Semantic Consistency Networks for 3D Object Detection

Wenwen Wei, Ping Wei*, Nanning Zheng
Xi'an Jiaotong University, Xi'an, China
wwwei@stu.xjtu.edu.cn, pingwei@xjtu.edu.cn, nnzheng@mail.xjtu.edu.cn

Abstract

Detecting 3D objects from point clouds is a significant yet challenging issue in many applications. While most existing approaches seek to leverage geometric information of point clouds, few studies accommodate the inherent semantic characteristics of each point and the consistency between the geometric and semantic cues. In this work, we propose a novel semantic consistency network (SCNet) driven by a natural principle: the class of a predicted 3D bounding box should be consistent with the classes of all the points inside this box. Specifically, our SCNet consists of a feature extraction structure, a detection decision structure, and a semantic segmentation structure. In inference, the feature extraction and the detection decision structures are used to detect 3D objects. In training, the semantic segmentation structure is jointly trained with the other two structures to produce more robust and applicative model parameters. A novel semantic consistency loss is proposed to regulate the output 3D object boxes and the segmented points to boost the performance. Our model is evaluated on two challenging datasets and achieves comparable results to the state-of-the-art methods.

Introduction

Detecting 3D objects from point clouds is a significant issue in numerous applications, such as autonomous driving, domestic robots, and augmented reality. Compared with 2D images, 3D point clouds provide more reliable geometric information. However, point clouds are usually sparse, irregular, and lacking in texture information, which makes 3D object detection a challenging problem.

For the decent effectiveness and accessibility, neural network models have been widely used in 3D object detection, such as DSS (Song and Xiao 2016), 3DSIS (Hou, Dai, and Niessner 2019), VoxelNet (Zhou and Tuzel 2018), MV3D (Chen et al. 2017), BirdNet (Beltrán et al. 2018), PointNet (Qi et al. 2017a) , and VoteNet (Qi et al. 2019). While these studies have made remarkable progress, most of them mainly pay attention to the geometric properties (e.g. center, edge, and corner) of point clouds but neglect the semantic characteristics of each point and the inherent consistency between geometric and semantic cues. These factors limit the improvement of object detection performance.

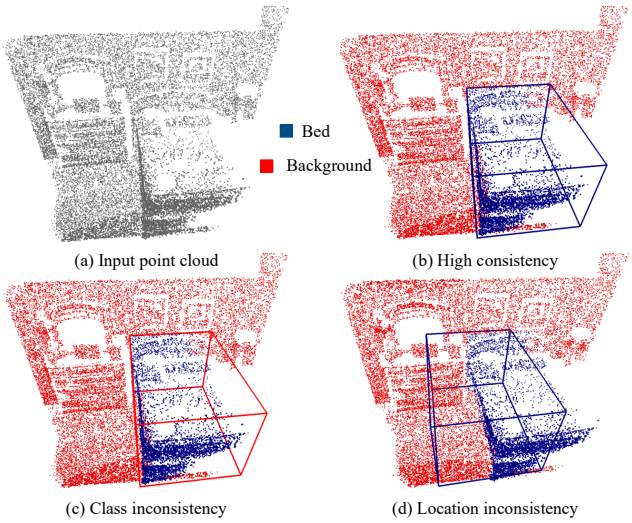

Figure 1: Illustration of the consistency between a 3D object box and the corresponding points inside. It is natural that the class of a correctly predicted bounding box should be consistent with the classes of all the points inside. (a) Input point clouds with an object *bed* to be detected. (b) High consistency between a bounding box and the points inside. (c) Class inconsistency. (d) Location inconsistency.

In fact, 3D object detection is closely related to semantic segmentation of point clouds since a 3D box's geometric attributes (size and location) and it's class are almost determined by the points inside the box. The more consistent the 3D box's semantic label and the points' labels inside the 3D box, the more accurate the 3D box is, as shown in Fig. 1 (b). On the other hand, the inconsistency, whether geometric or semantic, between the 3D box and the 3D points would mean incorrect 3D object detection, as shown in Fig. 1 (c) and (d). This phenomenon suggests that if a 3D object detection model was jointly trained with a semantic segmentation model and supervised by the consistency mechanism, the 3D object detection performance would be improved.

Inspired by this idea, in this paper, we propose a novel semantic consistency network (SCNet) for 3D object detection. This network contains three major parts: a feature ex-

*Ping Wei is the corresponding author.

traction structure, a detection decision structure, and a semantic segmentation structure. The feature extraction structure takes the point clouds as inputs and aims to extract deep features from the point clouds. The detection decision structure up-samples the deep features and combines the geometric and semantic cues to vote for the centers of 3D object boxes and output 3D proposals. The semantic segmentation structure is activated in training and jointly trained with the feature extraction and detection decision structures end-to-end. In this way, the learning of the network parameters is supervised not only by object detection but also by semantic segmentation, which makes the learned model more robust and applicative and therefore improves the performance.

We propose a novel semantic consistency loss which regulates the output 3D bounding boxes and the segmented semantic points. This loss formulates the semantic consistency mechanism that the class of a predicted 3D bounding box should be consistent with the classes of all the points inside this box. This inherent semantic consistency mechanism imposes effective constraints on both object classification and location, and therefore tends to improve the 3D object detection performance.

We test the proposed method on two challenging datasets: SUN RGB-D (Song, Lichtenberg, and Xiao 2015) dataset and ScanNetV2 (Dai et al. 2017) dataset. The experiments show that our method achieves comparable results to the state-of-the-art methods. Extensive ablation studies and analyses prove the effectiveness of each proposed module.

This paper makes four major contributions:

1. It proposes an architecture which incorporates semantic segmentation of point clouds to train the 3D object detection model and improves the model performance.
2. It proposes a semantic consistency mechanism and a corresponding loss function, which regulates the relations between the predicted 3D boxes and the inside points.
3. It proposes a novel voting module which combines geometric and semantic cues to make detection decision.
4. The model achieves comparable results to the state-of-the-art methods on two challenging datasets. Ablation studies prove the effectiveness of each proposed module.

Related Work

3D Object Detection

With the rapid progress in deep convolutional neural networks (Krizhevsky, Sutskever, and Hinton 2012), many state-of-the-art methods (Girshick 2015; Ren et al. 2015; He et al. 2017; Liu et al. 2016) achieve remarkable improvements for 2D object detection. However, 3D object detection task is more complicated due to the sparse, unevenly distributed, and irregular point clouds. From the perspective of how to organize the input point clouds, recent deep learning-based methods for 3D object detection can be divided into three categories: voxel-based methods, projection methods, and pointnet-based methods. Voxel-based methods (Song and Xiao 2016; Engelcke et al. 2017; Zhou and Tuzel 2018; Lang et al. 2019; Yan, Mao, and Li 2018; Chen et al. 2019; Song and Xiao 2014) convert point clouds into voxel grids and then apply 3D CNN directly. Projection based methods (Chen et al. 2017; Beltrán et al. 2018; Li et al. 2019; Liang et al. 2019; Xia et al. 2021) project point clouds into 2D format data (e.g. bird's eye view and front view) to reduce the computational complexity of the network. These two types of methods could not preserve all the information from raw point clouds because of the irreversible data conversion. Pointnet-based methods (Qi et al. 2017a,b, 2018, 2019; Yi et al. 2019; Shi, Wang, and Li 2019) extract features directly from raw point clouds to efficiently utilize the sparsity of 3D data and preserve the original geometry information. The PointNet++ model (Qi et al. 2017b) uses a hierarchical network to extract point cloud features for classification and segmentation.

The most recent work VoteNet (Qi et al. 2019) combines the traditional Hough voting (Hough 1959) with deep neural networks to vote for accurate object centers. It adopts the PointNet++ (Qi et al. 2017b) to extract base features and achieves quite good results on several 3D indoor datasets. However, the neglect of semantic characteristics of point clouds and inherent relations between geometry and semantic cues may limit the detection performance.

Semantic Segmentation

The development of semantic segmentation for point clouds is quite similar to 3D object detection. Voxel-based methods, projection methods, and pointnet-based methods are three mainstream types. SONet (Li, Chen, and Lee 2018) and Pointweb (Zhao et al. 2019) extract contextual features from point neighborhood for better representation of this local region. Some studies (Wang et al. 2019b; Shen et al. 2018; Wang, Samari, and Siddiqi 2018; Wang et al. 2019a) adopt graph convolutional network (GCNN) to learn points features jointly. ShapecontextNet (Xie et al. 2018) and PCAN (Zhang and Xiao 2019) aggregate local features of point clouds based on attention mechanism. SPLANet (Su et al. 2018) uses sparse bilateral convolutional layers as building blocks to maintain efficiency, and A-CNN (Komarichev, Zhong, and Hua 2019) designs a new convolution operator for better local geometry capture. These two are kernel-based convolution methods.

Considering that the semantic segmentation stream of our model is utilized as an auxiliary structure to assist the primary object detection task, we only adopt the plain PointNet++ (Qi et al. 2017b) as the backbone of the segmentation structure without using other complicated techniques.

Semantic Consistency Network

In this section, we will elaborate on our semantic consistency network model and its modules.

Overall Architecture

The input of the model is a 3D point cloud matrix of size $N \times 3$, where each row of the matrix is a 3D location of a point and N is the number of all points. Our semantic consistency network contains three major structures: feature extraction structure, detection decision structure, and semantic segmentation structure, as shown in Fig. 2. In inference, the

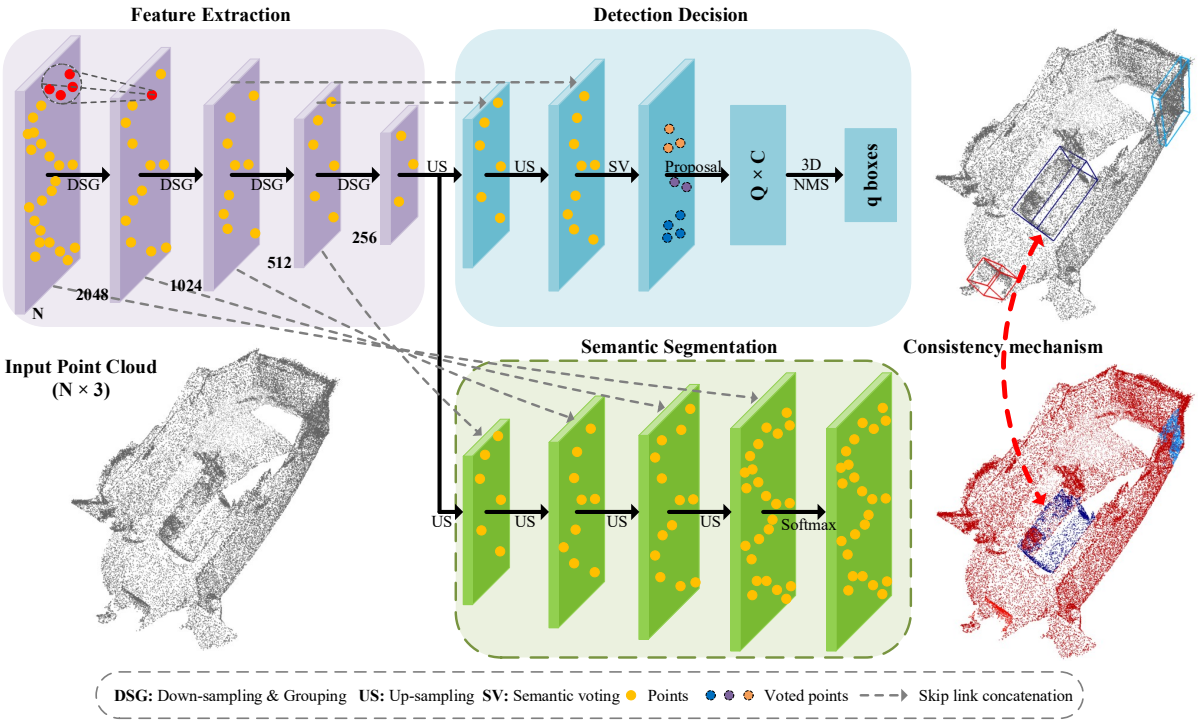

Figure 2: The architecture of our proposed SCNet. Given an input point cloud ($N \times 3$), the feature extraction structure down-samples and groups the points to learn deep representations. The detection decision structure utilizes the shared features to produce the final 3D bounding boxes. The semantic segmentation structure is activated only in training to output per-point semantic classes by processing these deep features through several up-sampling layers. In training, the semantic consistency between 3D bounding boxes and the corresponding semantic points is leveraged to guide the learning of object detection.

feature extraction structure and the detection decision structure are jointly used to detect 3D objects from point clouds; in training, all the three structures are jointly trained in an end-to-end way.

Feature Extraction Structure This structure takes the 3D point cloud matrix as inputs and aims to extract deep features of point clouds. The output features are fed to both the detection decision structure and the semantic segmentation structure. We use the base architecture of PointNet++ (Qi et al. 2017b) as this feature extraction structure. It is composed of five cascaded modules and each module contains three layers: down-sampling, grouping, and feature mapping (Qi et al. 2017b). The 3D points are down-sampled by the farthest point sampling method and all the points around each sampled point are grouped as a cluster, as shown in Fig. 2. In our model by down-sampling and grouping, the numbers of points in the five modules are N, 2048, 1024, 512, and 256, respectively. For each sampled point, a set of fully connected neural network layers are used to extract deep features, and the dimensions are 128, 256, 256, and 256 in the last four modules. Thus this feature extraction structure finally outputs a deep representation of $256 \times (256 + 3)$, where 3 indicates the dimension of the 3D location.

Detection Decision Structure This structure takes the learned base features from feature extraction structure as input and outputs predicted 3D bounding boxes. We adopt VoteNet (Qi et al. 2019) as the pipeline. It consists of five cascaded modules, and the first two modules up-sample the points to 512 and 1024 sequentially. Then these up-sampled points and their corresponding features are fed to the semantic voting module which outputs 1024 votes with geometric and semantic information. Feature dimensions of these three modules are 256, 256, and $256+K+1$, where K is the class number. The next proposal module groups these votes and produces Q initial proposals through a set of fully connected layers. We obtain the final q predicted bounding boxes after the 3D NMS module.

Semantic Segmentation Structure This structure is only used in model training to assist the object detection task by adjusting the model parameters. It receives the deep features from the feature extraction structure and outputs the semantic segmentation map of size $N \times (K + 1)$ representing the class scores of each point, where K is the class number. We adopt the base segmentation structure in PointNet++ (Qi et al. 2017b) as the pipeline. Four up-sampling layers are applied to recover the features to the original N points, and each up-sampling layer is linked to its corresponding down-sampling layer for the feature interpolation. We choose the inverse distance weighted average based on k Nearest Neighbors (kNN) as in PointNet++ (Qi et al. 2017b)

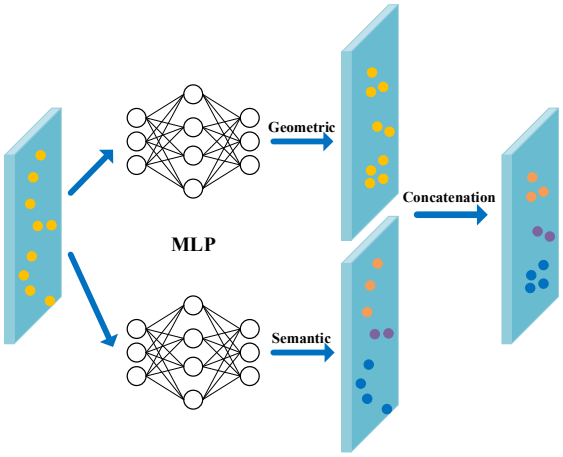

Figure 3: Structure of our proposed semantic voting module.

to interpolate the features. At the end of the semantic segmentation structure, a softmax layer is adopted to produce the class probabilities for each point.

The motivation of building this segmentation structure is that the relations between 3D object detection and semantic segmentation can be utilized to learn better base features from the point clouds. The produced high-level semantic features can be used in the consistency mechanism.

Semantic Voting

Traditional Hough voting (Hough 1959) is widely used for detecting simple patterns (e.g. lines). VoteNet (Qi et al. 2019) combines Hough voting with deep neural networks and utilizes geometric information (3D coordinates) of point clouds to vote for the object bounding box centers. However, the geometry-based voting method ignores the semantic information of each point which could be used to further improve the detection results. We propose a semantic voting method which combines geometric and semantic cues to make the decision. The proposed voting module is composed of the geometric voting sub-module and semantic voting sub-module, as shown in Fig. 3.

The input of the voting module is denoted as $\mathbf{f} = \{f_i \mid f_i = [x_i; a_i], i = 1, ..., M\}$, where M is the number of points. $x_i \in \mathbb{R}^3$ is the 3D coordinates of point f_i and $a_i \in \mathbb{R}^D$ is the corresponding deep features, where D is the dimension of the feature. The geometric voting sub-module predicts the coordinate offset Δx_i and the feature offset Δa_i for each f_i and finally outputs $\mathbf{g} = \{g_i \mid g_i = [x_i + \Delta x_i; a_i + \Delta a_i], i = 1, ..., M\}$ as in VoteNet (Qi et al. 2019). The semantic voting sub-module produces class scores $\mathbf{s} = \{s_i \mid s_i \in \mathbb{R}^{K+1}, i = 1, ..., M\}$ for each point. By concatenating these two sub-modules, the final voting result is formulated as $\mathbf{v} = \{v_i \mid v_i = [x_i + \Delta x_i; a_i + \Delta a_i; s_i], i = 1, ..., M\}$. We implement each sub-module using a three-layer MLP with ReLU activation function and batch normalization as in VoteNet (Qi et al. 2019).

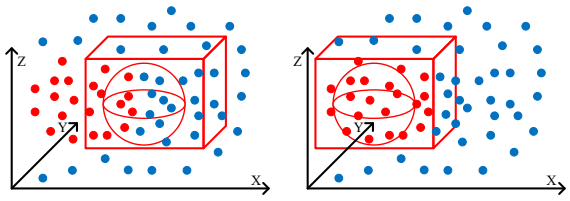

Figure 4: Illustration of the consistency computation. Red stands for object whereas blue denotes background. The 3D bounding box and the ball share the same center.

Loss Function

Our proposed SCNet is trained end-to-end with a multi-task loss function which includes three parts: semantic consistency loss, semantic segmentation loss, and detection loss. The loss function L is formulated as:

$$L = \mu_1 L_{\text{css}} + \mu_2 L_{\text{seg}} + \mu_3 L_{\text{det}}, \tag{1}$$

where L_{css}, L_{seg}, and L_{det} are the semantic consistency loss, semantic segmentation loss, and detection loss, respectively. μ_1, μ_2, and μ_3 are hyper-parameters weighting the different terms. In our experiment, $\mu_2 = 0.5$ and $\mu_3 = 1$. μ_1 increases linearly from 0 to 0.18 with the training epoch. The reason of designing μ_1 in this way is that both semantic segmentation and object detection produce poor outputs at the initial training stage.

Semantic Consistency Mechanism and Loss The motivation of the semantic consistency mechanism is that the inherent consistency between 3D bounding boxes and their corresponding semantic points could provide strong guidance for the learning of object classes and locations. As shown in Fig. 1, a well predicted 3D bounding box should be highly consistent with the predicted semantic points. Either class inconsistency or location inconsistency mean poor object detection results.

The semantic consistency loss is defined on top of the 3D box proposals and the corresponding semantic points. Given Q object box proposals from the detection decision structure, we filter out those proposals which are far away from all the ground truth centers (distance threshold is 0.3m). Then we map the remaining box proposals (Q_r) to the semantic points produced by the segmentation structure. As shown in Fig. 4, considering the numbers of points inside different boxes vary over a wide range, we use a ball with the same center (radius is 0.2m) to query the points that are close to the box center, which is more efficient than querying all points. The classes of the queried points and the bounding box are compared to compute the consistency loss L_{css}:

$$L_{\text{css}} = \frac{\sum_{i=1}^{Q_r} \sum_{j=1}^{n} \|p_i - s_j\|_1}{n \times Q_r}, \tag{2}$$

where Q_r is the number of remaining positive proposals and n is the number of the queried points. $\|\cdot\|_1$ denotes the L_1 norm. p_i and s_j are the probability vectors of bounding box and corresponding semantic point respectively.

By establishing the semantic consistency between 3D bounding boxes and corresponding semantic points, our model is able to properly accommodate the inherent relations between geometry and semantic cues, which is quite important for the 3D object detection task.

Detection Loss We follow the VoteNet (Qi et al. 2019) to define the detection loss as:

$$L_{\text{det}} = L_{\text{vot}} + \lambda_1 L_{\text{obj}} + \lambda_2 L_{\text{box}} + \lambda_3 L_{\text{cls}}, \tag{3}$$

where L_{vot}, L_{obj}, L_{box}, and L_{cls} are the semantic voting loss, objectness loss, box regression loss, and semantic classification loss, respectively. In our experiment $\lambda_1 = 0.5$, $\lambda_2 = 1$, and $\lambda_3 = 0.1$.

Semantic Segmentation Loss We use the weighted cross-entropy loss to define the semantic segmentation loss:

$$L_{\text{seg}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=0}^{K} 1\left[j = J_i\right] w_j \log p_i\left[j\right], \tag{4}$$

Where N is the total number of points, K is the number of categories, p_i and J_i denote the probability vector and semantic label of the i^{th} point respectively, and w_j is the weight of the j^{th} category. The weights are set to 0.2 for background points and 1 for other object points as default.

Experiments

Datasets

Our proposed SCNet is evaluated on two challenging datasets for 3D object detection task.

SUN RGB-D (Song, Lichtenberg, and Xiao 2015) is an indoor benchmark for 3D scene understanding. It consists of 10,335 RGB-D image pairs and is densely annotated with 64,595 amodal oriented 3D bounding boxes. Following the standard split, 5,285 RGB-D image pairs are used as the training samples and the rest for testing. We convert the original depth images to point clouds as the input of our model by using the provided camera parameters. We follow VoteNet (Qi et al. 2019) to evaluate our model on 10 classes with the standard evaluation protocol.

ScanNetV2 (Dai et al. 2017) is a richly annotated indoor dataset containing 2.5M views in 1,513 scenes. The training set contains 1,201 samples and the test set contains 312 samples. The input point clouds of our model are generated from the reconstructed meshes of ScanNetV2. Since the oriented bounding box annotations are not available, we perform the evaluation on 18 object categories with axis-aligned bounding boxes instead, as in (Qi et al. 2019; Hou, Dai, and Niessner 2019).

Implementation Details

The entire model is trained end-to-end for 180 epochs with a mini-batch size of 8. We adopt Adaptive Moment Estimation (Adam) (Kingma and Ba 2015) for optimization with the initial learning rate of 0.0015 for SUN RGB-D and 0.01 for ScanNetV2. The learning rate decay steps are set to $[100, 130, 160]$ and the corresponding decay rates are $[0.1, 0.1, 0.1]$.

We organize the input of our model by sub-sampling a certain number of points ($N = 40,000$) from original data. For data augmentation, the point clouds are randomly flipped in both horizontal directions, rotated by Uniform $[-5°, 5°]$, and scaled by Uniform $[0.9, 1.1]$ as in VoteNet (Qi et al. 2019). No RGB images or data are used in our method.

Comparison with State-of-the-art Methods

To prove the effectiveness of our model, we compare with different types of previous methods on both SUN RGB-D and ScanNet datasets. Deep sliding shapes (DSS) (Song and Xiao 2016) and 3D-SIS (Hou, Dai, and Niessner 2019) leverage the 3D CNN to extract features based on Faster R-CNN (Ren et al. 2015) pipeline. 2D-driven (Lahoud and Ghanem 2017) and F-PointNet (Qi et al. 2018) are cascaded detectors based on 2D detection. Cloud of gradients (COG) (Ren and Sudderth 2016) introduces an oriented gradient descriptor to link the 2D appearance and 3D pose. GSPN (Yi et al. 2019) is an instance segmentation based method. We also compare with several newest methods (Chen et al. 2020; Zhao, Chua, and Lee 2020; Najibi et al. 2020; Zhang et al. 2020).

The comparison results of the two datasets are shown in Table 1 and Table 2, respectively. All these listed methods adopt the same evaluation metrics. For SUN RGB-D dataset, our method achieves an mAP@0.25 of 59.9%, which is 2.2% higher than the VoteNet baseline. The heavy occlusion and points missing phenomenons of SUN RGB-D challenge the geometry-based methods (e.g. VoteNet) that ignore the semantic cues of point clouds. However, our proposed SCNet effectively utilizes the semantic information to guide the learning of detection task and thus improves the detection performance.

As for ScanNet dataset, our method achieves remarkable performance with an mAP@0.25 of 63.3% and an mAP@0.5 of 40.5%, which are 4.7% and 7% higher than the VoteNet baseline. It is observed that for some certain categories (e.g. bookshelf and shower curtain), the improvement is even much higher. Other similar objects (e.g. window, door, and picture) also benefit more from our model. We believe that these thin objects are under stronger supervision of the proposed consistency mechanism (a little translation along the width direction would cause huge inconsistency), which results in the performance boost.

Ablation Studies

Our model consists of three major modules for 3D object detection: semantic segmentation (SG), semantic voting (SV), and semantic consistency (SC). To prove the effectiveness of each module, we perform a set of ablation experiments with different combinations as shown in Table 3. These modules are added one by one to the baseline network (BaseNet). We adopt the VoteNet (Qi et al. 2019) as the BaseNet. Applying the semantic segmentation structure to the baseline alone contributes 1% and 2.1% increase on SUN RGB-D and ScanNet respectively, which indicates that joint training of the two tasks generates more robust and general base features. Adding the semantic voting module brings another

Methods	bath	bed	bkshf	chair	desk	drser	ntstd	sofa	table	toilet	mAP
DSS (Song and Xiao 2016)	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG (Ren and Sudderth 2016)	58.3	63.7	31.8	62.2	**45.2**	15.5	27.4	51.0	51.3	70.1	47.6
2D-driven (Lahoud and Ghanem 2017)	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
F-PointNet (Qi et al. 2018)	43.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	90.9	54.0
VoteNet (Qi et al. 2019)	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
H3DNet (Zhang et al. 2020)	73.8	85.6	31.0	**76.7**	29.6	33.4	65.5	66.5	50.8	88.2	60.1
SESS (Zhao, Chua, and Lee 2020)	76.9	84.8	35.4	75.8	29.3	31.3	**66.9**	66.4	**51.8**	**92.3**	61.1
HGNet (Chen et al. 2020)	78.0	84.5	**35.7**	75.2	34.3	**37.6**	61.7	65.7	51.6	91.1	**61.6**
SCNet (Ours)	**78.4**	**87.0**	33.5	74.8	25.9	29.1	64.5	**67.5**	50.1	88.5	59.9

Table 1: 3D object detection comparison (%) on SUN RGB-D V1 with 3D IoU threshold 0.25. The last five methods only use point cloud data for 3D object detection, without using RGB data.

Methods	mAP @0.25	mAP @0.5
DSS (Song and Xiao 2016)	15.2	6.8
MRCNN 2D-3D (He et al. 2017)	17.3	10.5
F-PointNet (Qi et al. 2018)	19.8	10.8
GSPN (Yi et al. 2019)	30.6	17.7
3D-SIS-5 (Hou, Dai, and Niessner 2019)	40.2	22.5
3D-SIS (Hou, Dai, and Niessner 2019)	25.4	14.6
VoteNet (Qi et al. 2019)	58.6	33.5
HGNet (Chen et al. 2020)	61.3	34.4
SESS (Zhao, Chua, and Lee 2020)	62.1	38.8
DOPS (Najibi et al. 2020)	63.7	38.2
H3DNet (Zhang et al. 2020)	**67.2**	**48.1**
SCNet (ours)	63.3	40.5

Table 2: 3D object detection performance (%) on ScanNetV2. The last seven methods only use point cloud data.

0.6% and 0.7% improvement, which demonstrates that combining geometry with semantic cues improves the original votes. Our model achieves the best performance with total 2.2% and 4.7% improvement by adding the last consistency module. The strong supervision provided by the consistency module effectively guides the learning of detection task and leads to better detection results.

The semantic consistency between 3D bounding boxes and corresponding points is an inherent property, which means the consistency loss will decrease automatically with the detection and segmentation results getting better, as the red line shown in Fig. 5. However, the loss can not converge to a low value without semantic consistency loss. Thus, we propose the semantic consistency mechanism to provide strong constraint, which leads to a lower converge value and a faster converge speed, as the blue line shown in Fig. 5.

Qualitative Results

Fig. 6 shows several representative detection results of VoteNet and our SCNet on ScanNet dataset. It is found that VoteNet misclassifies the bookshelf (the red arrow) and produces redundant or overlap boxes (bottom left) whereas our SCNet predicts more reasonable and accurate objects. We attribute this to the proposed consistency mechanism

Modules	mAP@0.25	
	SUN RGB-D	ScanNet
BaseNet	57.7	58.6
BaseNet + SG	58.7	60.7
BaseNet + SG + SV	59.3	61.4
BaseNet + SG + SV + SC (Our SCNet)	**59.9**	**63.3**

Table 3: Ablation studies on SUN RGB-D and ScanNet (%). SG: semantic segmentation; SV: semantic voting; SC: semantic consistency.

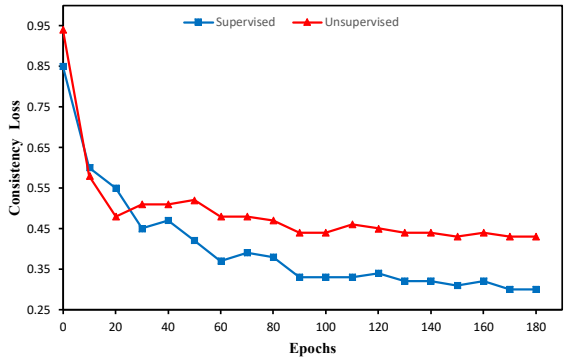

Figure 5: Consistency loss with (blue line) or without (red line) the semantic consistency regulation on ScanNet.

which effectively suppresses the production of these negative bounding boxes that lead to strong inconsistency. Furthermore, our SCNet could successfully detect some partially observed thin objects (e.g. door, the blue arrow).

Fig. 7 on SUN RGB-D also demonstrates the superiority of our method. Though some objects are partially observed (e.g. bed, the red arrow) or occluded (e.g. night stand, the blue arrow), our model could still provide amodal bounding boxes. Our method properly encodes the semantic information of point clouds to help predict such objects.

Conclusion

In this paper, we propose a semantic consistency network (SCNet) for 3D object detection based on an inherent se-

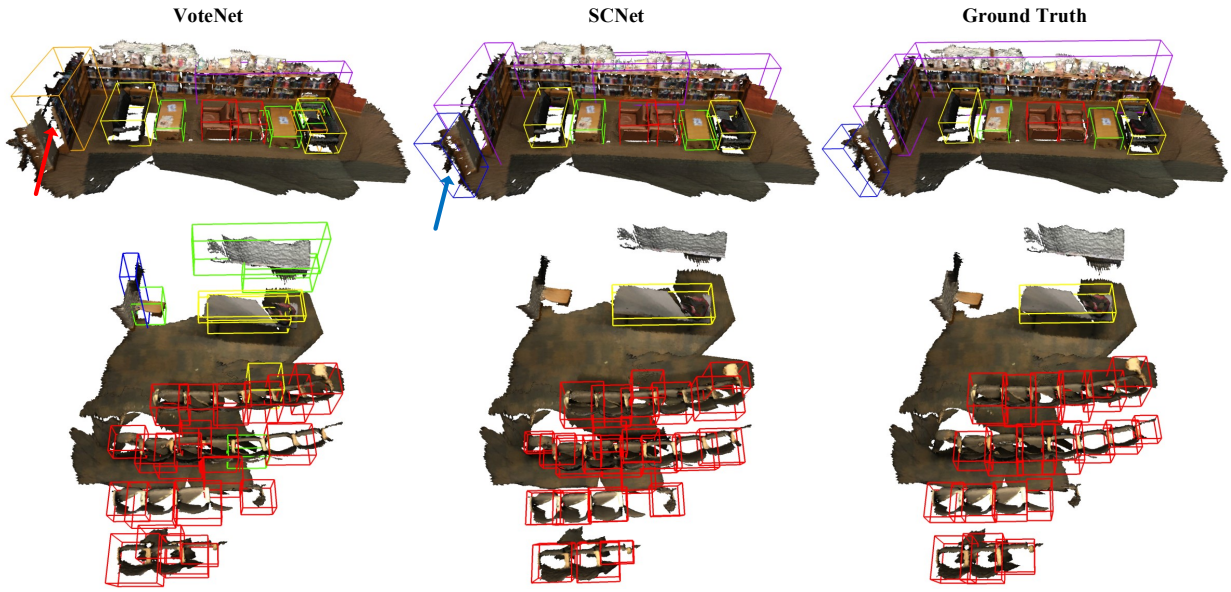

Figure 6: Qualitative example comparison between VoteNet and our proposed SCNet on ScanNetV2. Different colors stand for different classes. Top: library. Bottom: classroom.

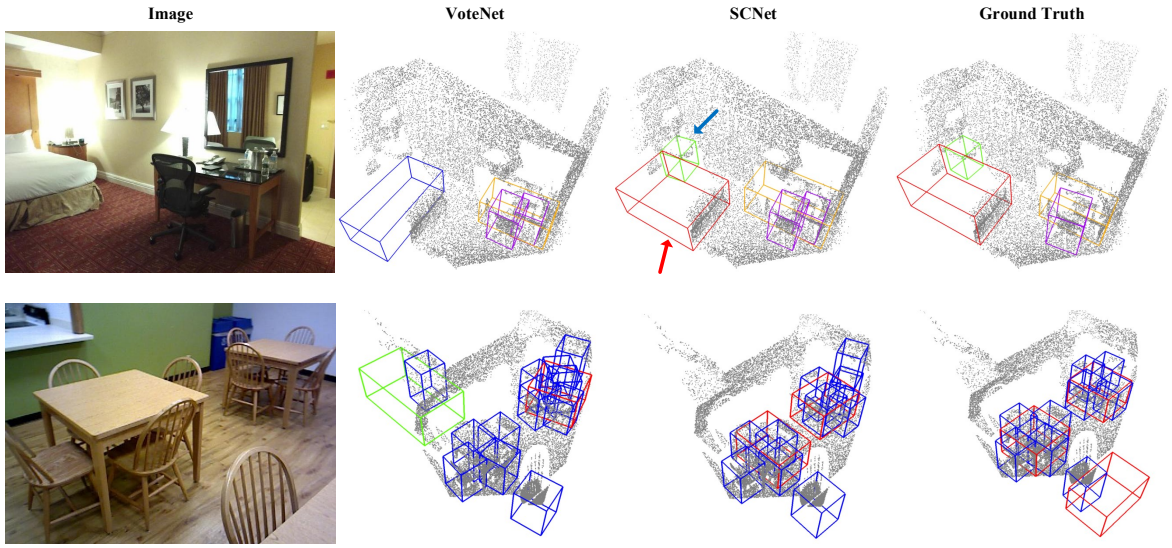

Figure 7: Qualitative example comparison between VoteNet and our proposed SCNet on SUN RGB-D V1. Top: bedroom. Bottom: restaurant.

mantic consistency between predicted 3D bounding boxes and corresponding semantic points. Specifically, we build an auxiliary semantic segmentation structure on top of the shared features to assist the base object detection task. We also propose a novel semantic voting module which utilizes the semantic characteristics of point clouds to improve the original geometry-based votes. We make full use of the inherent consistency between geometry and semantic cues to supervise the learning of our proposed model. SCNet achieves comparable performance to the state-of-the-art methods on two challenging 3D datasets. Extensive ablation experiments and qualitative analyses demonstrate the effectiveness of our model.

Acknowledgments

This research was supported by the National Key Research and Development Program of China (No. 2018AAA0102501) and National Natural Science Foundation of China (No. 61876149).

References

Beltrán, J.; Guindel, C.; Moreno, F. M.; Cruzado, D.; García, F.; and Escalera, A. D. L. 2018. BirdNet: A 3D Object Detection Framework from LiDAR Information. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*.

Chen, J.; Lei, B.; Song, Q.; Ying, H.; Chen, D. Z.; and Wu, J. 2020. A Hierarchical Graph Network for 3D Object Detection on Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-View 3D Object Detection Network for Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Y.; Liu, S.; Shen, X.; and Jia, J. 2019. Fast Point R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Niessner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Engelcke, M.; Rao, D.; Wang, D. Z.; Tong, C. H.; and Posner, I. 2017. Vote3Deep: Fast object detection in 3D point clouds using efficient convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*.

Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Hou, J.; Dai, A.; and Niessner, M. 2019. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hough, P. V. C. 1959. Machine Analysis of Bubble Chamber Pictures. *Conf. Proc. C* 590914: 554–558.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. *International Conference for Learning Representations (ICLR)* .

Komarichev, A.; Zhong, Z.; and Hua, J. 2019. A-CNN: Annularly Convolutional Neural Networks on Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Lahoud, J.; and Ghanem, B. 2017. 2D-Driven 3D Object Detection in RGB-D Images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. PointPillars: Fast Encoders for Object Detection From Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, B.; Ouyang, W.; Sheng, L.; Zeng, X.; and Wang, X. 2019. GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, J.; Chen, B. M.; and Lee, G. H. 2018. SO-Net: Self-Organizing Network for Point Cloud Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang, M.; Yang, B.; Chen, Y.; Hu, R.; and Urtasun, R. 2019. Multi-Task Multi-Sensor Fusion for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Najibi, M.; Lai, G.; Kundu, A.; Lu, Z.; Rathod, V.; Funkhouser, T.; Pantofaru, C.; Ross, D.; Davis, L. S.; and Fathi, A. 2020. DOPS: Learning to Detect 3D Objects and Predict Their 3D Shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep Hough Voting for 3D Object Detection in Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Qi, C. R.; Liu, W.; Wu, C.; Su, H.; and Guibas, L. J. 2018. Frustum PointNets for 3D Object Detection From RGB-D Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NIPS)*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Ren, Z.; and Sudderth, E. B. 2016. Three-Dimensional Object Detection and Layout Prediction Using Clouds of Oriented Gradients. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shen, Y.; Feng, C.; Yang, Y.; and Tian, D. 2018. Mining Point Cloud Local Structures by Kernel Correlation and

Graph Pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Song, S.; and Xiao, J. 2014. Sliding Shapes for 3D Object Detection in Depth Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Song, S.; and Xiao, J. 2016. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Su, H.; Jampani, V.; Sun, D.; Maji, S.; Kalogerakis, E.; Yang, M.-H.; and Kautz, J. 2018. SPLATNet: Sparse Lattice Networks for Point Cloud Processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, C.; Samari, B.; and Siddiqi, K. 2018. Local Spectral Graph Convolution for Point Set Feature Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; and Shan, J. 2019a. Graph Attention Convolution for Point Cloud Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019b. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph.* 38(5).

Xia, C.; Wei, P.; Wei, W.; and Zheng, N. 2021. A multilevel fusion network for 3D object detection. *Neurocomputing* 437: 107–117.

Xie, S.; Liu, S.; Chen, Z.; and Tu, Z. 2018. Attentional ShapeContextNet for Point Cloud Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors* 18(10): 3337.

Yi, L.; Zhao, W.; Wang, H.; Sung, M.; and Guibas, L. J. 2019. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, W.; and Xiao, C. 2019. PCAN: 3D Attention Map Learning Using Contextual Information for Point Cloud Based Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z.; Sun, B.; Yang, H.; and Huang, Q. 2020. H3DNet: 3D Object Detection Using Hybrid Geometric Primitives. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhao, H.; Jiang, L.; Fu, C.-W.; and Jia, J. 2019. PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, N.; Chua, T.-S.; and Lee, G. H. 2020. SESS: Self-Ensembling Semi-Supervised 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.