Computational Visual Media
https://doi.org/10.1007/s41095-021-0252-6 Vol. 8, No. 3, September 2022, 395–414

Research Article

ARM3D: Attention-based relation module for indoor 3D object detection

Yuqing Lan[1], **Yao Duan**[1], **Chenyi Liu**[1], **Chenyang Zhu**[1] (✉), **Yueshan Xiong**[1], **Hui Huang**[2], **and Kai Xu**[1] (✉)

Abstract Relation contexts have been proved to be useful for many challenging vision tasks. In the field of 3D object detection, previous methods have been taking the advantage of context encoding, graph embedding, or explicit relation reasoning to extract relation contexts. However, there exist inevitably redundant relation contexts due to noisy or low-quality proposals. In fact, invalid relation contexts usually indicate underlying scene misunderstanding and ambiguity, which may, on the contrary, reduce the performance in complex scenes. Inspired by recent attention mechanism like Transformer, we propose a novel 3D attention-based relation module (ARM3D). It encompasses object-aware relation reasoning to extract pair-wise relation contexts among qualified proposals and an attention module to distribute attention weights towards different relation contexts. In this way, ARM3D can take full advantage of the useful relation contexts and filter those less relevant or even confusing contexts, which mitigates the ambiguity in detection. We have evaluated the effectiveness of ARM3D by plugging it into several state-of-the-art 3D object detectors and showing more accurate and robust detection results. Extensive experiments show the capability and generalization of ARM3D on 3D object detection. Our source code is available at `https://github.com/lanlan96/ARM3D`.

Keywords attention mechanism; scene understanding; relational reasoning; 3D indoor object detection

1 Introduction

With the fast development of automatic and unmanned technology, 3D object detection has recently been brought to the fore. Nowadays, 3D object detection still remains challenging and plays an important role in 3D vision, including augmented reality, robot navigation, robot grasping, etc. Most current 3D object detection methods focus on point clouds, which are more readily available than before with the evolution of 3D scanning devices and reconstruction techniques. However, the orderless and unstructured nature of point clouds makes the detection in 3D more challenging than in 2D, as it is difficult to transfer widely used techniques for 2D object detection to 3D.

Recently, interests in point cloud have been on the rise to solve this challenge. With the boom of deep learning, more and more methods have been proposed to directly process 3D point clouds and use the extracted features for all kinds of 3D computer vision or graphics tasks [1–4]. Recent works [5–10] can effectively attain detected 3D objects in raw point clouds of indoor scenes. These methods mainly rely on the geometric features from deep backbones or contextual features from context encoding or relation reasoning.

Context has been shown to be informative and useful in scene understanding [11–13] and is intuitively present in reality theoretically and practically. Nowadays, relation reasoning is playing an essential part in context modeling, which is applied to both 2D and 3D indoor object detection [10, 14, 15]. However, there are still two main unsolved challenges. On the one hand, most 3D detectors rely on proposals (object candidates) for classification and bounding

1 College of Computer, National University of Defense Technology, Changsha 410073, China. E-mail: Y. Lan, lanyuqingkd@nudt.edu.cn; Y. Duan, duanyao16@nudt.edu.cn; C. Liu, liuchenyi_1013@nudt.edu.cn; C. Zhu, zhuchenyang07@nudt.edu.cn; Y. Xiong, ysxiong@hotmail.com; K. Xu, kevin.kai.xu@gmail.com (✉).

2 Shenzhen University, Shenzhen 518061, China. E-mail: hhzhiyan@gmail.com.

Manuscript received: 2021-07-26; accepted: 2021-08-25

box regression. Qualities of the proposals used in these methods are usually not satisfactory for extracting relation contexts, inevitably producing confusing or even improper contextual information. On the other hand, each proposal actually has its own specific needs for relation contexts from other proposals. Previous methods use equal weights for different relation contexts, which may ultimately result in more ambiguity or even misunderstanding (see Fig. 1).

In this paper, we propose an attention-based relation module for context modeling in 3D object detection to solve these two challenges. We argue that objects in indoor scenes are more or less relative to each other both semantically and spatially. As shown in Fig. 2, the core ideas of our novel method contain two parts which correspond to the two challenges respectively: object-aware relation reasoning among different proposal pairs; an attention module based on Transformer to take full advantage of the most useful ones to extract contextual relation features. The first part includes a simple but quite useful objectness module to select proposals with high qualities. Available with selected proposals, we reason about both of the pair-wise semantic and spatial relations for different proposal pairs. As for the second part, we leverage an attention module based on Transformer to model the importance towards contexts from different proposal pairs for each selected proposal and thus reduce the effects of confusing contexts. In this way, we can not only enhance understanding and mitigate the ambiguity towards various objects in manifold indoor scenes but also avoid being affected by confusing or even useless context information together with the useful ones. Different from previous works, our method does not depend on pre-defined templates for context modeling and pays more attention to the useful context information attained by relation reasoning instead of taking equal treatment. This mitigates the ambiguity and thus can boost the performance of detection.

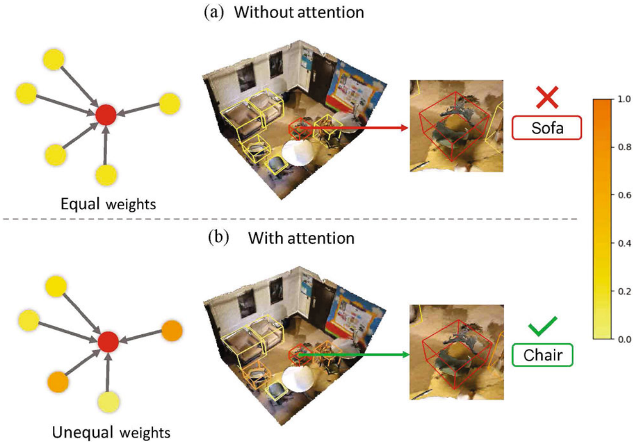


Fig. 1 We propose an attention-based relation module (ARM3D) to reason about the most useful semantic relation contexts in 3D object detection. For example, all the objects with boxes in this figure are chairs represented as dots on the left. (a) A chair with the red box is hard to detect due to noise in point clouds and is mistakenly classified as a sofa using equal attention towards other objects. The upper left chairs in this scene have untypical structures, resulting in unclear semantic relations. (b) With unequal attention, this chair can pay more attention to the semantic relationships with objects having similar structures to filter the confusing context and thus can be classified correctly and robustly. Darker orange indicates greater attention.

ARM3D is a plug-and-play module which can be conveniently applied to different 3D object detectors.

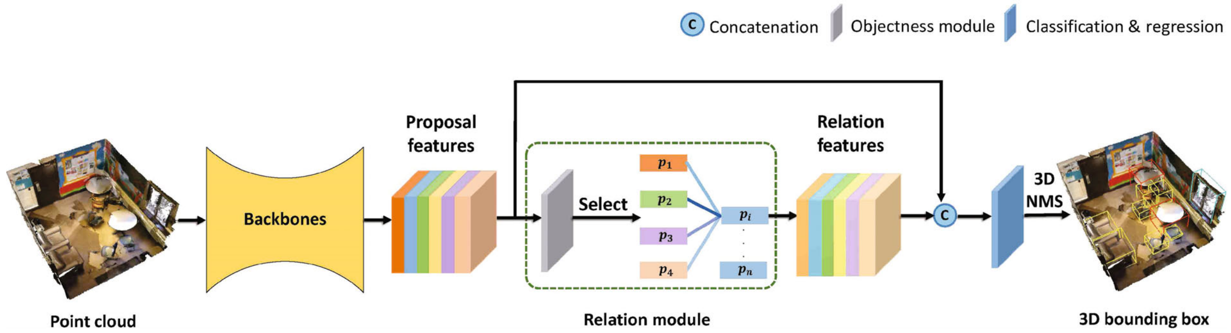


Fig. 2 3D detection pipeline equipped with our ARM3D. With point cloud as input, the backbone networks of current proposal-based 3D detectors produce numerous proposals. These proposals are then sent into our attention-based relation module to extract the fine-grained relation features. These proposals are first selected according to their objectness, and each proposal is matched with several selected proposals to reason about their specific relation contexts. Darker blue means greater attention and higher weights. The relation features are concatenated with the proposal features together. The combined features of different proposals are used by the detection heads to perform classification and regression. After 3D non-maximum suppression (NMS), the pipeline outputs the final detected bounding boxes.

It provides precise and useful relation contexts to help 3D detectors locate and classify objects more accurately and robustly. We apply ARM3D to two 3D object detectors and evaluate its improved performance on two challenging datasets. Extensive experiments demonstrate the effectiveness of ARM3D. Specifically, applying ARM3D to VoteNet [6] achieves **7.8%** improvement on ScanNetV2 [16] and **3.4%** on SUN RGB-D dataset [17]. As for MLCVNet [7], we achieve **3.4%** improvement on ScanNetV2.

In summary, the major contributions of this paper are:

- a novel attention-based 3D relation module, using a simple but useful objectness module to perform object-aware relation reasoning between selected proposals, which can extract reliable and rich semantic and spatial relation contexts for detection;
- an expressive attention module based on Transformer, intended to avoid the negative effects of confusing relation contexts and thereby enabling each object to take full advantage of the most useful context from others. Incorporated with the proposed objectness module and attention module, our method ARM3D can achieve more accurate and detection performance;
- extensive experiments demonstrating the benefits of our attention-based relation module. Using our relation module in two state-of-the-art detectors shows substantial improvements on ScanNetV2 and SUN RGB-D benchmarks indicating that our design is effective and can be widely applicable.

2 Related work

3D object detection in point clouds. 3D object detection has been investigated for decades with numerous applications [6–8, 18–24]. However, due to the orderless and sparse properties of point clouds, one of the main 3D representations, 3D object detection still remains challenging. Before the emergence of deep learning techniques on 3D point clouds [2, 25, 26], earlier attempts mainly turn to intermediate solutions such as using voxel grids [27–29], multi-view images [22, 30] or trying to transform 2D object candidates to 3D from existing 2D object detection methods [21, 31], which limits the applicability in certain situations.

Thanks to PointNet/PointNet++ [1, 3], in recent years 3D object detection has started to take point clouds directly as input. Inspired by Faster RCNN [32], PointRCNN [24] uses a two-stage 3D object detector for proposal generation and refinement. Yi et al. [5] proposed GSPN, a novel object proposal generation network by reconstructing shapes from noisy observations in a scene with an analysis-by-synthesis strategy. Motivated by Hough voting in 2D object detection, VoteNet [6] presents an end-to-end trainable 3D object detection framework and highlights the challenge and importance of directly predicting bounding box centers in point clouds because most surface points are far from the object centers. Extension works of VoteNet [7, 9, 20, 33, 34] make use of contextual information, graph neural networks with hierarchical structures, better reasonable sampling strategies, and back-tracing representative cluster points for better proposal generation. In fact, explicit relationships between objects provide abundant information for scene understanding, which are usually ignored. The significance of relation contexts between objects for 3D box estimation is also emphasized by Huang et al. [35].

Relational reasoning in 3D. With the emergence of the Relation Network [36], there have been a great number of methods that adapt the Relation Network [36] to various 2D image tasks [14, 14, 37, 37–46]. The successful applications of these works illustrate the importance of relation reasoning in visual tasks.

As a result of the successful applications of relational reasoning in 2D, various works began to explore its applications in 3D. For furniture layout in 3D, Ref. [47] defines five types of relations for modeling furniture in indoor scenes using a graph structure, which, however, is time-consuming for relations like *facing* and Ref. [48] measures the similarity between various furniture layouts with case-based reasoning. Duan et al. [49] took advantage of PointNet [1] to reason about the local structural dependencies with an additional relation network and attain improved performance in point cloud classification as well as part segmentation. Aimed at pose estimation, Ref. [50] proposes a joint object and relation network to analyze the relative poses between each pair of objects. For 3D object detection, Xie et al. [7] exploited self-attention to reason meaningful contextual information to generate better qualified proposals at three levels. GRNet [51] proposes a geometric relation network to leverage intra-object and inter-object features extracted by aggregation for 3D object detection. Ref. [10] proposes a relation module that explicitly defines the semantic and spatial relations between objects to get better relation contexts for 3D object detection. However, these works usually ignore the fact that part of the contextual information is misleading, and may degrade the performance in visual tasks when combined with correct information in complex environments.

Attention in 3D vision. Attention is an intelligent mechanism which can highlight what is important in a flexible manner. Recently, there have been numerous methods introducing attention to all kinds of 3D vision tasks. Refs. [52–54] intuitively leverage attention-based graph structures to capture the fine-grained features of 3D points for point cloud classification and segmentation. Ref. [55] proposes a skip-attention mechanism to bridge local region features and point features of the decoder for better point cloud completion. There are also applications of attention mechanism in point cloud registration [56, 57] and point cloud based retrieval [58, 59]. Moreover, Refs. [60, 61] adapt Transformer, which attracts much attention in natural language processing, to 3D point cloud learning, and obtains high performance. Inspired by these methods, we utilize an expressive attention module mainly based on Transformer to model the importance of relation contexts of different object pairs for more accurate and robust 3D object detection.

3 Method

3.1 Overview

Contextual relationships have been shown to be useful. However, there are still two main challenges when applying relational reasoning to 3D object detection. Firstly, most existing methods resort to object proposals first and rely on these proposals to perform bounding box classification and regression. Objectness of these raw proposals is usually represented as proposal quality, which actually makes a difference to relational reasoning. Proposals with low objectness, however, usually account for the majority, resulting in misleading context to some extent. Secondly, even for high-quality proposals, simply extracting the relation contexts between these proposals is not robust enough. Previous methods give relation contexts equal importance. This inevitably includes contradictory information with regard to a single object and may lead to ambiguity in 3D object detection.

To overcome these two challenges and utilize relation contexts better, we have designed an attention-based relation module, *ARM3D* for short, to distribute unequal attention towards relation contexts with different qualified object proposals. See Fig. 2: with point cloud as input, different backbones can be used to generate numerous object proposals. By taking features of these proposals as input, ARM3D first selects proposals with high objectness scores through MLP which in itself enhances reliability, and then each proposal is matched with other proposals in the same scene at random. Moreover, ARM3D uses an attention module to model the importance of different relation contexts for each selected proposal. For proposal p_i, darker blue indicates greater importance. Both semantic and spatial relational reasoning is performed to extract the contextual relation features for more robust and accurate detection.

In summary, we propose object-aware relational reasoning for the first challenge (see Section 3.2) and an attention module based on Transformer structures for the second challenge (see Section 3.3). Designs for loss function for ARM3D and its application to current 3D detectors are considered in Section 3.4. Extensive experiments show that our design can not only achieve more accurate and robust detection performance but also mitigate the ambiguity in 3D object detection.

3.2 Object-aware relational reasoning

Relation reasoning has been proven to be beneficial to 3D scene understanding [35, 47]. In fact, objects in the same scene are typically related to each other. For instance, only half of a chair beside a table may be visible in point cloud due to noise, but it is still likely to be recognized as a chair using human intuition. The reason why people can successfully understand this situation is that we know that chairs are often found beside tables in indoor scenes. This means that chairs are usually by the side of a table in indoor scenes. However, for neural networks, it is hard to model the correlation directly to provide prior information as used by humans. Thus, we need relational reasoning to model the high-level correlations between different objects involved in 3D object detection. We use two typical and intuitive relations including semantic and spatial relations to help the networks to learn the correlations. Furthermore, relational reasoning should be carried out for proposals with high objectness to avoid misleading contexts. As shown in Fig. 3, the upper part indicates the process of object-aware relation reasoning. Its components are as follows.

Objectness module. The reason why we need an objectness module to filter proposals is that poor quality proposals usually produce misleading contexts during relational reasoning. With $N \times C$ proposals in a given scene as input, the objectness module outputs $N \times 2$ binary labels demonstrating whether the proposals have high enough objectness to be qualified for relational reasoning. $C \in \mathbb{R}^d$ denotes the number of feature channels for each proposal generated by the backbones. Specifically, if the Euclidean distance d_i between the center of a proposal c_i and the center of its nearest ground-truth object is within a certain threshold ξ, the objectness label of this proposal is 1, and 0 otherwise:

$$d_i = \min(D(c_i, c_g)), g \in \{1, \cdots, N_{\mathrm{gt}}\} \tag{1}$$

where D denotes Euclidean distance, c_g is the center of a ground-truth object, and N_{gt} is the number of ground-truth objects in a scene.

The structure of objectness module H_φ is compromises of three MLPs including h_1, h_2, h_3, with $C/2, C/4, 2$ output feature channels respectively, and each convolution layer is followed by batch normalization and ReLU activation. The output of the objectness module is a binary label which indicates whether the proposal is a single object or not. It can be formulated as Eq. (2):

$$l_{\mathrm{obj}} = \operatorname{argmax}(H_\varphi(p_i)), i \in \{1, \cdots, N\} \tag{2}$$

where l_{obj} denotes the objectness prediction of proposal p_i, and $H_\varphi(p_i)$ indicates the binary logits of the last layer.

Matching and processing. Since the objectness module provides an objectness prediction for each proposal, it is simple to select $N_s \times C$ high-quality proposals. We argue that pair-wise relation contexts benefits the detection of each proposal to the full extent if the context is extracted from proposals with high objectness. Each proposal is matched with N_k proposals among the N_s ones selected by the objectness module in the same scene at random. The strategy of random matching is intended to increase the diversity of object-wise relation contexts, while the attention module in Section 3.3 is able to choose the more useful ones. Instead of using sampling strategies like farthest point sampling (FPS) or k-nearest neighbors (k-NN) to select proposals for matching, we argue that the selection results

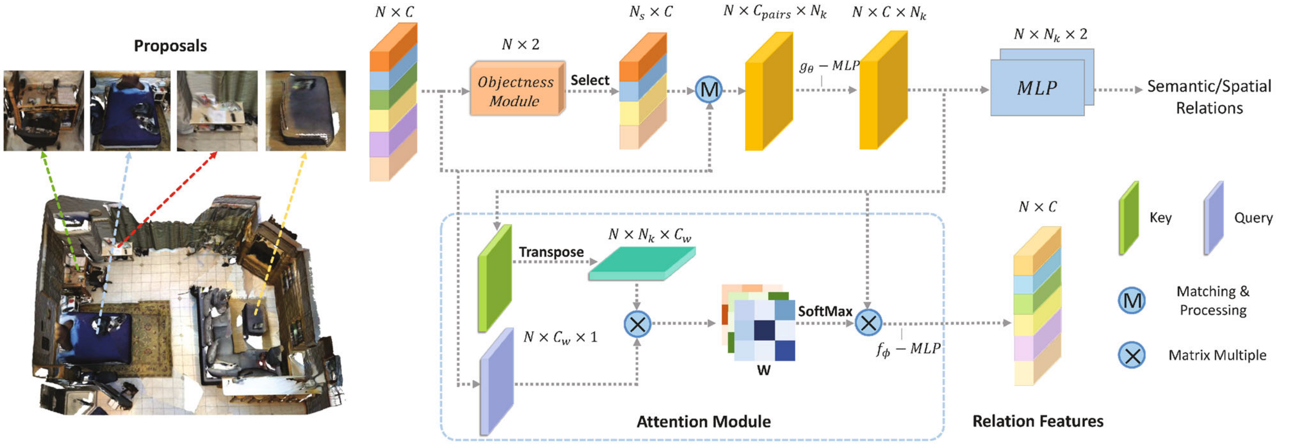


Fig. 3 Network architecture of ARM3D. With N proposals as input, the objectness module, mainly composed of MLPs, firstly outputs binary labels to select N_s proposals with high objectness. C indicates the feature channels. Each proposal is matched with a certain number of selected proposals at random, and further operations, including matrix subtraction and concatenation, are performed on these object pairs to obtain their differences. Pair-wise features corresponding to the same proposal go through the MLP labelled g_θ. The extracted pairs of features are then transposed and fed into other MLPs to reason about semantic or spatial relations: see Section 3.2. Moreover, the original $N \times C$ proposals and pair-wise proposals go through two MLPs named *Query* and *Key* MLPs which output the matrices that are multiplied to compute the attention matrix. SoftMax activation follows, which is then multiplied by pair-wise features. Processed by the f_ϕ MLP, the relation module outputs relation features for each proposal.

are relatively unchanged for these sampling methods. Modeling the accurate correlation between proposals is a great challenge and we use random sampling to increase the diversity for better understanding since we have an attention module to keep the relation contexts stable and useful. In our experiments, other sampling strategies work less well than random sampling. To process a pair of proposals (p_i, p_j), from the features of p_i we subtract those of p_j to obtain the difference, which is concatenated with p_i, and formulated as features of proposal pairs $N \times C_{\text{pairs}} \times N_k$.

To decide on the informativeness of context provided by proposal pairs, we leverage MLPs called g_θ to exploit the semantic or spatial relations within them. These pair-wise features are sent to the classification MLPs named R_θ to predict their semantic or spatial relation labels. For proposal p_i, the relation label l_{r} of itself and its matched proposal p_j can be formulated as follows:

$$l_{\text{r}} = R_\theta(g_\theta(C_\psi(p_i, \Delta(p_i, p_j)))), p_j \in P_k \tag{3}$$

where C_ψ denotes the concatenation of features, and Δ indicates subtraction. P_k denotes the randomly matched proposals for p_i.

Semantic and spatial relations. Motivated by Relation Networks proposed in Ref. [14], Ref. [10] adapts it to 3D object detection and explicitly performs relational reasoning on individual objects instead of on the entire scene. The main differences between our method and Ref. [10] are that we simplify the semantic relations to exclude relations between the same instance, and we use an attention module combined with an objectness module to make full use of the relation contexts to avoid redundant contexts and provide better performance. In this paper, the original relations presented in Ref. [10] including *group, same as, support, hang on* are simplified. Since only proposals with high objectness are selected for extracting relation contexts, relations like *same as* which indicates that two proposals belong to the same instance are in a minority and unsuitable in this case. Therefore, we believe that semantic relations and spatial relations are the most typical and beneficial pair-wise object relation types for indoor 3D object detection, which are exactly sufficient for gathering nontrivial relation contexts. As shown in Fig. 4, two types of relations indicate whether two objects are in the same category or not, and whether one is linked to the other horizontally or vertically.

With regard to semantic relations, there are usually various types of objects in indoor scenes. If two objects are in the same category, the semantic relation label is 1, and 0 otherwise. For each object, distinguishing semantic categories from other objects implies rich semantic information. The goal of semantic relations is to capture the semantic class-specific properties between objects. Objects in the same category usually have similar structures and parts, which helps the objects to better recognize themselves with semantic context. In contrast, for objects of different categories, an object can learn differences from their structures and appearance through the semantic relations. Although the principle of semantic relations is simple, it is useful and informative for 3D object detection.

As for spatial relations, we combine the relations of *support, hang on* proposed in Ref. [10] together as *spatial* relations. In this paper, spatial relations

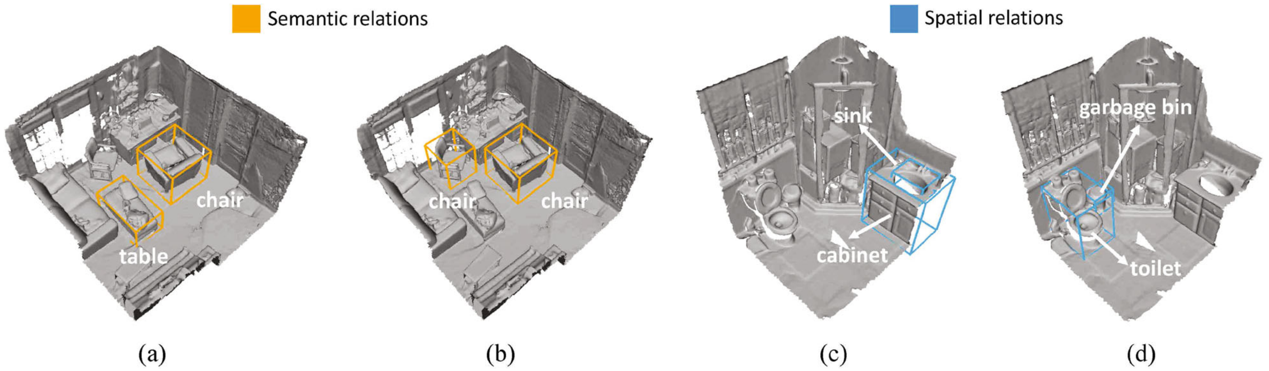


Fig. 4 Semantic and spatial relations. The orange bounding boxes indicate semantic relations, and the blue bounding boxes show spatial relations. (a) Semantic relations in different categories between a chair and the table. (b) Semantic relations in the same category between these two chairs. (c) Vertical spatial relations between the sink and the cabinet. (d) Horizontal spatial relations between the toilet and the garbage bin beside it. Best viewed on screen.

indicate that two objects are adjacent to each other horizontally or vertically, which can indicate that one is supporting or linked to the other one. In reality, objects are more or less spatially related, especially for those with 3D representations. For example, a chair is under a table, or a bookshelf is beside a wall. Such cases are typical spatial relations for indoor object pairs, which provide intuititive and meaningful contexts for object detection. We define that a spatial relation exists only if two proposals satisfy two conditions. First, the relative height H_r or horizontal distance L_r of two proposals should be lower than a threshold τ_d. This means that the nearest distance between points of two proposals should be small enough, either horizontally or vertically. Second, the overlap ratio of bounding boxes for two proposals should be larger than a threshold τ_r either on the $x-y$ plane, the $y-z$ plane and the $z-x$ plane with respect to the first condition. Take the $x-y$ plane as an example. The overlap ratio $r_{i,j}$ can be calculated as

$$r_{i,j} = \max\left(\frac{\varOmega_{xy}(p_i,p_j)}{\varphi_{xy}(p_i)}, \frac{\varOmega_{xy}(p_i,p_j)}{\varphi_{xy}(p_j)}\right) \tag{4}$$

where $\varOmega_{xy}(\cdot,\cdot)$ denotes the area of intersection in projection for two proposals on the horizontal plane, and $\varphi_{xy}(\cdot)$ is the projected area of a proposal on the horizontal plane.

If the overlap ratio $r_{i,j}$ is lower than τ_r, the pair of proposals (p_i,p_j) is supposed to have spatial relations, similarly for the $y-z$ plane and the $z-x$ plane. Such compact spatial relations are helpful for 3D object detection as well as scene understanding.

3.3 Attention module

Although relation contexts are generally beneficial for detection, not all contexts from other objects are essential and helpful for a single object. It is common and inevitable that some pair-wise relation contexts are misleading and even useless for specific objects (see Fig. 1). The attention mechanism, which has become a focus in 3D vision recently, is appropriate for solving this problem.

In order to make our relation module more expressive and robust, we adapt the attention module based on Transformer in Ref. [60] for analyzing the importance of different pair-wise relation contexts for every single object. Unlike Transformer in Ref. [60] which leverages self-attention to extract features of point clouds, our attention module is designed to weigh different pair-wise relation contexts. As shown in Fig. 4, to be specific, the $N\times C$ original proposals first go through MLPs named *Query* and the feature channel is downsampled to C_w. A similar operation is performed on $N\times C\times N_k$ pairs of proposals that are matched with the original N proposals. After the MLPs called *Key*, the pairs of proposals are transposed into $N\times N_k\times C_w$, and multiplied by the *Query* proposals to obtain the $N\times N_k$ attention matrix. Note that we use tanh activation to normalize the outputs before multiplication. Each row of the attention matrix corresponds to a proposal in the scene. Values in each row give the importance of relation contexts from different pairs of proposals, respectively. After SoftMax normalization, the attention matrix is used to assign different weights to the $N\times C\times N_k$ pairs of proposals; the sum of these values is used to compute the weighted average relation contexts. Last, the weighted relation contexts for each original proposal is fed into MLPs called f_ϕ to output the final relation features. The process of obtaining the relation features R_i of proposal p_i can be formulated as follows:

$$W = \Theta(\Gamma(K^\mathrm{T})\times\Gamma(Q)) \tag{5}$$

$$R_i = f_\phi\left(\sum_{\forall j} W_{i,j}\times(p_i,p_j)\right), j\in\{1,\cdots,N_k\} \tag{6}$$

where Q is the *Query* output matrix and K is the *Key* output matrix; Γ denotes the tanh activation function; Θ is SoftMax normalization; and W indicates the attention matrix of different pairs of proposals (p_i,p_j). Further details are provided in Algorithm 1.

Algorithm 1 Pseudo-code for attention-based relation features formulation

```
Input: Proposal $p_i$, $N_k$ pairs of proposals $(p_i,p_j)$, and MLPs $f_\phi, Key, Query$.
Output: Weighted relation features $R_i$.
Initialize: $R_i = 0$, $W = 0$.
for all $j \in \{1,\cdots,N_k\}$ do
    matrix $Q = Query(p_i)$, matrix $K = Key((p_i,p_j))$;
    $Q = \tanh(Q), K = \tanh(K)$;
    $W_{i,j} = K^\mathrm{T} \times Q$ and $W \leftarrow W_{i,j}$.
end for
Normalize $W$ matrix by SoftMax; $j = 0$.
while $j \leqslant N_k$ do
    $R_\mathrm{tmp} = f_\phi(W_{i,j} \times (p_i,p_j))$;
    $R_i = R_i + R_\mathrm{tmp}$; $j = j + 1$
end while
Return the weighted relation features $R_i$.
```

3.4 Application and loss function

To examine the effectiveness of our method, we have applied our attention-based relation module ARM3D to two state-of-the-art 3D object detectors: VoteNet [6] and MLCVNet [7]. Taking the grouped clusters as proposals, ARM3D predicts the pair-wise semantic or spatial relations between those with high objectness and outputs the beneficial relation features to boost the performance of 3D object detection.

The loss of ARM3D is simply made up of the objectness loss as well as the relation prediction loss, corresponding to Section 3.2 and Section 3.3 respectively. The objectness loss is formulated as $\mathcal{L}_{\mathrm{obj}}$, which is used to supervise the module to predict the accurate objectness of each proposal. The relation prediction loss is formulated as $\mathcal{L}_{\mathrm{rn}}$, which refers to the prediction loss of semantic or spatial relations between proposal pairs, using the binary cross entropy. For the better selection of objectness, we set different weights: w_0 for those proposals whose ground-truth objectness labels are false, and w_1 for the true ones. Similar strategies are adopted for $\mathcal{L}_{\mathrm{rn}}$ too. $\mathcal{L}_{\mathrm{rn}}$ represents the loss for a single type of relation (semantic or spatial relations). The final relation loss $\mathcal{L}_{\mathrm{r}}$ is the sum of these two losses. $\mathcal{L}_{\mathrm{rn}}$ is formulated as follows:

$$\mathcal{L}_{\mathrm{rn}} = -\frac{1}{N_{\mathrm{p}}}\sum_{i=1}^{N_{\mathrm{p}}} w_1 \cdot y_i \cdot \log(p(y_i)) + \\ w_0 \cdot (1-y_i) \cdot \log(1-p(y_i)) \tag{7}$$

where N_{p} is the number of proposal pairs with $N_{\mathrm{p}} = N \times N_k$ in this paper. y_i indicates the positive ground-truth semantic or spatial relation label of the proposal pair, and $p(y_i)$ is the predicted possibility of the relation of this pair to be positive. w_0 and w_1 are the weights as above.

Previous methods only calculate the objectness loss of proposals that are either within a small distance or beyond a large distance. Since the accuracy of objectness prediction makes a difference to our method, we calculate the objectness loss for all proposals and assign more weight to positive instances while training.

Following Refs. [6, 7], when using our ARM3D, the network is trained in an end-to-end manner by using a voting loss $\mathcal{L}_{\mathrm{vote}}$, a 3D bounding box regression loss $\mathcal{L}_{\mathrm{box}}$, and a semantic classification loss $\mathcal{L}_{\mathrm{cls}}$, in addition to the objectness loss $\mathcal{L}_{\mathrm{obj}}$ and relation prediction loss. The overall 3D object detection loss is formulated as

$$\text{loss} = \lambda_1\mathcal{L}_{\mathrm{vote}} + \lambda_2\mathcal{L}_{\mathrm{obj}} + \lambda_3\mathcal{L}_{\mathrm{box}} + \lambda_4\mathcal{L}_{\mathrm{cls}} + \lambda_5\mathcal{L}_{\mathrm{r}} \tag{8}$$

where in our experiments, we set $\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 1.0, \lambda_4 = 0.1, \lambda_5 = 0.1$.

4 Implementation details

In this section, we first describe the implementation details about the network architecture and the corresponding parameters for ARM3D. Then we explain how to apply our ARM3D to two 3D object detectors, VoteNet [6] and MLCVNet [7], as well as the overall training strategies.

Details in ARM3D. Proposals are object candidates for 3D object detection. Our ARM3D selects proposals with an objectness module, and relation contexts can be extracted from these relatively reliable ones. For the ground-truth objectness labels, we set the distance threshold $\xi = 0.3$. Objectness of proposals within ξ with respect to their ground-truth objects is set to 1. Unlike previous methods that only compute the objectness loss of proposals within the *near* distance threshold or the *far* threshold, we focus on the objectness of all proposals. For $\mathcal{L}_{\mathrm{obj}}$, we use the binary cross-entropy loss with different weights of $w_0 = 0.2$ and $w_1 = 0.8$ for the negative or positive cases respectively, since the backbone network initially produces few sufficiently good proposals. The same strategies and designs are applied to $\mathcal{L}_{\mathrm{rn}}$ since there are relatively fewer positive samples.

As for the strategies of matching different proposals to obtain pair-wise features, we randomly choose $N_k = 8$ proposals from the ones selected by the objectness module for the ScanNetV2 dataset and SUN RGB-D datasets. This strategy provides a good trade-off between speed and results. Moreover, random matching can diversify the relation contexts. For matched proposal pairs, we set the distance threshold $\tau_{\mathrm{d}} = 0.1$ and the ratio threshold $\tau_{\mathrm{r}} = 0.5$ to compute the ground-truth spatial relation labels.

For the attention mechanism used in ARM3D, the *Query* and *Key* MLPs are both composed of one convolutional layer to downsample the input feature from C to $C_w = C/4$ followed by a tanh activation function. Different from other methods, these two

MLPs do not share weights. The function f_{ϕ} is a fully connected layer with C channels as output.

The computational requirements of our method are indicated in Table 1. We compare VoteNet [6] to VoteNet using our ARM3D and the model size of our method is 14.2 MB. The inference time using our method is 0.14 s and 0.09 s on ScanNetV2 and SUN RGB-D datasets respectively, which is comparable to that for VoteNet alone. This demonstrates the efficiency of our method as a lightweight but useful plug-and-play module. The complexity of the calculation of semantic and spatial relations is relatively low-cost since we use matrix multiplication instead of loops in experiments.

Details in training. We apply our ARM3D relation module to VoteNet [6] and MLCVNet [7] to examine whether our method is effective and widely applicable. The number of feature channels C of proposals generated by these two methods is 128. Generally, we keep the same training strategies, including the base learning rates, decay steps, max training epochs, and so on, as in the original papers [6, 7]. The only difference is that, when applied to VoteNet on ScanNetV2, the maximal training epoch is 180, and the batch size is kept as 4 for the first 80 epochs while the batch size is changed to 8 for the remaining epochs. For MLCVNet, we keep the batch size as $b = 8$ from beginning to end. We implement our approach using PyTorch [62] on a single NVIDIA TITAN V. During training, we find that the mAP results fluctuate slightly, so the mAP results given here are mean results over three runs.

Table 1 Model size and processing time (per frame or scan) for VoteNet, and VoteNet with our method ARM3D

Method	Model size (MB)	ScanNetV2 (s)	SUN RGB-D (s)
VoteNet	11.2	0.12	0.08
VoteNet+ARM3D	14.2	0.14	0.09

5 Experiments

In this section, we evaluate the proposed attention-based relation module ARM3D applied to two 3D object detectors, VoteNet [6] and MLCVNet [7]. With point clouds of indoor scenes as input, the experiments are performed on two large 3D indoor scene datasets and evaluated on the corresponding detection benchmarks (see Section 5.1). The evaluation metric we use is demonstrated in Section 5.2. In Section 5.3, we analyze the performance improvement after applying our attention-based relation module ARM3D to the above two 3D object detectors. An ablation study for different components of our method is performed mainly with VoteNet on ScanNetV2 dataset (see Section 5.4). Note that VoteNet depends on Deep Hough Voting for object detection, while MLCVNet extends VoteNet with additional three-level useful contexts; it is challenging for the effectiveness of the relation contexts from our ARM3D. Experimental settings are the same when applying our ARM3D to these detectors. Both quantitative and qualitative results show the effectiveness and generalization ability of our ARM3D.

5.1 Dataset and benchmarks

We use two widely used datasets that provide 3D point clouds of indoor scenes to evaluate our methods: ScanNetV2 [16] and SUN RGB-D [17].

ScanNetV2 is a large RGB-D 3D indoor scene dataset with densely annotated 3D reconstructed meshes. There are approximately 1.5k scanned indoor scenes where both the semantic segmentation and bounding boxes of objects are given. The scanned indoor scenes are relatively complete, which makes it suitable for our method to extract the relation contexts.

SUN RGB-D is a well-known public single-view RGB-D dataset for scene understanding, which contains about 10k RGB-D images. The images are captured by four different sensors, providing accurately annotated oriented bounding boxes in 37 categories. Since it does not provide reconstructed point clouds, we convert the depth images to point clouds using known camera parameters. Most scenes are captured in household environments. Occlusion is common in the SUN RGB-D dataset, and there are fewer ground-truth objects in each scene, making it quite challenging for 3D object detection as well as relational reasoning.

5.2 Evaluation metric

The evaluation metric we take is the average precision of the detected object bounding boxes against those of ground-truth objects. We use two IoU thresholds of 0.25 and 0.5, in our experiments. The mean average precision mAP is the macro-average of the average precision across all test categories.

5.3 Evaluation on two detectors

5.3.1 Overview

We apply our ARM3D to two 3D object detectors, which are VoteNet [6] and MLCVNet [7]. These two methods are regarded as our baselines to examine the effectiveness and improvements of our method ARM3D. We also compare the effects of applying 3DRM [10] to these two detectors. We first analyze the improvement of VoteNet equipped with our ARM3D, which is denoted **VoteNet+ARM3D**. Then we analyze the increased performance after applying our ARM3D to MLCVNet, which is denoted as **MLCVNet+ARM3D**. A brief introduction to these two baselines and a pair-wise relation module for 3D object detection are given below.

- **VoteNet [6]:** An end-to-end trainable 3D object detection framework that takes advantage of deep Hough voting and aggregation to generate proposals for scenes. The aggregated clusters are used to perform classification and bounding box regression.
- **MLCVNet [7]:** A method that utilizes three levels of implicit contexts to enhance the performance of VoteNet, including patch-wise, object-wise, and global contexts.
- **3DRM [10]:** A pair-wise plug-and-play relation module for 3D object detection, which takes advantage of four types of relations to improve the performance of 3D object detectors.

5.3.2 Comparison to baselines

We evaluate our method against VoteNet, MLCVNet, and methods of applying 3DRM [10] to these detectors. Table 2 reports the average precision on the ScanNetV2 and SUN RGB-D datasets with mAP@0.5 and mAP@0.25 respectively. Our methods **VoteNet+ARM3D** and **MLCVNet+ARM3D** achieve the best performance on ScanNetV2 val set and SUN RGB-D val set both.

Table 2 Comparison of our approach against VoteNet and MLCVNet on 3D object detection on ScanNetV2 and SUN RGB-D val sets. VoteNet+3DRM and MLCVNet+3DRM* use 3DRM [10]. VoteNet+ARM3D and MLCVNet+ARM3D indicate VoteNet and MLCVNet equipped with our ARM3D

	ScanNetV2		SUN RGB-D	
	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
VoteNet	58.6	33.5	57.7	33.7
VoteNet+3DRM	59.7	37.3	59.1	35.1
VoteNet+ARM3D	**62.6**	**41.3**	**59.3**	**37.1**
MLCVNet	64.5	41.4	59.8	—
MLCVNet+3DRM*	63.6	40.2	58.4	34.3
MLCVNet+ARM3D	**64.8**	**44.8**	**60.1**	**35.8**

From the comparison in Table 2, our method significantly outperforms VoteNet by not only **4%** and **7.8%** on ScanNetV2 but also **1.6%** and **3.4%** on SUN RGB-D for mAP@0.25 and mAP@0.5 respectively. Note that MLCVNet+3DRM* means that we retrain 3DRM [10] on MLCVNet since 3DRM does not have this application. Compared to applying 3DRM to VoteNet, our method outperforms it by 2.9% and 4% on ScanNetV2 as well as by 0.2% and 2% on SUN RGB-D for mAP@0.25 and mAP@0.5 respectively. This shows that our attention-based relation module can extract more robust and accurate relation contexts to benefit the 3D object detectors for better classification and regression. Note that the increased performance on SUN RGBD val dataset is slightly lower than on the ScanNetV2 validation dataset, since SUN RGBD is a single-view RGB-D dataset. Most scenes in the SUN RGB-D dataset are in household environments, and have fewer objects. Occlusion is more common in SUN RGB-D dataset than in ScanNetV2, making it quite challenging for detection as well as extracting relation contexts for our method. However, Table 2 illustrates that our method ARM3D can reliably reason about the relational context even in challenging scenes and environments.

While MLCVNet uses three levels of contexts to boost its performance, our method ARM3D can still improve its performance on 3D object detection via fine-grained relation contexts from ARM3D. Equipped with ARM3D, our method improves MLCVNet by **0.3%** and **3.4%** on ScanNetV2 for mAP@0.25 and mAP@0.5 respectively. Our method also outperforms MLCVNet by **0.3%** on SUN RGB-D in terms of mAP@0.25. In contrast, applying 3DRM [10] to MLCVNet reduces the performance of MLCVNet due to the equal weights towards relation contexts from different proposal pairs, which may contain some misleading contexts. It is noteworthy that our method ARM3D still can improve the performance of MLCVNet which already fuse various contexts to help the detection while 3DRM cannot do this. This further explains the effectiveness and universal benefits of relation contexts extracted by our ARM3D.

5.3.3 Qualitative results and discussion

Qualitative results for different methods and ground truth for ScanNetV2 are shown in Fig. 5. We visualize the results of ground-truth (the first column), our method (the second column), VoteNet (the third column), and VoteNet+3DRM (the last column). Thanks to ARM3D, our method obviously detects the objects more accurately and robustly. For example, there are four chairs and a table in the scene of the third row. Our method can detect the ground-truth objects with almost the same bounding boxes, while other methods produce many redundant bounding boxes with even wrong category labels. Note that the results of VoteNet+3DRM are better than that of VoteNet while our method achieves the best results, showing the effectiveness of our method.

Figure 6 displays a qualitative comparison of results from our method and other methods on the SUN

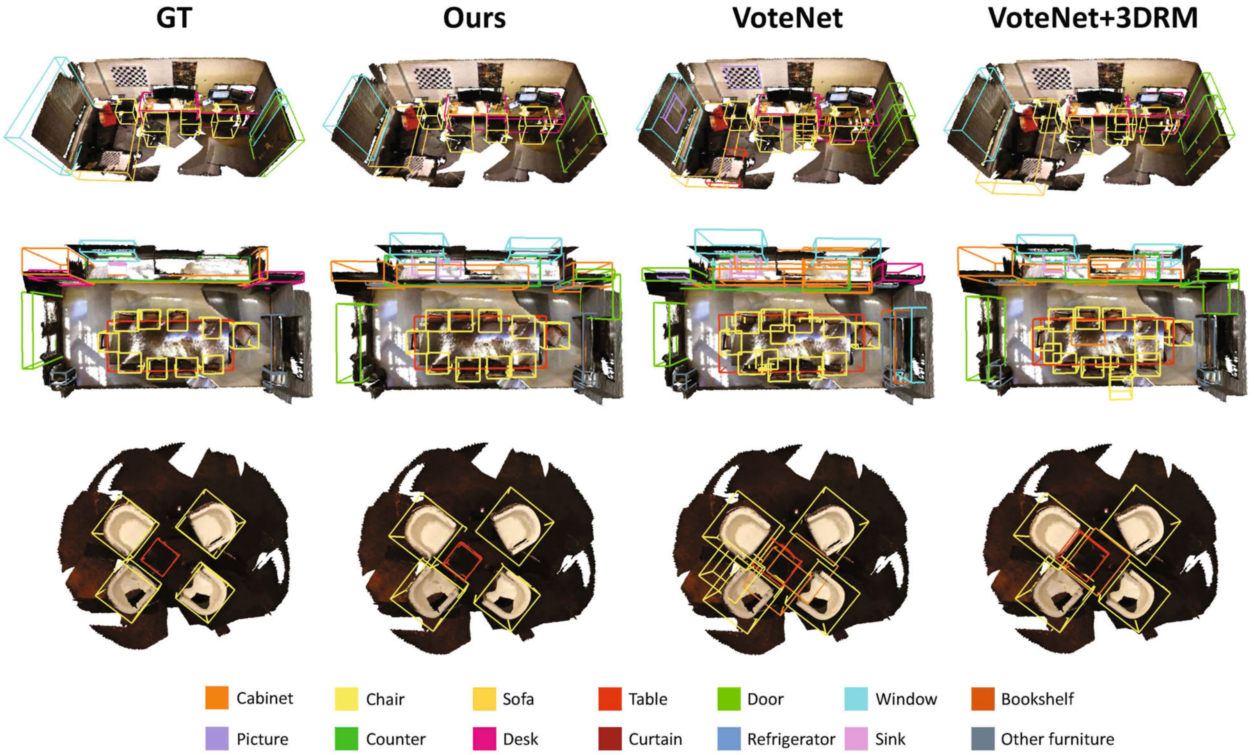


Fig. 5 Qualitative comparison results of 3D object detection on the ScanNetV2 val set. Columns left to right: ground-truth, our method, VoteNet, VoteNet+3DRM. The detailed comparison demonstrates that our method ARM3D enables more accurate and reasonable detection.

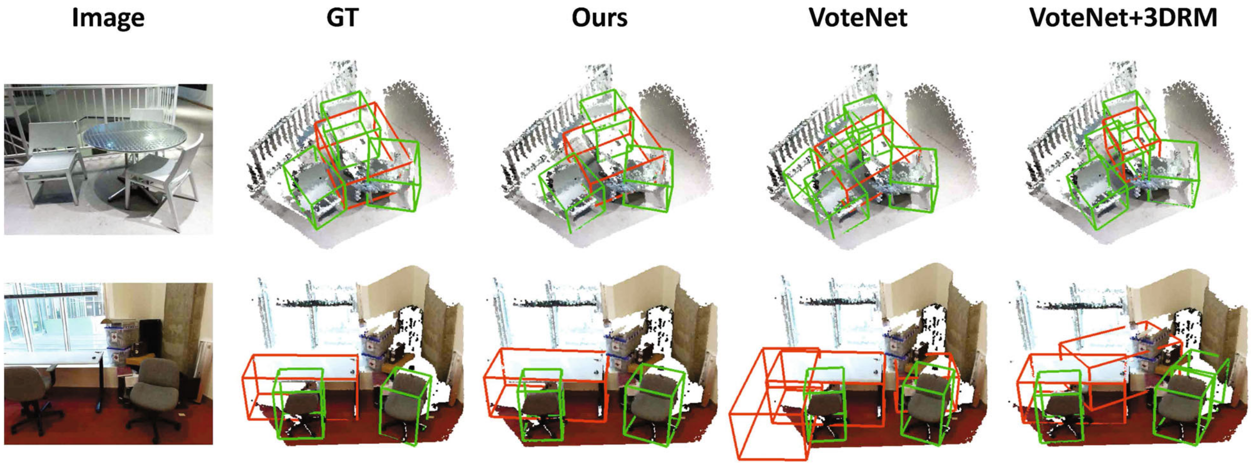


Fig. 6 Qualitative comparison results of 3D object detection on SUN RGB-D val set. Columns left to right: RGB image of the scene, ground-truth, our method, VoteNet, VoteNet+3DRM. Our method VoteNet+ARM3D provides better results. Color is for depiction, not used for detection.

RGB-D dataset. Using our method leads to better object detection with more accurate bounding boxes, while results from other methods are ambiguous or redundant. We argue that this is beneficial from our robust attention-based relation module ARM3D.

In Fig. 7, more comparison details are displayed, and it is clear that our method achieves more robust and accurate 3D object detection. Note that the green rectangles point out the main difference between these methods, which are shown in close up in the second row. Further qualitative comparisons can be found in the Appendix.

The visualization of the attention examples is shown in Fig. 8. On the left is the 8 × 8 attention matrix, and on the right are weights of different proposals (dots in different colors) towards the proposal (the red dot) in the second row of the matrix. It can be seen that the proposal of a chair (the red dot) pays more attention to the sofa (the green dot) and the desk (the blue dot), corresponding to the semantic (different categories) and spatial relations (horizontal adjacency) respectively.

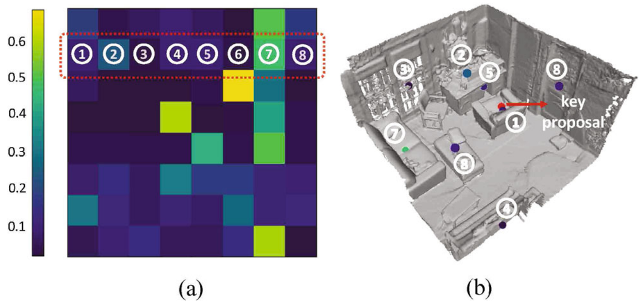


Fig. 8 Attention in VoteNet+ARM3D. (a) 8 × 8 attention matrix. Each row represents a proposal and corrsponding columns represent weights of other proposals towards it. (b) Visualization of the second row in (a), which is numbered 1–8 as for the key proposal (the red dot). The other eight proposals are shown in dots with weighted colors.

5.4 Ablation study

5.4.1 Effects of different components

We analyze the effects of the two main components of our method including the objectness module to select proposals and the attention module. The design of the objectness module is simple but useful. It aims to select proposals with high objectness and therefore our relation module can extract reliable and robust relation contexts among these proposals. The attention module is to distribute different weights towards the relation contexts extracted from the former part since not all relation contexts are useful for each single proposal and some context is confusing. The objectness module and the attention module is simplified as OBM and ATM respectively in Table 3.

Table 3 Comparison of our approach with different components against the baseline of VoteNet+3DRM on ScanNetV2 val set. We denote OBM as the objectness module and ATM as the attention module. VoteNet+ARM3D indicates applying our method ARM3D to VoteNet. Note that we only utilize the semantic relations in this experiment

Method	OBM	ATM	ScanNetV2	
			mAP@0.25	mAP@0.5
Baseline			59.7	37.3
VoteNet+ARM3D	√		60.9	38.7
VoteNet+ARM3D		√	61.5	37.8
VoteNet+ARM3D	√	√	**62.9**	**40.9**

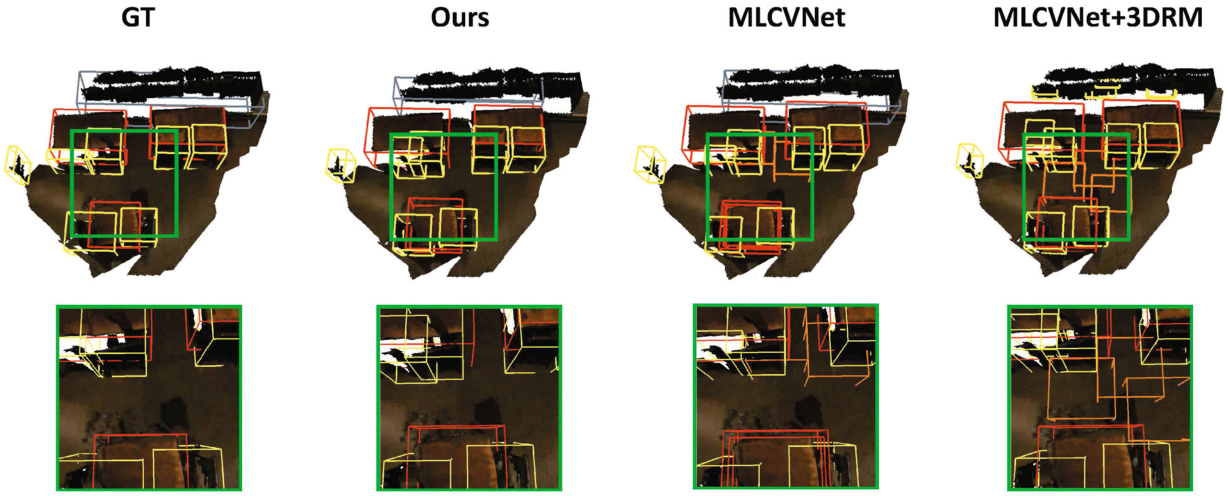


Fig. 7 Qualitative comparison results of 3D object detection on ScanNetV2 val set. The detailed comparison in the second row with green rectangles demonstrates that our ARM3D enables more accurate and reasonable detection. Color is for depiction, not used for detection.

The first row is the baseline of VoteNet+3DRM. The second row is the our method with only the objectness module and the third row is our method with only the attention module. The last row is our full method. It is noteworthy that using only OBM or ATM achieves a slight improvement. However, using both OBM and ATM, our method obtains a larger improvement. This is attributed to our novel designs which support each other and jointly boost the performance.

5.4.2 Comparison of different relations

The effects of different relation types we take on ScanNetV2 val dataset in terms of mAP@0.5 with regard to applying ARM3D to VoteNet are displayed in Table 4. We denote VoteNet+ARM3D as our method by applying our ARM3D to VoteNet. The third row is our method with semantic relations only and the fourth row is our method with spatial relations only. The last row is our method with these two types of relations both. Using both semantic and spatial relations achieve the best performance of **7.8%** improvement against VoteNet. Our method improves the categories of counters, showercurtains, sinks, tables, and chairs by a large margin. Moreover, the detailed average precision of each category shows that different categories of objects pay attention to different types of relations. For example, windows are more sensitive to spatial relations and refrigerators pay more attention to semantic relations with others, while challenging categories for detection like showercurtains and curtains need both of semantic and spatial relations for better detection. This illustrates the effectiveness and significance of both semantic and spatial relations. A comparison of different relations of applying our method to MLCVNet on ScanNetV2 val dataset in terms of mAP@0.25 is demonstrated in Table 5. Similarly, our method using both semantic and spatial relations achieves the highest performance. However, MLCVNet+3DRM* reduces the performance of MLCVNet. MLCVNet is a method with three levels of rich context. The improved mAP further demonstrates the benefits and robustness of our method. The comparison on SUN RGBD val dataset for the effects of different relations to VoteNet+ARM3D in terms of mAP@0.5 can be found in Table 6. Further comparative results are demonstrated in the Appendix.

Table 4 Comparison to VoteNet and VoteNet+3DRM with mAP@0.5 on ScanNetV2 val set for our method with different relations. We denote VoteNet+ARM3D as VoteNet equipped with our ARM3D

	wind	bed	cntr	sofa	tabl	showr	ofurn	sink	pic	chair	desk	curt	fridge	door	toil	bkshf	bath	cab	mAP
VoteNet	6.4	76.1	9.5	68.8	42.4	10.0	11.7	16.8	1.3	67.2	37.5	11.6	27.8	15.3	86.5	28.0	78.9	8.1	33.5
VoteNet+3DRM	12.3	80.6	14.6	71.8	41.3	10.4	13.4	29.5	0.1	67.7	34.7	17.0	37.8	15.7	90.0	**44.2**	**83.0**	8.0	37.3
VoteNet+ARM3D(semantic)	10.3	**82.4**	**32.0**	**76.2**	51.9	14.4	**20.9**	**32.3**	0.2	75.0	**48.4**	14.7	**40.2**	20.0	85.9	40.9	77.6	12.9	40.9
VoteNet+ARM3D(spatial)	**12.9**	80.7	24.1	73.5	**55.1**	11.7	20.6	28.1	2.1	**76.5**	43.7	17.7	36.2	20.3	85.3	36.6	77.5	14	39.8
VoteNet+ARM3D(all)	9.1	78.2	28.2	71.2	54.0	**24.4**	18.7	25.9	**2.8**	75.6	44.1	**23.9**	37.6	**21.9**	**92.0**	43.1	79.1	**13.4**	**41.3**

Table 5 Comparison to MLCVNet and MLCVNet+3DRM* with mAP@0.25 on ScanNetV2 val set for our method with different relations. We denote MLCVNet+ARM3D as MLCVNet equipped with our ARM3D. * denotes that we retrain MLCVNet with 3DRM since 3DRM has not been applied on MLCVNet

	wind	bed	cntr	sofa	tabl	showr	ofurn	sink	pic	chair	desk	curt	fridge	door	toil	bkshf	bath	cab	mAP
MLCVNet	47.0	88.5	63.9	87.4	63.5	65.9	47.9	59.2	11.9	90.0	76.1	**56.7**	60.9	56.9	98.3	**56.9**	87.2	42.5	64.5
MLCVNet+3DRM*	43.6	88.0	63.6	89.2	65.1	64.0	51.3	56.2	11.9	91.3	74.5	48.0	55.0	54.4	99.0	51.8	92.7	46.2	63.6
MLCVNet+ARM3D(semantic)	**48.7**	88.1	58.5	**90.9**	68.9	64.8	**51.7**	61.4	**13.5**	91.7	75.7	49.2	56.3	**58.0**	98.9	53.8	89.9	46.1	64.8
MLCVNet+ARM3D(spatial)	45.6	**90.1**	60.9	87.2	64.1	**75.3**	51.4	**66.0**	11.8	91.5	76.5	51.3	**62.3**	57.2	**99.4**	55.4	91.7	46.9	65.8
MLCVNet+ARM3D(all)	46.4	89.1	**67.2**	89.6	**69.7**	75.0	49.8	58.5	11.7	**92.3**	**78.7**	52.6	56.1	56.8	96.7	54.9	**92.9**	**47.7**	**65.9**

Table 6 Comparison of our approach VoteNet+ARM3D against VoteNet and VoteNet+3DRM with different relations on the SUN RGB-D val set with mAP@0.5

	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
VoteNet	47.0	50.1	7.2	53.9	5.3	11.5	40.7	42.4	19.5	59.8	33.7
VoteNet+3DRM	45.4	51.5	**8.5**	55.3	5.5	16.9	36.8	48.2	20.5	62.9	35.1
VoteNet+ARM3D(semantic)	38.7	51.8	6.4	57.9	**7.1**	15.9	38.4	**51.2**	22.8	**64.8**	35.5
VoteNet+ARM3D(spatial)	46.6	49.2	7.2	58.1	6.6	16.4	42.5	47.7	22.1	60.9	35.7
VoteNet+ARM3D(all)	**50.4**	**54.3**	8.4	**58.7**	6.4	**16.9**	**42.5**	50.0	**22.9**	60.9	**37.1**

5.4.3 Numbers of pairs

Table 7 shows the improved performance for differenet numbers of pairs for our method denoted as VoteNet+ARM3D on ScanNetV2. Sampling $N_k = 8$ proposal pairs for a proposal achieves the best improvement taking both mAP@0.25 and mAP@0.5 as well as computational efficiency into consideration. Further intuitive results are displayed in Fig. 9.

Table 7 Comparison of our ARM3D with different numbers N_k of proposals pairs for each one while relational reasoning on the ScanNetV2 val set. We denote VoteNet+ARM3D as our approach by applying our ARM3D on VoteNet

Method	ScanNetV2	
	mAP@0.25	mAP@0.5
VoteNet	58.6	33.5
VoteNet+ARM3D($N_k = 2$)	61.5	38.8
VoteNet+ARM3D($N_k = 4$)	61.3	39.8
VoteNet+ARM3D($N_k = 6$)	62.0	39.7
VoteNet+ARM3D($N_k = 8$)	**62.9**	**40.9**
VoteNet+ARM3D($N_k = 12$)	61.2	38.5
VoteNet+ARM3D($N_k = 16$)	60.6	37.9

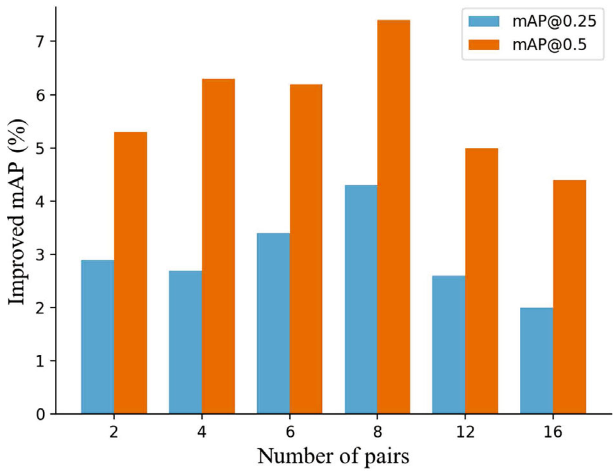


Fig. 9 Improved percentage of mAP for different numbers of proposal pairs for VoteNet+ARM3D over VoteNet.

6 Conclusions

We propose an attention-based relation module for indoor 3D object detection on large-scale scene datasets. Using an objectness module to select raw proposals generated by backbones, we reason about the weighted relation contexts among themselves. Thanks to our attention module based on Transformer, we extract the most useful relation features for each proposal, which enables the network to mitigate the ambiguity and filter out those less relevant or even confusing contexts. We apply our ARM3D to two 3D object detectors on two challenging datasets for more accurate and robust detection. The consistently improved 3D object detection performance illustrates the generalization ability and effectiveness of our method.

Future work. Two research directions are worth considering in future. On the one hand, it is worth trying to apply the attention-based relation module to other 3D visual tasks such as point cloud segmentation and layout arrangement. On the other hand, using a hierarchically designed relation module for reasoning about the relation contexts of sub-scenes or groups of objects is also a promising direction.

Appendix

A More analysis of experiments on ScanNetV2 and SUN RGB-D

More results of experiments for our method VoteNet+ARM3D against VoteNet and VoteNet+3DRM on ScanNetV2 and SUN RGB-D dataset are shown in Table 8 and Table 9. More comparison experiments of different relations of MLCVNet equipped with our ARM3D, denoted as MLCVNet+ARM3D are shown in Table 10.

Illustrated in Table 8, we compare VoteNet+ARM3D with different relations against VoteNet as well as VoteNet+3DRM on ScanNetV2 val set with mAP@0.25. The results show that VoteNet+ARM3D with only spatial relations achieves the best performance. To be specific, our method improves VoteNet by **4.9%**. We argue that objects of most categories like windows and beds are more sensitive to spatial relations in a lower threshold mAP@0.25. However, objects like toilets and refrigerators already have distinct structures and thus need various semantic contexts for better understanding. It is noteworthy that the improvement of our method (VoteNet+ARM3D) on mAP@0.25 is lower than on mAP@0.5. We argue that this can be attributed to the fact that our ARM3D helps the proposals which are within a threshold against the ground truth objects better than those that are far away from the centers and have poor qualities. In Table 9, and our method with only semantic relations performs the best on SUN RGB-D val dataset.

Table 10 shows the comparison results of

Table 8 Comparison to VoteNet and VoteNet+3DRM with mAP@0.25 on ScanNetV2 val set for our method with different relations. We denote VoteNet+ARM3D as VoteNet equipped with our ARM3D

	wind	bed	cntr	sofa	tabl	showr	ofurn	sink	pic	chair	desk	curt	fridge	door	toil	bkshf	bath	cab	mAP
VoteNet	38.1	87.9	56.1	89.6	58.8	57.1	37.2	54.7	7.8	88.7	71.7	47.2	45.4	47.3	94.9	44.6	92.1	36.3	58.6
VoteNet+3DRM	42.4	88.5	50.2	87.6	59.0	63.9	38.2	46.7	6.2	87.9	67.5	**49.2**	52.9	47.4	98.0	**58.8**	**92.3**	38.7	59.7
VoteNet+ARM3D(semantic)	41.4	89.0	61.2	**92.5**	63.6	67.7	43.9	56.9	8.9	90.5	70.7	47.1	**58.0**	52.2	**99.8**	54.2	90.9	44.2	62.9
VoteNet+ARM3D(spatial)	**42.8**	**89.5**	**67.3**	89.6	64.4	66.2	**49.5**	**60.5**	**10.8**	**91.7**	75.1	45.8	55.0	53.3	99.6	51.2	87.5	42.7	**63.5**
VoteNet+ARM3D(all)	41.3	88.9	57.4	90.3	**66.1**	**73.1**	44.0	50.7	9.2	90.9	**75.3**	43.5	55.3	**55.3**	97.1	57.9	86.1	**44.6**	62.6

Table 9 Comparison of our approach VoteNet+ARM3D against VoteNet and VoteNet+3DRM with different relations on SUN RGB-D val dataset with mAP@0.25

	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
VoteNet	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
VoteNet+3DRM	**77.5**	84.5	31.0	75.6	25.7	28.9	63.3	65.5	**50.1**	88.9	59.1
VoteNet+ARM3D(semantic)	76.8	**85.3**	28.9	**77.3**	**28.7**	**34.5**	62.0	66.3	49.0	90.1	**59.9**
VoteNet+ARM3D(spatial)	76.7	82.8	**31.7**	77.2	26.2	32.7	**64.3**	64.9	49.4	**91.0**	59.7
VoteNet+ARM3D(all)	74.0	85.1	28.4	77.3	27.7	32.2	63.4	**66.4**	49.1	89.7	59.3

Table 10 Comparison of differenet relations of MLCVNet+ARM3D with mAP@0.5 on ScanNetV2 val set. We denote MLCVNet+ARM3D as MLCVNet equipped with our ARM3D

	wind	bed	cntr	sofa	tabl	showr	ofurn	sink	pic	chair	desk	curt	fridge	door	toil	bkshf	bath	cab	mAP
MLCVNet+ARM3D(semantic)	**16.7**	78.0	29.4	**81.3**	55.0	**42.8**	**26.7**	27.4	**3.9**	76.4	45.7	**22.8**	34.5	**28.1**	89.4	42.3	86.1	**20.0**	**44.8**
MLCVNet+ARM3D(spatial)	15.2	78.7	17.0	74.1	51.2	23.1	25.0	**29.1**	2.1	76.6	**48.4**	20.8	**39.5**	21.9	**91.4**	**48.1**	82.7	16.6	42.3
MLCVNet+ARM3D(all)	11.4	**79.1**	**29.9**	74.8	**57.1**	17.5	22.0	25.7	1.9	**76.9**	45.7	13.8	37.1	24.0	86.5	47.4	**89.1**	19.0	42.2

our method (MLCVNet+ARM3D) with different relations on ScanNetV2 val dataset. Our method with only semantic relations performs the best with the improvement of **3.4%** towards MLCVNet. The improved results of MLCVNet+ARM3D indeed illustrate the generalization ability of our ARM3D, which can be widely applied on different 3D detection detectors and datasets.

B More visualization of detections on ScanNetV2

In Fig. 10, the comparison results on ScanNetV2 val set are shown. From the comparison of ours, MLCVNet, and MLCVNet+3DRM, it can be found that our methods can detect the objects more accurately and robustly than other methods. In

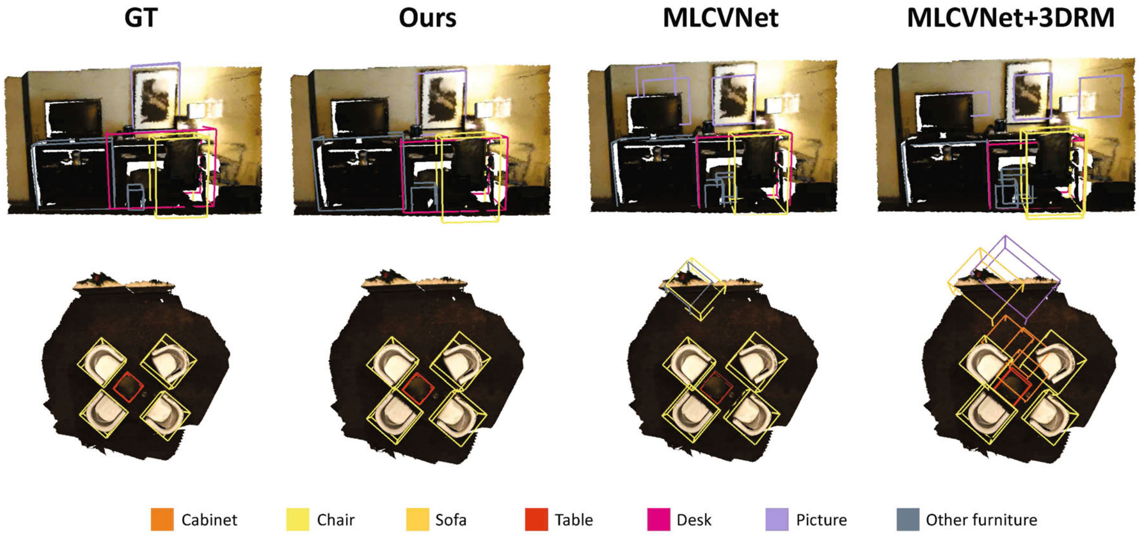


Fig. 10 Qualitative comparison results of 3D object detection on ScanNetV2 val set in terms of MLCVNet. The detailed comparison demonstrates that our ARM3D enables more accurate and reasonable detection

indoor scenes, there are usually many chairs and tables, which often results in redundant bounding boxes while in detection. Our ARM3D can alleviate this problem by utilizing reliable and robust relation contexts and thus achieve better detection. More qualitative results are shown in Fig. 11.

Acknowledgements

We thank Jiazhao Zhang for server management. This paper is supported in part by National Nature Science Foundation of China (62132021, 62102435, 62002375, 62002376), National Key R&D Program of China (2018AAA0102200), and NUDT Research Grants (ZK19-30).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Charles, R. Q.; Hao, S.; Mo, K. C.; Guibas, L. J. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 77–85, 2017.

[2] Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution on X-transformed points. In: Proceedings of the 32nd Conference on Neural Information Processing Systems, 820–830, 2018.

[3] Qi, C. R.; Yi, L.; Su, H.; Guibas, L. J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st Conference on Neural Information Processing Systems, 5099–5108, 2017.

[4] Wu, W. X.; Qi, Z. A.; Li, F. X. PointConv: Deep convolutional networks on 3D point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9613–9622, 2019.

[5] Yi, L.; Zhao, W.; Wang, H.; Sung, M.; Guibas, L. J. 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, 3942–3951, 2019.

[6] Qi, C. R.; Litany, O.; He, K. M.; Guibas, L. Deep Hough voting for 3D object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9276–9285, 2019.

[7] Xie, Q.; Lai, Y. K.; Wu, J.; Wang, Z. T.; Zhang, Y. M.; Xu, K.; Wang, J. MLCVNet: Multi-level context VoteNet for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10444–10453, 2020.

[8] Zhang, Z.; Sun, B.; Yang, H.; Huang, Q. H3DNet: 3D object detection using hybrid geometric primitives. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12357.* Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 311–329, 2020.

[9] Cheng, B. W.; Sheng, L.; Shi, S. S.; Yang, M.; Xu, D. Back-tracing representative points for voting-based 3D object detection in point clouds. In: Proceedings of

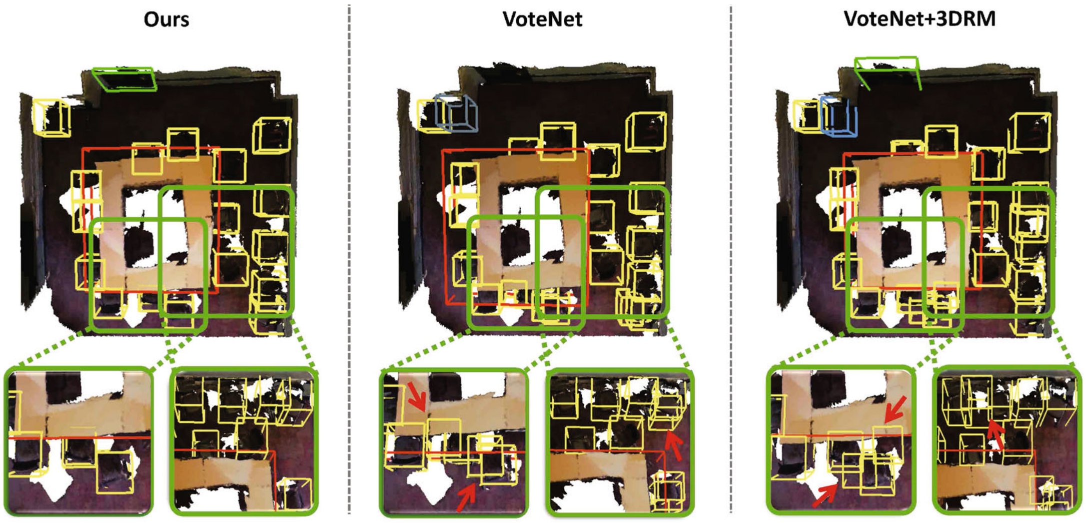


Fig. 11 Qualitative comparison results of 3D object detection on ScanNetV2 val set. The detailed comparison in the second row demonstrates that our ARM3D enables more accurate and reasonable detection.

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8959–8968, 2021.

[10] Lan, Y. Q.; Duan, Y.; Shi, Y. F.; Huang, H.; Xu, K. 3DRM: Pair-wise relation module for 3D object detection. *Computers & Graphics* Vol. 98, 58–70, 2021.

[11] Shi, Y. F.; Long, P. X.; Xu, K.; Huang, H.; Xiong, Y. S. Data-driven contextual modeling for 3D scene understanding. *Computers & Graphics* Vol. 55, 55–67, 2016.

[12] Qi, X. J.; Liao, R. J.; Jia, J. Y.; Fidler, S.; Urtasun, R. 3D graph neural networks for RGBD semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, 5209–5218, 2017.

[13] Zhang, Y.; Bai, M.; Kohli, P.; Izadi, S.; Xiao, J. DeepContext: Context-encoding neural pathways for 3D holistic scene understanding. In: Proceedings of the IEEE International Conference on Computer Vision, 1201–1210, 2017.

[14] Hu, H.; Gu, J. Y.; Zhang, Z.; Dai, J. F.; Wei, Y. C. Relation networks for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3588–3597, 2018.

[15] Xu, H.; Jiang, C. H.; Liang, X. D.; Li, Z. G. Spatial-aware graph relation network for large-scale object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9290–9299, 2019.

[16] Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; Niessner, M. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2432–2443, 2017.

[17] Song, S. R.; Lichtenberg, S. P.; Xiao, J. X. SUN RGB-D: A RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 567–576, 2015.

[18] Lin, D. H.; Fidler, S.; Urtasun, R. Holistic scene understanding for 3D object detection with RGBD cameras. In: Proceedings of the IEEE International Conference on Computer Vision, 1417–1424, 2013.

[19] Shi, Y. F.; Chang, A. X.; Wu, Z. L.; Savva, M.; Xu, K. Hierarchy denoising recursive autoencoders for 3D scene layout prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1771–1780, 2019.

[20] Chen, J. T.; Lei, B. W.; Song, Q. Y.; Ying, H. C.; Chen, D. Z.; Wu, J. A hierarchical graph network for 3D object detection on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 392–401, 2020.

[21] Qi, C. R.; Liu, W.; Wu, C. X.; Su, H.; Guibas, L. J. Frustum PointNets for 3D object detection from RGB-D data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 918–927, 2018.

[22] Chen, X. Z.; Ma, H. M.; Wan, J.; Li, B.; Xia, T. Multi-view 3D object detection network for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1907–1915, 2017.

[23] Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S. L. Joint 3D proposal generation and object detection from view aggregation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 1–8, 2018.

[24] Shi, S. S.; Wang, X. G.; Li, H. S. PointRCNN: 3D object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 770–779, 2019.

[25] Wang, P.-S.; Liu, Y.; Guo, Y.-X.; Sun, C.-Y.; Tong, X. O-CNN: Octree-based convolutional neural networks for 3D shape analysis. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 72, 2017.

[26] Atzmon, M.; Maron, H.; Lipman, Y. Point convolutional neural networks by extension operators. *arXiv preprint* arXiv:1803.10091, 2018.

[27] Yan, Y.; Mao, Y. X.; Li, B. SECOND: Sparsely embedded convolutional detection. *Sensors* (*Basel*) Vol. 18, No. 10, 3337, 2018.

[28] Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L. B.; Yang, J.; Beijbom, O. PointPillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12689–12697, 2019.

[29] Shi, S. S.; Wang, Z.; Shi, J. P.; Wang, X. G.; Li, H. S. 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* Vol. 43, No. 8, 2647–2664, 2021.

[30] Pang, G.; Neumann, U. 3D point cloud object detection with multi-view convolutional neural network. In: Proceedings of the 23rd International Conference on Pattern Recognition, 585–590, 2016.

[31] Lahoud, J.; Ghanem, B. 2D-driven 3D object detection in RGB-D images. In: Proceedings of the IEEE International Conference on Computer Vision, 4632–4640, 2017.

[32] Ren, S. Q.; He, K. M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the

28th International Conference on Neural Information Processing Systems, 91–99, 2015.

[33] Yang, Z. T.; Sun, Y. N.; Liu, S.; Jia, J. Y. 3DSSD: Point-based 3D single stage object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11037–11045, 2020.

[34] Engelmann, F.; Bokeloh, M.; Fathi, A.; Leibe, B.; NieBner, M. 3D-MPA: Multi-proposal aggregation for 3D semantic instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9028–9037, 2020.

[35] Huang, S.; Qi, S.; Xiao, Y.; Zhu, Y.; Wu, Y. N.; Zhu, S.-C. Cooperative holistic scene understanding: Unifying 3D object, layout, and camera pose estimation. In: Proceedings of the 32nd Conference on Neural Information Processing System, 207–218, 2018.

[36] Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; Lillicrap, T. A simple neural network module for relational reasoning. In: Proceedings of the 31st Conference on Neural Information Processing Systems, 4967–4976, 2017.

[37] Mou, L. C.; Hua, Y. S.; Zhu, X. X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12408–12417, 2019.

[38] Li, X.; Yang, Y. B.; Zhao, Q. J.; Shen, T. C.; Lin, Z. C.; Liu, H. Spatial pyramid based graph reasoning for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8947–8956, 2020.

[39] Chen, X. L.; Gupta, A. Spatial memory for context reasoning in object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 4086–4096, 2017.

[40] Cui, Q. J.; Sun, H. J.; Yang, F. Learning dynamic relationships for 3D human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6518–6526, 2020.

[41] Huang, Y. F.; Sugano, Y.; Sato, Y. Improving action segmentation via graph-based temporal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14021–14031, 2020.

[42] Krishna, R.; Zhu, Y. K.; Groth, O.; Johnson, J.; Hata, K. J.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* Vol. 123, No. 1, 32–73, 2017.

[43] Liu, C. C.; Jin, Y.; Xu, K. H.; Gong, G. Q.; Mu, Y. D. 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, 10837–10846, 2020.

[44] Cadene, R.; Ben-Younes, H.; Cord, M.; Thome, N. MUREL: Multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1989–1998, 2019.

[45] Sung, F.; Yang, Y. X.; Zhang, L.; Xiang, T.; Torr, P. H. S.; Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1199–1208, 2018.

[46] Wang, W. B.; Wang, R. P.; Shan, S. G.; Chen, X. L. Exploring context and visual pattern of relationship for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8180–8189, 2019.

[47] Huang, S. S.; Fu, H. B.; Hu, S. M. Structure guided interior scene synthesis via graph matching. *Graphical Models* Vol. 85, 46–55, 2016.

[48] Song, P.; Zheng, Y.; Jia, J. Web3d learning platform of furniture layout based on case-based reasoning and distance field. In: *E-Learning and Games. Lecture Notes in Computer Science, Vol. 10345*. Tian, F.; Gatzidis, C.; El Rhalibi, A.; Tang, W.; Charles, F. Eds. Springer Cham, 235–250, 2017.

[49] Duan, Y. Q.; Zheng, Y.; Lu, J. W.; Zhou, J.; Tian, Q. Structural relational reasoning of point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 949–958, 2019.

[50] Kulkarni, N.; Misra, I.; Tulsiani, S.; Gupta, A. 3D-RelNet: Joint object and relational network for 3D prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2212–2221, 2019.

[51] Li, Y.; Ma, L. F.; Tan, W. K.; Sun, C.; Cao, D. P.; Li, J. GRNet: Geometric relation network for 3D object detection from point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 165, 43–53, 2020.

[52] Wang, L.; Huang, Y. C.; Hou, Y. L.; Zhang, S. M.; Shan, J. Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10288–10297, 2019.

[53] Chen, C.; Fragonara, L. Z.; Tsourdos, A. GAPNet: Graph attention based point neural network for exploiting local feature of point cloud. *arXiv preprint* arXiv:1905.08705, 2019.

[54] Wen, C. C.; Li, X.; Yao, X. J.; Peng, L.; Chi, T. H. Airborne LiDAR point cloud classification with global–local graph attention convolution neural network. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 173, 181–194, 2021.

[55] Wen, X.; Li, T. Y.; Han, Z. Z.; Liu, Y. S. Point cloud completion by skip-attention network with hierarchical folding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1936–1945, 2020.

[56] Wang, Y.; Solomon, J. Deep closest point: Learning representations for point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 3522–3531, 2019.

[57] Yew, Z. J.; Lee, G. H. 3DFeat-Net: Weakly supervised local 3D features for point cloud registration. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11219.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 607–623, 2018.

[58] Zhang, W. X.; Xiao, C. X. PCAN: 3D attention map learning using contextual information for point cloud based retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12428–12437, 2019.

[59] Sun, Q.; Liu, H. Y.; He, J.; Fan, Z. X.; Du, X. Y. DAGC: Employing dual attention and graph convolution for point cloud based place recognition. In: Proceedings of the International Conference on Multimedia Retrieval, 224–232, 2020.

[60] Guo, M. H.; Cai, J. X.; Liu, Z. N.; Mu, T. J.; Martin, R. R.; Hu, S. M. PCT: Point cloud transformer. *Computational Visual Media* Vol. 7, No. 2, 187–199, 2021.

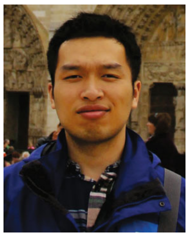
[61] Zhao, H.; Jiang, L.; Jia, J.; Torr, P.; Koltun, V. Point transformer. *arXiv preprint* arXiv:2012.09164, 2020.

[62] Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 8026–8037, 2019.

Yuqing Lan received his B.S. degree in network engineering from National University of Defense Technology, China, in 2019. He is now a postgraduate at the School of Computer, National University of Defense Technology, China. His research interests cover 3D object detection and 3D reconstruction.

Yao Duan received her master degree of computer science from National University of Defense Technology. She is now a Ph.D. student at the School of Computer, National University of Defense Technology, China. Her research interests include 3D object detection

Chenyi Liu received her B.S. degree in software engineering from Tianjin Normal University, China, in 2020. She is now a master student at the National University of Defense Technology, China. Her research interests cover 3D point cloud registration.



Chenyang Zhu is an assistant professor at the School of Computer, National University of Defense Technology. The current directions of interest include data-driven shape analysis and modeling, 3D vision and robot perception & navigation, etc.

Yueshan Xiong is a professor at the School of Computer, National University of Defense Technology. The current directions of interest include virtual surgery system, image and graphics processing, and intelligent computing.

Hui Huang is a Distinguished TFA Professor at Shenzhen University, where she directs the Visual Computing Research Center. Her research interests span computer graphics, 3D vision, and visualization. She is currently a senior member of IEEE/ACM/CSIG and a distinguished member of CCF.

Kai Xu is a professor at the School of Computer, National University of Defense Technology, where he received his Ph.D. degree in 2011. He serves on the editorial board of *ACM Transactions on Graphics*, *Computer Graphics Forum*, *Computers & Graphics*, and *The Visual Computer*. His research work can be found in his personal website: www.kevinkaixu.net.

Other papers from this open access journal are available free of charge from http://www.springer.com/journal/41095. To submit a manuscript, please go to https://www.editorialmanager.com/cvmj.