3DLG-Detector: 3D Object Detection via Simultaneous Local-Global Feature Learning

Baian Chen, Liangliang Nan, Haoran Xie, *Senior Member, IEEE,* Dening Lu, Fu Lee Wang, *Senior Member, IEEE,* and Mingqiang Wei, *Senior Member, IEEE*

***Abstract*—Capturing both local and global features of irregular point clouds is essential to 3D object detection (3OD). However, mainstream 3D detectors, e.g., VoteNet and its variants, either abandon considerable local features during pooling operations or ignore many global features in the whole scene context. This paper explores new modules to simultaneously learn local-global features of scene point clouds that serve 3OD positively. To this end, we propose an effective 3OD network via simultaneous local-global feature learning (dubbed 3DLG-Detector). 3DLG-Detector has two key contributions. First, it develops a Dynamic Points Interaction (DPI) module that preserves effective local features during pooling. Besides, DPI is detachable and can be incorporated into existing 3OD networks to boost their performance. Second, it develops a Global Context Aggregation module to aggregate multi-scale features from different layers of the encoder to achieve scene context-awareness. Our method shows improvements over thirteen competitors in terms of detection accuracy and robustness on both the SUN RGB-D and ScanNet datasets. Source code will be available upon publication.**

***Index Terms*—3D object detection, dynamic points interaction, multi-scale feature learning.**

I. Introduction

REAL-world complex scenes can be flexibly and efficiently represented by point clouds [1], [2]. 3D object detection (3OD) in scene point clouds is a prerequisite for supporting the tasks like autonomous driving and augmented reality. However, the captured point clouds of real-world scenes are natively irregular, compared to 2D (regular) images. Moreover, these point clouds are often sparse, incomplete, noisy, and contain outliers. Feature extraction from such irregularly-sampled yet degraded point clouds tends to weaken cutting-edging 3OD models to localize and recognize objects.

Representing the point clouds for effective processing in deep learning architectures is the first step for 3OD. Currently, two representations have been widely used: voxel-based or point-based. The voxel-based methods [3] divide a point cloud into regular 3D voxels and apply 3D CNNs to learn the high-level features. However, the memory and computational cost are growing exponentially with the increase in the resolution of voxels; it is also hard to trade-off between efficiency and accuracy for these methods [4].

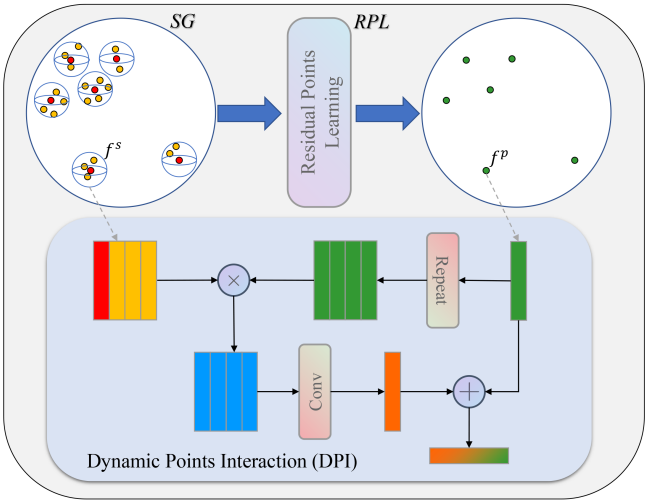


Fig. 1: Dynamic Points Interaction, which can preserve the local features during pooling. Each feature encoder includes three sub-modules, i.e., the Sample-and-Group (SG) module, the Residual Points Learning (RPL) module, and the Dynamic Points Interaction (DPI) module. SG first performs a down-sampling operation on the input point cloud and then groups the neighbor points near the sampled point to form point set features f^s. These point sets are fed into RPL for deep feature representation learning. The max-pooling operation is used to aggregate the point set features into the seed as point-wise features f^p. DPI takes f^s and f^p as input, where f^s is a supplement to make up the lost features of f^p due to the max-pooling operation.

The point-based methods directly take raw point clouds as input to learn feature representations. To handle the irregularity of point clouds without transformations, the seminal work of PointNet [5] and PointNet++ [6] apply multi-layer perceptrons (MLPs) independently on each point, which enables to directly process sparse 3D points. Inspired by PointNet/PointNet++, 3D detectors [7], [8], [9], [10] have achieved satisfactory performance by designing various detection heads.

The key to 3OD is to simultaneously learn different scales and types of features from scene point clouds, such that the learned features can effectively capture both local geometric details and global scene features (context). The local features contribute to regressing the size and orientation of object

B. Chen and M. Wei are with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China (e-mail: 2116068@nuaa.edu.cn; mingqiang.wei@gmail.com).

L. Nan is with the Urban Data Science Section, Delft University of Technology, Delft, Netherlands (e-mail: liangliang.nan@tudelft.nl).

H. Xie is with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong, China (e-mail: hrxie2@gmail.com).

D. Lu is with the Department of Systems Design Engineering, University of Waterloo, Waterloo, Canada (e-mail: d62lu@uwaterloo.ca).

F. L. Wang is with the School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China (e-mail: pwang@hkmu.edu.hk).

bounding boxes, and the global features enhance inferring the classification of objects. The existing point-based 3D detectors learn point features based on PointNet/PointNet++, which inherit several drawbacks of PointNet/PointNet++. First, utilizing PointNet/PointNet++ as backbones loses part of important local features. PointNet/PointNet++ utilizes simple symmetric functions, such as max-pooling, which is an indispensable component to deal with the permutation invariance for point cloud processing. However, the intrinsic character of max-pooling forces it to select the maximal value in each dimension as the representative feature. That means some equally important non-maximum features are lost in each dimension. *We attempt to preserve such local features by designing a Dynamic Points Interaction module.* Second, the global context can well describe the semantic information of the whole scene and the correlations between different objects in the scene. PointNet/PointNet++ only extracts the high-level feature representation by continuously expanding the receptive field while ignoring the global context. The lack of global contextual information hinders the performance of these point-based detectors. Although the recent Pointformer [11] resorts to Transformer [12] to learn the context-aware representation to capture the long-range dependency, it relies on plenty of data for long-term training, which is more difficult to train. *We attempt to mine the global features by designing a Global Context Aggregation module.*

We propose 3DLG-Detector, a 3D object detection network by simultaneously learning local and global features. Inspired by dynamic learning [13], [14], [15], we design a Dynamic Points Interaction (DPI) module to preserve local features during pooling (see Figure 1). In DPI, the input point cloud is first sampled and grouped to form a series of point sets. Then these point sets are fed to the Residual Points Learning module, which consists of several residual MLP blocks, to learn the deep feature representation and aggregate these point sets to seeds by the max-pooling operation. The pooled seeds have simplified local context-aware features, while the grouped point sets possess detailed and redundant local geometric features. The DPI allows a seed to interact with each point in the corresponding point set to preserve local features. Meanwhile, we observe that with the decreasing number of sampling points, the receptive field of each point in different encoder stages constantly increases. Hence, we design a Global Context Aggregate (GCA) module to concatenate the multi-level features together to represent the contextual guidance. The final extracted features by GCA are therefore aware of the global information.

We conduct experiments on two indoor datasets, $i.e.$, ScanNet [16] and SUN-RGBD [17]. Extensive experiments have demonstrated the effectiveness of improvement under several evaluation metrics.

In summary, our contributions are as follows:

- We propose a novel 3D object detection network, 3DLG-Detector, which has a strong ability to learn local and global context features simultaneously. Extensive experiments show clear improvements of our 3DLG-Detector over thirteen competitors in terms of both numerical and visual evaluations.
- We design three modules, among which the DPI and RPL modules extract rich local geometric information, and the GCA module captures the global scene context. Ablation experiments show the effectiveness of these modules in promoting detection performance.

II. Related Work

A. Feature Extraction for 3D Object Detection

Local features extraction. Local feature extraction strategies can be divided into two categories: voxel-based and point-based methods. Voxel-based methods mostly use 3D sparse convolution [18] on regular voxel grids. PointPillars [4] directly adopts the mature 2D convolution by compressing the voxels into a pillar from the vertical dimension. For point-based methods, PointNet [5] and PointNet++ [6] directly consume unorganized 3D points and utilize symmetric functions and Set Abstract (SA) layers to learn the point-wise features and the local features progressively. PCCN [19] exploits parameterized kernel functions to generalize convolution to learn the non-grid structured data. DGCNN [20] constructs a graph in the local region of sampled points and dynamically computes message propagation in each layer of the network. The above-mentioned strategies mostly use pooling operation for feature aggregation to progressively expand the receptive field of the sampled points, leading to the loss of local features. Similarly, our method extracts point-wise features from the local and global perspectives, respectively.

Multi-scale features learning. With the continuously sampling operation on the point clouds, the perception field of each sampled point is extended incessantly. Multi-scale features are concatenated together as the overall scene information to ensure the local features are aware of the global context. PV-RCNN [21] introduces the Voxel Set Abstraction (VSA) module to encode multi-scale voxel-wise features from the feature volumes to the key points. MLCVNet [22] incorporates the multi-level context information from local point patches to global scenes into VoteNet. HVPR [23] proposes an Attentive Multi-scale Feature Module (AMFM), which can refines the hybrid pseudo image to obtain scale-aware features.

B. Voxel-based Object Detection

Voxel-based detectors first convert the point clouds into regular and compact voxel grids to utilize the matured convolutional neural networks. The current approaches can be divided into two groups: one-stage [24], [25], [26], [27], [28] and two-stage detectors [29], [30], [31], [32]. The one-stage detectors focus on lightweight and efficiency, which usually lose the detailed structural information due to the voxelization and continuously down-sampling. SA-SSD [26] introduces an auxiliary network to transform the convolution features into the point-wise representations to exploit the structural information. HVNet [27] proposes a novel voxel feature encoder to attentively aggregate features at different levels and project the multi-scale feature maps to achieve accurate object localization. CIA-SSD [28] proposes a lightweight aggregation module to fuse the semantic and spatial features to predict with accurate confidence, which is subsequently rectified by

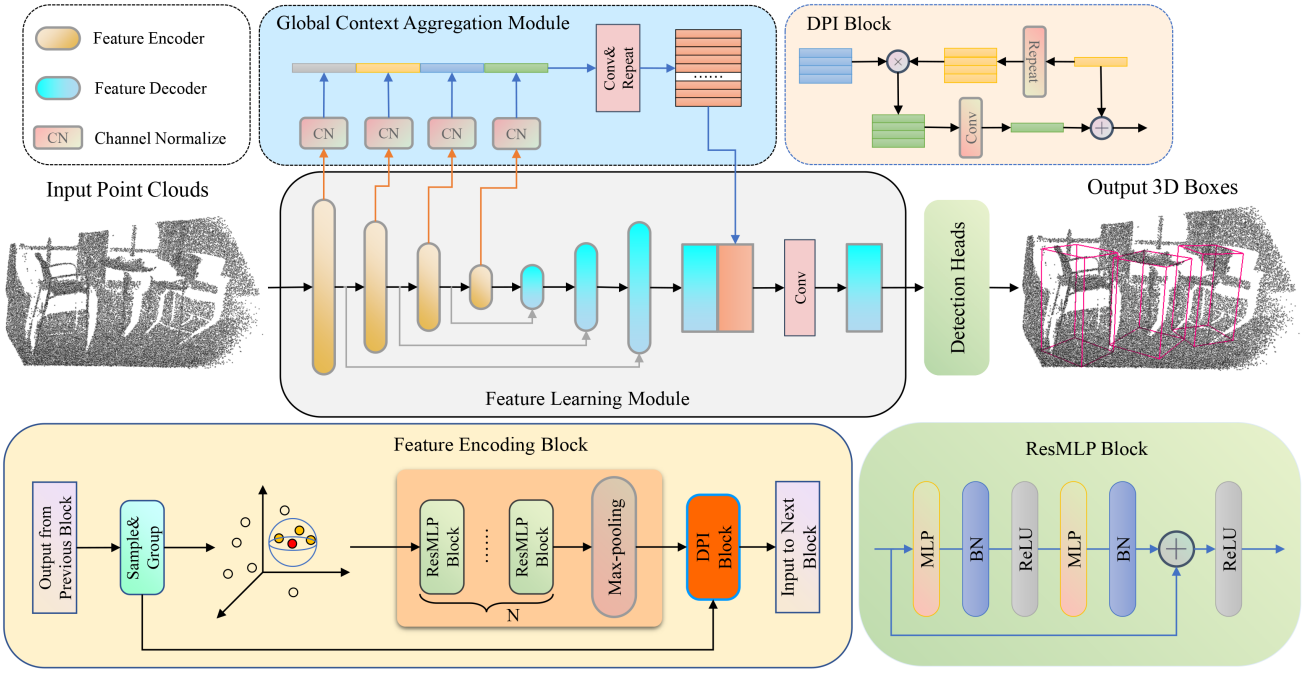


Fig. 2: The pipeline of our 3DLG-Detector. The input scene point clouds are first fed into the feature learning module, in which a feature encoding (FE) block learns the high-level feature representations, and a feature decoding (FD) block recovers the discarded foreground points for accurate prediction. In each FE block, a Sample-and-Group (SG) module samples the seed points and groups the local region features near the seeds to expand the receptive field of the sampled points. Then, a Residual Points Learning (RPL) module further learns and aggregates the deep features. Finally, a Dynamic Points Interaction (DPI) module recovers the pooled local features. The outputs of FE at different levels have various receptive fields, which are concatenated together by a Global Context Aggregation (GCA) module as the global context to incorporate the global information into point features.

an IoU-aware rectification module. In comparison, the two-stage detectors pay more attention to the accuracy of detection. These detectors rely on the post-processing stage to refine the candidate proposals from the previous stage, which often has high demands in computation and memory. Voxel-RCNN [31] exploits voxel RoI pooling to aggregate the voxel features within proposals for further refinement. Part-A^2 [32] proposes a network with part-aware and part-aggregation stages, in which the former predicts proposals and locations of intra-object parts by the part supervision using ground truth boxes, and the latter excavates the spatial relationship of intra-object part locations to refine the proposals.

C. Point-based Object Detection

Point-based methods [33], [34], [35], [8] directly take point clouds, which keeps the original geometric information without any quantitative loss. However, it is challenging to achieve feature extraction due to the sparse and irregular characteristics of point clouds. PointRCNN [7] is a two-stage 3D object detector, which first segments the foreground points and generates a small number of proposals. Then semantic features and local spatial cues are excavated from the proposals for further refinements. VoteNet [8] is a one-stage detector based on the Hough voting algorithm, which identifies instance centroids by voting from the points in a local region. Based on the VoteNet, MLCVNet [22] proposes three context learning modules, respectively Patch-to-Patch Context, Object-to-Object Context, and Global Scene Context to capture the long-range dependencies at different levels. 3DSSD [9] designs a novel fusion sampling strategy, which samples the farthest point according to the feature and Euclidean distance. Pointformer [11] designs a transformer backbone to learn the context-dependent local features and context-aware global representations for 3D object detection.

The aforementioned point-based object detection methods mostly use PointNet++ as the backbone to extract features. However, their insufficient feature learning capacity limits the performance of the detectors. In this paper, we propose a novel feature learning framework for 3D object detection, which excavates and retains the complete local geometric cues by a dynamic points interaction module and captures the global scene context from different-level feature encoders.

III. Methodology

A. Overview

We propose a novel 3D object detector by learning both local and global features. It effectively preserves the local features after the pooling operation by dynamic points interaction and meanwhile learns the global context from multi-scale encoder blocks. As shown in Figure 2, the feature learning

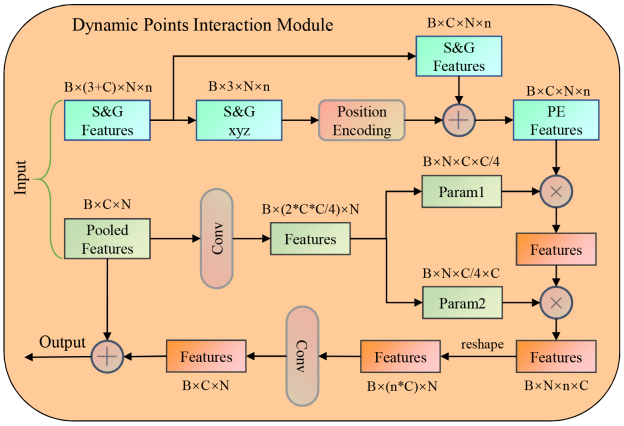


Fig. 3: Illustration of the Dynamic Points Interaction module. The input includes grouped features and pooled features. The grouped features conduct position encoding to embed the position information as the query, and the pooled features are equally split to generate key-value pairs for carrying out dot-product with the query.

module is an encoder-decoder structure. The feature encoder learns the high-level semantic features, in which the Sample-and-Group (SG) module first conducts points down-sampling and feature grouping. Then the Residual MLP (ResMLP) block learns a deeper feature representation from the grouped features. Lastly, the Dynamic Points Interaction (DPI) module takes the grouped features and pooled features as input and exploits grouped features to alleviate feature loss. The feature decoder follows the feature propagation module of PointNet++ to recover the discarded foreground points caused by downsampling. The outputs of the feature encoder blocks are concatenated together as global guidance, making point representations aware of the scene context.

B. Background

The VoteNet [8] is the baseline of our model, which consists of two components: the point features extraction module and the detection head. PointNet++ is the backbone network to extract high-level point features from the input point clouds. The detection head contains a voting module and a proposal module. The voting module takes the features from the previous component as input and regresses the offset from each seed point to the corresponding object center by MLPs, simulating the Hough voting process. The proposal module groups the predicted centers as object candidates to generate the 3D bounding boxes and classified labels.

In the succeeding works, MLCVNet [22] reveals that the contextual information between different objects plays an active role in object recognization. Hence it designs three levels of context modules to learn the contextual information in the voting and proposal stages of VoteNet, respectively are Patch-to-Patch Context(PPC), Object-to-Object Context (OOC), and Global Scene Context (GSC) modules. Besides, Pointformer [11] resorts to the popular Transformer to effectively learn context-aware feature representations. Specifically, a pointformer block consisting of the Local Transformer (LT) module and the Global Transformer (GT) module replaces the set abstract module of PointNet++ for feature extraction. However, these methods all neglect features loss during the pooling stage.

C. Residual Points Learning Module

The residual feed-forward MLPs have been proven to be effective for feature learning in PointMLP [36]. We introduce the Residual Points Learning (RPL) module by stacking the residual MLP blocks to learn the deeper point representations. As shown in Figure 2, the RPL module can be formulated as

$$g_i = \mathcal{A}(\phi(f_{i,j})|j = 1, ..., K), \tag{1}$$

where $f_{i,j}$ is the feature of the jth point near the ith sampled point, and $\phi(\cdot)$ denotes the residual MLP block used to capture the deep feature. Specifically, the residual MLP block includes the mapping function $MLP(x) + x$, in which $MLP(\cdot)$ is combined by full connection, normalization, and activation layers. The aggregation function $\mathcal{A}$ is the max-pooling operation conducted on the features from the last residual MLP block to aggregate the local region features into the sampled point. Similar to ResNet [37], benefiting from the residual connections, the MLPs can easily be extended to dozens of layers for deeper feature representations.

D. Dynamic Points Interaction Module

Although max-pooling operation leads to the loss of part of local geometric features, it is still an indispensable component in dealing with the permutation invariance for point cloud processing. Thus we design a Dynamic Points Interaction (DPI) module to compensate for the feature loss caused by max-pooling without bypassing the max-pooling operation. This dynamic interaction operation is similar to the attention function, which can be described as a mapping from the query term and key-value pairs to the output. The self-attention mechanism usually takes the same or similar features as input to perform QKV operations, focusing on excavating the inner relationship in features. In contrast, our dynamic points interaction (DPI) module takes grouped and pooled features as the query term and key-value pair. The QKV operation simulates the interaction process between pooled seeds and grouped sets and recovers the lost features progressively by continuous queries. Finally, the output is added with the pooled features to avoid pooled features being disturbed from the background point features in the grouped features.

The specific process is shown in Figure 3, where the input of the DPI module includes the previous grouped features $F^g \in R^{B\times(3+C)\times N\times n}$ and pooled features $F^p \in R^{B\times C\times N}$. $F^g = \{c_1, c_2, ..., c_N\}$, where $c_i = \{p_i, p_j, j = 1, ..., n-1\}$ is a grouped points set in the local region. p_i is a sampled point as the centroid of the set and p_j is a neighboring point of p_i within a given radius. Let $\{x_i, f_i\}_t$ denotes the point p_i in the t_{th} point set, where $x_i \in R^3$ represents the coordinates and $f_i \in R^C$ denotes the features of points. Subsequently, the Position Encoding (PE) module takes x_i as input to transform the dimension as the same of f_i and adds x_i to the f_i in

TABLE I
3D OBJECT DETECTION RESULTS ON THE SCANNET V2 VALIDATION SET (LEFT) AND THE SUN RGB-D VALIDATION SET (RIGHT). THE EVALUATION METRIC IS THE MEAN AVERAGE PRECISION WITH 3D IOU THRESHOLDS OF 0.25 AND 0.5. THE RESULTS OF THE COMPETING METHODS ARE QUOTED FROM THEIR PUBLISHED PAPERS OR THE RELEASED CODES.

ScanNet V2	Input	mAP@0.25	mAP@0.5
DSS [38]	Geo + RGB	15.2	6.8
F-PointNet [39]	Geo + RGB	19.8	10.8
GSPN [40]	Geo + RGB	30.6	17.7
3D-SIS [41]	Geo + 5 views	40.2	22.5
VoteNet [8]	Geo only	58.6	33.5
HGNet [42]	Geo only	61.3	34.4
DOPS [43]	Geo only	63.7	38.2
RGNet [44]	Geo only	48.5	26.0
MLCVNet [22]	Geo only	64.7	42.1
3DETR [45]	Geo only	65.0	47.0
PointFormer [11]	Geo only	64.1	42.6
Ours	Geo only	**66.3**	**48.0**

SUN RGB-D	Input	mAP@0.25
DSS [38]	Geo + RGB	42.1
2D-driven [46]	Geo + RGB	45.1
COG [47]	Geo + RGB	47.6
F-PointNet [39]	Geo + RGB	54.0
VoteNet [8]	Geo only	57.7
H3DNet [48]	Geo only	60.1
3DETR [45]	Geo only	59.1
RGNet [44]	Geo only	59.2
MLCVNet [22]	Geo only	59.8
PointFormer [11]	Geo only	61.1
BRNet [49]	Geo only	61.1
Ours	Geo only	**61.6**

an element-wise manner for generating the queries f_q. This process can be formulated as follows

$$f_q = p_f \oplus PE(p_{xyz}), \tag{2}$$

where p_{xyz} represents the coordinates and p_f denotes the features of points.

The pooled features F_p first carry out dimension extension from C to $2 * C * C/m$ by a convolution layer. Then these features are equally split into key-value pairs in the feature channel dimension, respectively are keys $f_k(:\ C * C/m)$ (Param1 in Figure 3) and values $f_v(C * C/m :)$ (Param2 in Figure 3). The reshape operation is adopted on the QKV features to change the arrangement of the feature dimension ($f_q \in R^{B\times N\times n\times C}$,$f_k \in R^{B\times N\times C\times (C/m)}$,$f_v \in R^{B\times N\times (C/m)\times C}$) to fit the succeeding Dot-Product function between queries and key-value pairs. To improve efficiency, we present a bottleneck structure between the key and value. We attempt to reduce the number of feature channels by a factor of m. In this paper, we set m to 4. The whole calculation process can be formulated as follows,

$$y = RB(RB(f_q \odot f_k) \odot f_v), \tag{3}$$

$$o = R(y + F_p), \tag{4}$$

where W_q, W_k, and W_v are reshape operations for query, key, and value, respectively. R and B denote the activation function and the normalization function, respectively.

The prior feature extraction modules in [11], [22] rely on the sophisticated feature extractor to excavate the local geometric information by using attention mechanisms. However, they do not design an effective strategy to preserve the extracted local features. The succeeding aggregation function (e.g., max-pooling) still inevitably deserts part important features. We give full consideration to this issue. The grouped set has redundant and comprehensive local features, in particular, including the part features lost by pooled seed. Hence we take the pooled seed continuously interacts with each point in the corresponding grouped set to acquire the completed local geometric information.

E. Fourier Position Encoding

Position encoding is an essential component of Transformer since it can embed relative or absolute position information of each entry in the input to the corresponding features. For 3D point clouds, the position information of each point still plays a crucial role in describing the local geometric structure of the point clouds.

Inspired by [50], [51], we introduce a Fourier Position Encoding to map the low-dimension coordinates to the higher frequency representations by the heuristic sinusoidal function. Specifically, the function γ maps the coordinates ($xyz \in [0, 1]$) of the input points to the higher dimensional hypersphere with a set of sine-cosine functions

$$\delta_i(v) = (a_i cos(2\pi b_i v), a_i sin(2\pi b_i v)), \tag{5}$$

$$\gamma(v) = [\delta_1(v), ..., \delta_m(v)], v \in \{x, y, z\}, \tag{6}$$

where b_i is the Fourier basis frequency and a_i is the corresponding Fourier series coefficient. For simplicity, we set a_i = 1 and generate b_i by a power function b_i = $T^{i/m}$, $i = 0, ..., m - 1$. The results from the Fourier embedding are concatenated together as position encoding with a dimension of $3m$, and they are further transformed such that their dimension is the same as the corresponding point features.

F. Global Context Aggregation Module

The global context describes the semantic information of the whole scene, which is considerable in inferring the classes of objects as there is a close connection between the scene and objects. Prior works, no matter point-based models using PointNet++ or voxel-based methods using sparse 3D convolution, only extract the high-level feature representation by continuously expanding the receptive field but neglect the global context.

We note that the high-level features include rich semantic information while the low-level features contain the local geometric cues. Hence we propose the global context aggregation (GCA) module to concatenate them together as the global

TABLE II
RESULTS COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE **SCANNETV2** VALIDATION SET. THE EVALUATION METRIC IS THE AVERAGE PRECISION WITH **0.5 IOU THRESHOLD.**

Methods	cab	bed	chair	sofa	table	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
VoteNet [8]	8.1	76.1	67.2	68.8	42.4	15.3	6.4	28.0	1.3	9.5	37.5	11.6	27.8	10.0	86.5	16.8	78.9	11.7	33.5
MLCVNet [22]	11.7	80.8	74.2	70.4	44.8	22.4	17.7	50.0	1.8	24.6	39.8	21.8	40.2	24.6	82.8	29.5	78.7	17.2	40.7
Pointformer [11]	19.0	80.0	75.3	69.0	50.5	24.3	15.0	**41.9**	1.5	26.9	45.1	30.3	41.9	25.3	75.9	35.5	82.9	26.0	42.6
3DETR [45]	**24.4**	79.4	76.5	67.8	53.0	25.7	15.7	41.8	6.1	20.8	**46.8**	26.7	37.8	**40.1**	**96.0**	30.2	84.4	28.3	44.5
Ours	20.0	**80.6**	**79.1**	**77.7**	**61.3**	**34.1**	**21.7**	41.2	**10.8**	**28.3**	39.0	**34.8**	**54.3**	34.6	90.4	**36.1**	**88.3**	**31.9**	**48.0**

context guidance, to promote the ability of feature representations for 3D bounding box regression and object classification. Specifically, we first conduct the channel normalization (CN) to the outputs of each feature encoding block. This operation is to compress the number of the feature channel to k for the succeeding concatenation. The formulation of CN can be summarized as follows:

$$CN(f) = Max - Pooling(MLP(f)), \tag{7}$$

To solve the problem of the inconsistent number of the sampled points from different encoders, the max-pooling function is applied to compress the features to a 1D vector. Subsequently, these vectors representing respective encoders are concatenated together as the global context,

$$g = MLP(Cat[CN(f_i), i = 1, 2, 3, 4]). \tag{8}$$

The global context representations not only promote the message propagation among different objects in the scene, but also benefit the inference in object classification.

IV. EXPERIMENTS

In this section, we conduct extensive experiments on two indoor datasets to evaluate the proposed 3DLG-Detector and compare it with the state-of-the-art 3D object detection methods. In Section IV-A, we introduce the details of datasets and the setup of the model. In section IV-B, we demonstrate the qualitative and quantitative comparison results on indoor datasets. In section IV-C, we analyze the effectiveness of each component in 3DLG-Detector through comprehensive ablation studies. In section IV-D, we introduce the limitation of our model by analyzing several failure cases.

A. Datasets and Implementation Details

We evaluate our method on two indoor datasets, SUN RGB-D [17] and ScanNet V2 [16].

SUN RGB-D [17] is a single-view RGB-D dataset for 3D scene understanding. It contains $\sim$ 5K indoor RGB and depth images annotated with amodal oriented bounding boxes of 37 object categories for training, and the rest $\sim$ 5K RGB-D images for testing. Before feeding the data into the network, depth images are first converted to point clouds by the provided camera parameters. The evaluation metric is the standard mean Average Precision (mAP), and the evaluation is conducted on the 10 most common categories.

ScanNet V2 [16] is a densely annotated dataset consisting of 3D reconstructed meshes, which has rich texture, semantic and geometric information. It contains 1513 indoor scenes captured from hundreds of different rooms, with semantic and instance labels for all the points, as well as 3D object bounding boxes. Compared to the fragmentary scan in SUN RGB-D, the scenes of ScanNet are larger and more complete, so local geometric details of objects are well captured. The vertices of the meshes in the dataset are sampled as point clouds.

Data augmentation. To reduce computational complexity, we randomly down-sample each point cloud as input, $i.e.$, 20,000 points for the SUN RGB-D dataset and 40,000 points for the ScanNet dataset respectively. The height attribute of each point is also included as an extra feature to feed into the network. To augment the training data, we apply randomly flipping, rotating, and scaling operations to the point clouds, following VoteNet [8].

Training details. Our model is implemented with PyTorch on an NVIDIA GeForce RTX 3060 GPU and optimized by the Adam optimizer in an end-to-end manner. For ScanNet V2, we set the initial learning rate to 1e-3 and weight decay to 1e-1. The total training epochs are 48, and the learning rate continuously decreases in the 12, 24, and 36 epochs by $5\times$. For the other dataset SUN RGB-D, we set the base learning rate to 1e-3 and weight decay to 5e-2. The total epochs are 36, and the learning rate continuously decreases in the 12 and 24 epochs by $5\times$.

B. Comparisons with the State-of-the-art Methods

We compare our method with the related works, which can be divided into three groups: early methods [41], [46], [39], [40], [47] that locate 3D objects via 3D-2D queries, voting-based methods that excavate informative local representation such as VoteNet [8] and its successors [42], [43], [49], [48], and attention-based methods [44], [45], [11], [22] that explore the relationships between the local objects and point clusters. The results are reported in Table I and Table II. The bold texts denote the best results under the corresponding metrics.

Quantitative results. All results of comparison experiments are summarized in Table I. We can observe that 3DETR [45] has the highest mAP among the competing methods on the ScanNet dataset, while our method still outperforms 3DETR in both metrics on the ScanNet V2 validation sets (+1.3% $mAP@0.25$, +1% $mAP@0.5$). Note that $mAP@0.5$ is quite a challenging metric since it requires more than 79% coverage area in each dimension of the bounding box. Some methods like HGNet [42] only achieve decent performance under the metric $mAP@0.25$ but perform exceptionally poorly in terms of the $mAP@0.5$ metric. Our model has the highest accuracy

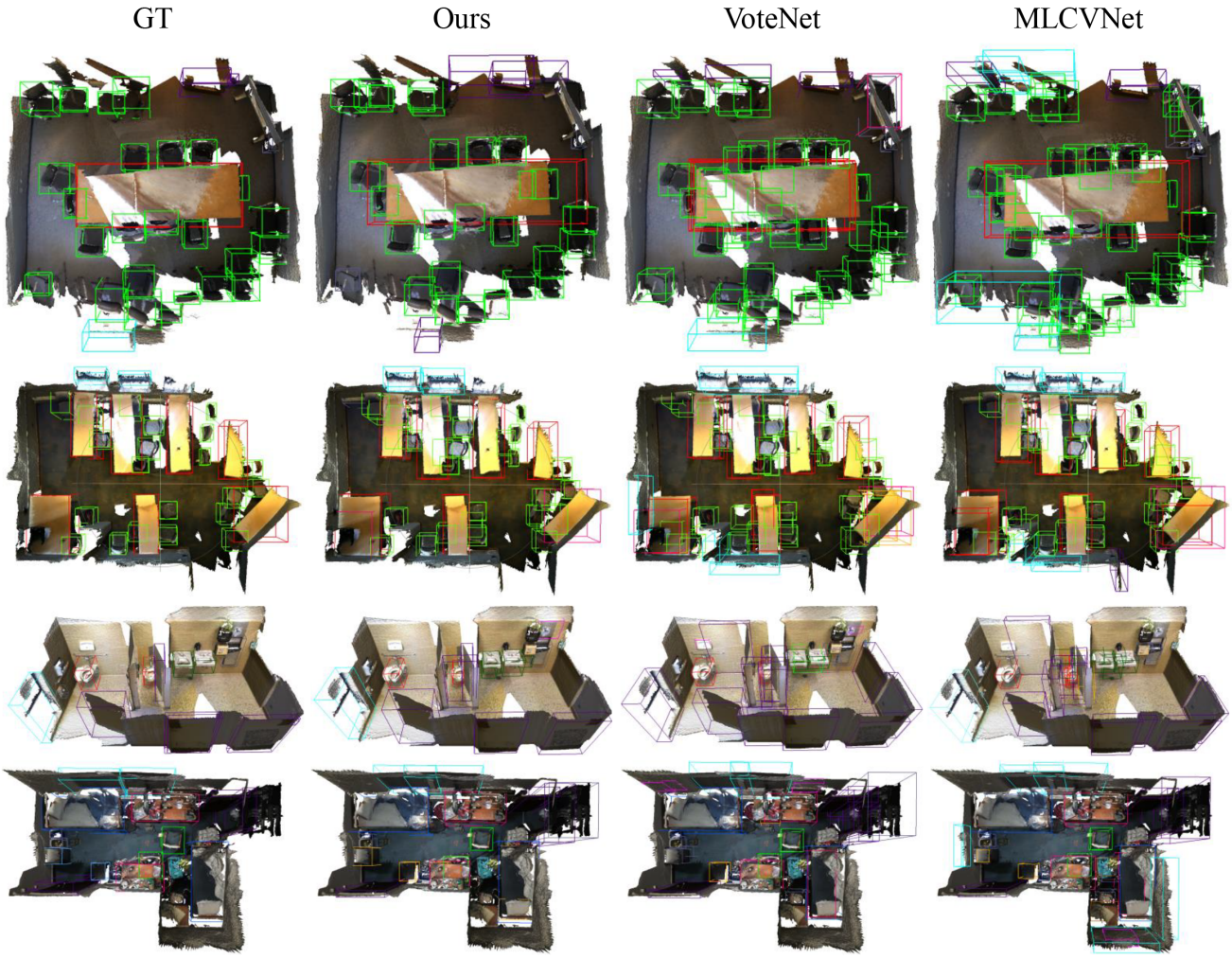


Fig. 4: The qualitative results of different 3D object detection methods on the ScanNet V2 validation sets.

in object location, so it retains significant performance. The ScanNet dataset consists of reconstructed meshes that cover complete objects in larger areas, while the SUN RGB-D dataset contains the single-view RGB-D images where point clouds projected from the depth map have fragmentary objects and smaller areas. The different characteristics result in the inability of many methods to perform consistently well on both datasets. For example, MLCVNet [22] performs well on the ScanNet dataset but achieves poor results on the SUN RGB-D dataset, while RGNet [44] is the opposite. Our method has also achieved impressive performance on the SUN RGB-D dataset, indicating it has a strong generalization capability to deal with different scenes.

The comparison results on the ScanNet dataset in terms of $mAP@0.5$ are shown in Table II. Our method achieves the best performance in 13 out of the 18 categories. Especially for the tabular objects like pictures and windows, whose neighborhoods mostly are background points. Other methods cannot detect them due to that the pooling operation aggregates too many background features but discards important object features. Nevertheless, our dynamic points interaction module preserves object features, which improves the detection accuracy by 4% AP and 4.1% on windows and pictures.

Qualitative results. We visualize the representative detection results from the ScanNet dataset and SUN RGB-D dataset in Figure 4 and Figure 5, from which we can observe that the VoteNet [8] and MLCVNet [22] have wrong detection regarding object number and category. For example, in the first row of Figure 4, VoteNet [8] and MLCVNet [22] recognize many wrong chairs on the table and in the wall. In contrast, enhanced by the proposed DPI and GCA modules, our model achieves more accurate bounding boxes in terms of both location and category.

C. Ablation Study

Residual points learning module. We first evaluate the effect of the number of ResMLP blocks in the Residual Points Learning (RPL) module on feature learning. We change the depth of the RPL module by setting the number of ResMLP blocks to 0, 1, 2, and 3, respectively. 0 block means using the

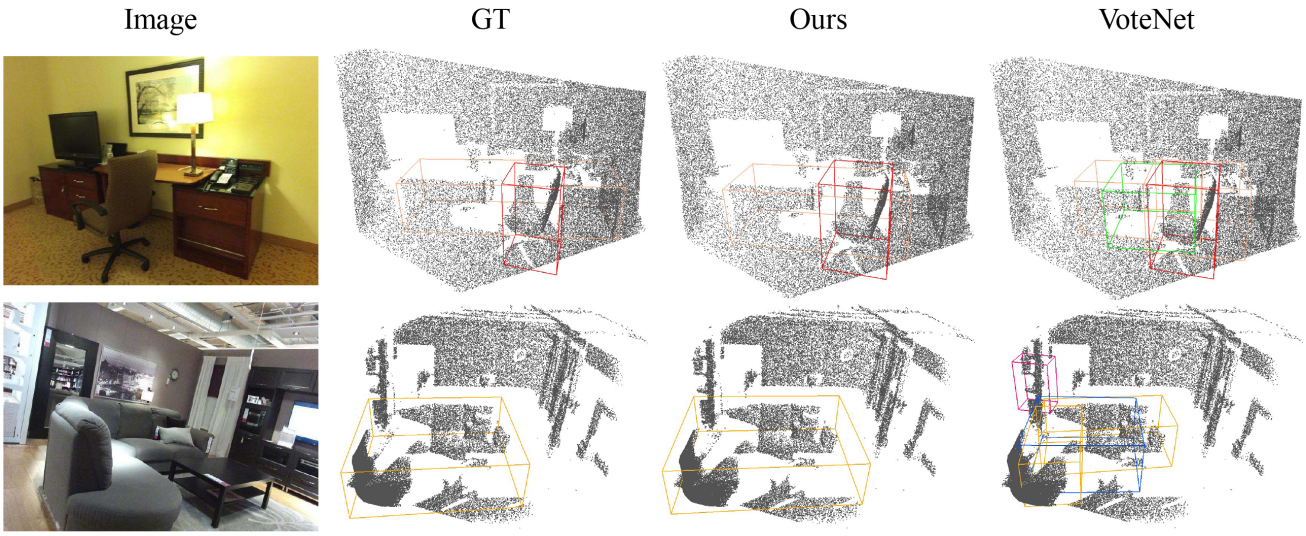


Fig. 5: The qualitative results of different 3D object detection methods on the SUN RGB-D validation sets.

TABLE III
ABLATION EXPERIMENTS REGARDING THE NUMBER OF RESMLP BLOCKS IN THE RPL MODULE. NOTE '0*' DENOTES USING THE TRADITIONAL MLPS.

ResMLP	ScanNet V2	
	mAP@0.25	mAP@0.5
0* block	63.9	45.4
1 block	64.9	47.1
2 blocks	**66.3**	**48.0**
3 blocks	65.2	47.0

TABLE IV
ABLATION EXPERIMENTS REGARDING THE NUMBER OF RESMLP BLOCKS IN THE RPL MODULE. '-DPI' MEANS THE 3DLG-DETECTOR WITHOUT THE DPI MODULE, '-GCA' INDICATES THE 3DLG-DETECTOR WITHOUT THE GCA MODULE.

ResMLP	ScanNet V2		SUN RGB-D	
	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
VoteNet	58.6	33.5	57.7	32.9
-DPI	64.7	45.4	58.5	35.1
-GCA	65.3	46.1	60.1	37.6
3DLG-Detector	**66.3**	**48.0**	**61.6**	**38.5**

traditional MLP layer for feature extraction. The experiment results are reported in Table III, from which we observe an increase in detection performance as the RPL module becomes deeper. However, merely increasing the number of ResMLP blocks would not always lead to better performance. When setting the number of ResMLP blocks to 3, the detection accuracy decreases $mAP@0.25$ by 1.1% and $mAP@0.5$ by 1.0%. In this work, two ResMLP blocks achieve the best performance.

Dynamic points interaction module. DPI module is the essential component in our model, which significantly improves detection accuracy. The quantitative results are reported in Table IV. We can see that without the DPI module, the performance drops 2.6% and 3.4% in terms of $mAP@0.5$ on the ScanNet and SUN RGB-D validation sets, respectively. The visualization of the object detection results is in the first two rows of Figure 7. After removing the DPI module,

TABLE V
ABLATION EXPERIMENTS ABOUT COMPARING DPI MODULE WITH SELF-ATTENTION.

ResMLP	ScanNet V2		SUN RGB-D	
	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
self-attention	64.7	45.0	60.1	36.3
DPI	**66.3**	**48.0**	**61.6**	**38.5**

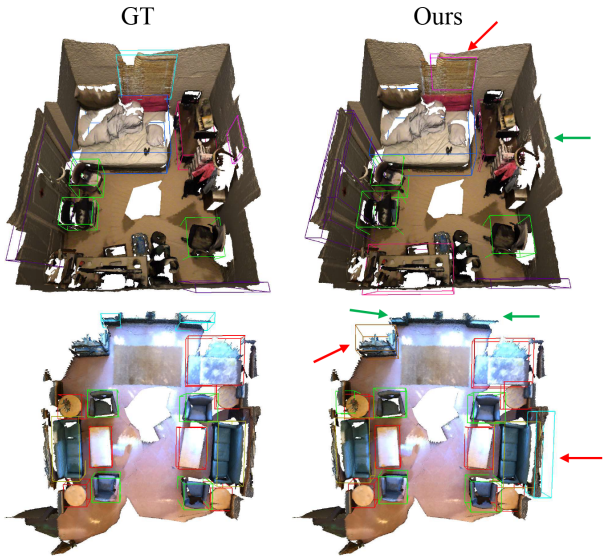


Fig. 6: Examples of failure cases on the ScanNet V2 dataset. The red arrows denote the false positive bounding boxes, and the green arrows indicate the missed objects.

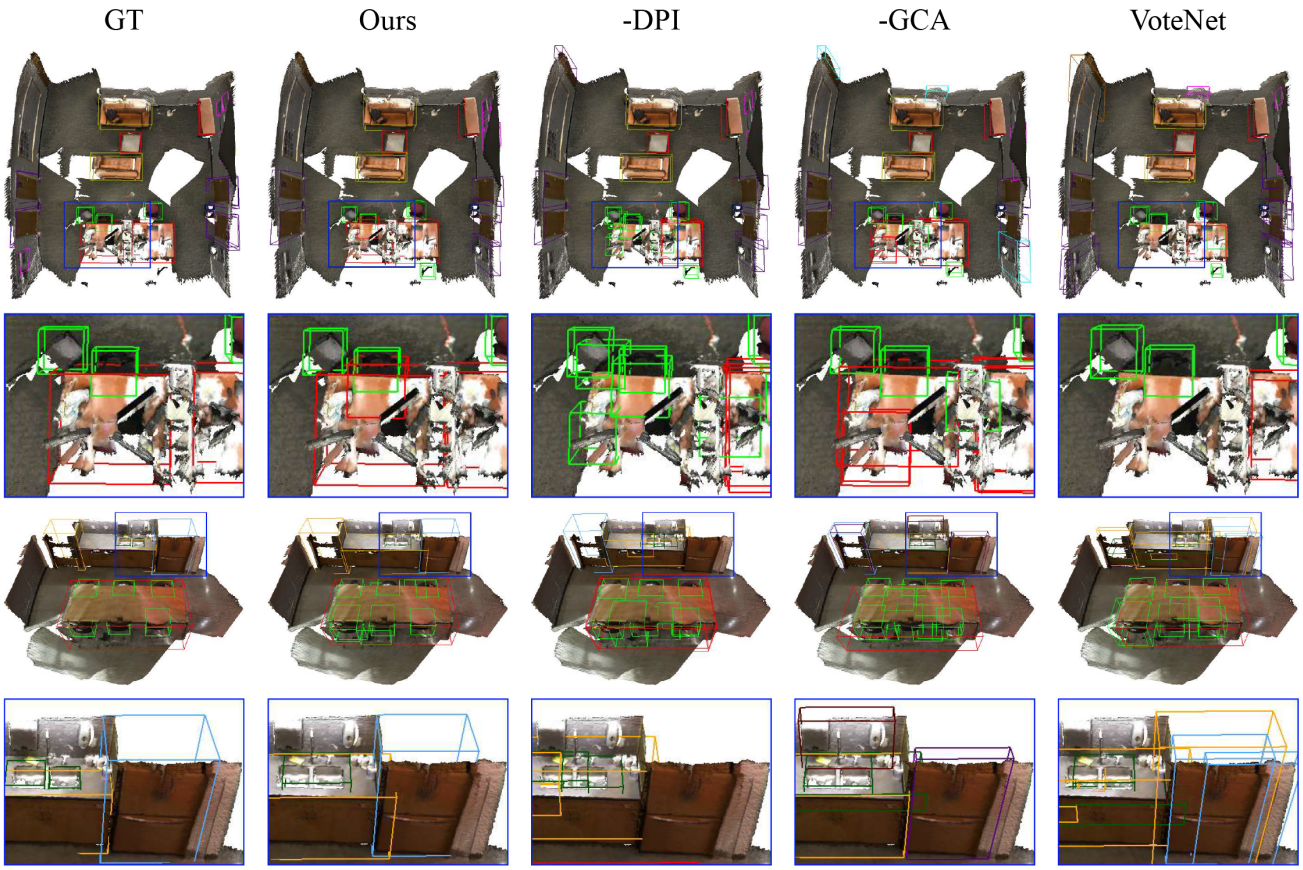


Fig. 7: The visual results of ablation experiments on the ScanNet V2 validation sets. '-DPI' denotes the 3DLG-Detector without the DPI module, '-GCA' indicates the 3DLG-Detector without the GCA module. The first and third rows demonstrate the whole scenes, and the second and fourth rows are close-up views.

several chairs (green boxes) instead of the table (red boxes) are incorrectly detected. This is also due to that the pooling operation aggregates features from the neighbor regions instead of object features. The sampled points of the table integrate with the points from the chairs beside the table, leading to the error of recognizing the table as several chairs. Our DPI module enables the grouped features to interact with the pooled features to preserve local features and thus ensures correct table detection.

Differences between DPI and self-attention. Although our dynamic points interaction module is similar to the self-attention in formulation, the inputs are completed different. Self-attention takes same or similar features as input while DPI takes grouped and pooled features as the query term and key-value pair. We apply the self-attention on the pooled features to replace the DPI module. We can find the performance has decreased by 3.0% and 2.2% in terms of $mAP@0.5$ on the ScanNet and SUN RGB-D validation sets, respectively. The reason is that self-attention only excavates the internal relationship of features while neglecting to introduce the external cues to compensate for the feature loss.

Global context aggregation module. GCA module plays a substantial role in learning the global contextual information for 3D object detection. As shown in Table IV, removing the GCA module causes the detection accuracy to decrease by 1.9% and 0.9% in terms of $mAP@0.5$ on the ScanNet and SUN RGB-D validation sets, respectively. The visualization results are shown in the last two rows of Figure 7. The fridge near the sink is wrongly detected as a door by the model without the GCA module. The global scene context encodes the multi-scale features to generate scene context information that helps to enhance object detection.

D. Limitations

Although 3DLG-Detector has demonstrated notable improvement on two indoor datasets, it still does not perform well for a few tricky scenes. Two such failure cases are presented in Figure 6. The common failures are false-positive bounding boxes of objects (red arrows in Figure 6) and the missed object detection (green arrows in Figure 6). We can see that the picture and the window in the smooth wall are the most challenging to be detected because they are too thin and clung to the wall. The false-positive bounding boxes also arise when several objects have similar shapes, for which the global context cannot distinguish between them. It is worth noting that these common failures are equally problematic for most SOTA methods. Additionally, we have not resolved the

over smoothing phenomenon caused by the residual points learning module. The residual connection can deepen the MLPs from several to dozens of layers to learn better deep feature representations. However, the performance may not be enhanced and may degrade when the number of layers increases largely.

V. Conclusion

We have presented a novel framework to improve voting-based 3D object detection networks. Our approach enhances the learning of both local and global features by introducing three different modules to the networks. The RPL module first learns the deep local feature representation, and then the DPI module captures the complete local geometric features. The GCA module constructs global contextual information from multi-scale feature encoders, thus enriching global features. Extensive experiments have demonstrated the effectiveness of the proposed approach.

Compared to prior works that propose sophisticated feature extractors to excavate detailed local geometric information, our work takes a different path to preserve the extracted features. The flourishing extractors have saturated performance in describing local geometric information while designing effective feature retention strategies has been rarely studied. We believe our work can promote the research in feature retention.

References

[1] C. Yi, D. Lu, Q. Xie, S. Liu, H. Li, M. Wei, and J. Wang, "Hierarchical tunnel modeling from 3d raw lidar point cloud," *Computer-Aided Design*, vol. 114, pp. 143–154, 2019.

[2] Q. Wu, H. Yang, M. Wei, O. Remil, B. Wang, and J. Wang, "Automatic 3d reconstruction of electrical substation scene from lidar point cloud," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 143, pp. 57–71, 2018.

[3] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[4] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 689–12 697.

[5] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[6] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[7] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 770–779.

[8] C. R. Qi, O. Litany, K. He, and L. Guibas, "Deep hough voting for 3d object detection in point clouds," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9276–9285.

[9] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3dssd: Point-based 3d single stage object detector," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 037–11 045.

[10] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1951–1960.

[11] X. Pan, Z. Xia, S. Song, L. E. Li, and G. Huang, "3d object detection with pointformer," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7459–7468.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[13] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.

[14] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 282–298.

[15] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 449–14 458.

[16] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 2432–2443.

[17] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 567–576.

[18] B. Graham, M. Engelcke, and L. v. d. Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9224–9232.

[19] S. Wang, S. Suo, W.-C. Ma, A. Pokrovsky, and R. Urtasun, "Deep parametric continuous convolutional neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2589–2597.

[20] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, oct 2019.

[21] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 526–10 535.

[22] Q. Xie, Y. Lai, J. Wu, Z. Wang, Y. Zhang, K. Xu, and J. Wang, "Mlcvnet: Multi-level context votenet for 3d object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 10 444–10 453.

[23] J. Noh, S. Lee, and B. Ham, "Hvpr: Hybrid voxel-point representation for single-stage 3d object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 600–14 609.

[24] L. Du, X. Ye, X. Tan, J. Feng, Z. Xu, E. Ding, and S. Wen, "Associate-3ddet: Perceptual-to-conceptual association for 3d point cloud object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 326–13 335.

[25] H. Yi, S. Shi, M. Ding, J. Sun, K. Xu, H. Zhou, Z. Wang, S. Li, and G. Wang, "Segvoxelnet: Exploring semantic context and depth-aware features for 3d vehicle detection from point cloud," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 2274–2280.

[26] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 870–11 879.

[27] M. Ye, S. Xu, and T. Cao, "Hvnet: Hybrid voxel network for lidar based 3d object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1628–1637.

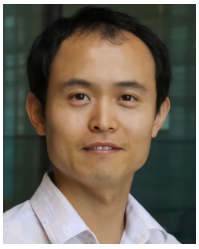
[28] W. Zheng, W. Tang, S. Chen, L. Jiang, and C.-W. Fu, "Cia-ssd: Confident iou-aware single-stage object detector from point cloud," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3555–3562.

[29] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[30] J. S. Hu, T. Kuai, and S. L. Waslander, "Point density-aware voxels for lidar 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8469–8478.

[31] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1201–1209.
[32] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2647–2664, 2021.
[33] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong, "Group-free 3d object detection via transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2929–2938.
[34] I. Misra, R. Girdhar, and A. Joulin, "An end-to-end transformer model for 3d object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2886–2897.
[35] Q. Xie, Y.-K. Lai, J. Wu, Z. Wang, D. Lu, M. Wei, and J. Wang, "Venet: Voting enhancement network for 3d object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3692–3701.
[36] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual mlp framework," in *International Conference on Learning Representations*, 2021.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[38] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 808–816.
[39] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.
[40] L. Yi, W. Zhao, H. Wang, M. Sung, and L. J. Guibas, "Gspn: Generative shape proposal network for 3d instance segmentation in point cloud," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3942–3951.
[41] J. Hou, A. Dai, and M. Nießner, "3d-sis: 3d semantic instance segmentation of rgb-d scans," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4416–4425.
[42] J. Chen, B. Lei, Q. Song, H. Ying, D. Z. Chen, and J. Wu, "A hierarchical graph network for 3d object detection on point clouds," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 389–398.
[43] M. Najibi, G. Lai, A. Kundu, Z. Lu, V. Rathod, T. Funkhouser, C. Pantofaru, D. Ross, L. S. Davis, and A. Fathi, "Dops: Learning to detect 3d objects and predict their 3d shapes," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 910–11 919.
[44] M. Feng, S. Z. Gilani, Y. Wang, L. Zhang, and A. Mian, "Relation graph network for 3d object detection in point clouds," *IEEE Transactions on Image Processing*, vol. 30, pp. 92–107, 2021.
[45] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.
[46] J. Lahoud and B. Ghanem, "2d-driven 3d object detection in rgb-d images," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4632–4640.
[47] Z. Ren and E. B. Sudderth, "Three-dimensional object detection and layout prediction using clouds of oriented gradients," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1525–1533.
[48] Z. Zhang, B. Sun, H. Yang, and Q. Huang, "H3dnet: 3d object detection using hybrid geometric primitives," in *European Conference on Computer Vision*. Springer, 2020, pp. 311–329.
[49] B. Cheng, L. Sheng, S. Shi, M. Yang, and D. Xu, "Back-tracing representative points for voting-based 3d object detection in point clouds," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8959–8968.
[50] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.
[51] C. He, R. Li, S. Li, and L. Zhang, "Voxel set transformer: A set-to-set approach to 3d object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8417–8427.

Baian Chen is now pursuing his PhD degree at Nanjing University of Aeronautics and Astronautics (NUAA), China. He received his B.Sc. degree from China University of Mining and Technology. His research interests include 3D vision and learning-based geometry processing.



Liangliang Nan received the B.S. degree in material science and engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2003, and the Ph.D. degree in mechatronics engineering from the Graduate University of the Chinese Academy of Sciences, Beijing, China, in 2009.

From 2009 to 2013, he was an Assistant and then an Associate Researcher at the Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences, Beijing. From 2013 to 2018, he worked at the Visual Computing Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, as a Research Scientist. He is currently an Assistant Professor with the Delft University of Technology (TU Delft), Delft, The Netherlands, where he is leading the AI Laboratory on 3D Urban Understanding (3DUU). His research interests include computer graphics, computer vision, 3D geoinformation, and machine learning.

Haoran Xie (Senior Member, IEEE) received the Ph.D. degree in Computer Science from City University of Hong Kong, Hong Kong SAR, and Ed.D. degree in Digital Learning from University of Bristol, UK. He is currently an Associate Professor at the Department of Computing and Decision Sciences, Lingnan University, Hong Kong SAR. His research interests include artificial intelligence, big data, and educational technology. He has published 300 research publications, including 159 journal articles such as IEEE TPAMI, IEEE TKDE, IEEE TAFFC, IEEE TCVST, and so on. He is the Editor-in-Chief of Natural Language Processing Journal, Computers & Education: Artificial Intelligence and Computers & Education: X Reality. He has been selected as The World Top 2% Scientists by Stanford University.

Dening Lu received his BSc and MSc degrees in Electrical Engineering, both from the Nanjing University of Aeronautics and Astronautics (NUAA), China in 2018 and 2021, respectively. He is currently pursuing his Ph.D. degree in Systems Design Engineering with the Geospatial Sensing and Data Intelligence Group at the University of Waterloo, Canada. His research interests include 3D point cloud processing and deep learning. He has published papers in the IEEE Transactions on Instrumentation and Measurement and ICCV.

Fu Lee Wang (SM'15) received the B.Eng. degree in computer engineering and the M.Phil. degree in computer science and information systems from the University of Hong Kong, Hong Kong, and the Ph.D. degree in systems engineering and engineering management from the Chinese University of Hong Kong, Hong Kong. Prof. Wang is the Dean of the School of Science and Technology, Hong Kong Metropolitan University, Hong Kong. He has over 250 publications in international journals and conferences and led more than 20 competitive grants with a total greater than HK$20 million. His current research interests include educational technology, information retrieval, computer graphics, and bioinformatics. Prof. Wang is a fellow of BCS and HKIE and a Senior Member of ACM. He was the Chair of the IEEE Hong Kong Section Computer Chapter and ACM Hong Kong Chapter.

Mingqiang Wei received his Ph.D degree (2014) in Computer Science and Engineering from the Chinese University of Hong Kong (CUHK). He is a full Professor at the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA). Before joining NUAA, he served as an assistant professor at Hefei University of Technology, and a postdoctoral fellow at CUHK. He was a recipient of the CUHK Young Scholar Thesis Awards in 2014. He is now an Associate Editor for ACM TOMM, The Visual Computer (TVC), Journal of Electronic Imaging, and a leading Guest Editor for IEEE Transactions on Multimedia, and TVC. He has published 140 research publications, including TPAMI, SIGGRAPH, TVCG, CVPR, ICCV, et al. His research interests focus on 3D vision, computer graphics, and deep learning.