network for 3d instance segmentation in point cloud," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3947–3956.

[21] H. Zhao, L. Jiang, C.-W. Fu, and J. Jia, "Pointweb: Enhancing local neighborhood features for point cloud processing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5565–5573.

[22] G. Li, M. Muller, A. Thabet, and B. Ghanem, "Deepgcns: Can gcns go as deep as cnns?" in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9267–9276.

[23] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in neural information processing systems*, 2017, pp. 4967–4976.

[24] 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.

[25] X. Wen, Z. Han, X. Liu, and Y.-S. Liu, "Point2spatialcapsule: Aggregating features and spatial relationships of local regions on point clouds using spatial-aware capsules," *arXiv preprint arXiv:1908.11026*, 2019.

[26] Y. Zhou, H. Chen, J. Xu, Q. Dou, and P.-A. Heng, "Irnet: Instance relation network for overlapping cervical cell segmentation," *arXiv preprint arXiv:1908.06623*, 2019.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[28] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "Masklab: Instance segmentation by refining object detection with semantic and direction features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4013–4022.

[29] G. Narita, T. Seno, T. Ishikawa, and Y. Kaji, "Panopticfusion: Online volumetric semantic mapping at the level of stuff and things," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4205–4212.

[30] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan, "Proposal-free network for instance-level object segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2978–2991, 2017.

[31] A. Fathi, Z. Wojna, V. Rathod, P. Wang, H. O. Song, S. Guadarrama, and K. P. Murphy, "Semantic instance segmentation via deep metric learning," *arXiv preprint arXiv:1703.10277*, 2017.

[32] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation with a discriminative loss function," *arXiv preprint arXiv:1708.02551*, 2017.

[33] Y. Liu, S. Yang, B. Li, W. Zhou, J. Xu, H. Li, and Y. Lu, "Affinity derivation and graph merge for instance segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 686–703.

[34] J. Lahoud, B. Ghanem, M. Pollefeys, and M. R. Oswald, "3d instance segmentation via multi-task metric learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9256–9266.

[35] B. Yang, J. Wang, R. Clark, Q. Hu, S. Wang, A. Markham, and N. Trigoni, "Learning object bounding boxes for 3d instance segmentation on point clouds," in *Advances in Neural Information Processing Systems*, 2019, pp. 6737–6746.

[36] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3d semantic parsing of large-scale indoor spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1534–1543.

[37] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.

[38] W. Zhuo, M. Salzmann, X. He, and M. Liu, "Indoor scene parsing with instance segmentation, semantic labeling and support relationship inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5429–5437.