DisARM: Displacement Aware Relation Module for 3D Detection

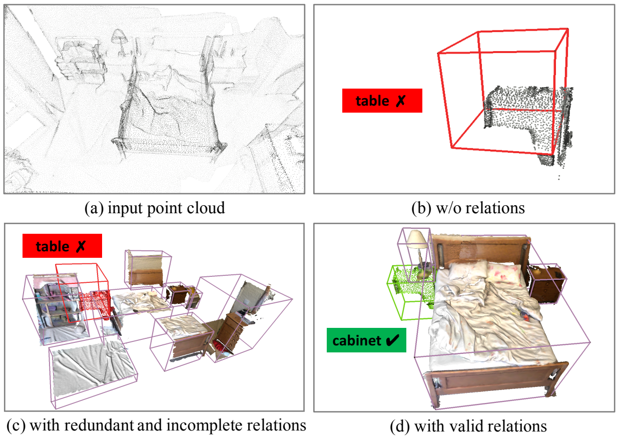
Yao Duan   Chenyang Zhu   Yuqing Lan   Renjiao Yi   Xinwang Liu   Kai Xu*
National University of Defense Technology

Abstract

We introduce Displacement Aware Relation Module (DisARM), a novel neural network module for enhancing the performance of 3D object detection in point cloud scenes. The core idea is extracting the most principal contextual information is critical for detection while the target is incomplete or featureless. We find that relations between proposals provide a good representation to describe the context. However, adopting relations between all the object or patch proposals for detection is inefficient, and an imbalanced combination of local and global relations brings extra noise that could mislead the training. Rather than working with all relations, we find that training with relations only between the most representative ones, or anchors*, can significantly boost the detection performance. Good anchors should be semantic-aware with no ambiguity and able to describe the whole layout of a scene with no redundancy. To find the anchors, we first perform a preliminary relation anchor module with an objectness-aware sampling approach and then devise a displacement based module for weighing the relation importance for better utilization of contextual information. This light-weight relation module leads to significantly higher accuracy of object instance detection when being plugged into the state-of-the-art detectors. Evaluations on the public benchmarks of real-world scenes show that our method achieves the state-of-the-art performance on both SUN RGB-D and ScanNet V2. The code and models are publicly available at* https://github.com/YaraDuan/DisARM.

1. Introduction

Detecting objects directly from the 3D point cloud is challenging yet imperative in many computer vision tasks, such as autonomous navigation, path planning for robotics, as well as some AR applications. The goal of 3D object detection is to localize all valid shapes and recognize their semantic label simultaneously, which puts forward high requirements for understanding the whole input scene.



(a) input point cloud

(b) w/o relations

(c) with redundant and incomplete relations

(d) with valid relations

Figure 1. Illustration of the importance of DisARM. (b) It is easy to mistake the cabinet as a table when the point cloud is incomplete and featureless. (c) Redundant relations are usually incomplete and lose the important displacement information of the target object. (d) The network can recognize and locate the cabinet easily with the help of DisARM which provides valid surrounding environment information.

With the rapid development of deep learning and the increasing scale of the online 3D dataset, data-driven methods such as CNN have been widely adopted for object detection. The critical observation of these methods is that the context is as important as the object itself for accurate detection. However, the extra information provided by 3D brings noise and irregularity, which makes it more challenging to apply convolution to gather the correct context for detection.

To avoid irregularity while applying convolution for 3D object detection, the community recently introduces two typical categories of methods. [17, 37, 45] are trying to project the raw point cloud onto aligned structures such as voxel grids which can apply 3D convolution naturally. In an alternative way, [25] adopts max-pooling to fuse information of an irregular point cloud directly. These methods can achieve good performance while the input scene is complete and clean. However, the real scanned data is usually incomplete and noisy, making it difficult to extract the key information through this intrinsic context fusion approach.

To further release the power of context, some methods try to adopt the context explicitly for object detection.

*Corresponding author: kevin.kai.xu@gmail.com

Building a relation graph between objects is a natural way to utilize the context. [32] leverages inference on scene graphs to enhance 3D scene understanding. However, it requires additional supervision for regression of a correct scene graph. Some methods intend to utilize all the possible relations among the scene to avoid this extra labeling labor. [34] introduces a multi-level framework to fuse all the local and global neighborhoods for 3D object detection. Even a hierarchical architecture is proposed to maintain the context, considering all the relations is still redundant. Furthermore, most methods that adopt context explicitly have their customized network architecture, making it difficult to enhance existing detection methods.

We believe that context fusion is critical for 3D understanding, which can improve object detection performance. We introduce a novel neural network module named Displacement Aware Relation Module (DisARM). It can be easily assembled with most existed object detection methods and achieves state-of-the-art performance on existing benchmarks. The key idea is that context should not only be a structure for information fusion. The relation itself is also a critical feature for 3D understanding. Unlike some previous methods, we try to encode the most critical relations explicitly for potential proposals to allow richer information to be included during the training.

To avoid the redundant relation features that mislead the training and extract the information that matters, we select and collect the most critical context from two aspects. First, we introduce a relation anchor module, which only samples the most representative and informative proposals as anchors through an objectness-aware *Furthest Point Sampling (FPS)* on feature space. The insight of this design is that the relation anchors for context encoding should distribute uniformly over the feature space while being complete and clean. Our experiments demonstrate that adopting these relation anchors instead of the whole set of relations for context fusion is more efficient and accurate. To maximize the utilization of the proposed relation anchors, we introduce a dynamic weighing mechanism depending on spatial and feature displacement. The key insight here is that the importance of each anchor should be variant regarding recognizing different objects. The importance should depend on the spatial layout and semantic relations between the object and the anchors since the object placement usually go with some specific organization pattern for indoor scenes. In summary, the contributions of this paper include:

- We propose a portable network module that can be assembled with most existing 3D object detection methods to further improve the performance, which can be easily implemented as a plug-in for widely used object detection toolbox like MMdetection3D [5].
- We introduce a method describing 3D context as a set of weighted representative anchors. This method can effectively extract valid information from the redundant relations in a complex scene.
- Our method is simple but effective, which achieves **state-of-the-art** performances on ScanNet V2 and mAP@0.25 on SUN RGB-D.

2. Related Work

2.1. 3D object detection on point clouds

3D object detection on point clouds is challenging due to the irregular and sparse distribution of points. Earlier attempts project the point clouds onto grids [2] and voxels [12,17,27,37,45], so that the convolutional networks can be directly applied. But these methods often suffer from the computational cost and quantization errors. Other methods localize the objects with the help of shape templates [39] or sliding shapes [30, 31]. As an alternative, some methods rely on the candidates from RGB-driven 2D proposal generation [16, 24] or segmentation hypotheses [15, 28].

PointNet [25] has pioneered the processing of irregular point clouds. Since then, point based detection methods have been proposed to directly compute features from point clouds for 3D object detection. PointRCNN [26] applies the idea of R-CNN [11] to 3D object detection which generates and refines proposals by the points within 3D boxes to obtain the final detection results. VoteNet [23] generates the points lying close to object centers by voting, which can be grouped and aggregated to compute proposal features by PointNet [25]. Some follow-up works further imporve the vote and point group generation procedure [4, 43] or the object box localization and recognition procedure [1]. GroupFree3D [19] computes object features from points by attention mechanism for more accurate detection results.

2.2. Relation information in 3D object detection

Contextual information has been demonstrated to be helpful in variety of computer vision tasks, including 2D object detection [13, 40], point cloud semantic segmentation [9, 38] and 3D scene understanding [18, 35, 41, 42]. Moreover, the relationships between objects can be treated as special contextual information which can help the network to improve the performance on computer vision tasks.

A line of works [33, 44] incorporate graph structures to describe the relationships or exploit the graph convolution networks for relation feature learning. [14] models the graph structure of furniture in indoor scenes by defining five types of relations, which, however, is time-consuming for computation of relations. [10] uses the pair-wise relationship information to construct 3D object-object relation graph but needs extra supervision. 3DSSG [32] defines a rich set of relationships and generates a graph to describe the objects in the scene as well as their relationships which

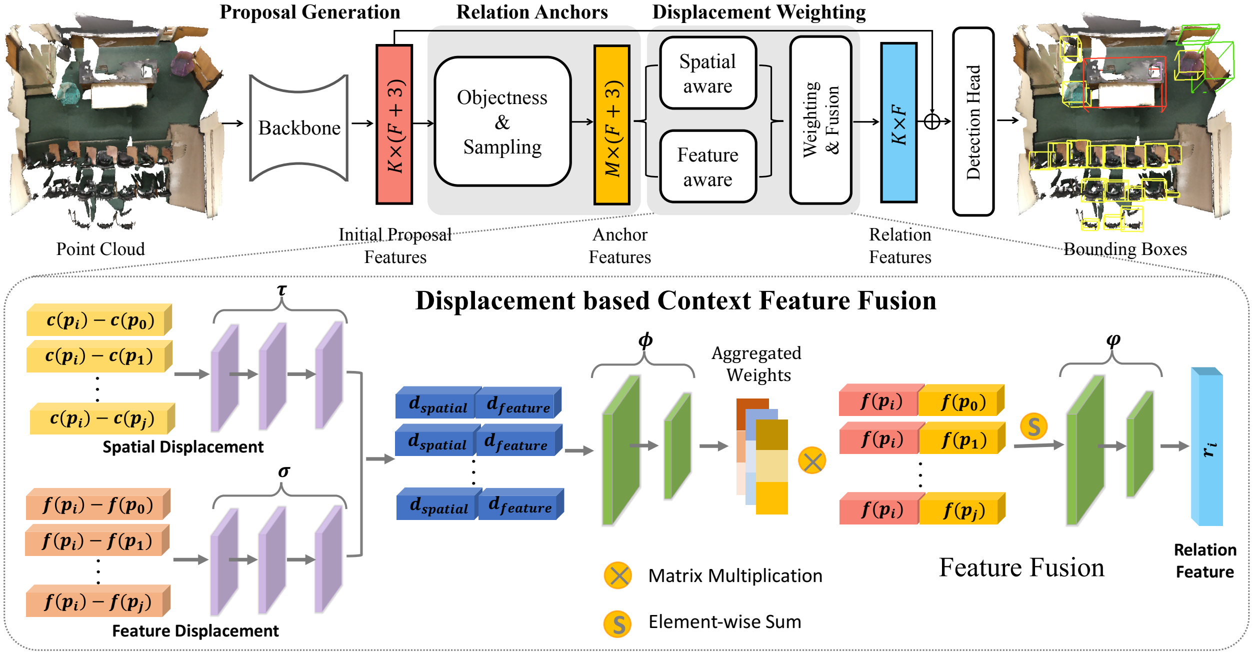


Figure 2. **DisARM network architecture.** Taking K proposals generated by backbone network as input, we first sample M relation anchors with rich information of the scene's layout. For each of proposals, we get the weights relative to the anchors by considering spatial-aware and feature-aware displacement . At last, the relation feature is obtained by fusing the weighted proposal-anchor pair features. Note that there is a skip connection operation of relation feature and proposal feature for final detection. In the bottom of figure, $c(p_i)$, $c(p_j)$ and $f(p_i)$, $f(p_j)$ indicate locations and features of proposals and anchors respectively ; τ, θ, ϕ and φ are the functions consisting of MLPs.

heavily depends on the ground-truth of instance segmentation. HGNet [1] leverages a graph convolution network to promote performance by reasoning on proposals, while it might be useless if the features for detecting an object had not been adequately learned.

Another line of works capture the relation information by incorporating features of objects in a variety of ways by neural networks which usually are accompanied by attention mechanisms. SRN [7] models the geometrical and locational relations of the local regions by considering the inner interactions of the object, which is not suitable for large indoor scenes understanding. MLCVNet [34] addresses the 3D object detection task by incorporating multi-level contextual information with the self-attention mechanism and multi-scale feature fusion by considering relations of all objects which results in information redundancy. MonoPair [3] learns the pair-wise relationship by only considering the spatial information.

3. DisARM module

3.1. Overview

Some cognitive psychology theories [9, 13, 41, 42] suggest that context can enhance the perception ability for detection. This paper proposes a portable network module, say DisARM, to utilize the 3D context effectively, which can be easily assembled with existing object detection methods to enhance performance.

In our case, we argue that useful contextual information for detection in indoor scenes needs to meet two criteria: it can reflect the intra-relationship between objects and implicitly represent the layout. Therefore, an end-to-end network framework is proposed to extract the context effectively. As demonstrated in Figure 2, the former module of DisARM samples the relation anchors between the learned deep feature of each potential object proposal and the following module takes the relative displacement of each proposal between anchors to encode the scene layout. More specifically, the core of former module is locating the most representative and informative proposals for relation feature construction. We denote these selected proposals as anchors (see Section 3.2). The following 2-way module calculates the weights for each anchor through the analysis of spatial and feature displacement (see Section 3.3). Our experiments demonstrate that the proposed framework can extract the context for detection effectively and improve the performance significantly over some state-of-the-art alternatives.

3.2. Relation anchors

Initial proposals Our DisARM requires initial object proposals $\mathcal{P} = \{p_0, p_1, ..., p_K\}$ to boost the relation analysis. VoteNet [23] is a widely used 3D detection network that

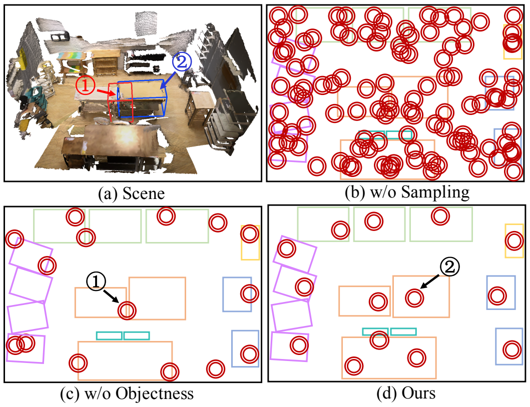


Figure 3. Anchors. In order to show the results more intuitively, we draw the anchors in red circles and objects in rectangles. Anchors without sampling in (b) are redundant. But some anchors sampled by FPS are incomplete and invalid(Figure 3, (c)), such as anchor ① which contains parts of the tables. Therefore, we sample on the high objectness anchors. Best view in (a).

can provide good object proposals. However, it lacks the consideration of the relationships between objects and surroundings. We adopt VoteNet [23] as the backbone to produce the input object proposals for DisARM. Note that DisARM can also aggregate with some other detection methods [4, 19, 22]. The evaluation are demonstrated in Table 5.

Each proposal p_i is represented with its center point. The feature encoder network has several multi-layer perception (MLP) layers and feature propagation layers with skip connections. The output feature $f(p_i)$ is an F dimension vector, which is the aggregation of the learned deep feature of each vote that supports the proposal p_i.

Proposal objectness As shown in Figure 3, the whole set of $\mathcal{P}$ is somehow redundant and contains massive incomplete and invalid proposals. Considering all the possible relations in a scene to formulate a context feature is ineffective and may introduce too much noisy information. Therefore, the key to designing a mechanism for utilizing these relations effectively is locating the most representative and informative ones. Figure 3 demonstrates only few proposals given by the backbone are complete. We introduce the concept of objectness to filter the incomplete and noisy ones.

Given a proposal p_i and its corresponded feature $f(p_i)$, we denote its objectness as $o(p_i)$. Then top-N proposals are selected by the objectness scores as candicate anchors $\mathcal{P}^{'}$. The network module calculating the objectness is a simple MLP network with fully connected layers, sigmoid activation and batch normalization. Since most datasets only label the valid objects $\mathcal{P}_{\text{gt}}$ in a scene, we define the objectness loss

$$\text{loss}_{\text{obj}} = \|o(p_i) - \chi_{\mathcal{P}_{\text{gt}}}(p_i)\| \tag{1}$$

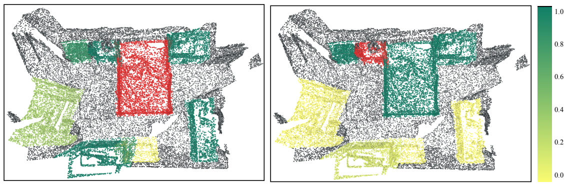


Figure 4. Displacement weights. We show the proposal in red points and anchors with different colors correspond to different weights. The bed has different perception of cabinets and shelves which is effected by spatial displacement. Cabinet in the same scene is more interested in another cabinet which is determined by feature displacement.

$$p_i \in \mathcal{P}_{\text{gt}} \iff \exists p \in \mathcal{P}_{\text{gt}} \rightarrow \text{IoU}(p_i, p) > 0.25 \tag{2}$$

where $\chi_{\mathcal{P}_{\text{gt}}}(p_i)$ is a function indicating whether a proposal p_i belongs to a ground-truth object. As demonstrated in Figure 3, $o(p_i)$ can indicate the completeness of a given proposal which is critical for locating the proposal anchors.

Anchor sampling Even we only focus on complete proposals, the aggregation of $\mathcal{P}$ appears to be redundant. Previous works such as *KPS* in [19] which only focus on high objectness proposals will still introduce redundant information. We find that conducting *Furthest Point Sampling (FPS)* on $\mathcal{P}^{'}$ with the assistance of objectness evaluation can help us locate the most representative proposal anchors.

In details, a proposal p_0 with the highest objectness score is first sampled in $\mathcal{P}^{(0)}_{\text{anchor}}$. The next sampling would be processed as below,

$$\mathcal{P}^{(k+1)}_{\text{anchor}} = \{\mathcal{P}^{(k)}_{\text{anchor}}, \underset{p_i \in \mathcal{P}}{\text{argmax}} \sum_{p_j \in \mathcal{P}^{(k)}_{\text{anchor}}} o(p_i)\|f(p_i) - f(p_j)\| + \Delta c\} \tag{3}$$

Eq. 3 indicates the metric adopting in our Furthest Point Sampling (FPS) is upon the feature space $f(\cdot)$ weighted by the objectness score $o(\cdot)$ and distance offsets Δc. Then the farthest proposal p_i to the already-chosen proposal set $\mathcal{P}^{(k)}_{\text{anchor}}$ is iteratively selected until the number of chosen proposals meets the candidate budget M, which is 15 for all our evaluations. Though it is simple, the finally selected anchors are representative and distributed in the whole scene.

3.3. Displacement based context feature fusion

Spatial displacement The proposal anchors $\mathcal{P}_{\text{anchor}}$ can effectively describe the context of the whole input scene. However, they should not contribute equally for detection of different objects as demonstrated in Figure 4. Adopting appropriate anchors is critical to utilize context in detection. Inspired by [36], spatial layout patterns can effectively describe the representative substructures in an indoor scene. Therefore, we think the context for detection should be weighted by layout-aware spatial displacement as well.

We argue that an object has different perceptions towards different proposal anchors regarding with different spatial displacement. For example, cabinets are usually placed next to bed and chairs are most commonly placed in front of a desk or table. These patterns can be reflected by spatial displacement among proposal-anchor pairs. Thus, we regard the importance of different displacement around proposals as displacement weights which encourages networks to pay different levels of attention. For details, given the target proposal p_i with location $c(p_i)$ and a proposal anchor p_j with location $c(p_j)$, the spatial displacement between them is formulated as $d_{\text{spatial}}(p_i, p_j) = \tau(c(p_i) - c(p_j))$, where τ is a perception function given by an MLP network.

Feature displacement Similar with the spatial displacement, the feature displacement $f(p_i) - f(p_j)$ given by the target proposal p_i and proposal anchor p_j should be also considered while measuring the importance of the proposal-anchor pair. The insight here is, layout patterns are sometimes semantic-aware. For example, the existence of a bathtub would always indicate a washbasin in the scene. This characteristic can be reflected by the pre-encoded features $f(p_i)$ and $f(p_j)$ since objects with similar semantic label would also be close on the feature space and vice versa. Therefore, given the target proposal p_i and a proposal anchor p_j, the feature displacement between them is formulated as $d_{\text{feature}}(p_i, p_j) = \sigma(f(p_i) - f(p_j))$, where σ is a perception function given by an MLP network.

Aggregated weights We concatenate spatial displacement $d_{\text{spatial}}(p_i, p_j)$ and feature displacement $d_{\text{feature}}(p_i, p_j)$ together to fuse the perceived information before putting them into an MLP network as shown in Figure 2. We can get final aggregated weights as below,

$$w(p_i, p_j) = \tanh(\phi[d_{\text{spatial}}(p_i, p_j); d_{\text{feature}}(p_i, p_j)]) \quad (4)$$

where ϕ is a perception function enabled by several MLP layers. To further normalize the weights between p_i and all the anchors in $\mathcal{P}_{\text{anchor}}$, we adopt softmax function and normalization operation $\mu(\cdot)$ in the end.

$$w(p_i, p_j) = \mu(\frac{w(p_i, p_j)}{\sum_{p_k \in \mathcal{P}_{\text{anchor}}} w(p_i, p_k)}) \quad (5)$$

Finally, We formulate the fused relation feature r_i of an object proposal p_i by a perception function φ for detection as below,

$$r_i = \varphi(\sum_{p_j \in \mathcal{P}_{\text{anchor}}} w(p_i, p_j) \cdot [f(p_i); f(p_j)]) \quad (6)$$

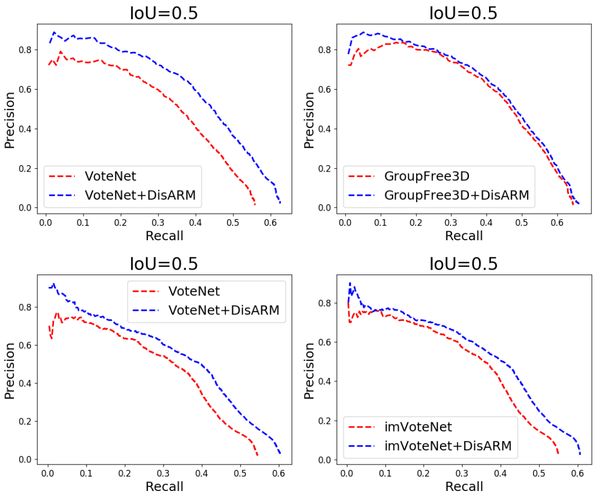


Figure 5. Precision-Recall Curves of different backbones equipped with our DisARM on mAP@0.5. We show the results on ScanNet V2 dataset in the first row and the results on SUN RGB-D dataset in the second row.

However, it is obvious that training $f(\cdot)$, $w(\cdot)$ and finding the optimal $\mathcal{P}_{\text{anchor}}$ are highly correlated which makes it a challenging optimization problem. The network finds the optimal r_i going through three stages during training. At the warm-up stage, $w(p_i, p_j)$ is inactive and the proposed module focus on locating the optimal $\mathcal{P}_{\text{anchor}}$ and training $f(p_i)$. The insight of this stage is $w(p_i, p_j)$ would only be functional while the network can already extract some reasonable proposal anchors. At the next stage, $\mathcal{P}_{\text{anchor}}$ and $f(p_i)$ are semantic enough and network focus on optimizing $w(p_i, p_j)$. This stage would fully utilize layout information extracted from the scene to measure the anchor importance. After these two stages, $w(p_i, p_j)$, $\mathcal{P}_{\text{anchor}}$ and $f(p_i)$ are fine-tuned together to achieve the final optimality.

4. Experiments

As our method can be applied to several backbones, we describe the implementation based on VoteNet [23] in brief. More details of other backbones are listed in the Supplementary. In our DisARM, we take the 256 output proposals of VoteNet [23] with 128-dimension features as input. And then we use a MLP network to predict objectness and subsample $N = 64$ candidate anchors according the scores. The MLP is realized with FC output sizes of 64, 32, 32, 1, where the final objectness scores are obtained by the output of last layer followed by sigmoid function. The function τ for spatial displacement has 3 layers of 8, 16, 32 hidden dimensions and function σ for feature displacement has 2 layers of 64, 32 hidden dimensions. The MLP hidden dimensions are 32, 1 of function ϕ for aggregated weights. The relation encoder φ for relation feature r_i has 4 layers of 256, 128, 128, 128 hidden dimensions.

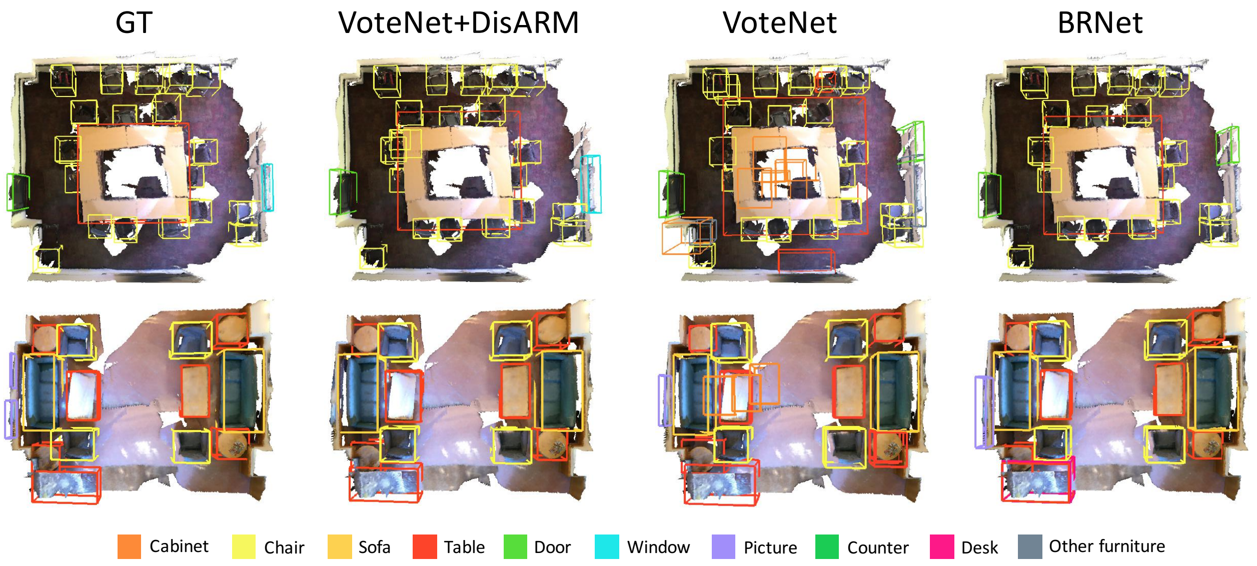


Figure 6. Qualitative results on ScanNet V2 dataset. We denote VoteNet+DisARM as applying our method to VoteNet. The first column is ground truth and the rest columns are detections of different methods. Best viewed on screen.

	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
VoteNet [23]	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
BRNet [4]	76.2	86.9	29.7	77.4	29.6	35.9	65.9	66.4	51.8	91.3	61.1
GroupFree3D [19]	**80.0**	**87.8**	32.5	79.4	32.6	36.0	66.7	70.0	**53.8**	91.1	63.0
imVoteNet [22]	75.9	87.6	41.3	76.7	28.7	**41.4**	**69.9**	70.7	51.1	90.5	63.4
VoteNet*+DisARM	76.7	86.2	35.4	78.4	31.0	34.6	66.3	68.1	51.2	86.9	61.5
imVoteNet*+DisARM	79.9	87.5	**43.7**	**80.7**	**33.3**	39.8	69.5	**74.1**	52.7	**91.6**	**65.3**

Table 1. 3D object detection results on SUN RGD-D val dataset with mAP@0.25. **Notations:** * denotes that the model is implemented on MMDetection3D. VoteNet*+DisARM and imVoteNet*+DisARM indicate applying our method to the 3D object detectors respectively.

We evaluate our method on two widely-used 3D object detection datasets: ScanNet V2 [6] and SUN RGB-D [29]. Standard data splits in [23] are adopted. Our network is end-to-end optimized with the batch size of 8. The initial learning rate is 0.008 and the network is trained for 220 epochs on both datasets. The Cosine Annealing [20] is adopted as the learning rate schedule. We implement our method on MMDetection3D [5] with one NVIDIA TITAN V GPU.

4.1. Comparisons

In this section, we compare our method with previous state-of-the-arts on ScanNet V2 and SUN RGB-D dataset, such as VoteNet [23] and its successors MLCVNet [34], HGNet [1], H3DNet [43], BRNet [4] and so on.

Quantitative results. The detection results of ScanNet V2 dataset are shown in Table 5. Applying our DisARM to VoteNet [23] achieves 66.1 on mAP@0.25 and 49.7 on mAP@0.5 over the implementation in MMDetection3D [5], which is **7.5** and **16.2** higher than the performance of VoteNet reported in [23].

Applying our DisARM to better 3D object detectors like H3DNet [43], BRNet [4], GroupFree3D [19], we obtain 0.4, 0.6, 0.7 improvement on mAP@0.25 and 0.8, 1.4, 2.9 improvement on mAP@0.5 respectively. Furthermore, DisARM applied to GroupFree3D [19] with the best-performance backbone achieves the **state-of-the-art** performance.

It is noteworthy that VoteNet*+DisARM outperforms GroupFree3D* using 12 attention modules on mAP@0.5, which indicates that our method is simple but more effective than those methods with complicated architectures. The results of more improved performance on mAP@0.5 which is a fairly challenging metric show that DisARM helps the backbones to detect the objects more accurately attributing success to our method eliminating ambiguity with the relational context information. We also draw the PR curves of different methods equipped with DisARM in Figure 5.

As shown in Table 1, we compare with previous state-of-the-arts on SUN RGB-D dataset. In the same way, we evaluate our method on VoteNet which outperforms the backbone on mAP@0.25 by 3.8 and mAP@0.5 by 5.5 (results in Supplementary). In particular, our DisARM applied to

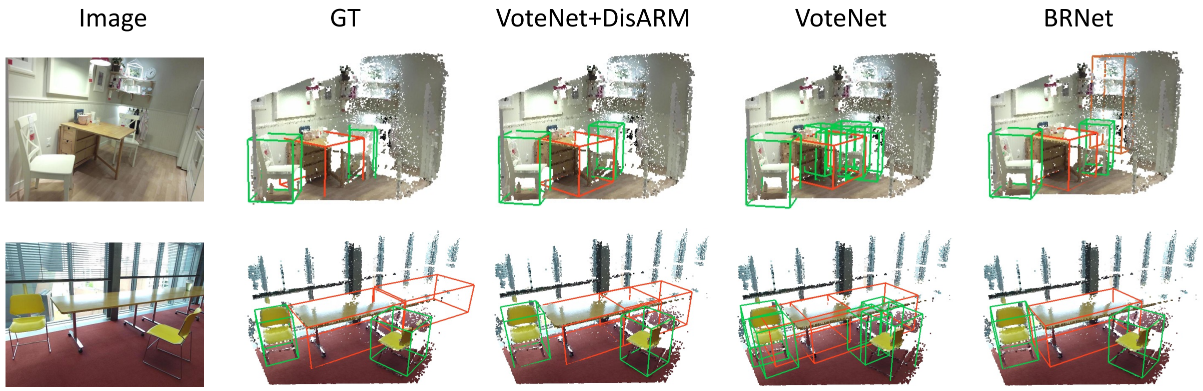


Figure 7. Qualitative results on SUN RGB-D dataset. We denote VoteNet+DisARM as applying our method to VoteNet. The first column is ground truth and the rest columns are detections of different methods. Best viewed on screen.

Settings	mAP@0.25	mAP@0.5
① Global	63.3	47.7
② Local	64.3	48.2
③ Random	65.0	48.7
④ D-FPS	65.1	48.7
⑤ F-FPS	65.3	**49.7**
⑥ D-FPS+F-FPS	65.0	48.8
⑦ F-FPS+D-FPS	65.3	48.4
⑧ K-means	65.2	48.2
⑨ K-means+D-FPS	65.0	48.7
⑩ K-means+F-FPS	64.4	48.3
⑪ Ours	**66.1**	**49.7**

Table 2. Ablation studies of sampling relation anchors strategies. Note that the experiments ① to ③ indicate selecting relation anchors by taking all proposals (Global), nearest 15 proposals (Local) and random 15 proposals (Random) as anchors. Experiments ④ to ⑦ use different combinations of FPS on distances(D-FPS) and features(F-FPS). Experiments ⑧ to ⑩ sample anchors on clusters generated by K-means. Experiment ⑪ conducts F-FPS on the anchors filtered by objectness score (OS).

DCFF	window	desk	showr	toil	sink	mAP$_{50}$
w/o	22.7	44.4	36.8	86.4	37.5	47.1
S-DW	26.2	46.3	31.4	89.7	37.3	47.9
F-DW	20.2	48.6	45.7	89.2	36.7	48.9
Ours	**27.5**	**55.1**	**49.8**	**91.4**	**44.5**	**49.7**

Table 3. DisARM with different components in displacement based context feature fusion (DCFF). The first row indicates fusing features of proposals and anchors without weighting. We denote the S-DW, F-DW as learning weights by spatial displacement and feature displacement respectively.

Method	Model size	time	GFLOPs	mAP@0.5
VoteNet*	11.6MB	0.095s	5.781	44.2
BRNet	13.2MB	0.132s	7.97	50.9
GroupFree3D*	113.0MB	0.170s	31.05	**52.6**
VoteNet*+DisARM	+1MB	+0.001s	+0.034	49.7
BRNet+DisARM	+1MB	+0.008s	+0.034	52.3

Table 4. Comparison of efficiency for different methods. * denotes the model implemented in MMDetection3D [5]. GroupFree3D [19] reported here is configured with best-performance setting.

imVoteNet [22] achieves **65.3** on mAP@0.25, which outperforms all previous state-of-the-arts. More quantitative results on ScanNet V2 and SUN RGB-D datasets can be found in Supplementary.

Qualitative results. In Figure 6 and Figure 7, we visualize the representative 3D object detection results from our method and the baseline methods. These results demonstrate that applying our method to baseline detector achieves more reliable detection results with more accurate bounding boxes and orientations. Our method also eliminates false positives and discovers more missing objects compared with the baseline methods. For example, the results in the second row of Figure 7 show that there are two tables in the scene, and the left one is complete while the right one is missing partially. Our method VoteNet+DisARM can basically detect two tables (red boxes), while BRNet misses the challenging one. This proves that our method can provide rich and effective context to boost the performance of 3D object detectors. More qualitative visualizations are shown in Supplementary.

4.2. Ablation Study

We conduct extensive ablation experiments to analyze the effectiveness of different components of DisARM. All experiments are trained and evaluated on the ScanNet V2 dataset and take VoteNet [23] as backbone method. The

Method	mAP$_{25}$	mAP$_{50}$
HGNet [1]	61.3	34.4
GSPN [39]	62.8	34.8
Pointformer+ [21]	64.1	-
3D-MPA [8]	64.2	49.2
MLCVNet [34]	64.7	42.1
VoteNet [23]	58.6	33.5
VoteNet*+DisARM	66.1 ↑	49.7 ↑
BRNet [4]	66.1	50.9
BRNet+DisARM	66.7 ↑	52.3 ↑
H3DNet* [43]	66.4	48.0
H3DNet*+DisARM	66.8 ↑	48.8 ↑
GroupFree3D*(L6, O256) [19]	66.3	47.8
GroupFree3D*(L12, O256) [19]	66.6	48.2
GroupFree3D*(w2×, L12, O512) [19]	68.2	52.6
GroupFree3D*(L6, O256)+DisARM	67.0 ↑	50.7 ↑
GroupFree3D*(L12, O256)+DisARM	67.2 ↑	52.5 ↑
GroupFree3D*(w2×, L12, O512)+DisARM	**69.3** ↑	**53.6** ↑

Table 5. 3D object detection results on ScanNet V2 dataset. **Notations:** We report the detection performance using mean Average Precision (mAP) at IoU thresholds of 0.25 and 0.5, denoted as mAP$_{25}$ and mAP$_{50}$. Pointformer+ indicates the VoteNet equipped with Pointformer and * denotes that the model is implemented on MMDetection3D. We denote VoteNet*+DisARM, BRNet+DisARM and GroupFree3D*+DisARM as enhanced versions with our method respectively, ↑ indicates the performance is improved with the equipment of DisARM.

network is implemented in MMDetection3D [5].

Strategies of sampling relation anchors. As shown in Table 2, applying DisARM to VoteNet using our sample strategy achieves the highest performance. Experiment ① and experiment ② shows that both global and local context can not provide effective information which introduce redundant information or limited information. We also find that conducting FPS on proposal features can keep the diversity of anchors which can provide more useful context through experiments ④⑤.

Clustering by K-means is a common way to aggregate information. Thus we try to conduct D-FPS and F-FPS on clusters generated by K-means as shown in experiments ⑧⑨⑩. Those strategies can not perform best on mAP@0.5 since the aggregated context of clusters loses the key information of objects for accurate detection. We argue that complete objects are more representative and informative, and experiments ⑤⑪ prove our argument.

Effects of displacement based context feature fusion. We evaluate the contribution of displacement weights in DisARM on ScanNet V2 dataset. The quantitative results are shown in the Table 3. It is clear that the proposed displacement weights are useful and can distribute accurate weights for context from different relation anchors, providing more helpful and robust context for better performance. We find that displacement weights are sensitive to the objects usually placed in special space or scenes, such as window, shower curtain, toil and sink. The large improved performance on mAP@0.5 also indicates the effectiveness of our displacement weights design. Therefore, DisARM can help the backbones to detect these hard ones more accurately.

Model size, speed and computational complexity. The comparison of efficiency is shown in Table 4. For a fair comparison, all experiments are running on the same workstation (a single Titan V GPU) and implemented with MMDetection3D. It is obvious that our proposed method is effective with increasing very few training parameters to backbone methods. The model size of BRNet equipped with DisARM is 10× smaller than that of GroupFree3D only with little performance dropping. Note that DisARM's computational complexity is 1000× faster than that of GroupFree3D. All the numbers demonstrate our lightweight model provides significant performance boosts over the backbone methods for 3D object detection.

5. Conclusion

In this paper, we present a simple, lightweight yet effective method for enhancing the performance of 3D object detection. Unlike previous methods detect objects individually or use context information inefficiently, our method samples representative relation anchors and captures the relation information with the contribution of each relation anchor weighted by the spatial-aware and feature-aware displacements. The proposed method achieves state-of-the-art performance on ScanNet V2 with both metrics and SUN RGB-D in terms of mAP@0.25.

Limitation Our approach is designed for the indoor scenes with some specific organization patterns, and it is not suitable for outdoor scenes with irregular displacement. However, we will explore more relation information for all kinds of scenes in the future.

6. Acknowledgements

This paper is supported in part by the Nation Key Research and Development Program of China (2018AAA0102200), National Nature Science Foundation of China (62132021, 62102435, 62002375, 62002376), NSF of Hunan Province of China (2021RC3071, 2021JJ40696) and NUDT Research Grants (ZK19-30).

References

[1] Jintai Chen, Biwen Lei, Qingyu Song, Haochao Ying, Danny Z Chen, and Jian Wu. A hierarchical graph network for 3d object detection on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 392–401, 2020. 2, 3, 6, 8

[2] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 2

[3] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 3

[4] Bowen Cheng, Lu Sheng, Shaoshuai Shi, Ming Yang, and Dong Xu. Back-tracing representative points for voting-based 3d object detection in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8963–8972, 2021. 2, 4, 6, 8

[5] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 2, 6, 7, 8

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6

[7] Yueqi Duan, Yu Zheng, Jiwen Lu, Jie Zhou, and Qi Tian. Structural relational reasoning of point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 949–958, 2019. 3

[8] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9031–9040, 2020. 8

[9] Francis Engelmann, Theodora Kontogianni, Alexander Hermans, and Bastian Leibe. Exploring spatial context for 3d semantic segmentation of point clouds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 716–724, 2017. 2, 3

[10] Mingtao Feng, Syed Zulqarnain Gilani, Yaonan Wang, Liang Zhang, and Ajmal Mian. Relation graph network for 3d object detection in point clouds. *IEEE Transactions on Image Processing*, 30:92–107, 2020. 2

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[12] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019. 2

[13] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3588–3597, 2018. 2, 3

[14] Shi-Sheng Huang, Hongbo Fu, and Shi-Min Hu. Structure guided interior scene synthesis via graph matching. *Graphical Models*, 85:46–55, 2016. 2

[15] Byung-soo Kim, Shili Xu, and Silvio Savarese. Accurate localization of 3d objects from rgb-d data using segmentation hypotheses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3182–3189, 2013. 2

[16] Jean Lahoud and Bernard Ghanem. 2d-driven 3d object detection in rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 4622–4630, 2017. 2

[17] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 1, 2

[18] Ligang Liu, Xi Xia, Han Sun, Qi Shen, Juzhan Xu, Bin Chen, Hui Huang, and Kai Xu. Object-aware guidance for autonomous scene reconstruction. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2

[19] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *arXiv preprint arXiv:2104.00678*, 2021. 2, 4, 6, 7, 8

[20] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[21] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7463–7472, 2021. 8

[22] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020. 4, 6, 7

[23] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 3, 4, 5, 6, 7, 8

[24] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 2

[25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1, 2

[26] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on com-*

puter vision and pattern recognition, pages 770–779, 2019. 2

[27] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[28] Yifei Shi, Angel X. Chang, Zhelun Wu, Manolis Savva, and Kai Xu. Hierarchy denoising recursive autoencoders for 3d scene layout prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[29] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 6

[30] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *European conference on computer vision*, pages 634–651. Springer, 2014. 2

[31] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 808–816, 2016. 2

[32] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 2

[33] Xiaogang Wang, Xun Sun, Xinyu Cao, Kai Xu, and Bin Zhou. Learning fine-grained segmentation of 3d shapes without part labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10276–10285, 2021. 2

[34] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10447–10456, 2020. 2, 3, 6, 8

[35] Kai Xu, Hui Huang, Yifei Shi, Hao Li, Pinxin Long, Jianong Caichen, Wei Sun, and Baoquan Chen. Autoscanning for coupled scene reconstruction and proactive object analysis. *ACM Transactions on Graphics (TOG)*, 34(6):1–14, 2015. 2

[36] Kai Xu, Rui Ma, Hao Zhang, Chenyang Zhu, Ariel Shamir, Daniel Cohen-Or, and Hui Huang. Organizing heterogeneous scene collection through contextual focal points. *ACM Transactions on Graphics, (Proc. of SIGGRAPH 2014)*, 33(4):35:1–35:12, 2014. 4

[37] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2

[38] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 403–417, 2018. 2

[39] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. 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*, pages 3947–3956, 2019. 2, 8

[40] Ruichi Yu, Xi Chen, Vlad I Morariu, and Larry S Davis. The role of context selection in object detection. *arXiv preprint arXiv:1609.02948*, 2016. 2

[41] Yinda Zhang, Mingru Bai, Pushmeet Kohli, Shahram Izadi, and Jianxiong Xiao. Deepcontext: Context-encoding neural pathways for 3d holistic scene understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 1192–1201, 2017. 2, 3

[42] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European conference on computer vision*, pages 668–686. Springer, 2014. 2, 3

[43] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 2, 6, 8

[44] Yawei Zhao, Kai Xu, En Zhu, Xinwang Liu, Xinzhong Zhu, and Jianping Yin. Triangle lasso for simultaneous clustering and optimization in graph datasets. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1610–1623, 2018. 2

[45] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 1, 2