Object Bounding Box-Aware Embedding for Point Cloud Instance Segmentation

Lixue Cheng, Taihai Yang, and Lizhuang Ma(✉)

East China Normal University, Shanghai, China
{clx_2021,thyang}@stu.ecnu.edu.cn,lzma@cs.ecnu.edu.cn

Abstract. In 2D image domain, recent researches have made significant progress in encoding context information for instance segmentation. While the counterpart in point cloud is still left far behind. Previous works mostly focus on leveraging semantic information and aggregating point local information through K-Nearest-Neighbor method. Such methods are unaware of object boundary information which is important to separating nearby objects. We propose a novel module to integrate object bounding box information into embedding for Point Cloud Instance Segmentation. The proposed module called Object Bounding Box-aware module (OBAM) boosts the instance segmentation performance by encoding Object Bounding Box information. Through attention mechanism, the module removes redundant boundary information. Comprehensive experiments on two popular benchmarks (S3DIS and ScanNetV2) show the effectiveness of our method. Our method achieves the State-of-the-art instance segmentation performance on S3DIS benchmark.

Keywords: 3D point cloud · Instance segmentation · Object bounding box-aware

1 Introduction

In computer vision, instance segmentation is a basic task for scene understanding. It is always regarded as an extension to semantic segmentation. The task of instance segmentation is to group pixels/points which have the identical semantic labels into different object instances. In 3D domain, instance segmentation has wild applications in robotics, autonomous driving. With the growth of 3D sensors, it has gained more researchers attention and some approaches have been proposed in some papers. However, it is far away from being solved.

Point cloud captured by 3D scanners is an important type of 3D data representation. It consists of collections of points in Euclidean space. In 3D point cloud, PointNet [5] is the pioneer deep-learning method directly using original point cloud as input. Subsequent method PointNet++ [6] abstracts local region information with PointNet to learn point features through a hierarchical structure. Methods like radius based ball query and K-Nearest-Neighbor are utilized

D. N. Pham et al. (Eds.): PRICAI 2021, LNAI 13033, pp. 182–194, 2021.
https://doi.org/10.1007/978-3-030-89370-5_14

for aggregating local region information. Our approach is building on PointNet++ network.

In 3D point cloud area, approaches for instance segmentation are mostly composed of clustering-based approaches and proposal-based approaches. To tackle the task of instance segmentation on point clouds, clustering-based approaches group points through clustering algorithm and proposal-based approaches are mostly based on object proposal. Semantic-aware instance segmentation is in ASIS [12]. They put two tasks (instance segmentation and semantic segmentation) together so the two tasks can help each other. While achieving competitive performance, global information and object boundary information are not encoded into embedding. To address the problem, we notice the approach Bonet [13]. Yang proposed a new end-to-end network framework Bonet to learn the coarse object bounding box information for point cloud instance segmentation. Object bounding box information is crucial for separating adjacent objects. Bonet directly regresses coarse bounding box vertexes and corresponding scores from global features.

As object boundary information is important to separate nearby objects, we combine two kinds approaches through proposing object bounding box-aware module. Our backbone network PointNet++ maintains an encoder-decoder architecture. After abstracting point features, semantic segmentation branch, instance segmentation branch and bounding box prediction branch compose our network. With our proposed OBAM module, bounding box information is encoded into our instance discriminative embedding. Our approach outperforms previous approaches. As object bounding box information is supervised, our network gains more information about the scene.

Extensive experiments on popular benchmarks S3DIS and ScanNetV2 are conducted to validate the effectiveness of our approach. To summarize, our main contributions are as follows:

1) We propose a novel framework which combines clustering-based approaches and proposal-based approaches. Our approach successfully encodes object bounding box information for point cloud instance segmentation.
2) We propose object bounding box-aware module (OBAM). The module successfully encodes object boundary information. Redundant object boundary information is removed through attention element-wise manipulation in OBAM.
3) Extensive experiments demonstrate the effectiveness of our network. With the proposed module, our network outperforms previous approaches.

2 Related Work

Instance segmentation on point cloud has attracted the attention of researchers in recent years. In this section, we briefly review previous approaches related to this field.

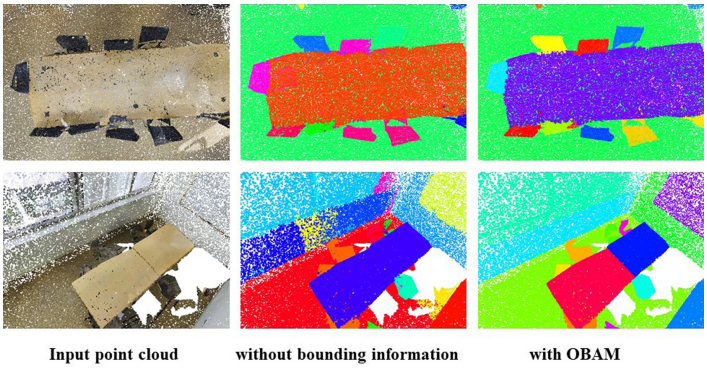

Fig. 1. Comparison of the instance segmentation results. Our proposed OBAM model successfully encodes object boundary information which is crucial to separate adjacent object instances.

2.1 Deep Learning Methods on Point Cloud

Deep learning methods on point cloud are mostly divided into multi-view-based methods, voxel-based methods and point-based methods (Fig. 1).

As 2D convolution neural networks have gained considerable success, multi-view-based methods projected 3D point clouds onto 2D images and process with 2D CNNs. MVCNN [9] recognizes 3D shapes from different views of the shapes. Through a view-pool-layer, information can be accumulated into a single, compact descriptor. However, such multi-view-based methods may lose geometric details.

Voxel-based methods voxelize point cloud into spatial grids and utilize standard 3D convolution neural network framework to extract point features. In order to improve the voxelization efficiency of point clouds, Riegler [8] proposes a novel representation which uses a set of unbalanced Octrees. Pooled features representation is stored in the leaf nodes of Octrees. Their methods enable the network to be deep and high resolution. While achieving promising results, lower running speed still effects because of spatial sparsity of point cloud. Graham [3] proposes submanifold sparse convolution network to address the problem. With a hash table storing point features, their networks avoid nonsense computation cost and memory occupation of vacent voxels. Although achieving leading performance, Voxel-based methods are still limited by heavy computation cost when processing large-scale point clouds.

Unlike voxel-based and multiview-based methods, point-based methods directly process point cloud. The pioneer work PointNet [5] learns per-point encoding with Multilayer Perceptron. PointNet++ [6] is proposed to hierarchically extract local point features and maintains an encoder-decoder architecture. Hu comes up with a novel framework called Randlanet [4] to address the problem of efficient semantic segmentation. Instead of complex point selection algorithms, Random point sampling is utilized for its remarkable memory and computation

efficiency. As random point sampling may discard key geometric details, They propose a novel local feature aggregation module to overcome the problem. In our work, we leverage PointNet++ as our backbone network to verify the validity of our approach.

2.2 Instance Segmentation on Point Cloud

Comparing with its counterpart on 2D images, the task of instance segmentation on point cloud is left far behind. Deep-learning approaches to the task can be divided into clustering-based approaches and proposal-based approaches.

SGPN [11] is the first work using deep learning technique in this field. With PointNet++ extracting global features and point features, the network learns feature space where points belonging to the same object have a close distance. They predict a similarity matrix yielding point-wise group proposals and a corresponding similarity map. They prune group proposals and generate point cloud groups through applying Non-Maximum Suppression. Due to the pair wise similarity matrix, the approach is heavily limited by computation and memory. In order to overcome the problem, clustering-based method ASIS [12] proposed by Wang removes the similarity matrix. Wang endorses that associative segmenting instances and semantics in point cloud are mutually beneficial to semantic segmentation task and instance segmentation task. Wang comes up with a method named mutual aid which enables the embedding of instance segmentation to benefit from point-level features of semantic segmentation. Semantic-aware embedding of instance segmentation achieves a huge breakthrough while it is unaware of the object bounding information. 3D Bonet [13] proposed by Yang directly predicts object bounding boxes. Better performance than ASIS is obtained through shared multi-layer perceptron without Non-Maximum Suppression algorithm. In our experiments, competitive performance are achieved through combining clustering-based methods and proposal-based methods.

3 Method

In this section, except semantic segmentation branch we mainly describe the other two branches (Bounding box branch and Instance segmentation branch). Details of our Object Bounding Box-aware module (OBAM) are presented below.

3.1 Network Framework

As shown in Fig. 2, our network is composed of a shared encoder and three parallel decoder branches. We apply PointNet++ as our backbone network to extract point features and global features. One of the branches handles semantic segmentation through decoding from point features. Another branch is to directly learn object bounding boxes from global features as 3D Bonet [13]. The other branch is to generate per-point embedding for instance segmentation. Backbone network encodes the input point cloud $P \in R^{N_p \times D}$ into point feature

$F_p \in R^{N_p \times D_f}$ matrix. Global point feature $F_g \in R^{D_f}$ is obtained by aggregation. N_p refers to the total number of input points. D denotes the dimension of input point cloud feature dimension and D_f is the point feature dimension. Subsequently, per-point semantic results are generated by semantic segmentation branch. Bounding box coordinates $B_c \in R^{N_b \times 6}$ and corresponding score B_s are obtained through bounding box branch. Two diagonal points coordinates refer to the rectangular bounding box. N_b is a predefined hyper-parameter denoting the number of object bounding boxes. The instance segmentation branch outputs per-point instance embedding $E_{ins} \in R^{N_p \times D_e}$. D_e is the embedding dimension. The embedding of points belonging to the same object should be close while the embedding of points belonging to different objects should be far away. Clustering algorithm mean-shift [7] is utilized to generate final group results during the inference.

To achieve the object bounding box-aware embedding, our proposed model OBAM is applied to encode the output of bounding box branch into instance segmentation branch. Besides, redundant object bounding box information is removed through an attention mechanism.

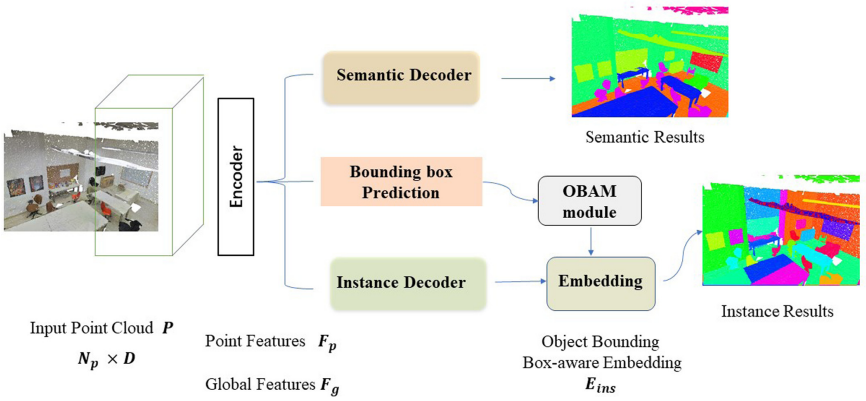

Fig. 2. The framework of our proposed method. Obviously, It is an encoder-decoder architecture. Point feature F_p and global feature F_g are obtained through a shared encoder. Three parallel decoders are applied. Semantic segmentation branch decodes from shared point features and classical cross entropy L_{sem} is used to supervise. Bounding box prediction branch predicts object bounding box and corresponding score. Output of bounding box prediction branch is integrated into instance branch through OBAM module. Final instance embedding are generated from instance segmentation branch.

3.2 Bounding Box Prediction Branch

We utilize bounding box prediction branch in 3D bonet as it is lightweight and effective. It takes global vector F_g as input. Bounding box $B_c \in R^{N_b \times 2 \times 3}$ and

its corresponding score $B_s \in R^{N_b}$ are generated by the branch. For simplicity, the rectangular bounding boxes are parameterized as follows:

$$b_c = (x_{min}, y_{min}, z_{min} x_{max}, y_{max}, z_{max}) \in B_s \tag{1}$$

The corresponding score b_{score} ranges from 0 to 1. As the number of object instance is variable, the bounding box prediction branch generates predefined number N_b of bounding boxes. We assume $N_b >= N_t$ where N_t refers to the number of ground truth object bounding boxes.

Although there is no fixed order for ground truth bounding boxes. We formulate it as an optimal assignment problem to learn one-to-one match between predicted bounding box and ground truth bounding box. Boolean matrix A denotes assignment where $A_{i,j} = 1$ refers to assign predicted box b_i to the ground truth box g_j. Cost matrix C is conducted where $C_{i,j}$ represents the cost between predicted bounding box b_i and ground truth bounding box g_j. The more similar the two boxes, the less the cost $C_{i,j}$. Optimal problem is solved through the existing Hungarian algorithm [14]. We formulate the problem as follows:

$$A = \arg\min_{A} \sum_{i=1}^{N_b} \sum_{j=1}^{N_t} A_{i,j} C_{i,j} \quad \text{subject to} \quad \sum_{i=1}^{N_b} A_{i,j} = 1 \ , \ \sum_{j=1}^{N_t} A_{i,j} \leq 1 \tag{2}$$

$$C_{i,j} = C_{i,j}^{Ecu} + C_{i,j}^{SIou} + C_{i.j}^{cro} \tag{3}$$

The association cost $C_{i,j}$ consists of three parts: Euclidean distance $C_{i,j}^{Ecu}$, soft intersection-over-union $C_{i,j}^{SIou}$ and point soft encoding cross-entropy $C_{i.j}^{cro}$ proposed in [13].

b_{score} lies in the range (0, 1) which indicates the validity of predicted bounding box. After bounding box assignment, N_t predicted bounding boxes of N_b are assigned to the ground truth. The scores b_{score}^t for the N_t ground truth bounding boxes are all 1 while the remaining $N_b - N_t$ scores are '0'. b_{score}^t refers to the scores for predicted bounding boxes which are assigned to the ground truth boxes while b_{score}^f refers to the antithesis. The loss function of bounding box prediction branch is defined as follows:

$$L_{asso} = \frac{1}{N_t} \left(\sum_{i=1}^{N_b} \sum_{j=1}^{N_t} A_{i,j} C_{i,j} \right) \tag{4}$$

$$L_{bscore} = -\frac{1}{N_b} \left(\sum_{1}^{N_t} \log b_{score}^t + \sum_{N_t+1}^{N_b} \log b_{score}^f \right) \tag{5}$$

$$L_{box} = L_{asso} + L_{bscore} \tag{6}$$

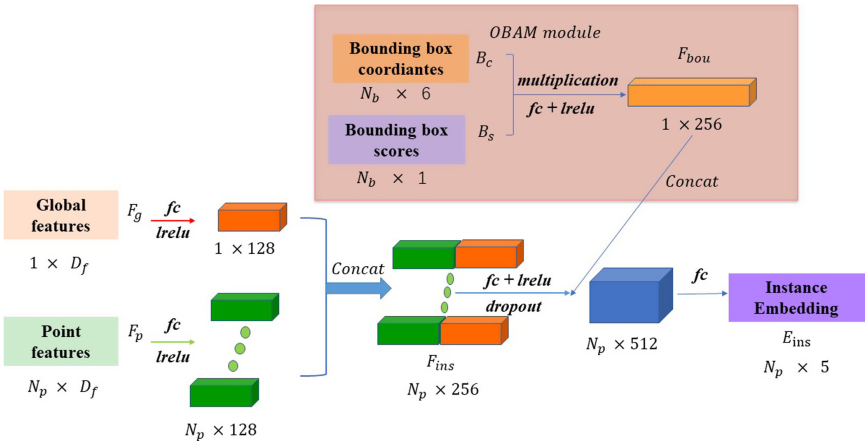

Fig. 3. The pipeline of our instance segmentation branch.

3.3 Instance Segmentation Branch

Pipeline. The instance segmentation branch fetches point feature matrix extracted by the backbone network and processes with the following predictions. Using efficient multilayer perceptrons (MLPs), the branch decodes the shared point feature matrix and global point feature is concatenated to the matrix. Applying Leaky Relu activation and Dropout technique, the intermediate feature matrix F_{ins} is obtained. Object boundary information obtained by the object bounding box branch is integrated and fused through OBAM module. The final embedding E_{ins} for instance segmentation can be represented as:

$$E_{ins} = \gamma_1 \left(F_{ins} \bigoplus F_{bou} \right) \tag{7}$$

Where $\gamma_1 : R^{D_m} \rightarrow \mathbb{R}^{D_e}$ (D_m refers to the intermediate point feature dimension) and $\bigoplus$ means concatenating the features. Mixing up object instance bounding information generates more informative instance embedding. The pipeline is illustrated in Fig. 3.

Object Bounding Box-Aware Module. Benefiting from the bounding box branch described above, object bounding information is integrated into our instance branch through our proposed Object Bounding Box-aware Module. It takes the outputs R_c, B_s of the bounding box branch as input. The higher validity of the predicted bounding box $b_c \in R_c$, the higher the corresponding score $b_{score} \in B_s$. Our module can be formulated as:

$$F_{bou} = \gamma_2 \left(\bar{R}_c \otimes B_s \right) \tag{8}$$

where $R_c \in R^{N_b \times 2 \times 3}$ is reshaped to $\bar{R}_c \in R^{N_b \times 6}$. $\otimes$ denotes element-wise multiplication, and γ_2 is a translation $R^{N_b \times 6} \rightarrow R^{N_b \times D_m}$ implemented by MLP. D_m refers to the dimension of the intermediate point feature F_{ins}.

As described above, B_s is supervised by L_{bscore}. Redundant predicted bounding boxes have scores close to '0'. Clearing up the effect of redundant predicted object bounding information, B_s is also supervised by L_{ins} (will be discussed in next section) through element-wise multiplication. The network selects helpful object bounding boxes information for instance embedding.

Loss Function. The informative embedding E_{ins} for instance segmentation is to learn a distance metric that can measure the probability of points belonging to the same object. Intra-instance embedding should be pulled toward the corresponding cluster center and different instance centers should be pushed far away from each other. The loss function can be formulated as:

$$L_{push} = \frac{1}{N_t\left(N_t-1\right)} \sum_{i=1}^{N_t} \sum_{j=1 j!=i}^{N_t} \left[2\sigma_d - \|\mu_i - \mu_j\|_1\right]_+^2 \tag{9}$$

$$L_{pull} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{1}{I_i} \sum_{j=1}^{I_i} \left[\|\mu_i - e_j\|_1 - \sigma_v\right]_+^2 \tag{10}$$

$$L_{ins} = L_{push} + L_{pull} \tag{11}$$

Where μ_i is the mean point embedding of instance i. I_i is point number of instance i and e_j refers to an embedding of a point belonging to instance i. σ_d and σ_v are loose margins. $\|x\|_1$ is defined as the l_1 distance. $[x]_+$ denotes $[x]_+ = max\left(0, x\right)$

During the training, L_{push} aims to make different instances repel each other and L_{pull} is designed to pull point embedding toward the mean embedding of instance. During the inference, we adopt existing clustering algorithm mean-shift on instance embedding to obtain instance labels. As our instance embedding is class-agnostic, the semantic label of the points having the same instance label is assigned as the final instance category.

To summarize, our network is end-to-end trainable and supervised by three branches losses. The loss weights are all equals to 1 in our experiment.

$$L = L_{ins} + L_{sem} + L_{box} \tag{12}$$

4 Experiments

In this section, we conduct quantitative and qualitative experiments to evaluate the effectiveness of our proposed approach. Ablation study and comparison with other approaches are reported below.

4.1 Experiment Settings

Dataset. We evaluate our approach on two public datasets: Stanford 3D Indoor Semantics Dataset (S3DIS) [1] and ScanNetV2 [2]. S3DIS consists of 3D scans in

6 large-scale indoor areas, covering total 272 rooms. S3DIS is a large-scale real indoor dataset containing more than 215 million points. Each point of S3DIS is associated with an instance label and a semantic label from 13 common semantic categories. Besides S3DIS, we further evaluate our approach on ScanNetV2. ScanNetV2 [2] contains about 1500 scans, divided into 1201, 300 and 100 scans, for training, validation and testing. We carry out our experiments on ScanNetV2 validation dataset.

Evaluation Metrics. We follow the 6-fold-cross-evaluation on S3DIS. Similar to ASIS [12] and Bonet [13], the performance on area 5 is also reported. Our instance segmentation performance is evaluated by four metrics: mean instance-wise coverage ($mCov$), mean weighted instance-wise coverage ($mWcov$), mean instance precision ($mPrec$), and mean recall ($mRec$). The experiments results are presented with IOU threshold of 0.5. For ScanNetV2, results on validation set are presented below.

Implement Details. For both S3DIS and ScanNetV2, each Scan contains a great deal of points, which makes it difficult to process all the points at one time. Each scene is split into 1 m $\times$ 1 m overlapped blocks. Each block contains 4096 points. Our experiment settings strictly follow Bonet [13], ASIS [12] and IAE [10]. N_b is set as 24. The margins σ_d, σ_v are set as $\sigma_d = 0.5$ and $\sigma_v = 1.5$. The embedding dimension D_e is 5. The learning rate is set to 0.01 (0.001 for S3DIS) and divided by 2 every 20 epochs. We train the network 50 epochs for PointNet++. We adopt Adam optimizer with its default hyper-parameters to optimize the network. At test time, mean-shift [7] clustering with bandwidth 0.6 is used for inference. We use Blockmerging algorithm [11] to merge object instances from different blocks.

Table 1. Instance segmentation results on ScanNetV2 dataset (validation set). We report the metric of mAP@0.25. Categories of Sink, Sofa, Table, Toilet, and Window are not presented in the table.

Method	mAP	bat	bed	she	cab	cha	cou	cur	des	doo	oth	pic	ref	shc
MaskRCNN [16]	26.1	33.3	0.2	0.0	5.3	0.2	0.2	2.1	0.0	4.5	2.4	23.8	6.5	0.0
SGPN [11]	35.1	20.8	39.0	16.9	6.5	27.5	2.9	6.9	0.0	8.7	4.3	1.4	2.7	0.0
ASIS [12]	47.4	57.3	52.1	1.4	18.5	46.1	19.2	20.3	13.3	13.8	18.8	6.6	17.6	33.1
Ours	51.2	64.7	61.3	0.3	23.1	69.7	13.6	16.9	15.4	14.7	24.0	11.5	18.3	60.7

4.2 Ablation Study

We firstly build a baseline without OBAM module. The baseline is made up of two decoder branches: the semantic segmentation branch and the instance segmentation branch. The baseline is supervised by cross-entropy loss L_{sem} for

Table 2. Instance segmentation results on the S3DIS. Experiment results on Area 5 and 6-fold are reported. **mCov**: average instance-wise coverage. **mWcov**: weighted average instance-wise coverage. **mPre**: mean precision. **mRec**: mean recall. Experiment performance is reported with IOU threshold of 0.5. For fair comparison, we carefully train the vanilla PointNet++ (without multi-scale grouping) as our backbone.

Method	Year	mCov	mWcov	mPre	mRec
Test on area 5					
SGPN [11]	2018	32.7	35.5	36.0	28.7
ASIS [12]	2019	44.6	47.8	55.3	42.4
3D-BoNet [13]	2019	–	–	57.5	40.2
JSNet [15]	2020	48.7	51.5	62.1	46.9
IAE [10]	2020	49.9	**53.2**	61.3	48.5
Ours	–	**50.3**	52.8	**65.3**	**49.2**
Test on 6-fold					
SGPN [11]	2018	36.0	28.7	31.2	38.2
MV-CRF [15]	2019	–	–	36.3	–
ASIS [12]	2019	51.2	55.1	63.6	47.5
3D-BoNet [13]	2019	–	–	65.5	47.6
PartNet[17]	2019	–	–	56.4	43.4
Ours	–	**54.7**	**57.1**	**68.4**	**52.9**

semantic task and discriminative loss L_{ins} for instance grouping. All ablation experiments we carry out are on the largest area 5 of S3DIS. The experiment results are shown in Table 3.

OBAM. We study the influence of our proposed OBAM and its components. Our proposed OBAM module with l_{score} can improve the results by 9.1 for mPre and 3.4 for mRec. It indicates that encoding boundary information indeed boosts instance segmentation performance by a large margin.

Manipulation. We find out that bounding box scores B_s supervised by l_{score} are whether close to 1 or close to 0. As our OBAM module is based on multiple layer perceptron, we design the attention element-wise manipulation to remove the redundant boundary information and it further benefits instance segmentation performance. Comparing with the pipeline without attention element-wise manipulation, the full pipeline of our method improves the result by 5 for both $mPre$ and $mRec$.

The Loss of Bounding Box Score. Presented in [13], bounding box scores B_s serve as a regularizer for bounding box prediction branch. After removing

Table 3. Ablation studies on the Area 5 of S3DIS. Both **mPre** and **mRec** metrics are reported. **OBAM**: using our proposed OBAM module. **Manipulation**: attention element-wise manipulation in Eq. (8). l_{score}: using l_{score} to supervise the bounding boxes prediction branch.

Method	**OBAM**	**Manipulation**	l_{score}	mPre	mRec
Baseline				51.2	40.7
Ours1	✓		✓	60.3	44.1
Ours2	✓	✓		60.5	46.1
Ours3	✓	✓	✓	**65.3**	**49.2**

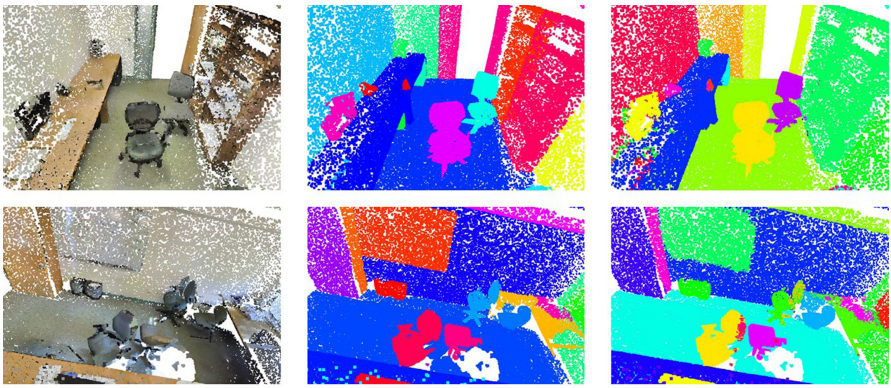

Fig. 4. Visualization of instance segmentation results on S3DIS. There are input point cloud, instance segmentation ground truth and our results from left to right. Through our proposed OBAM and discriminative embedding, our methods achieve sterling results of distinguishing adjacent objects.

bounding box score loss l_{score} supervision, bounding box scores B_s are only determined by attention element-wise manipulation. The instance segmentation performance drops significantly, primarily because of the difficulty to directly learn the score through attention mechanism.

4.3 Comparison with State-of-the-Art Approaches

In this section, Our comparison with other approaches is made on two popular benchmarks. Results on S3DIS and ScanNetV2 show the superiority of our approach.

Quantitative Results on S3DIS. Instance segmentation results testing on the area 5 of S3DIS and 6-fold validation are reported in Table 2. Our method

is compared with other state-of-the-art methods which are also based on PointNet++. Equipped with instance boundary information, our method achieve obvious improvement with metric mPre. Comparing with existing state-of-the-art methods, our method outperforms IAE [10] and JSNET [15], but not significantly. Both IAE and JSNET make a full use of point semantic information. IAE utilizes point semantic information and selects points from the instance to encode geometric information and instance context. JSNET jointly processes point cloud for Instance and Semantic Segmentation. Without leveraging semantic information, our approach achieves competitive performance. The effectiveness of our method and the importance of boundary information to instance segmentation are demonstrated. However, our approach is heavily affected by the bounding box prediction branch. We figure that more accurate bounding box prediction may boost the performance. Respectively, Fig. 4 shows our results of instance segmentation on the S3DIS dataset.

Quantitative Results on ScanNetV2. We conduct experiments on ScanNetV2 validation set and the performance are reported in Table 1. Comparing with the previous state-of-the-art approach ASIS [12], our method achieves a significant improvement of metric mAP@0.25, by 3.8 from 47.4 to 51.2. Our bounding box-aware embedding shows great superiority on some categories. The instance segmentation results on ScanNetV2 demonstrate the superiority of our method.

5 Conclusion

In this paper, We presented a novel framework combining clustering-based and proposal-based approaches. Our proposed module OBAM integrates bounding box information into instance segmentation branch. Through OBAM, redundant bounding information is removed. Extensive experiments indicate the effectiveness of our method. Our bounding box-aware embedding indeed boots the instance segmentation performance on S3DIS and ScanNetV2. Our method achieves state-of-the-art performance on S3DIS dataset.

However, our method is limited by bounding box prediction. The limitation that directly learning object boundary information may lead to the future work.

Acknowledgement. The research is funded by National Natural Science Foundation of China (No. 61972157), National Key Research and Development Program of China (No. 2019YFC1521104) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

1. Armeni, I., et al.: IEEE: 3D semantic parsing of large-scale indoor spaces. In: Computer Vision & Pattern Recognition (2016)

2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Niener, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes (2017)
3. Graham, B., Engelcke, M., Maaten, L.: 3D semantic segmentation with submanifold sparse convolutional networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
4. Hu, Q., Yang, B., Xie, L., Rosa, S., Markham, A.: RandLA-Net: efficient semantic segmentation of large-scale point clouds. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
5. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
6. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: deep hierarchical feature learning on point sets in a metric space. In: NIPS, pp. 5105–5114 (2017)
7. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. (2002)
8. Riegler, G., Ulusoy, A.O., Geiger, A.: OctNet: learning deep 3D representations at high resolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
9. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3D shape recognition. IEEE (2015)
10. He, T., Liu, Y., Shen, C., Wang, X., Sun, C.: Instance-aware embedding for point cloud instance segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 255–270. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_16
11. Wang, W., Yu, R., Huang, Q., Neumann, U.: SGPN: similarity group proposal network for 3D point cloud instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
12. Wang, X., Liu, S., Shen, X., Shen, C., Jia, J.: Associatively segmenting instances and semantics in point clouds. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
13. Yang, B., et al.: Learning object bounding boxes for 3D instance segmentation on point clouds. In: Advances in Neural Information Processing Systems, pp. 6737–6746 (2019)
14. Yaw, H.: The Hungarian method for the assignment problem. Naval Res. Logist. Q. **2**, 83–97(1955)
15. Zhao, L., Tao, W.: JSNet: joint instance and semantic segmentation of 3D point clouds. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 12951–12958 (2020)
16. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 2017
17. Mo, K., Zhu, S., Chang, A.X., Li, Y., Hao, S.: PartNet: a large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)