

3DRM: Pair-wise Relation Module for 3D Object Detection

Yuqing Lan, Yao Duan, Yifei Shi
National University of Defense Technology

lanyuqingkd, duanyao16@nudt.edu.cn, yifei.j.shi@gmail.com

Hui Huang
Shenzhen University
hhzhiyan@gmail.com

Kai Xu
National University of Defense Technology
kevin.kai.xu@gmail.com

Abstract

Context has proven to be one of the most important factors in object layout reasoning for 3D scene understanding. Existing deep contextual models either learn holistic features for context encoding or rely on pre-defined scene templates for context modeling. We argue that scene understanding benefits from object relation reasoning, which is capable of mitigating the ambiguity of 3D object detections and thus helps locate and classify the 3D objects more accurately and robustly. To achieve this, we propose a novel 3D relation module (3DRM) which reasons about object relations at pair-wise levels. The 3DRM predicts the semantic and spatial relationships between objects and extracts the object-wise relation features. We demonstrate the effects of 3DRM by plugging it into proposal-based and voting-based 3D object detection pipelines, respectively. Extensive evaluations show the effectiveness and generalization of 3DRM on 3D object detection. Our source code is available at <https://github.com/lanlan96/3DRM>.

1. Introduction

3D scene understanding involves the detection of 3D objects and the inference of their spatial layouts. It is one of the most fundamental problems in graphics, vision and robotics. Recently, the fast development of 3D data acquisition and reconstruction techniques has made the collection of large-scale 3D real-world scene data more accessible than ever. Nowadays, the reconstructed real-world 3D scene datasets (e.g. S3DIS [2] and ScanNet [9]) usually contain a lot of various objects distributing in multiple areas or rooms. This makes 3D object detection quite challenging.

Context in 3D scenes refers to the spatial or semantic relations between different objects, which is critical to scene understanding (see Figure 1). It has proven to be extremely

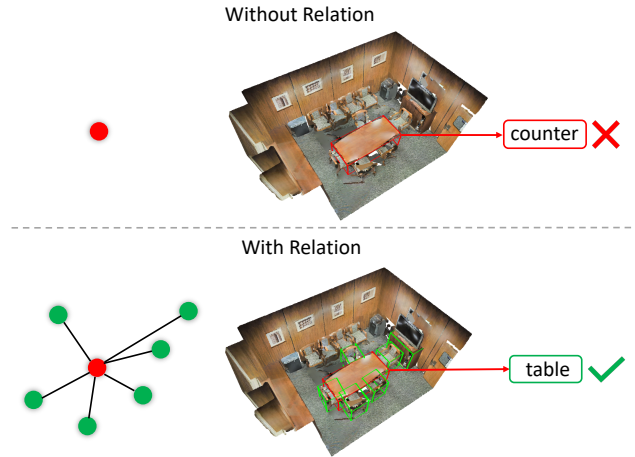


Figure 1. We propose 3DRM which reasons about object relations in 3D object detection. For example, a single object can usually not be correctly identified without knowing the context around it. The proposed 3DRM is able to boost 3D object detection by reasoning the relations between the surrounding objects.

useful in 3D object detection [41, 45, 58]. In the era of deep learning, contextual modeling continues to play an important role in scene analysis [12, 35, 44, 57]. Existing context-based deep learning approaches either extract a holistic feature encoding contextual information via 3D or graph convolutional networks [15, 28, 40] or learn contextual information from pre-defined templates of object layout [59]. However, these approaches require large amount of training data with complete scene geometry or object layout, which limits their flexibility.

Qi et al. [39] propose PointNet for learning 3D representations directly from point cloud data to perform classification and segmentation, yielding many follow-up works. These 3D geometric features have so far been the mainstream in scene understanding. However, a powerful

method calls for more diverse features, inspired by the success of multi-modal object detection in 2D images [14]. In fact, objects in a particular scene are functionally related or have correlation in structure. Such inherent relations can supply a new type of high-level 3D features which may fill in the gap in 3D object detection.

In this paper, we propose to model object context through reasoning about their relations. The proposed 3D Relation Module, or *3DRM* for short, operates directly on features of 3D point cloud and outputs the relational features which can be used to boost the performance of various object detection frameworks for 3D scenes (see Figure 2). The core of our method is a pair-wise relation reasoning module which is not only capable of predicting relational attributes of object pairs, but also mitigates the ambiguity of 3D objects that are hard to detect. Different from previous works, our method does not rely on pre-defined scene templates for contextual features extraction.

3DRM adapts the Relation Network [42] to reasoning about relations between object pairs in 3D representations. Objects in indoor scenes are typically semantically and spatially related. Given integrated features extracted by different backbones with scene point cloud as input, 3DRM performs pair-wise object relation reasoning with a relation module. Objects in the same scene are paired using specific matching strategies. Pair-wise object features are then processed by the proposed 3DRM and prediction of relations is performed with extracted relation features which will be leveraged to help the task of detection.

3DRM is a plug-and-play module which can be applied to different 3D detection frameworks to detect 3D objects more accurately and robustly. We apply 3DRM to two 3D object detection backbones, and evaluate its performance on three challenging datasets. Extensive experiments demonstrate the effectiveness of 3DRM. Specifically, applying 3DRM to different detection backbones achieves **30%** improvement on S3DIS [2], **3.8%** on ScanNetV2 [9] and **1.4%** on SUN RGB-D dataset [48].

In summary, we make the following contributions:

- We propose a 3D relation module which reasons about the relations between 3D objects. Different from other methods which only extract geometry or location features for individual objects, our method is able to capture relation features. This diversifies the feature palette of 3D point cloud and can be combined with other features to boost the performance of object detection.
- We design four dedicated relationships of semantic and spatial properties between objects which can be computed in real time instead of manual annotation.
- Extensive experiments demonstrate the benefits of relation information. We plug our relation module into

two popular detection backbones. The results show substantial improvements on the S3DIS, ScanNetV2 and SUN RGB-D datasets which demonstrates that our design is effective and can be widely applicable.

2. Related Work

3D object detection. 3D object detection in point cloud is now common in indoor scene understanding [5, 15, 16, 19, 28, 30, 36, 39, 40, 44, 52, 53, 61, 62] and autonomous driving [7, 24, 29, 38, 43, 55].

Yang *et al.* [56] directly predict object bounding boxes from a learned global feature vector and obtain instance masks by segmentation points inside a bounding box. VoteNet [37] highlights the challenge of directly predicting bounding box centers in sparse 3D data as most surface points are far away from object centers. Shi *et al.* [44] also generate the objects proposals by graph cuts for an over-segmentation of the point cloud based on point normal differences to create the initial set of segments and leverages the features from PointCNN [28] to explore the hierarchy structure of objects and context. 3D-MPA [11] adapts the object-center approach, extends it with a branch for instance mask prediction and replaces NMS with a grouping mechanism of jointly-learned proposal features. However, all these methods take PointNet, PointNet++ or PointCNN as their backbone to extract geometry features which is insufficient. Relationships between objects provide abundant information for scene understanding which is usually ignored. Huang *et al.* [18] also emphasize the importance of context relations among objects for 3D box estimation.

Relation reasoning in 3D. Since the Relation Network [42] has been proposed, there has been an explosion of methods that apply the Relation Network [42] in various tasks on 2D image, such as object detection [13, 17, 34, 54, 64], semantic segmentation [27], object recognition [6, 60], action recognition [8, 22, 46], object relationship detection [23, 31, 33], VQA [1, 4, 26, 42], few-shot learning [49], scene graph generation [51] etc. [17] uses a relation module to reason object relations and improves the recognition accuracy. [34] applies relation modules on features extracted from VGG-16 for semantic segmentation in Aerial Scenes. All these work demonstrate the importance of relation reasoning in visual tasks.

As the result of the great success of relation reasoning in the 2D domain, some work has already explored the relationships in 3D data. [10] equips the PointNet++ [40] with relation network to reason about the structural dependencies of local regions in 3D point clouds and get a big boost on the tasks of 3D point cloud classification and part segmentation. Liu *et al.* [32] propose a convolution operator which encodes geometric relations of points by reasoning about the spatial layout of points for point cloud analysis. [56]

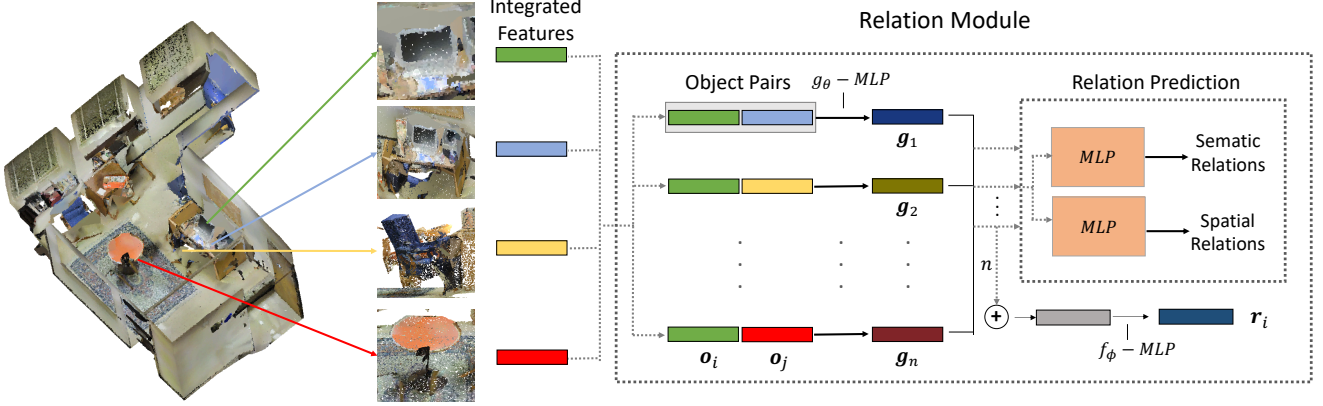


Figure 2. The network architecture of 3DRM. Input of Relation Module comes from features of object candidates extracted by different backbones. Integrated features of different objects are matched by pairs and go through MLP called g_θ . Pair-wise Features corresponding to the same object like the green one will be added together and go through another MLP called f_ϕ to obtain relation feature r_i . Features extracted by g_θ are fed into different MLPs to reason different relations of object pairs. In summary, the Relation Module outputs predictions of semantic and spatial relations as well as relation features r_i .

improves the performance on both 3D object recognition and retrieval tasks, which reinforces the information for individual view by modeling the relationships between its inside regions and the corresponding regions in other views, and then integrates the information from multiple views by modeling the inter-relationships together. [25] reasons about the relative pose between each pair of objects to improve 3D pose prediction. [21, 63] use relation graphs or specific relations like *support* to perform relation reasoning towards different components of an object. [47] leverages case-based reasoning to measure similarity between different furniture layouts. [20] designs five types of relations, which however are dependent on handcrafted labeling as well as time-consuming for relations like *facing*, to build structure graphs of furniture in different scenes respectively, and performs scene matching for novel scene synthesis.

However, there are few work reasoning about the relationships between 3D objects pairs in the indoor scenes by automatic computation and taking advantage of the relationships to capture the relation feature for improving the 3D object detection performance.

3. Method

3.1. Overview

Traditional networks for 3D object detection mainly leverage geometric features of objects to regress the bounding box and conduct classification. Different from that, our Relation Module is aimed for learning the pair-wise objects relationships to extract relation features, which fills a gap in features of 3D data representations. The goal of this paper is to apply the proposed 3DRM to the existing popular detection pipelines with point clouds as input and improve the

final performance.

In our method, we leverage two detection frameworks to generate object candidates and extract their features: proposal-based methods and voting-based methods. With features of objects extracted by the backbones as input, our 3DRM can build up pair-wise object relations and extract the comprehensive relation features. As a result, the relation feature will be concatenated with input feature to help the task of detection (Section 3.2). Strategies about application of our Relation Module on different backbones are demonstrated in Section 3.3. Designs for loss function about different backbones are illustrated in Section 3.4.

3.2. Relation module

Different from existing work on 3D object detection which extract contextual information by taking the entire scene as input, our method learns the object-level relational context features and infers attributes between object pairs. We argue that the relation between object pairs is beneficial to object reasoning for 3D scene understanding. Motivated by the Relation Networks proposed in [17], we adapt the relation module to 3D object detection task. Unlike the strategy that applies relation prediction on the cells of feature map, we perform the relation prediction on individual objects, so object relations are explicitly obtained. On the other hand, our goal of relation reasoning is not to aggregate the global context feature for predicting the attributes of the entire scene. Since we aim to detect individual objects, our relation reasoning module is essentially learning the relation-related feature for individual objects.

The architecture of 3DRM is shown in Figure 2. The input of our relation module is the object candidates with their features $\mathbf{o}_i \in \mathbb{R}^d$ generated by backbone methods. We

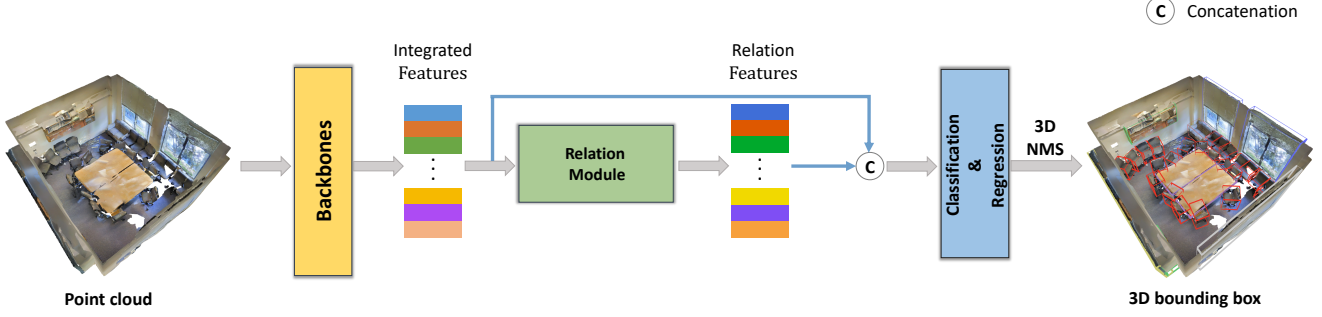


Figure 3. The 3D detection pipeline utilizing our 3DRM. Input point clouds go through the backbone feature extraction networks to gain integrated features which are then sent into Relation Module to get the relation features. Both of integrated features and relation features are used to perform classification and regression. 3D non-maximum suppression (NMS) is followed to output the final 3D bounding box.

propose to use a deep network to extract relational features and predict the relations of a pair of objects. Specifically, for each object \mathbf{o}_i , we randomly choose k objects \mathbf{o}_j in the same scene, which then compose several pairs. In Eq. (1), the pairwise function g_θ aims to exploit the semantic or spatial relations between \mathbf{o}_i and \mathbf{o}_j , and then f_ϕ fuses the relations followed by an element-wise sum for all \mathbf{o}_j . As a result, \mathbf{r}_i is the learned relational feature of \mathbf{o}_i .

$$\mathbf{r}_i = f_\phi\left(\sum_{\forall j} g_\theta(\mathbf{o}_i, \mathbf{o}_j)\right), j \in \{1, \dots, k\} \quad (1)$$

where both i and j are the indexes of the objects in the same scene; g_θ and f_ϕ are the functions implemented by MLPs.

At the same time, we also predict the relationships of each pairs of objects. The output of the pairwise function g_θ will be sent to the classification MLPs to predict the relation label l_{rn} . Note that h_φ is the function implemented by MLPs.

$$l_{rn} = h_\varphi(g_\theta(\mathbf{o}_i, \mathbf{o}_j)), j \in \{1, \dots, k\} \quad (2)$$

We design different classifiers for different relations. There are four types of relations which include semantic and spatial information: *group*, *same as*, *support* and *hang on*. Relations of different components within an object are omitted since we argue that pair-wise object relations are more significant to indoor object detection. The reason why we choose these four types of relations is that these relations are typical in indoor scenes and sufficient for attaining useful relation features for detection intuitively. For the sake of efficiency and practicability, relations like *facing* or *contain* are not under consideration in this paper because indoor 3D object detection is usually aimed for representative objects like chairs and tables instead of bottles of wine in the cabinet. In what follows, we describe how to compute the relation labels of each pair.

Semantic relations. Generally, there are many objects belonging to the same category in the same scene. For ex-

Algorithm 1: Pseudo code for spatial relation formulation.

```

for all object pairs  $(p_i, p_j)$  do
  compute axis-distance  $\Psi_x, \Psi_y, \Psi_z$ .
  compute plane-wise IoU  $\Omega_{xy}, \Omega_{xz}, \Omega_{yz}$ .
   $label_{rn} \leftarrow 0$ 
  if  $\Psi_z \leq \tau_z$  and  $\Omega_{xy} > \tau_{xy}$  then
    | relation  $\leftarrow support$ ,  $label_{rn} \leftarrow 1$ 
  end
  else if  $\Psi_y \leq \tau_y, \Omega_{xz} > \tau_{xz}$  or
     $\Psi_x \leq \tau_x, \Omega_{yz} > \tau_{yz}$  then
    | relation  $\leftarrow hang\ on$ ,  $label_{rn} \leftarrow 1$ 
  end
end

```

ample, in most cases, couples of chairs often simultaneously exist in a conference room. We argue that semantic information is beneficial to detection tasks. In this paper, semantic information covers two relations: *group* and *same as*.

Group relations learn the potential connection between objects that have the same categorical class label and learn the diversity of the different types of objects. We try to capture the relations between objects in terms of the semantic class-specific properties. Distinguishing semantic class relations from various objects benefits the classification, which is equal to answer the question that whether two objects have the same categorical label or not.

Same as relations indicate that a pair of objects may belong to the same instance even if they cover different parts of the object. This type of relation gets the candidate objects belonging to same instance closer while keeps other objects away. Actually this is an exploration of the instance’s intrinsic property.

Spatial relations. Objects in the same scene are potentially connected in the field of space, especially for those with 3D representations. Spatial information indeed implies

abundant and useful information, which is helpful for better understanding of the scene with our relation network. We divide spatial relations into two canonical relations: *support* and *hang on*. The algorithm for spatial relation computation is illustrated in Algorithm 1.

Support relations describe the spatially adjacent relations between the objects. In other words, it emphasizes that objects are functionally close and connected with this relation. Automatically extracting relations of *support* is quite challenging due to the noisy and partial occlusion of real 3D scans. We define that two objects have relations of *support* only when they are close enough on the z-axis and IoU between their projection in the horizontal plane is large enough. Furthermore, if object A is on top of object B and ground projection of object A against the one of object B is higher than a certain threshold, we argue that object A and object B have *support* relations. For example, a flower vase standing on the table describes the relation of *support*.

Specifically, when deciding the *support* relations of two proposals, we first check if 1) their relative height $\Psi_z(p_i, p_j)$ between the lower surface of one object and the upper surface of the other object is smaller than the threshold τ_z , and 2) the overlapping ratio in xy-plane $\Omega_{xy}(p_i, p_j)$ surpasses the threshold τ_{xy} . If so, they will have relations of *support*.

$$\psi_z(p_i, p_j) = |\nu_{max}^z(p_i) - \nu_{min}^z(p_j)| \quad (3)$$

$$\Psi_z(p_i, p_j) = \min(\psi_z(p_i, p_j), \psi_z(p_j, p_i)) \quad (4)$$

where p_i and p_j are two objects, $\nu^z(p_i)$ and $\nu^z(p_j)$ denote their position on z-axis for points of them. $\Psi_z(p_i, p_j)$ is their minimum distance on the z-axis.

$$\Omega_{xy}(p_i, p_j) = \max\left(\frac{\delta_{xy}(p_i, p_j)}{\beta_{xy}(p_i)}, \frac{\delta_{xy}(p_i, p_j)}{\beta_{xy}(p_j)}\right) \quad (5)$$

where $\delta_{xy}(\cdot)$ computes the intersection area of projection for two objects. $\beta_{xy}(\cdot)$ denotes the size of projection area. $\Omega_{xy}(p_i, p_j)$ indicates the larger overlapping ratio of the IoU towards these two projection areas.

Hang on relations imply that two proposals are horizontally adjacent and one hangs on the other. Similar to relations of *support*, if object A is horizontally close to object B and perpendicular projection (parallel to xz-plane or yz-plane) of object A against the object B is higher than a certain threshold, we argue that object A and object B have *hang on* relations. For example, one is classified as wall and the other is an object that can hang on the wall, like board, lamp, curtain, etc. The relations of these kinds of object pairs can be regulated as *hang on*. Such compact spatial relations are helpful for detection and understanding of the scene.

3.3. Application of 3DRM

In order to apply our 3DRM to existing popular detection pipelines and verify its effectiveness and generalization, we design specific strategies to plug our 3DRM into two mainstream methods: proposal-based method and voting-based method. The detection pipeline including our Relation Module is shown in Figure 3. It is composed of two main parts, backbones for processing raw point cloud, and Relation Module for reasoning pair-wise object relations. Both the integrated and relation features are concatenated together to perform the classification and regression towards numerous candidate bounding boxes. Followed by 3D non-maximum suppression (NMS), the pipeline outputs classified and qualified 3D bounding box. We introduce how we apply 3DRM to different 3D detection frameworks as following.

Proposal-based methods. Lots of detection methods establish a baseline system by introducing region proposals as object candidates and classifying the objects as well as regressing the bounding box which we called proposal-based methods. These methods have achieved promising results but ignore the relation information between object candidates, so we aim to equip these methods with our 3DRM to improve the performance. Since there is no widely used framework for proposal-based methods, we design the whole backbone by ourselves. We choose over-segmentation method described in [44] for raw proposal generation, object hypothesis generation module for filtering low quality proposals, PointCNN [28] as feature extractor for point clouds and contextual features for enriching features of objects. Above all is the baseline for proposal-based methods in this paper.

With 3D point clouds as input, we first perform an over-segmentation on the input point cloud as described in [44]. The over-segmented patches are then merged recursively by a bottom-up fashion. The output is a binary hierarchy in which each node is a potential object (segment). We take the node segments as the initial proposals. To improve the quality of raw proposals, we first start from an object hypothesis generation module by filtering proposals with low objectness. This is achieved by using a deep neural network based on PointCNN and MLPs, predicting the objectness labels of the proposals.

With selected and reliable proposals, each proposal can be recognized as a candidate object. Moreover, for improving the baseline performance of detection, like many of the previous works on object detection [12, 35], we add the context information by exploring the points around the object candidate with radius R to extract enriched contextual feature.

Both of the original geometry features and enriched features, which extract corresponding information from the object itself and its surroundings, consist of the integrated fea-

ture for each object candidate. After that, we apply our 3DRM to predict the pair-wise object relations and extract relation features.

Then we perform concatenation of multiple features learning from different aspects including geometric features, context features and relation features. After the concatenation, we feed the concatenated features into MLPs to predict the categorical label and regress the bounding of detection box. Finally, a 3D non-maximum suppression (NMS) is used to remove the redundant proposal candidates and obtain the final 3D objects with their bounding boxes.

Voting-based methods. VoteNet [37] is an end-to-end 3D object detection network based on a synergy of deep point set networks and Hough voting. Similar to classical Hough voting, VoteNet generates votes that lie close to objects centers, and then these votes are grouped and aggregated as clusters to generate box proposals. Each cluster can be regarded as an object candidate which can also be utilized to capture the relation information by our 3DRM.

Specifically, with $N \times 3$ point cloud as input, VoteNet first subsamples $M \times (3 + C)$ seed points by a backbone network. Note that C is the extended feature dimensions and M is the sample number. Each seed point then goes through a voting module, predicts the offset to its object center and thus becomes a vote point for potential clusters. Furthermore, all the votes will be grouped into K clusters each with dimension $(3 + C)$.

At this stage, clusters with their features are sent to our 3DRM to extract the enriched C_r -dimensional relation feature vector. Similar to proposal-based methods, we take the concated features $K \times (3 + C + C_r)$ as the input for the following detection modules. In this way, the inference of the final 3D bounding boxes and the object classes will consider the compatibility with the relations, which makes the final prediction more reliable. In the following steps, We keep the same proposal and classify module as VoteNet to generate final 3D bounding boxes. More details will be described in the Section 3.4.

3.4. Loss function

The loss for our 3DRM is simply formulated as \mathcal{L}_{rn} using the binary cross entropy. In this way, it is judged independently whether an object pair should have a certain relation. As for the detection pipeline, different backbones lead to diverse designs for the final loss function.

Proposal-based methods. The network can be trained in an end-to-end manner with a multi-task loss including a semantic classification loss of the object candidate, a 3D bounding box regression loss and a classification loss of relations. We weigh the losses with the parameter $\lambda_1, \lambda_2, \lambda_3$ to make sure they are in similar scales. In our experiments,

we set $\lambda_1 = 1.0, \lambda_2 = 10, \lambda_3 = 0.5$.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{reg} + \lambda_3 \mathcal{L}_{rn} \quad (6)$$

Voting-based methods. The network is trained in an end-to-end manner with a multi-task loss including a voting loss, an objectness loss, a 3D bounding box estimation loss, a semantic classification loss and a relation loss. It is worthy to note that the objectness loss is designed to help the proposal module to generate good enough proposals. The component losses except the voting loss are weighted by $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. In our experiments, we set $\lambda_1 = 0.5, \lambda_2 = 1.0, \lambda_3 = 0.1, \lambda_4 = 0.1$.

$$\mathcal{L}_{total} = \mathcal{L}_{vote} + \lambda_1 \mathcal{L}_{objectness} + \lambda_2 \mathcal{L}_{box} + \lambda_3 \mathcal{L}_{cls} + \lambda_4 \mathcal{L}_{rn} \quad (7)$$

4. Implementation Details

In this section, we describe some implementation details of the network architectures of our 3DRM, the relevant parameters in previous methods and strategies for the application of our 3DRM to different backbones in training and testing.

Details in 3DRM. As mentioned in Section 3.2, there are some differences in Relation Module between [42] and our architecture. We process a pair of objects' features at a time instead of the whole features of all objects. There are thousands of objects and the number of object pairs' permutation is too large to train. In order to obtain a stable relation feature and a faster convergence speed, we sample fixed number $k = 8$ object pairs for each object in the same scene by two ways. One is random sampling, and the other is nearest sampling. Different sampling of object pairs results in slightly different performance (see Section 5.3 and Section 5.4). For relation label computation, we set the axis-wise threshold of distance $\tau_x = \tau_y = \tau_z = 0.1$ and $\tau_{xy} = \tau_{xz} = \tau_{yz} = 0.5$ for IoU threshold of bounding boxes projection onto planes. The function g_θ and f_ϕ in Relation Module are different for two backbones and detailed in the following paragraphs.

Details in proposal-based detection. We pre-train the object hypothesis generation module to get the object candidates. Then, we train the object relation module and detection module for final classification and regression end-to-end. For this backbone, we use 4 fully connected layers for MLP g_θ , and 2 fully connected layers for MLP f_ϕ in Relation Module to extract relation features and four fully convolutional (FC) layers to predict the four types of relations. The context points around object candidates are obtained by KDTree [3] with $R = 0.5m$. Finally, the geometric features, context feature and relation feature of the object are leveraged to perform the final prediction. Furthermore, there are also some differences of using data in training and test. We leverage train data whose $IoU \geq 0.5$

against ground truth. At test time, we use all object candidates filtered by the object hypothesis generation module.

Note that it is hard to train object hypotheses generation module with the unbalanced training data. We use Cross Entropy Loss and two training strategies in our method: one is data balancing and the other is hard negative mining. In data balancing, we randomly choose negative samples with the same number of positive samples to form the training data. In hard negative mining, we keep the same procedure as original paper.

We implement our approach using TensorFlow. The Adam optimizer is leveraged in our experiments with a base learning rate of 0.001. We train the model with the maximum training epoch number as 50 and batchsize as 8 on one NVIDIA TITAN V GPU.

Details in voting-based detection. The Relation Module for VoteNet backbone is slightly different from the module for proposal-based backbone. The inputs for our Relation Module are features of clusters with 128-dim. The g_θ layer is realized through a multi-layer perceptron with FC and the output channel size is 256. The features are further processed by the MLP f_ϕ to get the 128-dim relation features after the channel-wise mean operation. At the same time, the outputs of g_θ are sent to classifiers to predict the relation labels. Note that we combine four types of relations into semantic and spatial relations in voting-based detection. Each classifier predicts one type of relation and is implemented with two FCs, output of which is 128 and 2 respectively. As a result, the combined features of input features and relation features with dimension $128+128$ will be sent to the following modules.

We train the entire network end-to-end and use the same optimizer, batch size, initial learning and learning rate decay steps as VoteNet. It takes around 180 epochs for the model to converge on one NVIDIA TITAN V GPU while training.

5. Experiments

In this section, we evaluate the proposed 3DRM applied on proposal-based methods and voting-based methods respectively, in the field of 3D object detections with point cloud of indoor scenes as input. Experiments are performed on three large 3D indoor scene datasets and evaluated on the detection benchmarks. (Section 5.1). The evaluation metric is described in Section 5.2. We analyze the improved performance after applying our Relation Module on the two mentioned detection pipelines (OSegNet in Section 5.3 and VoteNet in Section 5.4). Note that since we plan to verify the effectiveness and generalization of 3DRM on detection pipelines with low or relatively high performance, we choose to evaluate our 3DRM on OSegNet and VoteNet respectively. Experiments settings including evaluation on detection and ablation studies are the same for these two

pipelines. Further discussion is illustrated in Section 5.5. Both of the quantitative and qualitative results demonstrate the effectiveness and generalization of the proposed Relation Module.

5.1. Dataset and benchmarks

We leverage a widely used dataset that provides 3D point clouds of indoor scenes: Stanford large-scale 3D Indoor Spaces Dataset S3DIS [2] for proposal-based pipeline. S3DIS is from real scans of indoor environments which contains 3D scans from Matterport scanners in 6 areas including 271 rooms. The objects in this dataset are divided into 13 categories. We perform a k-fold cross validation across areas [50].

Both of ScanNetV2 [9] and SUN RGB-D [48] are leveraged to evaluate the voting-based pipeline. ScanNetV2 is an RGB-D video indoor scene dataset with richly annotated 3D reconstructed meshes. It contains about 1.5K scans annotated with both semantic segmentation and object instance labels for 18 categories. Since it doesn't provide reconstructed point clouds and oriented bounding boxes, we sample the reconstructed meshes and predict axis-aligned bounding boxes in the same way as VoteNet.

SUN RGB-D is a large single-view RGB-D dataset for scene understanding. It contains about 10K RGB-D images captured by four different sensors with accurately annotated oriented bounding boxes for 37 object categories. Note that since it doesn't provide point cloud data, we first convert the depth images to point clouds using known camera parameters.

5.2. Evaluation metric

We use average precision as our evaluation metric of the detected object bounding boxes against the ground truth bounding boxes. We use two *IoU* thresholds as 0.5 and 0.25 respectively in our experiments. The mean average precision (mAP) is the macro-average on average precision across all test categories.

5.3. Evaluation on proposal-based framework

In this section, we denote the detection framework proposed in Section 3.3 as baseline named OSegNet which utilizes proposals generated by over-segmentation and PointCNN [28] as backbones. Applying our 3DRM to OSegNet is denoted as **OSegNet+RM**. We first compare our method with OSegNet on 3D object detection. We also compare our method to state-of-the-art methods and analyze the difference and gap. After that, we conduct the ablation studies to evaluate the impact of each component in our approach. Last, we demonstrate the qualitative results of our method.

Comparison to baseline and State-of-the-art methods. We evaluate our method against several prior works

Table 1. Comparison of our approach against prior works and the framework OSegNet on 3D object detection. We denote OSegNet+RM as OSegNet equipped with our 3DRM. Values report average precision at mAP@0.5 on S3DIS dataset evaluated with 6-fold cross-validation on Area1~Area6.

	chair	board	table	sofa	mAP
Sliding PointCNN	0.36	0.07	0.39	0.23	0.26
PointNet	0.34	0.12	0.47	0.05	0.25
SGPN	0.41	0.13	0.50	0.07	0.28
VDREA	0.41	0.14	0.53	0.43	0.39
OSegNet	0.20	0.01	0.11	0.25	0.14
OSegNet+RM	0.74	0.01	0.47	0.51	0.43

and our baseline OSegNet which detects 3D object in indoor scenes with point cloud as input:

- **Sliding PointCNN [28]:** A baseline which contains a PointCNN backbone and detect objects in a 3D sliding window fashion.
- **PointNet [39]:** A method that first predicts the category of all points and then uses a breadth-first search to group nearby points with the same category.
- **SGPN [52]:** A semantic instance segmentation approach for point clouds by using an embedding learning network for point pairs.
- **VDRAE [44]:** A variational auto-encoder that detects 3D objects in indoor scene by using a hierarchical structure.

Table 1 reports the average precision on the S3DIS dataset using 6-fold cross validation across six areas with mAP@0.5. Compared to the baseline OSegNet, our method **OSegNet+RM** obtains **54%**, **36%**, **26%** and **29%** increase on chair, table, sofa and mAP respectively, which proves the efficiency of our Relation Module. Note that, limited to the performance of over-segmentation method for proposal generation, there are very few proposals with good quality for board objects, resulting in low performance on this category. While all methods are learning-based methods and there is a lack of valid proposals on board category, our method still achieves the best performance. Especially, our method get a huge improvement (**33%** increase) compared to the state-of-the-art on chair category and **4%** increase on mAP, thanks to our 3DRM proposed in Section 3.2 and illustrated in Figure 1. Moreover, without relation prediction and using only the relation features, OSegNet+RM still surpasses OSegNet by a large margin, proving the effectiveness of the relation features extracted by our method.

Ablation study. We evaluate the impact of each component of our approach to investigate the efficiency of dif-

Table 2. Comparison of different relations of our 3DRM applied on OSegNet framework on S3DIS dataset. Experiments are trained on Area2~Area6 and tested on Area1. We denote OSegNet+RM as OSegNet equipped with our 3DRM and OSegNet+RM- as OSegNet+RM without relation prediction. Note that board category is eliminated from comparison due to poor quality of proposals.

	chair	table	sofa	mAP
OSegNet	0.21	0.07	0.28	0.19
OSegNet+RM(<i>group</i>)	0.72	0.36	0.60	0.56
OSegNet+RM(<i>same as</i>)	0.69	0.34	0.71	0.58
OSegNet+RM(<i>support</i>)	0.69	0.34	0.72	0.58
OSegNet+RM(<i>hang on</i>)	0.70	0.29	0.69	0.56
OSegNet+RM(<i>all</i>)	0.69	0.39	0.77	0.62
OSegNet+RM-	0.70	0.32	0.67	0.56

Table 3. Comparison of different selection modes of object pairs on S3DIS dataset. Experiments are trained on Area2~Area6 and tested on Area1.

	chair	table	sofa	mAP
OSegNet+RM(random)	0.69	0.39	0.77	0.62
OSegNet+RM(nearest)	0.69	0.37	0.66	0.57

ferent relations and relation features. Note that, all experiments for the ablation studies are trained on Area2~Area6 and tested on Area1. Board category is eliminated from comparison due to poor quality of proposals generated by OSegNet.

Specifically, we first analyze the contribution of different relations to our 3DRM. The ablation results are shown in Table 2. While our method **OSegNet+RM(*all*)** which predicts all relations at one time achieves the best performance on table and sofa with mAP of **0.62** on three categories. OSegNet+RM using only *group* relations achieves the highest AP **0.72** on category of chair. We argue that, on S3DIS dataset, spatial structures of objects like table and sofa are various and complex to distinguish, leading to the dependence of both semantic and spatial relations. Shapes of chairs are relatively fixed, which relies more on semantic relations like *group*.

As for the selection mode of object pairs, we also conduct the ablation study about random mode and nearest mode. Experiments are trained on Area2~Area6 and tested on Area1. Table 3 illustrates that random selection outperforms selecting object pairs according to the nearest euclidean distance on all terms. Specifically, random selection surpasses the nearest selection by **5%** on mAP. The improvement shows that object pairs randomly selected provide more information for the network to learn, while the pairs selected nearestly can be regarded as a kind of local

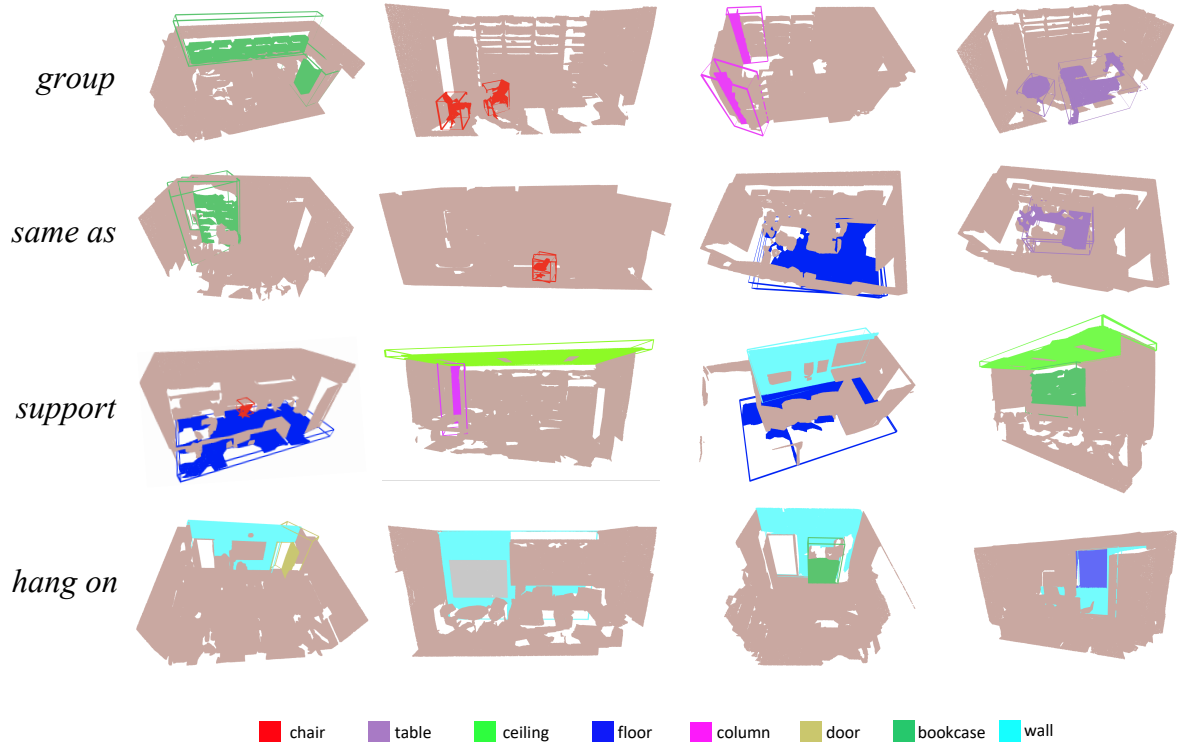


Figure 4. Visualization of the objects with semantic and spatial relationships. The first row shows the two objects with *group* relation. The rest rows are for *same as*, *support* and *hang on* relations respectively.

Table 4. Comparison of our approach against VoteNet on 3D object detection on ScanNetV2 val set and SUN RGB-D val set. We denote VoteNet+RM as our approach with applying our 3DRM on VoteNet.

	mAP@0.25		mAP@0.5	
	ScanNet	SUN RGB-D	ScanNet	SUN RGB-D
VoteNet	58.6	57.7	33.5	33.7
VoteNet+RM	59.7	59.1	37.3	35.1

information.

Visualization of relation prediction. Figure 4 visualizes the relations between objects in the same room. Both of semantic and spatial relations provide rich context information to help detection. Different objects belonging to the same categories implies that they should have similar shapes. If objects belonging to the same instance have overlapping bounding boxes, this helps the network to regress the box correctly. For spatial relations, such explicit information provides extra knowledge to accomplish the task of detection.

Qualitative examples. Figure 5 shows several qualitative results of object detection on S3DIS dataset. The proposed Relation Module leverages multiple features from

relations to help detection module classify the objects in 3D bounding boxes. We visualize the result of OSegNet (first column and fourth column), our method OSegNet+RM (second column and fifth column) and ground truth (third column and last column). It is demonstrated that our method is capable of detecting objects in cluttered scenes and regressed bounding boxes are more accurate and distinct than OSegNet.

5.4. Evaluation on voting-based framework

We first compare our method with the baseline VoteNet on 3D object detection in 3D point clouds. Results justify the effectiveness and practicality of the proposed 3DRM. After that, we conduct extensive ablation studies to evaluate the impact of each component in our approach. Lastly, we demonstrate the qualitative results of our method.

Comparison to VoteNet. We evaluate our method against VoteNet in 3D object detection. Quantitative results on ScanNet and SUN RGB-D are summarized in Table 4. We apply the proposed Relation Module to the representative VoteNet and denote the network as **VoteNet+RM** as our method. Note that we take the performance VoteNet+RM with semantic relations only as the final results since it performs the best on these two datasets. Our method significantly outperforms VoteNet by not only **1.1%** and **3.8%** on

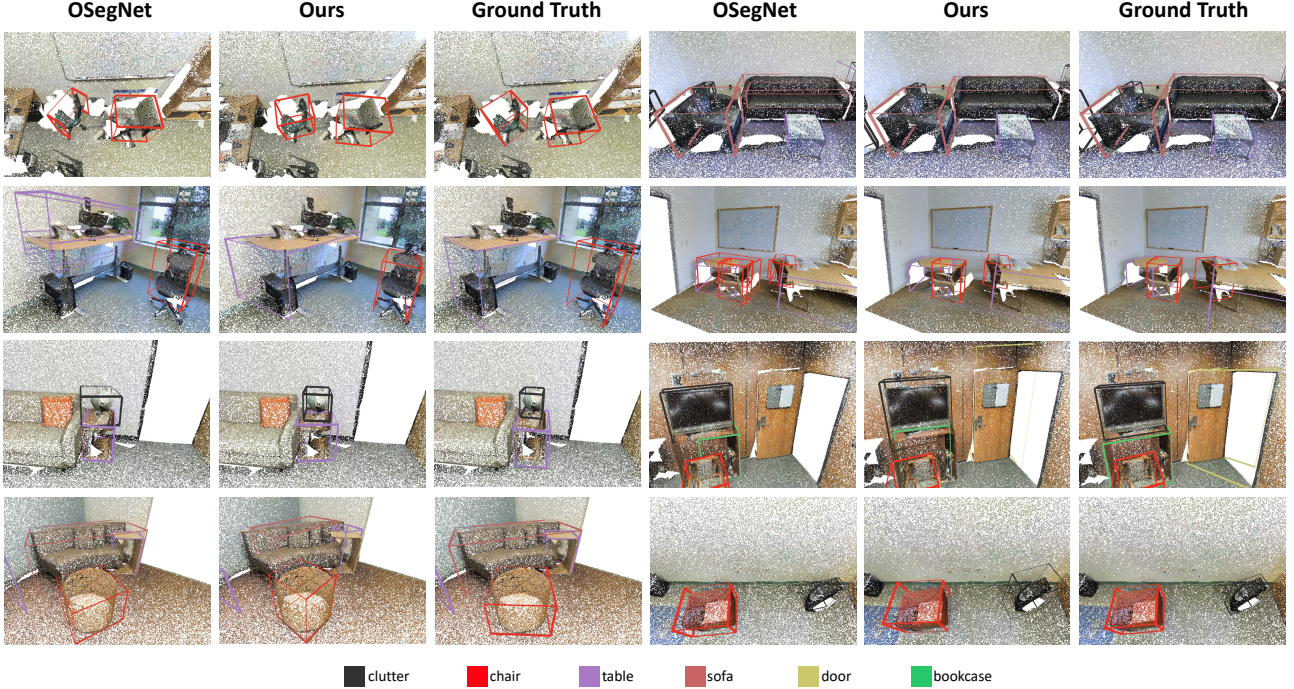


Figure 5. 3D detection using our 3DRM with OSegNet on the S3DIS test set. The first/fifth column shows the bounding boxes for the OSegNet. The second/fourth column shows the qualitative detections with our 3DRM called OSegNet+RM and the third/sixth column shows the ground truth of that. Our method is capable of detecting objects in cluttered scenes and regressed bounding boxes are more accurate and distinct than OSegNet.

Table 5. Comparison to VoteNet with mAP@0.5 on ScanNetV2 val set for our method with different relations. We denote VoteNet+RM as VoteNet equipped with our 3DRM and VoteNet+RM- as VoteNet+RM without relation prediction.

	wind	bed	cntr	sofa	tabl	showr	ofurn	sink	pic	chair	desk	curt	fridge	door	toil	bkskf	bath	cab	mAP
VoteNet	7.89	76.70	20.11	69.04	41.80	7.75	14.05	21.06	0.76	67.30	32.52	10.58	28.89	14.68	82.07	27.86	79.31	9.36	33.99
VoteNet+RM(semantic)	12.29	80.63	14.59	71.79	41.28	10.41	13.35	29.46	0.14	67.73	34.74	16.95	37.79	15.70	89.96	44.22	82.95	8.03	37.33
VoteNet+RM(spatial)	10.48	81.17	22.07	68.95	42.02	4.20	16.56	26.08	1.57	68.83	36.49	13.30	33.02	17.67	84.37	39.66	89.43	10.79	37.04
VoteNet+RM(all)	9.68	77.97	21.72	67.65	41.89	15.11	14.56	26.58	0.22	69.20	40.30	26.30	30.37	13.67	89.89	32.41	78.94	9.63	37.01
VoteNet+RM-	10.33	81.40	18.97	66.57	42.94	9.33	15.45	25.55	0.42	69.13	37.72	18.83	29.75	15.20	87.40	40.66	82.26	6.61	36.58

Table 6. Comparison of different selection modes of object pairs on ScanNetV2 val set. We denote VoteNet+RM as our approach with applying our 3DRM on VoteNet.

	mAP@0.25	mAP@0.5
VoteNet	58.6	33.5
VoteNet+RM(random)	59.73	37.33
VoteNet+RM(nearest)	58.44	36.79

ScanNet, but also **1.4%** and **1.4%** on SUN RGB-D in terms of mAP with IoU=0.25 and IoU=0.5 respectively. Note that, our method increase the performance of VoteNet by 3.8% on mAP@0.5 which illustrates that our 3DRM can not only mitigate ambiguity but also increase accuracy of the detection. Furthermore, we argue that this benefits from

the enriched relation features from our 3DRM, which provides comprehensive understanding to the object and its surrounding environment. More quantitative results are shown in appendix.

Ablation study. Extensive ablation studies are performed to verify the increased accuracy of our approach. Note that we combine *group* and *same as* relations as semantic relations and *support* and *hang on* as spatial relations. To prove the efficiency of the proposed 3DRM, we first study how the semantic and spatial relations help the task of 3D object detection for different categories. Applying our 3DRM to VoteNet without predicting relation labels denoted as VoteNet+RM- is also considered. Results on ScanNet are shown in Table 5.

The results show that VoteNet+RM with only semantic relations achieves the best performance with an increase of

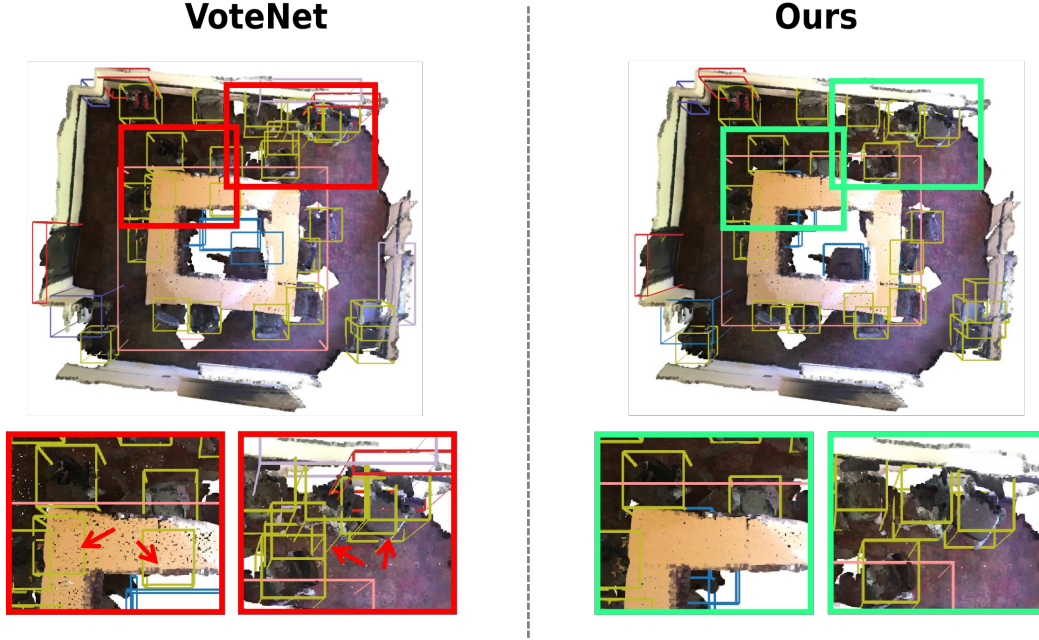


Figure 6. Qualitative comparison results of 3D object detection on ScanNetV2 val set. Left: VoteNet, Right: Ours. The detailed comparison demonstrates that our 3DRM enables more accurate and reasonable detection. Color is for depiction, not used for detection.

3.34% in terms of mAP@0.5. The reason is that objects of most categories are sensitive to semantic relations, and can obtain more context information from semantic relations. For example, objects, such as windows, sofas, fridges, toilets and so on, have simple and clear spatial structure, and thus need various objects and context to help understanding. Objects, such as counters, pictures, doors, baths and cabs, are usually placed in a complex environment where semantic information is rich enough and spatial information are critical to them since their structures are more complicated. Moreover, some categories of objects benefit from both semantic and spatial relations like shower, chair, desk, etc. Note that applying our 3DRM to VoteNet without predicting relation labels and only use the relation features also outperforms VoteNet, which proves that relation features extracted by our 3DRM do help detect 3D objects better.

As for the mode of selecting object pairs, we compare two ways of selection: random mode and nearest mode. Comparison results are illustrated in Table 6. It is clear that random selection of object pairs achieves higher performance than selecting several nearest objects to form relation pairs. This is because random selection can provide various object pairs distributed in the whole scene and thus enrich the information around objects to improve the detection quality.

Qualitative results and discussion. The qualitative results on ScanNet are shown in Figure 6. Ours method detects the objects more accurately and robustly, which is ben-

eficial from our 3DRM. It is noteworthy that detection results of VoteNet are confused with other objects and ambiguous in some areas with noisy point cloud, while ours can classify and locate the objects precisely and clearly without redundant bounding boxes. We argue that this is attributed to the pair-wise relation reasoning. Details are shown in the second row of Figure 6 where red rectangles on the left refer to the ambiguous and wrong detections on chairs by VoteNet. Green rectangles on the right demonstrate the accurate detections by ours.

Figure 7 shows the detection results on ScanNetV2 val set. From the comparison of Ours and VoteNet, we can detect the objects accurately and robustly with less ambiguity. Specifically, in some cluttered areas, ours can distinguish different objects and regress the bounding boxes precisely. There are usually many chairs in scenes like offices and it is quite common to misunderstand the chairs as other categories due to noise and their various appearance. Our 3DRM can help alleviate this problem by using relation reasoning and thus achieve better detection results.

5.5. Further discussion

Complexity and computational efficiency. Our 3DRM is consisted of semantic relations and spatial relations including *group*, *same as*, *support* and *hang on*. Among these, relations like *group*, *same as* are easy and fast to compute since we only need to compare semantic or instance labels. The computational time for semantic relations can be ig-

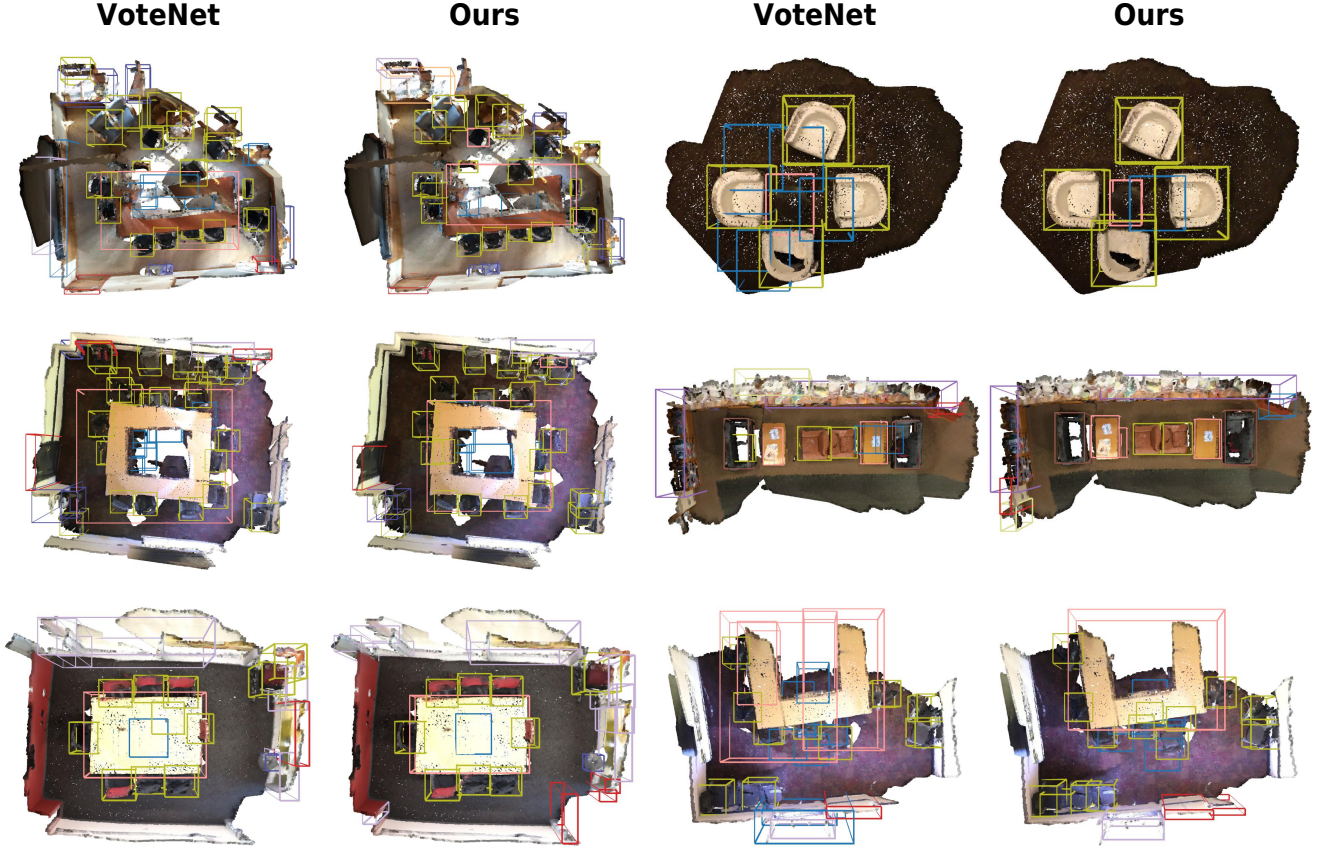


Figure 7. 3D detection using our 3DRM with VoteNet on the ScanNetV2 val set. The first/third column shows the bounding boxes for the VoteNet. The second/fourth column shows the qualitative detections with our 3DRM called VoteNet+RM.

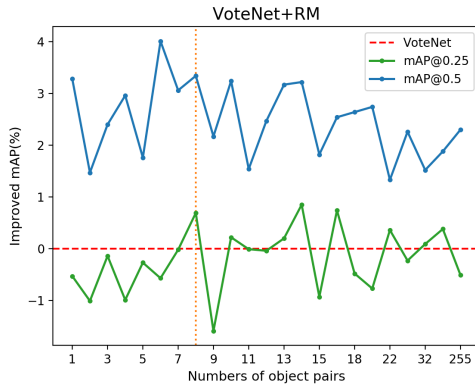


Figure 8. Improved percentage of mAP for different numbers of object pairs. We denote VoteNet+RM as VoteNet equipped with our 3DRM. Sampling 8 object pairs for an object achieves the most improvement taking both mAP@0.25 and mAP@0.5 as well as computational efficiency into consideration.

nored. As for spatial relations, Algorithm 1 can illustrate the complexity. Note that although we use loops for simpli-

cation in the algorithm, we actually use matrix multiplication for further acceleration in experiments. Moreover, the computational time and complexity grows as the number of object pairs is larger. Numerically, we test the computational efficiency of spatial relations for 2048 proposals in ScanNet dataset. We sample 8 object pairs for each proposal, which means $8 * 2048$ object pairs in total. The total time is around $0.047s$ and the average time is approximately $3 \times 10^{-6}s$ for each object pair, which proves the high efficiency of our 3DRM.

With regard to the number of object pairs for each object which may have influence on the complexity of relation computation, we perform experiments by comparing the contributions of different pair numbers to the detection with VoteNet+RM on ScanNet dataset. Results are shown in Figure 8. Explicitly, sampling 8 or 14 object pairs for an object achieves the most improvement considering both mAP@0.25 and mAP@0.5. Since these contributions of these two numbers of pairs are almost the same, we build 8 object pairs for each one for the balance of computational efficiency and performance.

Improvement on different pipelines. We have evalu-

ated the improved performance of detection on mAP for both OSegNet and VoteNet to justify the generalization of our 3DRM on detection pipelines with different basic performance. Explicitly, applying 3DRM to OSegNet obtains **29%** improvement on mAP@0.5 on S3DIS while VoteNet gets **3.8%** on ScanNet and **1.4%** on SUN RGB-D. OSegNet is an intuitively simple baseline implemented mainly with over-segmentation for proposal generation and PointCNN as backbones. Initial proposals generated by over-segmentation in OSegNet are relatively less organized and accurate compared to VoteNet which relies on deep hough voting. The network architecture of OSegNet is much simpler than VoteNet. Less organized proposals and simpler architecture result in lower basic performance for OSegNet. However, this actually demonstrates the generalization and effectiveness of our 3DRM by being able to help attain a huge improvement on detection for simple pipelines, and comparable improvement even for comprehensive pipelines.

6. Conclusions

We presented a Relation Module for 3D object detection on large-scale scene datasets. With the object candidates generated from backbones, we predict object relations and capture relation features by our 3DRM, which is capable of mitigating the ambiguity of 3D object detection, thus helping locate and classify the 3D objects more accurately and robustly. We applied our 3DRM to both proposal-based methods and voting-based methods. Improved detection results demonstrate the effectiveness and generalization of our method.

Limitations. Although experiments verify the effectiveness of our 3DRM, our method can only predict pair-wise object relations explicitly. High-level relations that may help scene understanding are not considered.

Future work. There are several directions worth trying for the future work. First, it is worth trying to add more relation types to the relation module. Second, we only perform relation reasoning on object pairs. To explore the possibility to apply relation networks for analyzing sub-scenes is an interesting direction. Finally, to perform relation reasoning in more complicated 3D task (such as Vision Question Answering on 3D scenes) is also a promising direction.

References

- [1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017. [2](#)
- [2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016. [1](#), [2](#), [7](#)
- [3] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, Sept. 1975. [6](#)
- [4] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019. [2](#)
- [5] 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 (CVPR)*, pages 392–401, June 2020. [2](#)
- [6] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4086–4096, 2017. [2](#)
- [7] 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](#)
- [8] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6519–6527, 2020. [2](#)
- [9] 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. [1](#), [2](#), [7](#)
- [10] Yueqi Duan, Yu Zheng, Jiwen Lu, Jie Zhou, and Qi Tian. Structural relational reasoning of point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 949–958, 2019. [2](#)
- [11] 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. [2](#)
- [12] 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. [1](#), [5](#)
- [13] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020. [2](#)
- [14] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 2020. [2](#)

- [15] 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. 1, 2
- [16] 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 (CVPR)*, pages 4421–4430, June 2019. 2
- [17] 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
- [18] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In *Advances in Neural Information Processing Systems*, pages 207–218, 2018. 2
- [19] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 187–203, September 2018. 2
- [20] Shi-Sheng Huang, Hongbo Fu, and Shi-Min Hu. Structure guided interior scene synthesis via graph matching. *Graphical Models*, 85:46–55, 2016. 3
- [21] Shi-Sheng Huang, Hongbo Fu, Ling-Yu Wei, and Shi-Min Hu. Support substructures: Support-induced part-level structural representation. *IEEE transactions on visualization and computer graphics*, 22(8):2024–2036, 2015. 3
- [22] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14024–14034, 2020. 2
- [23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia Li, and David A. Shamma. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2016. 2
- [24] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018. 2
- [25] Nilesh Kulkarni, Ishan Misra, Shubham Tulsiani, and Abhinav Gupta. 3d-relnet: Joint object and relational network for 3d prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2212–2221, 2019. 3
- [26] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9972–9981, 2020. 2
- [27] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8950–8959, 2020. 2
- [28] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in neural information processing systems*, pages 820–830, 2018. 1, 2, 5, 7, 8
- [29] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. *Deep Continuous Fusion for Multi-sensor 3D Object Detection*. 2018. 2
- [30] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgb-d cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1417–1424, December 2013. 2
- [31] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10840–10849, 2020. 2
- [32] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8895–8904, 2019. 2
- [33] Li Mi and Zhenzhong Chen. Hierarchical graph attention network for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2020. 2
- [34] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12416–12425, 2019. 2
- [35] Jake Porway, Kristy Wang, Benjamin Yao, and Song Chun Zhu. A hierarchical and contextual model for aerial image understanding. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008. 1, 5
- [36] 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 (ICCV)*, pages 9277–9286, October 2019. 2
- [37] 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 International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 6
- [38] 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
- [39] 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, 8
- [40] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30:5099–5108, 2017. 1, 2

- [41] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgb-d semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5199–5208, 2017. 1
- [42] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017. 2, 6
- [43] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019. 2
- [44] 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 Conference on Computer Vision and Pattern Recognition*, pages 1771–1780, 2019. 1, 2, 5, 8
- [45] Yifei Shi, Pinxin Long, Kai Xu, Hui Huang, and Yueshan Xiong. Data-driven contextual modeling for 3d scene understanding. *Computers & Graphics*, 55:55–67, 2016. 1
- [46] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27:568–576, 2014. 2
- [47] Peihua Song, Youyi Zheng, and Jinyuan Jia. Web3d learning platform of furniture layout based on case-based reasoning and distance field. In *International Conference on Technologies for E-Learning and Digital Entertainment*, pages 235–250. Springer, 2017. 3
- [48] 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. 2, 7
- [49] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 2
- [50] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017. 7
- [51] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring context and visual pattern of relationship for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2019. 2
- [52] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2569–2578, 2018. 2, 8
- [53] 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 (CVPR)*, pages 10447–10456, June 2020. 2
- [54] Hang Xu, Chenhan Jiang, Xiaodan Liang, and Zhenguo Li. Spatial-aware graph relation network for large-scale object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9298–9307, 2019. 2
- [55] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. 2
- [56] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. In *Advances in Neural Information Processing Systems*, pages 6737–6746, 2019. 2
- [57] 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. 1
- [58] 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. 1
- [59] Yinda Zhang, Mingru Bai, Pushmeet Kohli, Shahram Izadi, and Jianxiong Xiao. Deepcontext: Context-encoding neural pathways for 3d holistic scene understanding. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1201–1210, 10 2017. 1
- [60] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020. 2
- [61] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 311–329, Cham, 2020. Springer International Publishing. 2
- [62] 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
- [63] Youyi Zheng, Daniel Cohen-Or, Melinos Averkiou, and Niloy J Mitra. Recurring part arrangements in shape collections. In *Computer Graphics Forum*, volume 33, pages 115–124. Wiley Online Library, 2014. 3
- [64] Zhuo Zheng, Yanfei Zhong, Junjie Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2020. 2