International Journal of Computer Vision (2021) 129:1857–1874
https://doi.org/10.1007/s11263-021-01456-w

ORIGINAL PAPER

Vote-Based 3D Object Detection with Context Modeling and SOB-3DNMS

Qian Xie[1] · Yu-Kun Lai[2] · Jing Wu[2] · Zhoutao Wang[1] · Yiming Zhang[1] · Kai Xu[3] · Jun Wang[1]

Received: 14 July 2020 / Accepted: 5 March 2021 / Published online: 1 April 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
Most existing 3D object detection methods recognize objects individually, without giving any consideration on contextual information between these objects. However, objects in indoor scenes are usually related to each other and the scene, forming the *contextual information*. Based on this observation, we propose a novel 3D object detection network, which is built on the state-of-the-art VoteNet but takes into consideration of the contextual information at multiple levels for detection and recognition of 3D objects. To encode relationships between elements at different levels, we introduce three contextual sub-modules, capturing contextual information at patch, object, and scene levels respectively, and build them into the voting and classification stages of VoteNet. In addition, at the post-processing stage, we also consider the spatial diversity of detected objects and propose an improved 3D NMS (non-maximum suppression) method, namely Survival-Of-the-Best 3DNMS (SOB-3DNMS), to reduce false detections. Experiments demonstrate that our method is an effective way to promote detection accuracy, and has achieved new state-of-the-art detection performance on challenging 3D object detection datasets, i.e., SUN RGBD and ScanNet, when only taking point cloud data as input.

Keywords Object detection · Point cloud processing · 3D deep learning

1 Introduction

3D object detection is becoming an active research topic in both computer vision and computer graphics. Compared to 2D object detection in RGB images, predicting 3D bounding boxes in real world environments captured by point clouds is more useful and essential for many tasks such as indoor robot navigation (McCormac et al. 2018), robot grasping (Wang et al. 2019), etc. However, unstructured point cloud data makes the detection more challenging than in regular 2D images. In particular, the popular convolutional neural networks (CNNs), which are highly successful in 2D object detection, are difficult to be applied to point clouds directly.

Growing interests have been attracted to tackle this challenge. With the emergence of deep 3D point processing networks, such as PointNet (Qi et al. 2017a) and PointNet++ (Qi et al. 2017b), several deep learning based 3D object detection works have been proposed recently to detect objects directly from 3D point clouds (Hou et al. 2019; Qi et al. 2019). The recent popular work VoteNet Qi et al. (2019) proposed an end-to-end 3D object detection network on the basis of Hough voting. VoteNet transfers the Hough voting procedure into a regression problem implemented by a deep network, and samples a number of seed points from the input point cloud to generate patches voting for potential object centers. The voted centers are then used to estimate the 3D bounding boxes. The voting strategy enables VoteNet to significantly reduce the search space and achieve the state-of-the-art results in several benchmark datasets. However, treating every point patch and object individually, VoteNet lacks the consideration of the relationships between different objects and between objects and the scene they belong to, which limits its detection accuracy.

An example can be seen in Fig. 1. Point clouds, captured by e.g. depth cameras, often contain noisy and missing data. This together with indoor occlusions makes it difficult

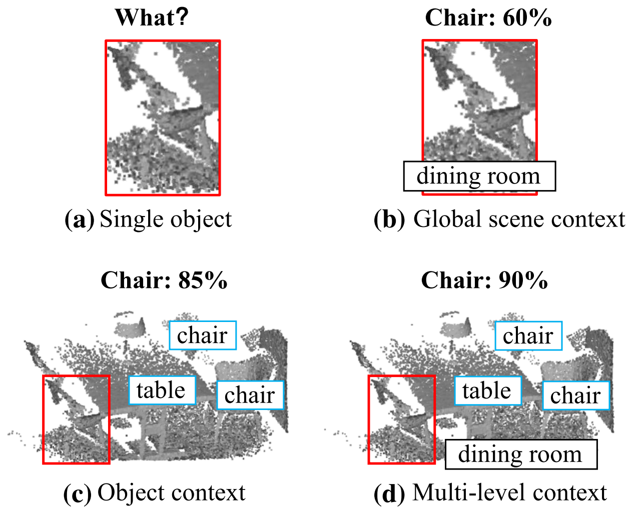
Communicated by Dima Damen.

✉ Jun Wang
wjun@nuaa.edu.cn

[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China

[2] Cardiff University, Cardiff, Walse, UK

[3] National University of Defense Technology, Changsha, China



(a) Single object **(b)** Global scene context

(c) Object context **(d)** Multi-level context

Fig. 1 Illustration of the importance of multi-level contextual information for 3D object detection from point cloud data. **a** It is hard to recognize the object when the point cloud is shown independently. **b–d** When the surrounding environment information is given, we can then recognize the chair easily. *In fact, unlike general object detection in open scenes, indoor scenes usually contain strong contextual constraints, which can be utilized in indoor scene understanding tasks such as 3D object detection*

even for humans to recognize what and where an object is in Fig. 1a. Nevertheless, considering the surrounding contextual information in Fig. 1b–d, it is much easier to recognize it is a chair given the surrounding chairs and the table in the dining room scene. Actually, the representation of a scanned point set could be ambiguous when it is presented individually, due to lack of color appearance and data missing problems. Therefore, we argue that indoor depth scans are often so occluded that contexts could even play a more important role in recognizing objects than the point data itself. This contextual information has been demonstrated to be helpful in a variety of computer vision tasks, including object detection (Hu et al. 2018a; Yu et al. 2016), image semantic segmentation (Zhang et al. 2019; Fu et al. 2019) and 3D scene understanding (Zhang et al. 2014, 2017). In this paper, we show how to leverage the contextual information in 3D scenes to boost the performance of 3D object detection from point clouds.

In our view, contextual information for 3D object detection consists of multiple levels. At the lowest is the patch level where the data missing problem is mitigated with a weighted sum over similar point patches to assist more accurate voting of object centers. At the object level, coexistence of objects provides strong hints on detection of certain objects. For example, as shown in Fig. 1d, the detected table can give a tendency for chairs to be detected at surrounding points. At the scene level, global scene clues can also help prevent the detection of inappropriate objects in a given scene. For example, we will not expect to detect a bed in a kitchen. The contexts at different levels complement each other and are utilized together to assist the correct inference of objects in noisy and cluttered environments.

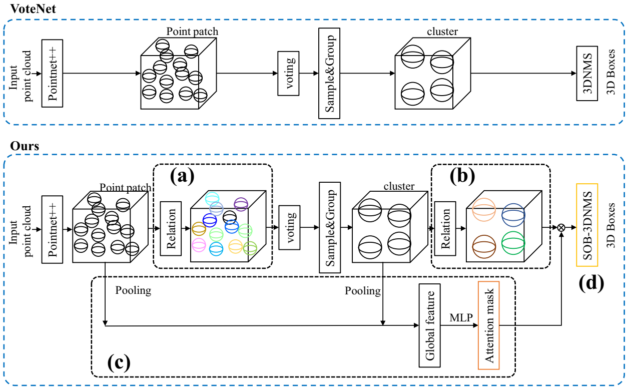


Fig. 2 Comparison of architectures between VoteNet (Qi et al. 2019) and the proposed network. Three sub-modules are integrated to capture the multi-level contextual information in point cloud data. **a** patch level context sub-module; **b** object level context sub-module; **c** global scene context sub-module; **d** SOB (Survival-Of-the-Best)-3DNMS (Non-Maximum Suppression) module

We thus propose a novel 3D object detection framework to incorporate into VoteNet multi-level contextual information for 3D object detection. Specifically, we propose a unified network to model the multi-level contexts, from local patches to global scenes. The difference between VoteNet and the proposed network is highlighted in Fig. 2. To model the contextual information, three sub-modules are proposed in the framework, i.e., patch-to-patch context (PPC) module, object-to-object context (OOC) module and the global scene context (GSC) module. In particular, similar to Zhang et al. (2019), we use the self-attention mechanism to model the contextual information in both PPC and OOC modules. According to relation networks (Hu et al. 2018a), the contextual information in object detection can be interpreted as the relations between objects. And there have been many works (Hu et al. 2018a; Yang et al. 2018; Zambaldi et al. 2018; Cao et al. 2019; Liu et al. 2019) using self-attention mechanism to encode relations between objects in scene understanding tasks. Specifically, we use the Compact Generalized Non-Local block (CGNL) proposed in Yue et al. (2018) as our self-attention operation, which is an extension work from non-local networks (Wang et al. 2018). CGNL uses Taylor expansion to optimize the original non-local module, and reduces the quadratic complexity to linear with respect to the number of channels. Thus, it requires light computation and little additional parameters, making it more practical. The above two sub-modules aim at adaptively encoding contextual information at the patch and object levels, respectively. For the scene-level, we design a new branch as shown in Fig. 2c to fuse multi-scale features to equip the network with the ability of learning global scene context.

In addition to capturing contextual information for better detection, we also improve the removal of overlapping detections by proposing an adaptive 3D NMS (Non Maximum Suppression) method which better considers the spatial relations between objects in 3D space. The traditional 3D NMS is inherited from 2D NMS. As 2D reveals the scene from a single view, overlapping objects are common. However, this is not the case for 3D, as objects in 3D space are naturally separated. Ideally, there should be no overlapping between the detected 3D bounding boxes. In implementation, we thus propose a more strict overlapping suppression strategy, namely Survival-Of-the-Best 3D NMS (SOB-3DNMS) which improves the traditional 3D NMS by adaptively adjusting the threshold to strictly suppress overlapping for confident detections.

This paper is an extended version of Xie et al. (2020), where a Multi-Level Context VoteNet (MLVCNet) is proposed. The major extensions in this journal paper include: (1) a novel Survival-Of-the-Best (SOB) 3D NMS is proposed to replace the traditional 3D NMS in post-processing, which is the first work on 3D NMS improvement for 3D object detection in indoor scenes, to the best of our knowledge; (2) an enhanced GSC sub-module compared to the original one in Xie et al. (2020) is proposed, to improve the global scene information integration; (3) the related work is extended to review literature on NMS in object detection; and (4) more experiments are carried out to holistically verify the effectiveness of the proposed components. For simplicity and clarity, we refer to our new, extended approach MLVCNet++.

In summary, the contributions of this paper include:

- We propose the first 3D object detection network that exploits *multi-level* contextual information at patch, object and global scene levels.
- We design three contextual sub-modules, including two self-attention modules and a multi-scale feature fusion module, to capture the contextual information at multiple levels in 3D object detection, and integrate the new modules into the state-of-the-art VoteNet framework.
- We design a novel SOB-3DNMS algorithm to eliminate redundant 3D bounding boxes, which is more suitable for 3D object detection in indoor scenes considering the object layout in 3D space.
- Extensive experiments demonstrate the benefits of using multi-level contextual information and the SOB-3DNMS. The proposed network outperforms state-of-the-art methods on both SUN RGB-D and ScanNetV2 datasets, when only taking point cloud data as input.

2 Related Work

2.1 3D Object Detection from Point Clouds

With the development of deep learning on 3D point clouds (Wang et al. 2017; Li et al. 2018; Atzmon et al. 2018), a large number of deep learning based 3D object detection methods from point cloud have emerged (Qi et al. 2018; Hou et al. 2019; Lang et al. 2019; Shi et al. 2019a; Chen et al. 2020; Qi et al. 2020; Shi and Rajkumar 2020; He et al. 2020; Najibi et al. 2020; Yang et al. 2020; Shi et al. 2020; Li et al. 2020b).

Among them, some are devoted to detecting objects in outdoor scenes. F-PointNet (Qi et al. 2018) is a milestone model which first generates 2D bounding boxes in images and then uses a frustum to locate the object in the point cloud. Dividing the point cloud into 3D voxels, VoxelNet (Zhou et al. 2018) introduces a voxel feature encoding (VFE) layer and stacks several VFE layers to learn complex features for each voxel. Instead of voxels, PointPillars (Lang et al. 2019) utilizes pillar shape to generate point-wise features. Recently, PointRCNN (Shi and Rajkumar 2020) introduces a two-stage 3D object detector. Their method first generates several 3D bounding box proposals, and then refines these proposals to obtain the final detection results.

More pertinent to our work are the works on 3D object detection in indoor scenes. Compared to outdoor, indoor scenes have more variety of objects and heavier occlusions, which make the detection more challenging. DSS (Deep Sliding Shapes) (Song and Xiao 2016) proposes the first 3D Region Proposal Network (RPN) which takes a 3D volumetric scene as input and outputs 3D object proposals. Similar to F-PointNet, PointFusion (Xu et al. 2018) also uses a 2D detector to detect 2D boxes in RGB images. However, this kind of methods heavily depends on the performance of 2D object detectors. Instead of treating 3D object proposal generation as a direct bounding box regression problem, Yi et al. (2019) proposed a novel 3D object proposal approach called GSPN (Generative Shape Proposal Network) which takes an analysis-by-synthesis strategy and reconstructs 3D shapes from point clouds. Recently, by virtual of PointNet/PointNet++, Qi et al. proposed end-to-end trainable 3D object detection networks (Qi et al. 2019, 2020) which handle 3D point clouds directly. They are inspired by the Hough voting strategy in 2D object detection and form the baseline of our work.

Although a lot of methods have been proposed recently, there is still large room for improvement especially for real-world challenging cases. Previous works largely ignored contextual information, i.e., relationships within and between objects and scenes. In this work, we show how to leverage the contextual information to improve the accuracy of 3D object detection.

2.2 Contextual Information

The work in Mottaghi et al. (2014) has demonstrated that contextual information has significant positive effect on 2D semantic segmentation and object detection. Since then, contextual information has been successfully employed to improve performance on many tasks such as 2D object detection (Yu et al. 2016; Hu et al. 2018a; Liu et al. 2018), 3D point matching (Deng et al. 2018), point cloud semantic segmentation (Engelmann et al. 2017; Ye et al. 2018), and 3D scene understanding (Zhang et al. 2014, 2017). The work in Hu et al. (2018c) achieves reasonable results on instance segmentation of 3D point clouds by analyzing point patch context. The work (Shi et al. 2019b) proposes a recursive auto-encoder based approach to detecting 3D objects via exploring hierarchical context priors in 3D object layout. Inspired by the self-attention idea in natural language processing (Vaswani et al. 2017), recent works connect the self-attention mechanism with contextual information mining to improve scene understanding tasks such as image recognition (Hu et al. 2018b), semantic segmentation (Fu et al. 2019) and point cloud recognition (Xie et al. 2018). As to 3D point data processing, the work in Zhang et al. (2019) proposes to utilize the attention network to capture the contextual information in 3D points. Specifically, it presents a point contextual attention network to encode local features into a global descriptor for point cloud based retrieval. In Paigwar et al. (2019), an attentional PointNet is proposed to search regions of interest instead of processing the whole input point cloud, when detecting 3D objects in large-scale point clouds. Different from previous works, we are interested in exploiting the combination of *multi-level* contextual information for 3D object detection from point clouds. In particular, we embed two self-attention modules and one multi-scale feature fusion module into a deep Hough voting network to learn multi-level contextual relationships between patches, objects and the global scene.

2.3 Non-maximum Suppression

In object detection algorithms, it is prone to generate redundant bounding boxes in the initial detection results to avoid missing true positives. Thus, non-maximum suppression (NMS) is usually used as a post-processing to remove these redundant detections. Dalal and Triggs (2005), for the first time, introduce the greedy NMS into human detection in 2D images, and achieve promising performance. Since then, NMS has become a basic component for most deep learning based 2D object detectors (Ren et al. 2015; Redmon et al. 2016; Liu et al. 2016). It is worth mentioning that several methods (Hu et al. 2018a; Engelmann et al. 2020; Carion et al. 2020) have been proposed in recent years as alternatives to avoid using NMS operations. For instance, relation networks (Hu et al. 2018a) propose to replace NMS by formulating duplicate removal as a binary classification problem. However, NMS is still the mainstream post-processor due to its simplicity. Considering different characteristics between 2D and 3D bounding boxes, there is still room for improvement of NMS in 3D cases, which previous methods rarely considered. Thus, we focus on the improvement of NMS operation in 3D object detection in this paper. The core idea of NMS is to keep good bounding boxes, such as those with high classification or confidence scores, while suppressing those which overlap too much with the good ones. However, traditional NMS has drawbacks in dealing with complicated cases, such as overlaps and occlusions, that are common in 2D object detection. Recently, several improved variants based on the traditional NMS have been proposed to tackle these issues, such as soft-NMS (Bodla et al. 2017), softer-NMS (He et al. 2018), Adaptive NMS (Liu et al. 2019a) and FeatureNMS (Salscheider 2020). Nevertheless, 3D NMS has not been formally studied before, which could be a new research direction to improve 3D object detection. In 3D object detection, most 3D detectors (Song and Xiao 2016; Qi et al. 2019) directly employ the 3D version of 2D NMS by simply replacing the 2D IOU (Intersection Over Union) calculation with 3D. However, this straightforward conversion is not suitable in 3D space where 3D bounding boxes overlap much less than 2D boxes in 2D space. Considering this spatial relationship between 3D objects, we propose a novel 3D NMS algorithm to remove redundant 3D bounding boxes as many as possible, while preserving the best ones.

3 Approach

As shown in Fig. 3, our network contains three main components: a fundamental 3D object detection framework based on VoteNet which follows the architecture in Qi et al. (2019), the multi-level context module and the SOB-3DNMS module. The multi-level context module consists of three context encoding sub-modules. The PPC (patch-patch context) sub-module combines the point groups to encode the patch correlation information, which helps to vote for more accurate object centers. The OOC (object-object context) sub-module is for capturing the contextual information between object candidates. This module helps to improve the results of 3D bounding box regression and classification. The GSC (global scene context) sub-module is to integrate the global scene contextual information. In brief, the proposed three sub-modules are designed to capture complementary contextual information in 3D object detection at multiple levels, with the aim to improve the detection performance in 3D point clouds. The subsequent SOB-3DNMS is further proposed to improve the typical 3DNMS on removing redundant detections during post-processing.

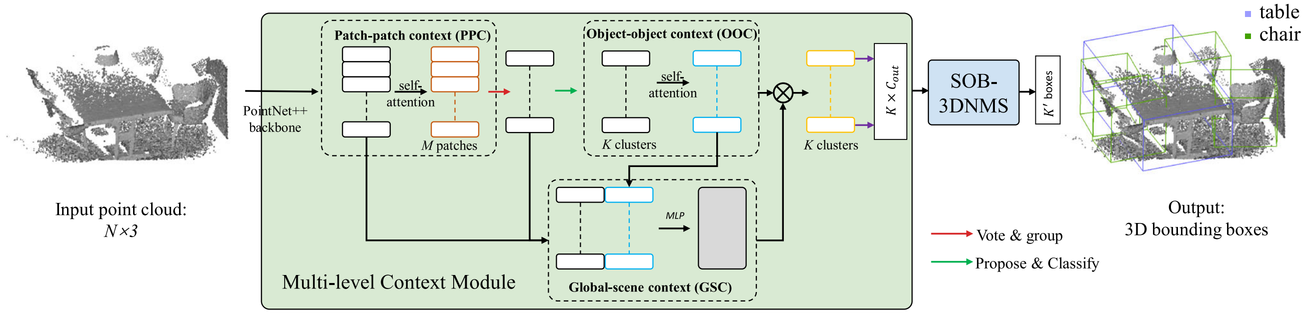


Fig. 3 Architecture of the proposed network MLCVNet++ for 3D object detection in point cloud data. Three new sub-modules are proposed to capture the multi-level contextual information in 3D indoor scene object detection. Please see Fig. 4 for the details of network

3.1 VoteNet

VoteNet (Qi et al. 2019) is the baseline of our work. As illustrated in Fig. 2, it is an end-to-end trainable 3D object detection network consisting of three main blocks: *point feature extraction*, *voting*, and *object proposal and classification*.

To extract point features, PointNet++ (Qi et al. 2017b) is used as the backbone network for seed sampling and extracting high dimensional features for the seed points from the raw input point cloud. The features of each seed point contain information from its surrounding points within a radius as illustrated in Fig. 4a. Analogous to regional patches in 2D, we thus call these seed points *point patches* in the remaining of this paper. The voting block takes the point patches with extracted features as input and regresses object centers. This center point prediction is performed by a multi-layer perceptron (MLP) which simulates the Hough voting procedure. Clusters are then generated by grouping the predicted centers, and form object candidates, from which the 3D bounding boxes are then proposed and classified through another MLP layer.

Note that in VoteNet, both the point patches and the object candidates are processed *independently*, ignoring the surrounding patches or objects. Thus, we introduce our MLCVNet++ network to encode context information in VoteNet with three context related sub-modules. Moreover, we also replace the typical 3DNMS with a novel SOB-3DNMS, considering the nature of non-overlapping spatial layout of objects in 3D space.

3.2 Context Module

(1) PPC sub-module We take relationships between point patches as the first level of context, i.e., patch-patch context (PPC), as shown in Fig. 4a. At this level, contextual information between point patches, on the one hand, helps relieve the data missing problem via gathering supplementary information from similar patches. On the other hand, it considers inter-relationships between patches for voting (Wang et al. 2013) by aggregating voting information from both the current point patch and all the other patches. We thus propose a sub-network, PPC module, to capture the relationships between point patches. The basic idea is, for each point patch, to employ a self-attention module to aggregate information from all the other patches before sending it to the voting stage.

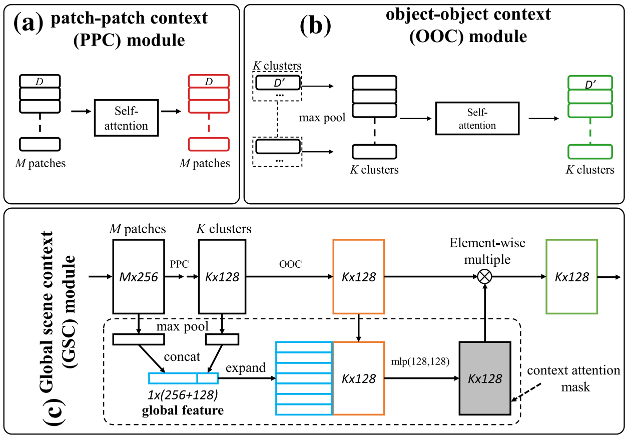


Fig. 4 Architecture details of the proposed three sub-modules. CGNL (Yue et al. 2018) is adopted as the self-attention module in our paper

As shown in Fig. 4a, after feature extraction using PointNet++, we get a feature map $\mathbf{A} \in \mathbb{R}^{1024\times D}$, where 1024 is the number of point patches sampled from the raw point cloud, and D is the dimension of the feature vector. We intend to generate a new feature map $\mathbf{A}'$ that encodes the correlation between any two point patches, and it can be formulated as a non-local operation:

$$\mathbf{A}' = f(\theta(\mathbf{A}), \phi(\mathbf{A}))g(\mathbf{A}) \tag{1}$$

where $\theta(\cdot)$, $\phi(\cdot)$, $g(\cdot)$ are three different transform functions, and $f(\cdot, \cdot)$ encodes the similarities between any two positions of the input feature. Moreover, as shown in Hu et al. (2018b), channel correlations in the feature map also contribute to the contextual information modeling in object detection tasks, we thus make use of the compact generalized non-local network (CGNL) (Yue et al. 2018) as the attention module to explicitly model rich correlations between any pair of point patches and of any channels in the feature space. CGNL requires light computation and little additional parameters, making it more practically applicable. After the attention module, each row in the new feature map still corresponds to a point patch, but contains not only its own local features, but also the information associated with all the other point patches.

(2) OOC sub-module Most object detection frameworks detect each object individually. Each cluster in VoteNet is independently fed into the MLP layer to regress its object class and bounding box. However, combining features from other objects gives more information on the object relationships, which has been demonstrated to be helpful in image object detection (Chen et al. 2018). Intuitively, objects will get weighted messages from those highly correlated objects. In such a way, the final predicted object result is not only determined by its own individual feature vector but also affected by object relationships. We thus regard the relationships between objects as the second level contextual information, i.e., object-object context (OOC).

We get a set of vote clusters $\mathbf{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ after grouping the voted centers. K is the number of clusters in this work. A cluster $\mathcal{C} = \{v_1, v_2, \ldots, v_n\}$, where v_i represents the ith vote in $\mathcal{C}$, and n is the number of votes in $\mathcal{C}$. Each cluster is fed into an MLP followed by a max pooling to form a single vector representing the cluster. Then comes the difference from VoteNet. Instead of processing each cluster vector independently to generate a proposal and classification, we consider the relationships between objects. Specifically, we introduce a self-attention module before the proposal and classification step. Figure 4b shows the details inside the OOC module. Specifically, after max pooling, the cluster vectors $\mathbf{C} \in \mathbb{R}^{K \times D'}$ are fed into the CGNL attention module to generate a new feature map to record the affinity between all clusters. The encoding of object relationships can be summarized as:

$$\mathcal{C}_{OOC} = Attention(\max_{i=1,\ldots,n} \{MLP\ (v_i)\}) \tag{2}$$

where $\mathcal{C}_{OOC}$ is the enhanced feature vector in the new feature map $\mathbf{C}_{OOC} \in \mathbb{R}^{K \times D'}$, and $Attention(\cdot)$ is the CGNL attention mapping. By doing so, the contextual relationships between these clusters (objects) are encoded into the new feature map.

(3) GSC sub-module The whole point cloud usually contains rich scene contextual information which can help enhance the object detection accuracy. For example, it would be highly possible that a chair rather than a toilet is identified when the whole scene is a dining room rather than a bathroom. Therefore, we regard the information about the whole scene as the third level context, i.e., global scene context (GSC). Inspired by the idea of scene context extraction in Liu et al. (2018), we propose the GSC module (the green module in Fig. 3) to leverage the global scene context information to improve feature representation for 3D bounding box proposal and object classification, without explicit supervision of scenes. Note that we improve the global scene information integration procedure via replacing the simple addition operation in Xie et al. (2020) with the attention mask weighting, which is proven to be a more efficient way to capture the global scene contextual information through experiments.

The GSC module is designed to capture the global scene contextual information by introducing a global scene feature extraction branch. Specifically, we create a new branch with the input from the patch and object levels, concatenating the features at layers before applying self attention in PPC and OOC. As shown in Fig. 4c, at the two layers each row represents a point patch $\mathcal{P} \in \mathbf{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_M\}$ or an object candidate $\mathcal{C} \in \mathbf{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$, where M and K are the numbers of the sampled point patches and clusters, respectively. Max-pooling is first applied to get two vectors (i.e., the patch vector and the cluster vector), combining information from all the point patches and object candidates. Following the idea of multi-scale feature fusion in the contextual modeling strategy of 2D detectors, these two vectors are then concatenated to form a global feature vector F_g, which is formulated as:

$$F_g = [\max(\mathbf{P}); \max(\mathbf{C})] \tag{3}$$

The global feature vector is expanded to the same size as $\mathbf{C}$, and then concatenated with $\mathbf{C}$. An MLP layer is applied to further aggregate global information by generating a context attention mask A_g, which is formulated as:

$$A_g = MLP([expand(F_g); \mathbf{C}]) \tag{4}$$

We then apply this weight mask A_g on the cluster feature map $\mathbf{C}$ to embed the global contextual information. The integration procedure can be summarized as:

$$\mathbf{C}_{new} = \mathbf{C}_{OOC} \otimes (1 + A_g) \tag{5}$$

where $\otimes$ is element-wise multiplication. To retain the original proposal information, we adopt the residual connection strategy by adding the refined features to the original feature map, hence 1 is added to A_g.

In the original version, the integration is done by simply adding a global scene-level feature to the object-level features. In this version, we revisited the integration strategy, and took the more direct weight multiplication strategy instead of feature addition. It is based on the motivation that we should give low weights to incompatible detections and high weights to compatible detections. The integration of global scene context is designed to improve the detection results by suppressing false detection (objects not compatible with the scene type) and enhancing correct detection (objects compatible with the scene type), accordingly, the weight multiplication step in our paper is to assign different weights to the detections according to the global scene information, which is a more direct way to embed the global scene context. Moreover, as the weight measures the compatability of the object and the scene, it should be co-determined by both the global scene-level information and the object-level information. Thus, we concatenate the output of OOC (i.e., object-level information) to the global scene features to generate the weights. Our experiments also demonstrate that the feature fusion and the weight multiplication help achieve better detection results compared with the original version.

3.3 SOB-3DNMS

Most of the existing 3D object detection methods post process the box proposals by using a typical 3DNMS which is a direct 3D version of a simple 2DNMS. Specifically, the 3D box $\mathcal{M}$ with the highest classification score is first selected, and then the IOU values between $\mathcal{M}$ and the remaining 3D boxes are computed. Then, for a specific box b_i, it is retained only when the IOU value, which measures the overlapping between $\mathcal{M}$ and b_i, is less than a threshold, i.e.,

$$s_i = \begin{cases} s_i, & \text{iou}\,(\mathcal{M}, b_i) < N_t \\ 0, & \text{iou}\,(\mathcal{M}, b_i) \geq N_t \end{cases} \tag{6}$$

where N_t is the pre-defined IOU threshold, and s_i is the classification score of box b_i. Setting s_i to 0 means b_i will be deleted.

However, as shown in Fig. 5, many false positives still exist after applying the typical 3DNMS. The reason lies in the difference between 2D and 3D layouts of objects. Overlaps are unavoidable (and plausible) for 2D boxes due to the projection onto a single view. However, in 3D space, large overlaps between 3D objects of the same class are rare, which means the IOU threshold should be set to a much lower value to suppress false positives.

How about setting a lower fixed NMS threshold? It seems that setting a lower threshold may be a solution to remove the redundant boxes in 3D. However, there are two main problems. First, it is hard to determine a proper fixed threshold for all the objects. Second, a too low threshold, due to the imperfect box proposal, is prone to remove true positives as well.

In all, the typical 3DNMS, as the direct conversion of 2DNMS, is inappropriate for 3D object detection in indoor scenes. Thus, we propose to use an adaptive IOU threshold which depends on the confidence of the boxes. That is, for the current box $\mathcal{M}$, the IOU threshold is adaptively determined by the confidence score of $\mathcal{M}$, i.e.,

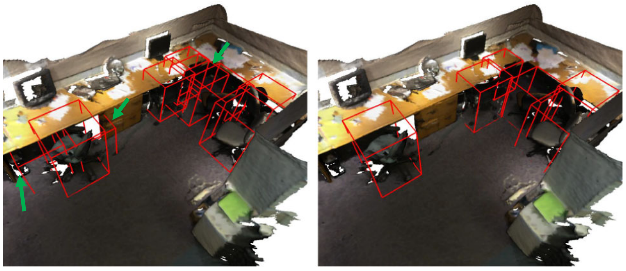

(a) Typical 3DNMS **(b)** Groundtruth

Fig. 5 Typical 3DNMS results with VoteNet (Qi et al. 2019) on ScanNetV2 dataset. As can be seen, there are still boxes overlapping with each other after typical 3DNMS. Visually, there appear to be large overlaps between these boxes; however, the computed IOUs between these boxes are below the given NMS threshold. That is, these remaining boxes still satisfy the filter criteria of typical 3DNMS and thus they are retained, *which is exactly the unreasonable point of typical 3D NMS and cannot be resolved by changing the threshold*

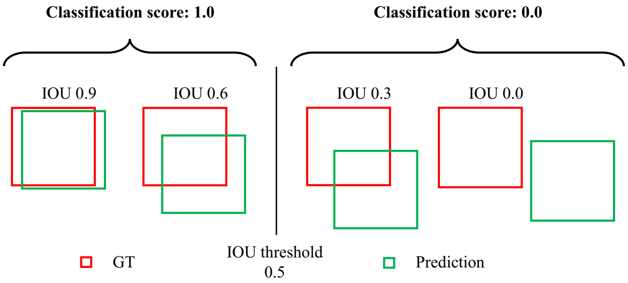


Fig. 6 Illustration of classification score in box quality measurement. As can be seen, the 0/1 classification score representation is so coarse that it ignores so much localization information

$$\begin{aligned} N_{\mathcal{M}} &:= \min\,(N_t, 1 - s_{\mathcal{M}}) \\ s_i &= \begin{cases} s_i, & \text{iou}\,(\mathcal{M}, b_i) < N_{\mathcal{M}} \\ 0, & \text{iou}\,(\mathcal{M}, b_i) \geq N_{\mathcal{M}} \end{cases} \end{aligned} \tag{7}$$

where $N_{\mathcal{M}}$ is the adaptive IOU threshold for the current box $\mathcal{M}$. From the definition, our IOU threshold $N_{\mathcal{M}}$ incorporates both the fixed threshold N_t and the confidence score $s_{\mathcal{M}}$ of the current box. The larger the confidence score of the box, the lower the IOU threshold. And lower IOU thresholds in NMS mean stricter criteria with surrounding overlapping boxes, i.e., more boxes will be suppressed. Moreover, instead of directly taking the classification score as the confidence score $s_{\mathcal{M}}$, we use a more reasonable scoring scheme to measure the quality of boxes, as explained below.

As is known, the box ranking is also an important factor for the performance of NMS. We argue that the classification score is not sufficient to measure the quality of 3D boxes. The classification score measures the likelihood of a box being an

```
Input :$\mathcal{B} = \{b_1, .., b_N\}, \mathcal{S} = \{s_1, .., s_N\}, N_t$
        $\mathcal{B}$ is the list of initial detection boxes
        $\mathcal{S}$ contains corresponding confidence scores
        $N_t$ is the NMS threshold
begin
  $\mathcal{D} \leftarrow \{\}$
  while $\mathcal{B} \neq empty$ do
    $m \leftarrow \text{argmax}\, \mathcal{S}$
    $\mathcal{M} \leftarrow b_m$
    $N_{\mathcal{M}} \leftarrow min(N_t, 1 - s_{\mathcal{M}})$
    $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{M}; \mathcal{B} \leftarrow \mathcal{B} - \mathcal{M}$
    for $b_i$ in $\mathcal{B}$ do
      if $iou(\mathcal{M}, b_i) \geq N_t$ then
        $\mathcal{B} \leftarrow \mathcal{B} - b_i; \mathcal{S} \leftarrow \mathcal{S} - s_i$
      end                Typical 3DNMS
      if $iou(\mathcal{M}, b_i) \geq N_{\mathcal{M}}$ then
        $\mathcal{B} \leftarrow \mathcal{B} - b_i; \mathcal{S} \leftarrow \mathcal{S} - s_i$
      end                SOB-3DNMS
    end
  end
  return $\mathcal{D}, \mathcal{S}$
end
```

Fig. 7 The pseudocode in red is replaced by that in green in SOB-3DNMS, which adaptively suppresses the detections by scaling their NMS threshold according to their box confidences. Note that NMS is only performed within bounding boxes of the same class

object, but neglects the location of the object, and therefore cannot predict the accuracy of the box well. As illustrated in Fig. 6, taking the two boxes on the left for example, they both have a classification score of 1.0 (showing that they are likely to contain an object), but the first box is of much higher quality than the second, as it well overlaps with the ground truth box. This similarly applies to other boxes. So in addition to the classification score, the location of an object is a more important factor determining the quality of the proposed box. Thus, inspired by Jiang et al. (2020), we add one more element to the final output vector for each proposal for predicting a single value to evaluate the box, which is supervised by the IOU between the predicted box and the ground truth during training, and therefore predicts the IOU value $s_{\mathcal{M}}$ when it is trained. Specifically, instead of directly using the classification score as the ranking criterion in NMS, our SOB-3DNMS adopts the newly predicted IOU scores to rank 3D boxes, considering that the new scores can reflect the quality of the boxes more accurately.

A formal algorithm description of the SOB-3DNMS is shown in Fig. 7 with highlighted difference from the typical 3DNMS.

4 Results and Discussions

4.1 Datasets

We evaluate our approach on SUN RGB-D (Song et al. 2015) and ScanNet (Dai et al. 2017) datasets. SUN RGB-D is a well-known public RGB-D image dataset of indoor scenes, consisting of 10,335 frames with 3D object bounding box annotations. Over 64,000 3D bounding boxes are given in the entire dataset. As described in Zhang et al. (2017), these scenes were mostly taken from household environments with strong context. The occlusion problem is quite severe in SUN RGB-D dataset. Sometimes, it is even difficult for humans to recognize the objects in the scene when merely a 3D point cloud is given without any color information. Thus, it is a challenging dataset for 3D object detection.

ScanNet dataset contains 1513 scanned 3D indoor scenes with densely annotated meshes. The ground-truth 3D bounding boxes of objects are also provided. The completeness of scenes in ScanNet makes it an ideal dataset for training our network to learn the contextual information at multiple levels.

4.2 Training Details

Our network is trained end-to-end using an Adam optimizer and batch size 8. The base learning rate is set to 0.01 for ScanNet dataset and 0.001 for SUN RGB-D dataset. The network is trained for 220 epochs on both datasets. The steps that the learning rate decays are set to be {120, 160, 200} for ScanNet, {100, 140, 180} for SUN RGB-D, and the decay rates are $\{0.1, 0.1, 0.1\}$. Training the model until convergence on one RTX 2080 ti GPU takes around 4 h on ScanNetV2 and 11 h on SUN RGB-D. During training we found the mAP result fluctuates within a small range on different runs. To accommodate the difference, the mAP results reported in the paper are the mean results over three runs.

For parameter size, the stored PyTorch model size of our network is 13.9 MB, compared reasonably with the 11.2 MB of VoteNet. For training time, VoteNet takes around 40 s for 1 epoch with batch size of 8, while ours is around 42 s. For inference time, measuring for one batch, VoteNet takes around 0.13 s, while ours takes 0.14 s. The time reported here are all tested on ScanNet dataset. These show that our method only slightly increases the complexity of VoteNet.

4.3 Qualitative Results

Figure 8 shows the predicted bounding boxes using our method and VoteNet on the validation set of ScanNetV2. It is observed that the proposed network detects more reasonable objects (red arrows), and predicts more precise boxes (blue arrows). The pale blue box detected by VoteNet is clas-

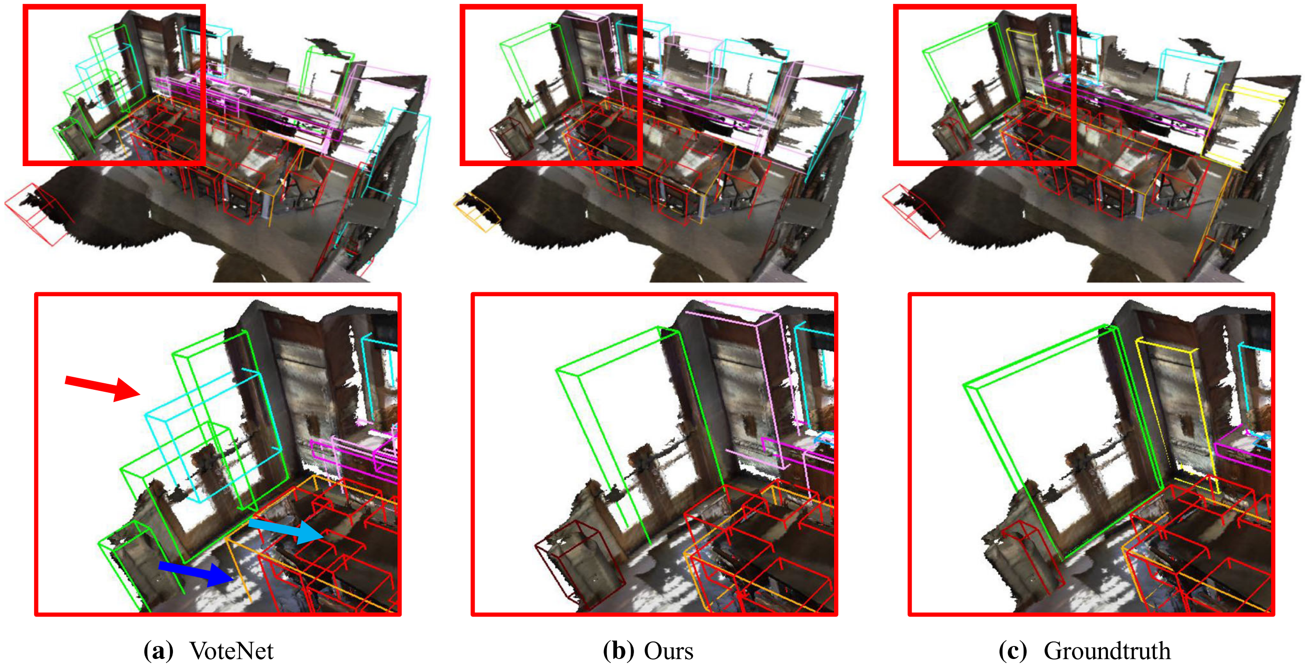


Fig. 8 Qualitative comparison results of 3D object detection in ScanNetV2. Our multi-level contextual information analysis strategy enables more reasonable and accurate detection. *Color is for depiction only, and not used for detection*

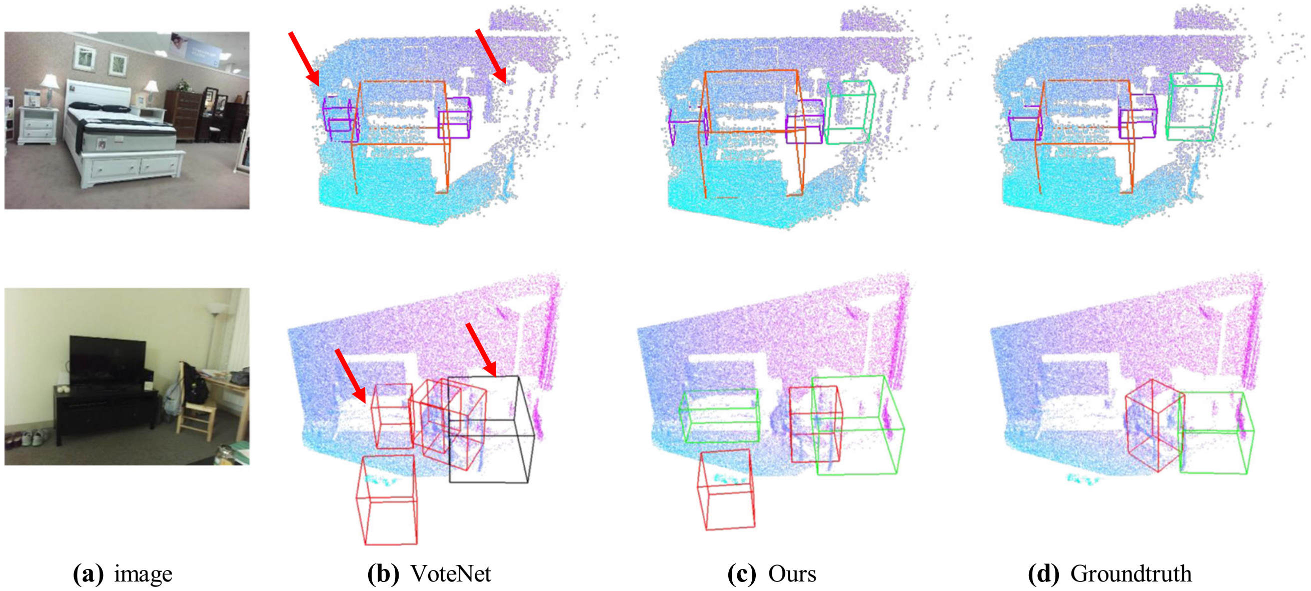


Fig. 9 Qualitative results of 3D object detection on SUN RGB-D

sified as a window, but improperly overlaps with a detected door (green box). The boxes detected by our method are with less overlaps and more accurate localization. The qualitative results on SUN RGB-D are shown in Fig. 9. As shown, our model is able to produce high-quality boxes even though the scenes are much occluded and less informative. As shown in the bedroom example in Fig. 9, there are overlaps and missing detections (red arrows) using VoteNet, while our model successfully detects all the objects with good precision compared to the ground-truth. For the second scene in Fig. 9, VoteNet misclassifies the table, produces overlaps, and predicts inaccurate boxes (red arrows), while our model produces much cleaner and more accurate results. However, it is worth noting that our method may still fail in accurate localization of some predictions when most of the data is missing, such as the door (green) in the red square in Fig. 8b.

Table 1 Performance comparison on ScanNetV2 validation set

	Input	mAP@0.25	mAP@0.5
DSS Song and Xiao (2016)	Geo+RGB	15.2	6.8
MRCNN 2D-3D He et al. (2017)	Geo+RGB	17.3	10.5
F-PointNet Qi et al. (2018)	Geo+RGB	19.8	10.8
GSPN Yi et al. (2019)	Geo+RGB	30.6	17.7
3D-MPA Engelmann et al. (2020)	Geo+RGB	64.2	**49.2**
3D-SIS Hou et al. (2019)	Geo+5views	40.2	22.5
3D-SIS Hou et al. (2019)	Geo only	25.4	14.6
VoteNet Qi et al. (2019)	Geo only	58.6	33.5
HGNet Chen et al. (2020)	Geo only	61.3	34.4
DOPS Najibi et al. (2020)	Geo only	63.7	38.2
MLCVNet Xie et al. (2020)	Geo only	64.5	41.4
MLCVNet++ (Ours)	Geo only	**66.2**	45.3

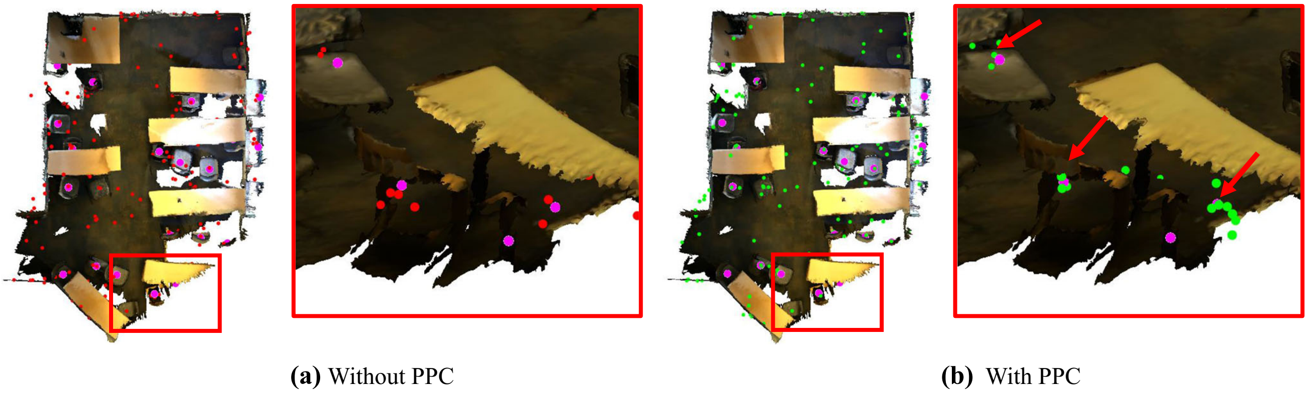
There is still room for improvements on 3D bounding box prediction, especially when dealing with complicated scenes.

4.4 Comparison with the State-of-the-Art

We first compare the detection results on ScanNet using our method and the state-of-the-art methods on this benchmark, including MRCNN 2D-3D (He et al. 2017), GSPN (Yi et al. 2019), 3D-SIS (Hou et al. 2019), 3D-MPA (Engelmann et al. 2020), HGNet (Chen et al. 2020) and DOPS (Najibi et al. 2020). The results are shown in Table 1. Both mAP@0.25 and mAP@0.5 are used for evaluation. As seen, our method achieves the best performance on mAP@0.25 among all the compared methods. Specifically, the proposed network reaches 66.2% making 7.5 absolute points improvement over the baseline VoteNet, and 1.7 points improvement over the original MLCVNet which ranked the second best. On mAP@0.50, even higher improvements over VoteNet and MLCVNet are observed. The improvements confirm the effectiveness of the proposed method. Note that 3D-MPA has better performance than ours on mAP@0.50. However, 3D-MPA utilizes the point-wise segmentation label to supervise their network. Moreover, 3D-MPA makes use of RGB and normal information in addition to geometry. Our method outperforms all the methods with geometry only input. Table 2 shows the detailed results at mAP@0.25 for each object category in ScanNetV2 dataset. As can be seen, for some specific categories, such as shower curtain and sink, the improvements exceed 10 points. It is found that plane-like objects,

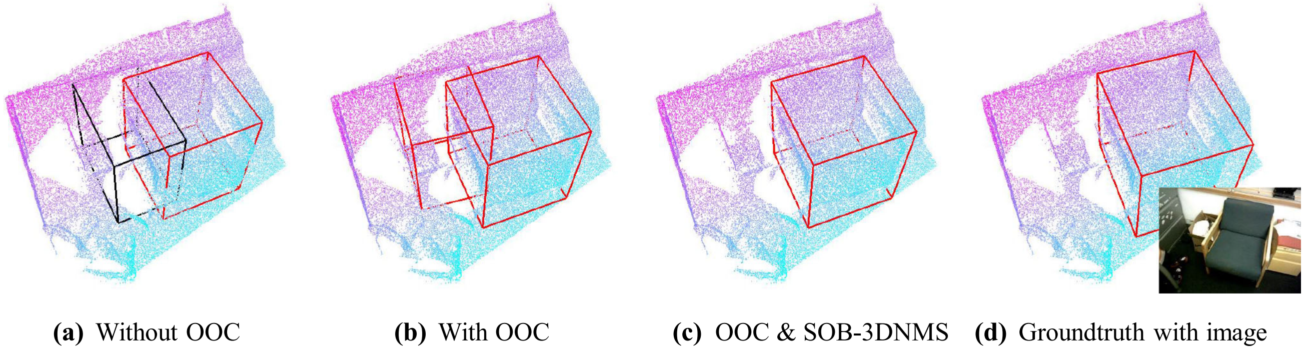
Table 2 Per-category evaluation on ScanNetV2, evaluated with mAP@0.25 IoU

	wind	bed	cntr	Sofa	tabl	showr	ofurn	sink	pic	chair	desk	curt	fridge	door	toil	bkshf	bath	cab	mAP
3DSIS5views	10.88	69.71	10.00	71.81	36.06	35.96	16.2	42.98	0.00	66.15	46.93	14.06	53.76	30.64	87.6	27.34	84.3	19.76	40.23
3DSISGeo	2.79	63.14	6.92	46.33	26.91	12.17	7.05	22.87	0.00	65.98	33.34	2.47	10.42	7.95	74.51	2.3	58.66	12.75	25.36
VoteNet	38.1	87.92	56.13	89.62	58.77	57.13	37.2	54.7	7.83	88.71	71.69	47.23	45.37	47.32	94.94	44.62	92.11	36.27	58.65
MLCVNet	**46.98**	**88.48**	**63.94**	87.4	63.50	65.91	47.89	59.18	11.94	89.98	**76.05**	**56.72**	**60.86**	56.93	**98.33**	56.94	87.22	42.45	64.48
MLCVNet++ (Ours)	44.34	88.12	62.97	**90.88**	**66.38**	**71.44**	**53.15**	**66.63**	**13.60**	**92.45**	72.84	55.70	56.76	**57.68**	95.33	**62.70**	**92.33**	**47.62**	**66.16**



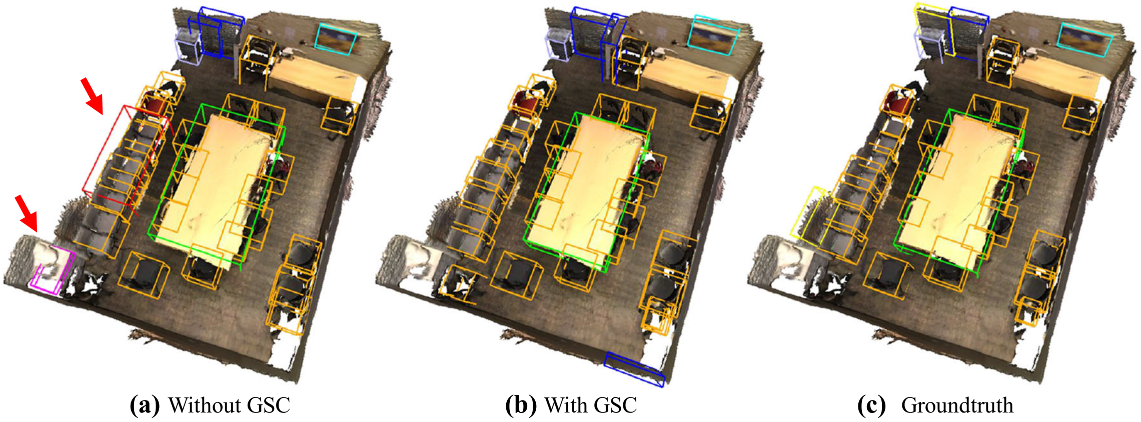
(a) Without PPC **(b)** With PPC

Fig. 10 A voting example for our method with or without the PPC sub-module. Compared to our network without PPC, the whole model generates more accurate voting centers. Pink dots are object center ground truth



(a) Without OOC **(b)** With OOC **(c)** OOC & SOB-3DNMS **(d)** Groundtruth with image

Fig. 11 A detection example using our method with or without the OOC sub-module. Compared to without OOC, the complete model generates more desirable result. Black: toilet; Red: chair



(a) Without GSC **(b)** With GSC **(c)** Groundtruth

Fig. 12 A detection example for our method with or without the GSC sub-module. Compared to our network without GSC, the whole model generates less unreasonable boxes which are inconsistent with the global scene type. Red: sofa; Pink: counter. Scene type: conference room

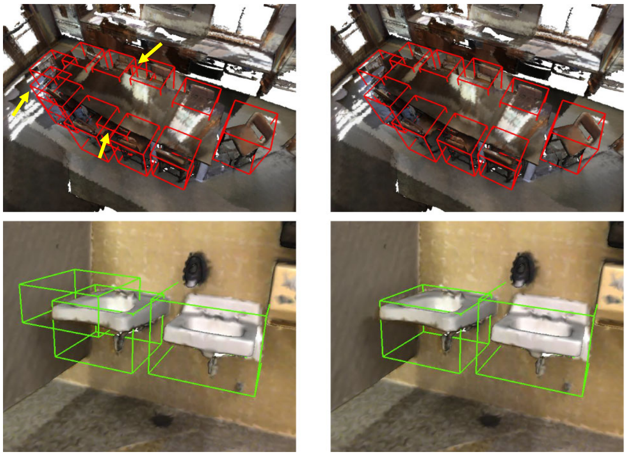
such as doors, windows, pictures and shower curtains, usually get higher improvements. A possible reason is that these objects contain more similar point patches, which, via the attention module, complement each other to a great extent.

We also evaluate our network on SUN RGB-D dataset using the same 10 most common object categories as in Qi et al. (2019). Table 3 gives a quantitative comparison of our method with DSS (Song and Xiao 2016), cloud of gradients (COG) (Ren and Sudderth 2016), 2D-driven (Lahoud and Ghanem 2017), F-PointNet (Qi et al. 2018) and VoteNet (Qi et al. 2019).

Remarkably, our method achieves better overall performance than all the other methods on SUN RGB-D dataset. The overall mAP (mean average precision) of the proposed

Table 3 Performance comparison with state-of-the-art 3D object detection networks on SUN RGB-D V1 validation set

	Input	Table	Sofa	Booksh	Chair	Desk	Dresser	Nightst	Bed	Bathtub	Toilet	mAP@0.25
DSS Song and Xiao (2016)	Geo+RGB	50.3	53.5	11.9	61.2	20.5	6.4	15.4	78.8	44.2	78.9	42.1
2D-driven Lahoud and Ghanem (2017)	Geo+RGB	37.0	50.4	31.4	48.3	27.9	25.9	41.9	64.5	43.5	80.4	45.1
COG Ren and Sudderth (2016)	Geo+RGB	**51.3**	51.0	31.8	62.2	**45.2**	15.5	27.4	63.7	58.3	70.1	47.6
F-PointNet Qi et al. (2018)	Geo+RGB	51.1	61.1	33.3	64.2	24.7	**32.0**	58.1	81.1	43.3	**90.9**	54.0
VoteNet Qi et al. (2019)	Geo-only	47.3	64.0	28.8	75.3	22.0	29.8	**62.2**	83.0	74.4	90.1	57.7
MLCVNet Xie et al. (2020)	Geo only	50.4	66.3	31.9	75.8	26.5	31.3	61.5	**85.8**	79.2	89.1	59.8
MLCVNet++ (Ours)	Geo only	50.7	**68.3**	**36.5**	**77.1**	28.7	31.6	61.4	85.3	**79.3**	90.0	**60.9**



(a) Typical 3DNMS **(b)** SOB-3DNMS

Fig. 13 Two examples on comparison of typical 3DNMS and the proposed SOB-3DNMS. As seen, the false positive boxes retained by typical 3DNMS are successfully eliminated by SOB-3DNMS

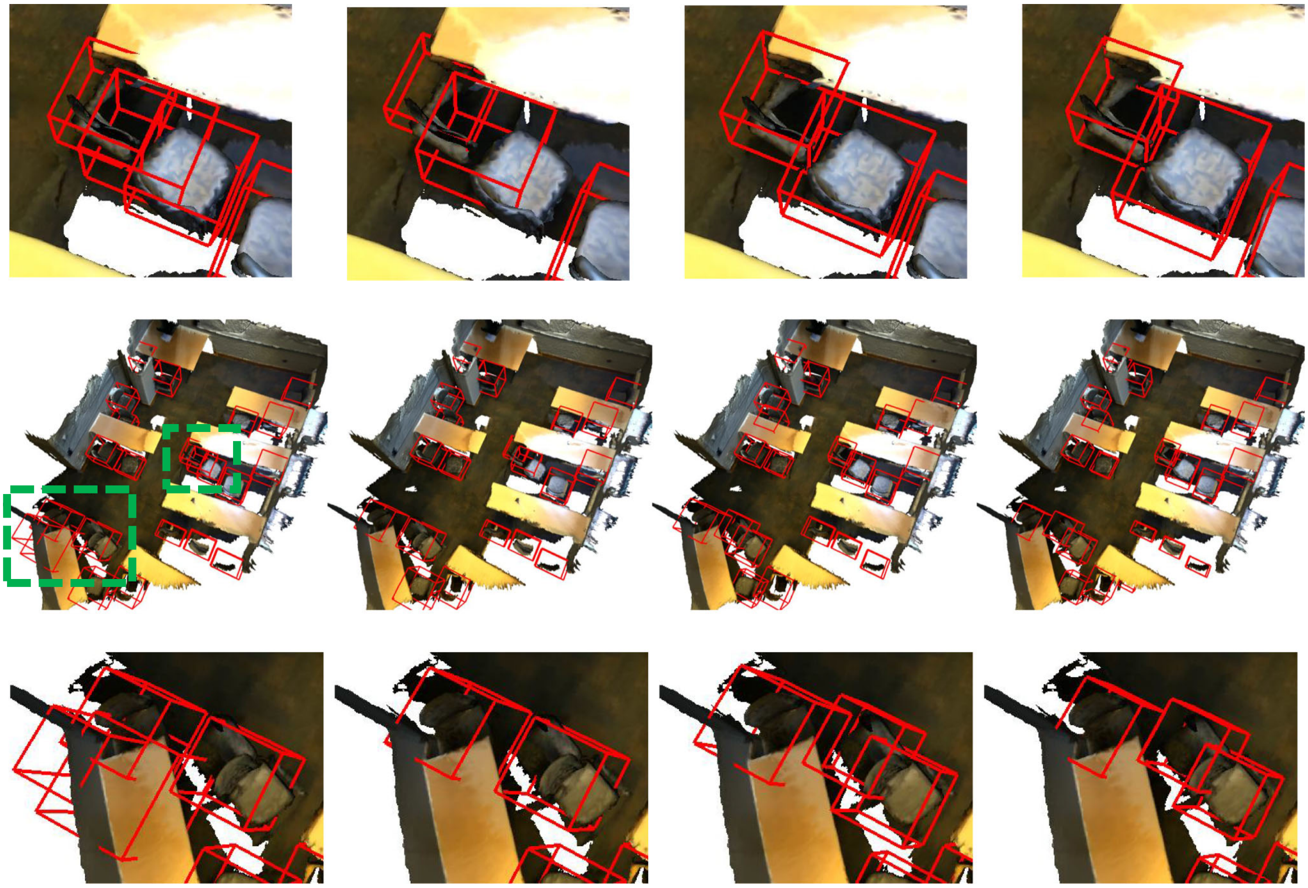
network reaches 60.9% on SUN RGB-D validation set, 3.2% higher than the current state-of-the-art, VoteNet. The heavy occlusion presented in SUN RGB-D dataset is a challenge for methods (e.g., VoteNet) that consider point patches individually. However, the use of contextual information and improved NMS in our method helps with the detection of occluded objects with missing parts, which we believe is the reason for the improved detection accuracy.

4.5 Effectiveness of PPC

As described in Sect. 3.2, patch-to-patch context (PPC) sub-module is integrated in our network to exploit the useful information between these point patches (i.e., seed points). In this group of experiments, we perform an ablation study to evaluate the effectiveness of the PPC sub-module, which is visualized in Fig. 10. As shown, with the PPC module, the voted centers (green) are more meaningful with more of them appearing on objects rather than on non-object regions. Moreover, the voted centers are more closely clustered compared to those without the module (red). The results demonstrate that our self-attention based weighted fusion over local point patches can indeed enhance the performance of voting for object centers.

4.6 Effectiveness of OOC

Figure 11 shows a detection example using the proposed network with and without the OOC sub-module. It is found that without OOC, boxes with wrong labels are detected. As can be seen in Fig. 11a, a toilet (black box) that is close to a chair (red box) is wrongly detected without OOC. Actually, the point cloud in the black box could be mis-classified as



(a) Typical 3DNMS

(b) SOB-3DNMS Cls Score

(c) SOB-3DNMS IOU Score

(d) Groundtruth

Fig. 14 Qualitative results of 3D object detection on SUN RGB-D. Cls (classification) score chooses the wrong box, while our IOU score keeps the correct box and eliminates the wrong one, which has a relatively high classification score but a low IOU score

toilet even by humans, if we look at it separately. So, it is reasonable that the false detection, i.e., the toilet, is generated, since the feature of each box is individually processed, without consideration of the surrounding boxes. When the OOC sub-module is integrated, the surrounding red boxes are taken into consideration. Then, as shown in Fig. 11b, the wrong label of toilet (black) is changed to the label of chair (red). Finally, we can get the very similar detection result (c) to the groundtruth (d) after the post-processing of SOB-3DNMS. This example demonstrates that the OOC sub-module enables communication between object proposals and provides more comprehensive information to improve 3D object detection.

Table 4 Quantitative comparison between typical NMS and the proposed SOB-3DNMS on ScanNetV2, evaluated with mAP@0.25 IoU

Method	Typical NMS		SOB-3DNMS	
	Cls score	IoU score	Cls score	IoU score
mAP	65.0	65.1	65.5	66.2

Cls means classification. The IOU score is predicted by the network, similar to 3D IOU-Net Li et al. (2020a)

4.7 Effectiveness of GSC

As described in Sect. 3.2, the GSC sub-module is proposed to learn the contextual information at the global level. In this way, the inference of the final 3D bounding boxes and the object classes will consider the compatibility with the scene context, making the final predictions more reliable under the global cues. As shown in Fig. 12, the given scene is a conference room. However, without consideration of global scene context, VoteNet generates detections of a sofa (colored in red) and a counter (colored in pink), which rarely happens in the training data. In contrast, the GSC module in our method effectively reduces false detections in the scene, which implies that the integration of GSC sub-module to capture global context is beneficial for the object labeling task.

Table 5 Quantitative comparison on two kinds of ranking scores under the proposed SOB-3DNMS framework

Ranking scores	mAP@0.25		mAP@0.5	
	ScanNet	SUN RGB-D	ScanNet	SUN RGB-D
IOU score	66.2	60.9	45.3	39.7
Cls score	65.5	60.3	43.1	38.5

It is obvious that using the predicted IOU score is more effective than Cls (classification) score

Table 6 Ablation study on the validation dataset. The baseline model is trained by ourselves using the official VoteNet code released by the authors

Method	Multi-level context			SOB-3DNMS	mAP@0.25	
	PPC	OOC	GSC		SUN RGB-D	ScanNet
Baseline					57.8	59.6
Ours				✓	58.9	62.5
Ours	✓				58.6	62.2
Ours	✓	✓			59.1	63.4
Ours	✓	✓	✓		60.1	65.0
Ours	✓	✓	✓	✓	60.9	66.2

4.8 Effectiveness of SOB-NMS

As described in Sect. 3.3, the proposed SOB-3DNMS can further refine the duplication removal in post-processing. Figures 13 and 14 compare the final results using SOB-3DNMS and the typical 3DNMS after the box proposal and classification.

From Fig. 13, we can see that false positive 3D boxes still remain using the typical 3DNMS, while using SOB-3DNMS, the results are much cleaner with all the false positives removed. Specifically, in the first scene (first row), three redundant chairs (pointed by yellow arrows) are still detected. Because their overlaps (i.e., 3D IOU) with any other boxes are below the fixed NMS threshold, which is set to be 0.25. It may look as if their overlaps with other boxes are large visually, but the calculation of their 3D IOU could be much smaller. Using our method, as shown in Fig. 13b, the above three redundant boxes are all successfully removed with the adaptive threshold. The reason is that the surrounding boxes are detected with higher confidence scores, resulting lower thresholds to suppress the redundant boxes even harder.

We also evaluate the new box ranking strategy. Figure 14b and c compare the detection results using the classification score for ranking and using the predicted IOU score for ranking. It can seen that using the latter generates boxes with more accurate locations. This is because using the predicted IOU scores specifically takes object location into consideration, and thus provides a measure that better reflects the detection quality.

Table 4 reports the quantitative results on the typical NMS and the proposed SOB-3DNMS. Similar to 3D IOU-Net (Li et al. 2020a), we adopt the typical NMS strategy, and use the network-predicted IOU score to replace classification confidence as the ranking metric. As seen, when using typical NMS, the IOU score can merely get marginal improvement. However, the combination of IOU score and the proposed SOB-3DNMS achieves the best performance. Table 5 quantitatively compares the results using the two different ranking scores. Consistent with the visual results in Fig. 14, using the proposed IOU score as the ranking criterion achieves better performance than using the traditional classification score for both mAP@0.25 and mAP@0.50. Moreover, it is noticed that the improvement on mAP@0.50 is more significant on both datasets. The reason is that when using the IOU score for ranking, the selected boxes tend to have more accurate locations. The improved accuracy makes the results more robust to changes of evaluation criterion, and thus the performance is less decreased when the criterion becomes stricter from mAP@0.25 to mAP@0.50. These results further demonstrate the effectiveness of SOB-3DNMS, especially the IOU ranking score as a better quality measure of boxes and its suitability in 3DNMS for ranking.

Table 7 Ablation study on the validation dataset of ScanNet

Method	PPC	OOC	GSC	mAP@0.25
Baseline				61.5
Ours	✓			62.3
Ours	✓	✓		63.6
Ours	✓	✓	✓	65.0

Baseline: VoteNet with a Sparse Residual U-Net (Choy et al. 2019) as backbone network

4.9 Ablation Study

To quantitatively evaluate the effectiveness of the proposed contextual sub-modules and the SOB-3DNMS, we conduct experiments with different combinations of these components. The quantitative results are shown in Table 6. The baseline method is VoteNet. We then add the proposed contextual sub-modules one by one into the baseline model. Applying the PPC module leads to improvements in mAP@0.25 of 0.8 and 2.6. The combination of PPC and OOC modules further improves the evaluation scores to 59.1 and 63.4 respectively. As expected, when equipped with all the three sub-modules, the mAP@0.25 of our network is boosted up to higher scores on both datasets. It can be seen that contextual information captured by the designed sub-modules indeed brings notable improvements over the state-of-the-art method. In terms of performance improvement for the SOB-3DNMS, it alone improves the baseline model by 1.1 and 2.9 on the two datasets. When it is further combined with the contextual model, we achieve the best final results of 66.2 and 60.9 on mAP@0.25, compared to the baseline at 57.8 and 59.6.

To further evaluate our multi-level context encoding strategy with a stronger backbone, we replace the PointNet++ backbone in VoteNet with a Sparse Residual U-Net, which is proposed in (Choy et al. 2019) and achieves promising results on ScanNet benchmark in the instance segmentation task. We evaluate the detection performance on the ScanNet dataset. As shown in Table 7, using the new backbone can increase mAP@0.25 to 61.5. It is noticed that the improvement brought in by PPC using the new backbone is smaller than using PointNet++ backbone, while the improvements brought in by OOC and GSC are almost the same. We reckon the reason is that the Sparse Residual U-Net can also capture the context between seed points (local point patch level) to some extent. However, the object and global scene level context cannot be captured in the Sparse Residual U-Net backbone. In semantic segmentation, the dominant convolution and deconvolution operations benefit the most from sparse convolutional layers, which leads to the huge success. However, unlike semantic segmentation, there are two key components subsequent to PointNet++/sparse convolutional backbone in the architecture of VoteNet, the Voting and the Proposal&Classification steps. These two steps are not as directly influenced by the sparse convolutional backbone as in semantic segmantation. This may be a reason why the improvement brought in by a better backbone feature representation in the VoteNet-based 3D object detection is not as significant as in semantic segmentation. Moreover, the PPC sub-module in our paper is also an enhancement of the PointNet++ backbone, as it helps get more representative features of point patches by enabling the information communication between seed points. That means the PPC sub-module has overlapped advantages with the sparse convolutional backbone, and thus the combination of the two sees insignificant further improvements. As can be seen in Tables 6 and 7, the improvement brought in by the PPC sub-module with the sparse convolutional backbone is smaller than with PointNet++.

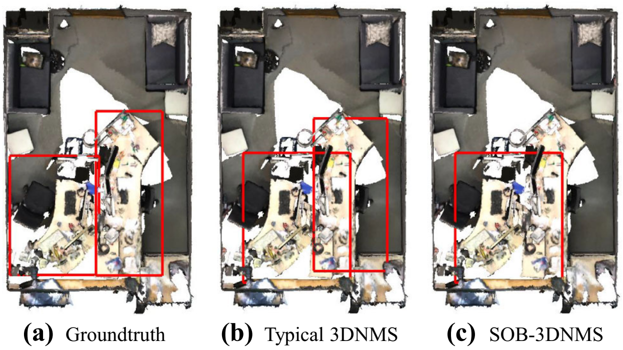


Fig. 15 A failure case of our method. There are two connected desks in the scene (**a**). However, our SOB-3DNMS removes the right desk in (**b**) since it overlaps “too much” with the left one

4.10 Limitation

While our method improves the accuracy performance of deep Hough voting based 3D object detector via introducing contextual information, it is not without limitation. As shown in Fig. 15, our method may get the wrong detection result when two desks are connected together. In that case, our model may detect two desks as one whole desk (Fig. 15c). As seen in Fig. 15b, the left box incorrectly covers most part of the two connected desks. Then, our SOB-3DNMS would only keep one box with the highest ranking score, as it treats the overlapping box (the right box in (b)) as a duplication. The reason is that SOB-3DNMS assumes that 3D bounding boxes of the same category should not overlap too much with each other in a 3D scene. However, we think it is reasonable to some extent that our method tends to detect these two desks as a united one in (c), since these two desks are closely connected to each other, and the left box coincidentally covers almost the two desks. Moreover, we realize that the assumption of our SOB-3DNMS is still not strictly true, as, although not observed in our experiments, some objects of the same class may still overlap. For example, a small table may be placed under a large table.

5 Conclusions

In this paper, we propose a novel network that integrates contextual information at multiple levels into 3D object detection. We make use of self-attention mechanism and multi-scale feature fusion to model the multi-level contextual information, and propose three sub-modules. The PPC module encodes the relationships between point patches, the OOC module captures the contextual information of object candidates, and the GSC module aggregates the global scene context. Moreover, an enhanced 3DNMS, i.e., SOB-3DNMS, is proposed to improve the filtering of boxes in post-processing by considering the spatial locations of objects in 3D space. Ablation studies demonstrate the effectiveness of the proposed contextual sub-modules and the SOB-3DNMS. Quantitative and qualitative experiments further demonstrate that our architecture successfully improves the performance of 3D object detection.

Future work Contextual information analysis in 3D object detection still offers huge space for exploration. For example, to enhance the global scene context constraint, one possible way is to use the global feature in the GSC module to predict scene types as an auxiliary learning task, which can explicitly supervise the global feature representation. Another direction would be a more effective mechanism to encode the contextual information as in Hu et al. (2018a). Moreover, the reasoning behind the intuition that self-attention may be learning spatial context is not yet perfectly demonstrated by the experimental validation. We plan to make a much deeper study of this point in the future. Apart from contextual information analysis, another promising direction is about the NMS post-processing. First, there is room to improve the current 3D IOU calculation. Several works have been proposed based on the typical IOU formula, such as Generalized-IoU (Rezatofighi et al. 2019) and Distance-IoU (Zheng et al. 2020) in 2D field, which may inspire better 3D IOU calculation in future work. Second, traditional NMS methods are usually considered to be not efficient enough, owing to their complex designs. Non-NMS based methods are becoming increasingly popular both in 3D (Yin et al. 2020; Wang et al. 2020) and 2D (Hu et al. 2018a; Carion et al. 2020). Thus, how to design the 3D detectors that eliminate NMS is another promising direction of future work.

Acknowledgements This work is funded by the National Key Research and Development Program of China (2020YFB2010702, 2018AAA0102200), National Natural Science Foundation of China under Grant 61772267, Aeronautical Science Foundation of China (No. 2019ZE052008), and the Natural Science Foundation of Jiangsu Province under Grant BK20190016.

References

Atzmon, M., Maron, H., & Lipman, Y. (2018). Point convolutional neural networks by extension operators. arXiv preprint arXiv:1803.10091.

Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017). Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision* (pp. 5561–5569).

Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE international conference on computer vision workshops*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. arXiv preprint arXiv:2005.12872.

Chen, Z., Huang, S., & Tao, D. (2018). Context refinement for object detection. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 71–86).

Chen, J., Lei, B., Song, Q., Ying, H., Chen, DZ., & 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* (pp. 392–401).

Choy, C., Gwak, J., & Savarese, S. (2019). 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3075–3084).

Dai, A., Chang, AX., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5828–5839).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886–893). IEEE.

Deng, H., Birdal, T., & Ilic, S. (2018). Ppfnet: Global context aware local features for robust 3D point matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 195–205).

Engelmann, F., Bokeloh, M., Fathi, A., Leibe, B., & Nießner, M. (2020). 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9031–9040).

Engelmann, F., Kontogianni, T., Hermans, A., & Leibe, B. (2017). Exploring spatial context for 3D semantic segmentation of point clouds. In *Proceedings of the IEEE international conference on computer vision* (pp. 716–724).

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3146–3154).

He, C., Zeng, H., Huang, J., Hua, XS., & Zhang, L. (2020). Structure aware single-stage 3d object detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11,873–11,882).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

He, Y., Zhang, X., Savvides, M., & Kitani, K. (2018). Softer-nms: Rethinking bounding box regression for accurate object detection. arXiv preprint arXiv:1809.08545

Hou, J., Dai, A., & Nießner, M. (2019). 3d-sis: 3D semantic instance segmentation of RGB-D scans. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4421–4430).

Hu, H., Gu. J., Zhang. Z., Dai. J., & Wei, Y. (2018a). Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3588–3597).

Hu, J., Shen, L., & Sun, G. (2018b). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Hu, S. M., Cai, J. X., & Lai, Y. K. (2018c). Semantic labeling and instance segmentation of 3D point clouds using patch context analysis and multiscale processing. *IEEE Transactions on Visualization and Computer Graphics, 26*, 2485–2498.

Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, CW., & Jia, J. (2020). Pointgroup: Dual-set point grouping for 3d instance segmentation. arXiv preprint arXiv:2004.01658.

Lahoud, J., & Ghanem, B. (2017). 2D-driven 3D object detection in RGB-D images. In *Proceedings of the IEEE international conference on computer vision* (pp. 4622–4630).

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 12,697–12,705).

Li, J., Luo, S., Zhu, Z., Dai, H., Krylov, A. S., Ding, Y., & Shao, L. (2020a). 3d iou-net: Iou guided 3d object detector for point clouds. arXiv preprint arXiv:2004.04962.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., & Chen, B. (2018). PointCNN: Convolution on x-transformed points. In *Advances in neural information processing systems* (pp. 820–830).

Li, Y., Ma, L., Tan, W., Sun, C., Cao, D., & Li, J. (2020b). Grnet: Geometric relation network for 3d object detection from point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing, 165*, 43–53.

Liu, S., Huang, D., & Wang, Y. (2019a). Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6459–6468).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21–37). Springer.

Liu, Y., Fan, B., Xiang, S., & Pan, C. (2019b). Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8895–8904).

Liu, Y., Wang, R., Shan, S., & Chen, X. (2018). Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6985–6994).

McCormac, J., Clark, R., Bloesch, M., Davison, A., & Leutenegger, S. (2018). Fusion++: Volumetric object-level SLAM. In *2018 international conference on 3D vision (3DV)* (pp. 32–41). IEEE.

Mottaghi, R., Chen, X., Liu, X., Cho, N. G., Lee, S. W., Fidler, S., Urtasun, R., & Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 891–898).

Najibi, M., Lai, G., Kundu, A., Lu, Z., Rathod, V., Funkhouser, T., Pantofaru, C., Ross, D., Davis, L. S., & 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* (pp. 11,913–11,922).

Paigwar, A., Erkent, O., Wolf, C., & Laugier, C. (2019). Attentional PointNet for 3D-object detection in point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*.

Qi, C. R., Chen, X., Litany, O., & Guibas, L. J. (2020). Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4404–4413).

Qi, C. R., Litany, O., He, K., & Guibas, L. J. (2019). Deep Hough voting for 3D object detection in point clouds. arXiv preprint arXiv:1904.09664.

Qi, C. R., Liu, W., Wu, C., Su, H., & 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* (pp. 918–927).

Qi, C. R., Su, H., Mo, K., & 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* (pp. 652–660).

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017b). PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems* (pp. 5099–5108).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Ren, Z., & 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* (pp. 1525–1533).

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 658–666).

Salscheider, N. O. (2020). Featurenms: Non-maximum suppression by learning feature embeddings. arXiv preprint arXiv:2002.07662.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., & Li, H. (2020). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*

Shi, S., Wang, X., & Li, H. (2019a). PointRCNN: 3D object proposal generation and detection from point cloud. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–779).

Shi, W., & Rajkumar, R. (2020). Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1711–1719).

Shi, Y., Chang, AX., Wu, Z., Savva, M., & Xu, K. (2019b). Hierarchy denoising recursive autoencoders for 3D scene layout prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1771–1780).

Song, S., Lichtenberg, S. P., & 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* (pp. 567–576).

Song, S., & 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* (pp. 808–816).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., & Savarese, S. (2019). DenseFusion: 6D object pose estimation by iterative dense fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3343–3352).

Wang, G., Tian, B., Ai, Y., Xu, T., Chen, L., & Cao, D. (2020). Centernet3d: An anchor free object detector for autonomous driving. arXiv preprint arXiv:2007.07214.

Wang, P. S., Liu, Y., Guo, Y. X., Sun, C. Y., & Tong, X. (2017). O-CNN: Octree-based convolutional neural networks for 3D shape analysis. *ACM Transactions on Graphics (TOG)*, *36*(4), 72.

Wang, T., He, X., & Barnes, N. (2013). Learning structured Hough voting for joint object detection and occlusion reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1790–1797).

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803).

Xie, Q., Lai, YK., Wu, J., Wang, Z., Zhang, Y., Xu, K., & Wang, J. (2020). Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10,447–10,456).

Xie, S., Liu, S., Chen, Z., & Tu, Z. (2018). Attentional shapecontextnet for point cloud recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 4606–4615).

Xu, D., Anguelov, D., & Jain, A. (2018). Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 244–253).

Yang, J., Lu, J., Lee, S., Batra, D., & Parikh, D. (2018). Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 670–685).

Yang, Z., Sun, Y., Liu, S., & Jia, J. (2020). 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11,040–11,048).

Ye, X., Li, J., Huang, H., Du, L., & Zhang, X. (2018). 3D recurrent neural networks with context fusion for point cloud semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 403–417).

Yi, L., Zhao, W., Wang, H., Sung, M., & Guibas, L. J. (2019). GSPN: Generative shape proposal network for 3D instance segmentation in point cloud. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3947–3956).

Yin, T., Zhou, X., & Krähenbühl, P. (2020). Center-based 3d object detection and tracking. arXiv preprint arXiv:2006.11275.

Yu, R., Chen, X., Morariu, V. I., & Davis, L. S. (2016). The role of context selection in object detection. arXiv preprint arXiv:1609.02948.

Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., & Xu, F. (2018). Compact generalized non-local network. In *Advances in neural information processing systems* (pp. 6510–6519).

Zambaldi, V., Raposo, D., Santoro, A., Bapst, V., Li, Y., Babuschkin, I., Tuyls, K., Reichert, D., Lillicrap, T., Lockhart, E., et al. (2018). Relational deep reinforcement learning. arXiv preprint arXiv:1806.01830.

Zhang, W., & Xiao, C. (2019). PCAN: 3D attention map learning using contextual information for point cloud based retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 12,436–12,445).

Zhang, Y., Bai, M., Kohli, P., Izadi, S., & Xiao, J. (2017). Deepcontext: Context-encoding neural pathways for 3D holistic scene understanding. In *Proceedings of the IEEE International conference on computer vision* (pp. 1192–1201).

Zhang, Y., Song, S., Tan, P., & Xiao, J. (2014). Panocontext: A whole-room 3D context model for panoramic scene understanding. In *European conference on computer vision* (pp. 668–686). Springer.

Zhang, H., Zhang, H., Wang, C., & Xie, J. (2019). Co-occurrent features in semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 548–557).

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-iou loss: Faster and better learning for bounding box regression. In *AAAI* (pp. 12,993–13,000).

Zhou, Y., & 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* (pp. 4490–4499).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.