# Spatial and Semantic Information Enhancement for Indoor 3D Object Detection

Chunmei Chen
School of Information Engineering, Southwest University of Science and Technology, China
ccm@swust.edu.cn

Zhiqiang Liang
School of Information Engineering, Southwest University of Science and Technology, China
lzq1799883013@163.com

Haitao Liu
School of Information Engineering, Southwest University of Science and Technology, China
2316938781@qq.com

Xin Liu
School of Information Engineering, Southwest University of Science and Technology, China
2690961460@qq.com

**Abstract:** *Object detection technology is one of the key technologies for indoor service robots. However, due to the various types of objects in the indoor environment, the mutual occlusion between the objects is serious, which increases the difficulty of object detection. In view of the difficult challenges of object detection in the indoor environment, we propose an indoor three-dimensional object detection based on deep learning. Most existing 3D object detection techniques based on deep learning lack sufficient spatial and semantic information. To address this issue, the article presents an indoor 3D object detection method with enhanced spatial semantic information. This article proposes a new (Edge Convolution+) EdgeConv+, and based on it, a Shallow Spatial Information Enhancement module (SSIE) is added to Votenet. At the same time, a new attention mechanism, Convolutional Gated Non-Local+ (CGNL+), is designed to add Deep Semantic Information Enhancement module (DSIE) to Votenet. Experiments show that on the ScanNet dataset, the proposed method is 2.4% and 2.1% higher than Votenet at mAP@0.25 and mAP@0.5, respectively. Furthermore, it has strong robustness to deal with sparse point clouds.*

**Keywords:** *3D object detection, edge convolution, attention mechanism, spatial information, semantic information.*

*Received February 10, 2023; accepted May 11, 2023*
*https://doi.org/10.34028/iajit/20/5/15*

## 1. Introduction

Due to the intensification of the ageing population and the shortage of labour, more service robots have been c reated to service people. The indoor environment is an important environment to consider when designing ser vice robots.

When people are in an indoor environment, they often use objects to describe the spatial environment and take corresponding actions based on them. Referring to the people's environmental perception method, the article studies object detection technology, to improve the robot's perception ability. Object detection technology for robots is of research importance in areas such as localization and navigation of indoor mobile robots, assisted navigation for visually impaired groups and security robots [9, 21].

Traditional object detection methods generally traverse the region to be detected through sliding windows of different scales and aspect ratios and then use an exhaustive strategy to frame all the positions of the area to be detected that may contain the detection object. Artificially designed feature extraction operators such as Scale-Invariant Feature Transform (SIFT) [15], Histogram of Oriented Gradients (HOG) [4], Local Binary Pattern (LBP) [16], etc., are used to extract features. Finally, classification is performed using classifiers Deformable Part Model (DPM) [7], Support Vector Machine (SVM) [2], etc.,) based on the extracted feature. However, artificially designed feature operators are very limited for indoor environments, and their performance is very poor in terms of detection accuracy and robustness.

Recently, object detection methods based on deep learning have become the dominant approach in object detection tasks. Object detection methods based on 2D images by deep learning are mainly divided into two categories. One category is the two-stage approach based on region proposal, represented by the R-CNN series [8, 10, 23], etc., Instead of directly applying feature extraction and classification on the entire image, they first generate a set of region proposals through selective search or other techniques. These region proposals are then individually processed to extract features and classify the presence of objects. The other is a regression-based single-stage method represented by the YOLO series [13, 22, 26], Single Shot MultiBox Detector (SSD) [14], etc. These methods take a different approach by formulating the object detection problem as a regression task. They divide the input image into a grid or anchor boxes and directly predict the presence, class labels, and bounding box coordinates of objects within each grid or anchor box. However, the application of 2D object detection in robotics is greatly limited by the fact that 2D images themselves are projections of the 3D world onto a 2D plane, lacking some structural information, and that they are heavily influenced by various factors in the environment. Therefore, researchers have proposed a 3D object detection method based on deep learning. 3D point cloud data is widely used in 3D object detection. Compared to 2D images, 3D point cloud data has richer structural information and is less influenced by external factors in the environment.

The input to a general neural network model is a fixed serialised data format, while 3D point cloud data are disordered, making it difficult to process the point cloud data. Early 3D object detection methods based on point cloud data are mostly indirect processing of point cloud data. For example, Complex-YOLO [24] and VeloFCN [12] convert 3D point clouds into 2D images by projection, then achieve object detection through the above 2D object detection network, and finally recover the geometric pose of the object in 3D space. Another common indirect processing method is to convert point cloud data into 3D voxels, such as Vote3Deep [6] and VoxelNet [32], which convert point cloud data into 3D voxels, and then realize object detection by 3D convolution. However, the indirect processing of point clouds, whether by conversion to 2D images or 3D voxels, can lead to the loss of part of the 3D feature information during network training and even introduce errors.

In 2017, PointNet [19], a point cloud neural network was proposed, opened up the direct processing of point cloud data. PointNet solves the rotation problem of the point cloud by learning a spatial transformation network, which generates a spatial transformation matrix to transform the point cloud to a direction that is more conducive to classification and segmentation. At the same time, a maximum pooling function is used to solve the disorder of point clouds. PointNet++ [20] is an extension of PointNet that introduces additional layers, including a Sampling Layer (SL), Grouping Layer (GL), and Feature Extraction Layer (FEL). These layers aim to improve the extraction of point cloud information compared to the original PointNet. SL selectively downsamples the point cloud. The GL combines neighboring points to form locally regions. FEL utilizes fully connected layers to extract feature representations from the point cloud, enhancing its representation capability. Neither Pointnet nor pointnet++ performs well in the extraction of point cloud information. Dynamic Graph Convolutional Neural Network (DGCNN) [28], proposed by Wang, is a neural network based on dynamic graph convolution. It computes edge features between each point and its neighboring points to capture local information between points. This network effectively captures local features of point clouds. Votenet [17], a 3D object detection network based on deep learning and Hough voting, using only point cloud information as network input. Xie proposed Multi-Level Context VoteNet (MLCVNet) [29], which successfully introduces multi-level contextual information into Votenet. By considering contextual cues at multiple levels, MLCVNet improves the understanding and representation of the surrounding environment, leading to more accurate object detection results. Object DGCNN [27] proposes a 3D object detection architecture on point clouds. The method models 3D object detection as message passing on a dynamic graph, generalizing the DGCNN framework to predict a set of objects. We have studied them and found that there is more room for their improvement in terms of detection accuracy and robustness.

The article proposes an indoor 3D object detection method with spatial semantic information enhancement. The main contributions of the method are summarized as follows:

- A new edge convolution method, Edge Convolution+ (EdgeConv+), and a new attention mechanism, Convolutional Gated Non-Local+ (CGNL+), are proposed, which are more suitable for indoor object detection.
- A Shallow Spatial Information Enhancement module (SSIE) and a Deep Semantic Information Enhancement module (DSIE) are added to Votenet.
- The experimental results show that the EdgeConv+ and CGNL+ proposed have better performance than the original EdgeConv and CGNL. At the same time, SSIE and DSIE based on them improve Votenet's ability to extract spatial information and semantic information. On the ScanNet dataset, the proposed method is 2.4% and 2.1% higher than Votenet at mAP@0.25 and mAP@0.5, respectively.

The article is structured as follows. Section 1 reviews the relevant methods for object detection. Section 2 introduces the method presented in detail. Section 3 presents the experimental results. And section 4 is the conclusion of the article.

## 2. Method

### 2.1. Votenet

The proposed method uses the Votenet as the underlying framework, which is an end-to-end trainable 3D object detection network. Votenet consists of three main modules: feature extraction, vote, and object proposal and classification, as shown in Figure 1.

- Feature Extraction: votenet utilizes PointNet++ as the backbone network for sampling and feature extraction of the input point cloud. PointNet++ is a deep neural network architecture designed for processing unordered point clouds. It hierarchically samples points using a set abstraction operation to capture local and global context information.
- Vote: the voting module in Votenet simulates the

hough voting process. It takes the seed points from the feature extraction as input. These seed points are fed into Multi-Layer Perceptron (MLP) that regresses object centers, generating voting points.

- Object Proposal and Classification: the voting points are then sampled and grouped to form voting clusters. These voting clusters are then fed into an MLP layer for final bounding box regression and classification.

## 2.2. Improved Votenet

In object detection, shallow features have stronger spatial information, more detailed features and clear contours, but their semantic information is lower. Deep features have stronger semantic information, but the resolution is low, losing the detailed features of the object and the perception of details is poor. Considering the advantages and disadvantages of the deep and shallow features, the proposed method adopts the idea of cross-region fusion. As shown in Figure 1, SSIE, DSIE and a feature fusion module are added to Votenet.

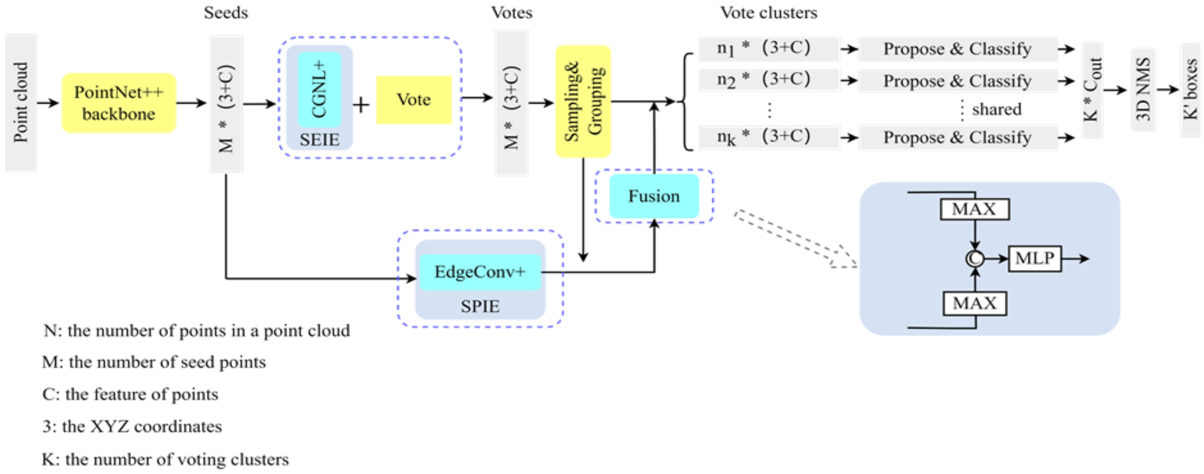


Figure 1. The overall flow of the network based on Votenet, SSIE, and DSIE module are added, and spatial semantic information is efficiently fused.

SSIE closely follows the feature extraction block pointnet++, which first constructs the local point graph G of the seed points processed by pointnet++ shown in Figure 1, and then learns the spatial information by EdgeConv+, a new edge convolution the article designed. In this way, SSIE introduces rich spatial information to the network while solving the problem that PointNet++ ignores the local geometric spatial relationships between points during the processing of point clouds. DSIE consists of CGNL+, a new attention mechanism the article designed, which models the semantic correlation between each seed point and other seed points. The DSIE greatly optimizes the voting module of Votenet, and improves the network's ability to extract semantic information from point cloud.

To achieve an efficient fusion of shallow spatial information and deep semantic information, the proposed method designs a feature fusion module. The fusion module in Figure 1, in which the spatial information extracted by SSIE is fused with the semantic information extracted by DSIE through a skip branch. The output of the fusion module is $F_{fusion}$

$$F_{fusion}=MLP(max(G)+max(S)) \tag{1}$$

Where $G$ is the spatial information, and $S$ is the semantic information. By using maximum pooling, $G$ and $S$ are added to obtain a new global feature vector. Finally, the MLP layer is applied to further aggregate the global feature.

Subsequently, $F_{fusion}$ is combined with the original output of Votenet. Therefore, the feature information entered into the object proposal and classification module contains fusion information of spatial and semantic information, as well as fusion information of global and local information. In this way, the accuracy and robustness of the network are improved.

## 2.3. Shallow Spatial Information Enhancement Module

The shallow spatial information enhancement module consists of an edge convolution EdgeConv+. Traditional network models deal with each point in the point cloud individually, ignoring the local geometric spatial relationships between the points. To address this problem, DGCNN proposes an edge convolution EdgeConv. Wang and Solomon [28] of DGCNN construct a local point graph G of the point cloud, as shown in Figure 2. $x_i$ is a point in the point cloud, $x_{ij}$ is the $K$ neighbours of $x_i$ calculated by K-Nearest Neighbors (KNN) [1], and $e_{ij}$($e_{ij}$ =$x_{ij}$-$x_i$)is the edge feature from $x_{ij}$ to its central node $x_i$. The DGCNN considers both $x_i$ and $e_{ij}$ through EdgeConv ($x_i$ considers the global information of the point itself, and $e_{ij}$ considers the local geometric spatial information between points). The fusion of global and local information, which improves the accuracy of point cloud classification and segmentation.

Although EdgeConv has excellent performance in point cloud classification and segmentation tasks, EdgeConv does not show high performance in point cloud object detection. It is found that in addition to $e_{ij}$ information and $x_i$ information, $x_{ij}$ information is valuable in object detection. To improve the performance of EdgeConv in object detection, the article proposes a new edge convolution EdgeConv+ based on EdgeConv. Similar to DGCNN, a local directed graph $G$ is constructed. The difference is that EdgeConv+ learns a more complete shape representation of the local point cloud than EdgeConv, as shown in Figure 3. For a local directed graph $G$ in point cloud, EdgeConv only considers the blue parts $x_i$ and $e_{ij}$ in feature extraction, while EdgeConv+ considers $x_i$, $e_{ij}$ and $x_{ij}$ simultaneously. By the way, the geometric position relationship between neighborhood points is used to describe the shape characteristics of the object, which can effectively gather the spatial information in the local point graph to achieve the effect of spatial information enhancement.

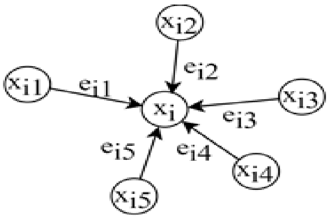


Figure 2. Local point graph G.

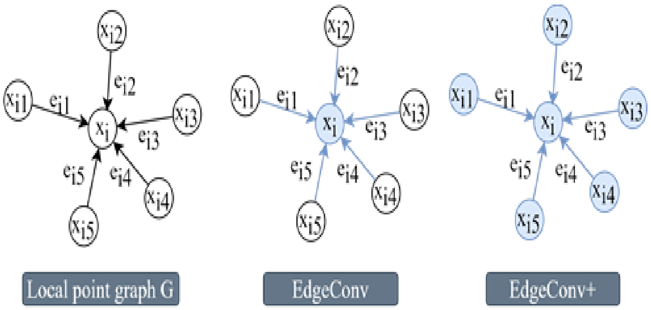


Figure 3. Shape modelling of local point clouds.

EdgeConv+ largely improves EdgeConv's shape perception. It can improve the ability of the object detection network to learn the spatial information of the point cloud. Table 4 demonstrates the effectiveness of EdgeConv+.

Algorithm (1) describes in detail the pseudocode of the EdgeConv+, Where RBC represents a combination of layers with specific functions: Rectified Linear Unit (ReLU) activation function (R) for introducing non-linearity and enhancing feature expression, BatchNorm (B) for normalizing and stabilizing the intermediate outputs, and Convolutional (C) layers for extracting local patterns and capturing spatial dependencies. The Maximum Pooling Function (MAX) is a symmetric aggregation function, a maximum pooling function, which can extract the most important features in all feature vectors.

*Algorithm 1: A new edge convolution: EdgeConv+*

```
Input: local directed graph G
Output: Point cloud feature information $X_i$
Concatenate $x_i$ and $e_{ij}$: $cat(x_i, e_{ij})$
Calculate the fused point cloud feature information $X_i=RBC\left(cat(x_i, e_{ij})\right)$
Calculate the sum of $X_i$ and $x_{ij}$: $X_i+x_{ij}$
Calculate the fused point cloud feature information $X_i=RBC(X_i+x_{ij})$
Aggregate point cloud information $X_i=MAX\{RBC\left(X_i+x_{ij}\right)\}$
return $X_i$
```

## 2.4. Deep Semantic Information Enhancement Module

Figure 4 shows the structure of CGNL+, which makes up the deep semantic information enhancement module.The excellent performance of Laplace matrice is well demonstrated by graph convolutional networks [3]. The proposed method extends Laplace matrices to the attention mechanism CGNL [31] to enhance the expressiveness of CGNL and enable it to achieve better performance. Following the rules of CGNL, all the space (width W and length H), and time (video length T) are stacked into one dimension, i.e., $N=H\times W$ or $N=T\times H\times W$. So, the input feature $X\in N\times C$, $C$ is the number of channels of $X$.

$$\begin{cases} \theta=\mathrm{vec}(XW_\theta)\in R^{NC} \\ \emptyset=\mathrm{vec}(XW_\emptyset)\in R^{NC}\ (W_\theta,\ W_\emptyset,\ W_g\in R^{CC}) \\ \mathrm{g}=\mathrm{vec}(XW_g)\in R^{NC} \end{cases} \tag{2}$$

Where θ, ∅ and g are obtained by linear transformation of the input features $X$ through 1×1 or 1×1×1 convolution layers whose kernel size and stride both equal 1 (k=1, s=1) respectively.

To improve the reusability of the model, CGNL applies the idea of channel grouping, dividing the linearly transformed features into G groups along the channel dimension, so that the number of channels in each group becomes $C/G$. To capture dependencies across the whole feature map, the original nonlocal operation computes the response $Y\in R^{NC}$ as the weighted sum of the features at all positions, as shown in Equation (3), CGNL assumes that $f$ is a general kernel function (e.g., RBF, bilinear, etc.,) that computes a $NC\times NC$ matrix. Then, $vec(Y)$ is approximated by a Taylor series. The article calculates the semantic similarity between objects by CGNL.

$$vec(Y)=f(vec(XW_\theta),vec(XW_\emptyset))vec(XW_g)\approx\theta\emptyset^T\mathrm{g} \tag{3}$$

Based on CGNL, CGNL+ introduces the Laplacian matrix. It expresses the Laplacian matrix $L=D\text{-}E$ by adding the offset ($X$-$Y$) between the input feature and the feature $Y$ calculated by CGNL, as shown by the red line in Figure 4, which enriches the expressive ability of CGNL.

$$Z=concat(BN((X—Y)W_z))+concat(BN(YW_z))+X \tag{4}$$

Where $W_z \in R^{CC}$, is a 1×1 or 1×1×1 convolution layer, and BN (batch normalization) layer is applied to each group. Finally, the output feature $Z$ is obtained by summing all the grouping information. Table 5 demonstrates the effectiveness of the CGNL+.

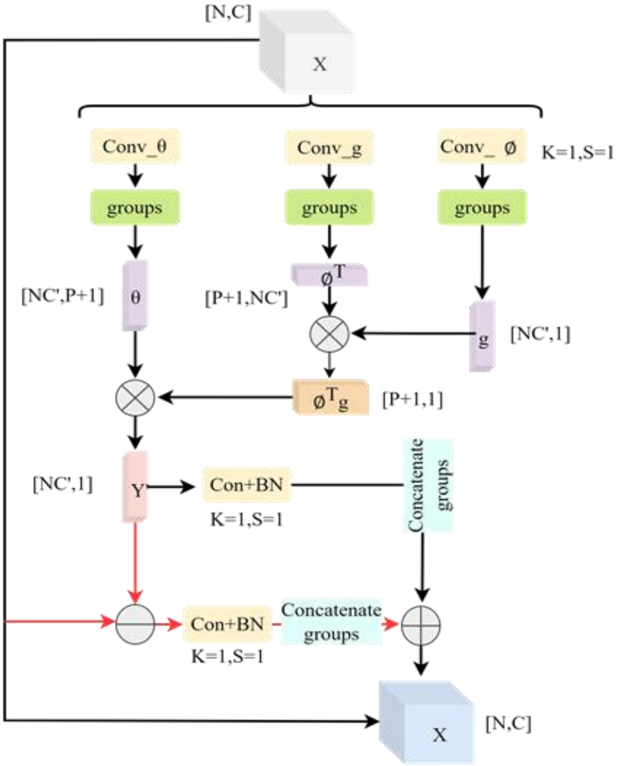


Figure 4. The architecture of the CGNL+. The black line indicates the CGNL. The red line indicates the addition of the Laplace matrix to the CGNL.

## 3. Experiments

Table 1 shows the experimental parameters. The article conducts experiments on the ScanNet [5] dataset. N is the number of random samples for each input data, and K is the number of nearest neighbours queried by the KNN when constructing a local point graph G in the SSIE. The network was trained for a total of 200 epochs with an Adam optimizer, batch size 4 and an initial learning rate of 0.001. The learning rate decay steps are set to {80, 120, 160}, and the decay rates are {0.1, 0.1, 0.1}.

The Graphics Processing Unit (GPU) model configured is GeForce Ray Tracing Technology (RTX) 2070 Super, and the experimental software environment is Ubuntu system 18.04, python 3.7, Compute Unified Device Architecture (CUDA) 10.0. The network is implemented through the PyTorch framework under the Python platform.

Table 1. Experimental parameters.

| Datasets | N | K | Optimizer | Batch Size | Epoch |
|---|---|---|---|---|---|
| ScanNet | 40000 | 20 | Adam | 4 | 200 |

### 3.1. Dataset Preparation

ScanNet is a large indoor environment dataset containing 1513 3D scans of different indoor scenes. The dataset itself also provides ground-truth 3D bounding boxes for indoor objects. The object detection categories are 18 (the remaining categories after removing unlabeled categories, floors, and walls from the 21 categories).

To augment the training data, the article randomly flips the point cloud in two horizontal directions and rotates the point cloud in the vertical coordinate system (rotation angle is controlled between -5° and 5°), scales point cloud (scaling factor is controlled between 0.9 and 1.1).

### 3.2. Comparative Experiments

Table 2 compares the performance of the proposed method with other methods on the ScanNet dataset. The evaluation is based on mean Average Precision (mAP) at two Intersection over Union (IoU) thresholds, specifically 0.25 and 0.5.

mAP is one of the metrics used to evaluate the accuracy of object detection algorithms. It is based on the concepts of Precision and Recall. Precision measures the ratio of correctly detected positive samples to the total number of detected positive samples, while Recall measures the ratio of correctly detected positive samples to the total number of true positive samples. mAP calculates the Precision-Recall (PR) curve at different confidence thresholds and computes the area under the curve as the final evaluation metric. A higher mAP value indicates higher accuracy in detecting objects. IoU is a metric used to measure the degree of overlap between the predicted bounding box and the ground truth bounding box. It calculates the ratio of the intersection area between the two boxes to the union area of the two boxes. If the IoU value between a predicted box and a ground truth box exceeds a threshold (0.25 or 0.5), it is considered a match.

Table 2 compares the proposed method with other methods on the ScanNet dataset. The article shows mAP when IoU is 0.25 and 0.5, respectively. The mAP@0.25 of the proposed method reaches 61.0%, which is higher than others, and mAP@0.5 is also higher than others.

Table 2. 3D object detection results on the ScanNet dataset.

| Method | Input | mAP@0.25 | mAP@0.5 |
|---|---|---|---|
| DSS [25] | Geo + RGB | 15.2 | 6.8 |
| MRCNN 2D-3D [10] | Geo + RGB | 17.3 | 10.5 |
| F-PointNet [18] | Geo + RGB | 19.8 | 10.8 |
| GSPN [30] | Geo + RGB | 30.6 | 17.7 |
| 3D-SIS [11] | Geo + 1 view | 35.1 | 18.7 |
| 3D-SIS | Geo + 3 views | 36.6 | 19.0 |
| 3D-SIS | Geo + 5 views | 40.2 | 22.5 |
| 3D-SIS | Geo only | 25.4 | 14.6 |
| VoteNet [17] | Geo only | 58.6 | 36.1 |
| MLCVNet [29] | Geo only | 59.0 | 36.4 |
| Ours | Geo only | 61.0 | 38.2 |

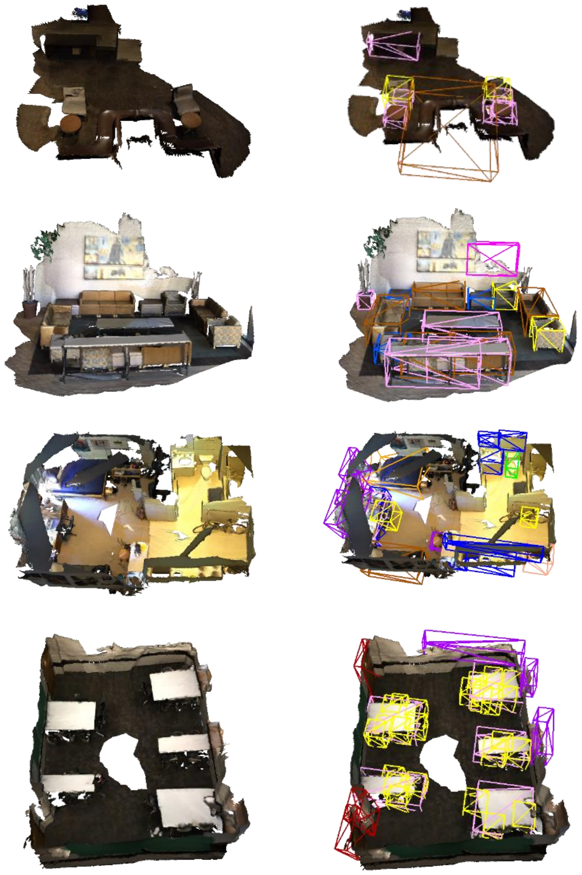
Table 3. 3D object detection results on 10 common objects of the ScanNet dataset.

| Method | bed | table | sofa | chair | toilet | sink | desk | door | bookshelf | bathtub |
|---|---|---|---|---|---|---|---|---|---|---|
| 3DSIS-Geo [11] | 63.1 | 51.3 | 46.3 | 66.0 | 74.5 | 22.9 | 33.3 | 8.0 | 2.3 | 58.7 |
| 3DSIS-5views | 69.8 | 36.1 | 71.8 | 66.2 | 87.6 | 43.0 | 46.9 | 30.6 | 27.3 | 84.3 |
| VoteNet [17] | 86.7 | 59.4 | 87.7 | 88.0 | 90.9 | 49.5 | 62.7 | 47.1 | 51.0 | 89.8 |
| MLCVNet [29] | 87.2 | 59.9 | 84.8 | 88.7 | 96.5 | 49.5 | 71.0 | 49.0 | 48.8 | 88.5 |
| Ours | 89.3 | 61.8 | 90.9 | 88.2 | 99.7 | 59.1 | 71.1 | 49.3 | 49.8 | 91.2 |

The article selects 10 common objects from the ScanNet dataset for comparison, as shown in Table 3, the evaluation metric is mAP@0.25. Our method performs better than the previous methods on 8/10 objects, and the mAP@0.25 reaches more than 90% among the sofa, toilet and bathtub.

Figure 5 shows a qualitative demonstration of the results of 3D bounding box prediction using the proposed method. The method can detect objects well in both complex indoor environments and simple indoor environments.

From the qualitative results, it can be observed that there is a certain lack of performance in the orientation angle of the 3D bounding box in the predicted results. Specifically, for objects with a poor sense of depth, such as the murals or windows shown in Figure 5, there is a displacement of the bounding boxes.



a) The actual scenes.

b) The predicted results.

Figure 5. Qualitative results of our approach on the ScanNet dataset.

## 3.3. Validity Experiments

The key to the proposed method is the design of a new EdgeConv+ and a new attention mechanism CGNL+.

Tables 4 and 5 show the superiority of EdgeConv+ and CGNL+ compared to the original EdgeConv and CGNL models, respectively.

Table 4 shows the experimental results of adding the SSIE with EdgeConv and EdgeConv+ as the main components respectively on Votenet. The proposed method endows EdgeConv with a more complete point cloud representation to improve its perception of shape, which is effective for improving the performance of object detection.

Table 5 shows the experimental results of adding the DSIE with CGNL and CGNL+ as the main components respectively on Votenet. It is not difficult to find that the Laplace matrix is useful for improving CGNL.

Table 4. EdgeConv+ validity experiments.

| Method | mAP@0.25 | mAP@0.5 |
|---|---|---|
| VoteNet | 58.6 | 36.1 |
| EdgeConv | 58.6 | 37.7 |
| EdgeConv+ | 59.6 | 37.7 |

Table 5. CGNL+ validity experiments.

| Method | mAP@0.25 | mAP@0.5 |
|---|---|---|
| VoteNet | 58.6 | 36.1 |
| CGNL | 59.8 | 37.4 |
| CGNL+ | 60.2 | 38.1 |

## 3.4. Ablation Experiments

To quantitatively evaluate the effectiveness of the proposed modules, the article conducts experiments on different combinations of these modules, as shown in Table 6. At mAP@0.25 and mAP@0.5, when SSIE is only used, it is 1.0% and 1.6% higher than Votenet, respectively; when DSIE is only used, it is 1.6% and 2.0% higher than the original Votenet, respectively; when using SSIE and DSIE at the same time, it is 2.4% and 2.1% higher than Votenet, respectively. Therefore, the two modules designed are effective.

Table 6. Module ablation experiments.

| | SSIE | DSIE | mAP@0.25 | mAP@0.5 |
|---|---|---|---|---|
| VoteNet | | | 58.6 | 36.1 |
| VoteNet | √ | | 59.6 | 37.7 |
| VoteNet | | √ | 60.2 | 38.1 |
| VoteNet | √ | √ | 61.0 | 38.2 |

## 3.5. Robustness Testing

Considering that indoor service robots will cause the number of scene point clouds to decrease due to jittering or being obscured during movement.

To test the robustness of the problem of the sparse point cloud, the article tests the proposed method on the ScanNet dataset, using the same experimental parameters as before, selecting 40,000 and 20,000 points as input, respectively. As shown in Figure 6, even when the number of points decreases by one-half, the proposed method is higher than Votenet at mAP@0.25. The experimental results fully prove the feasibility of the dual fusion idea of spatial information and semantic information fusion, global information and local information fusion proposed by the feature fusion module in this article to improve the robustness of the network.

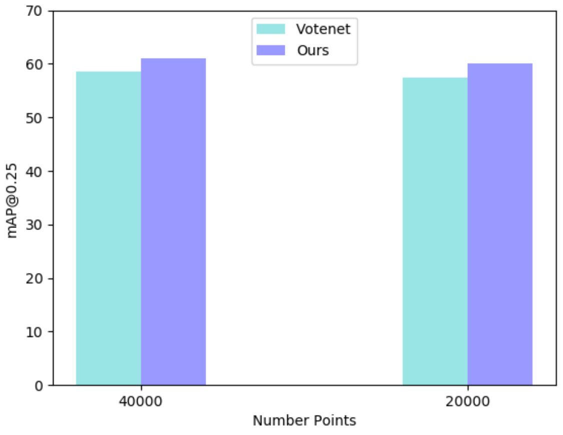


Figure 6. Robustness experiments when sparse point clouds are used as input. When the number of points drops from 40000 to 20000, the mAP@0.25 of our method decreases from 61.0% to 60.0%, and the mAP@0.25 of Votenet decreases from 58.6% to 57.4%.

## 4. Conclusions

In this article, we propose a 3D object detection method with enhanced spatial semantic information. A new edge convolution EdgeConv+ and a new attention mechanism CGNL+ are proposed. SSIE captures the spatial information of the point cloud through EdgeConv+, and DSIE acquires the semantic information of the point cloud under the guidance of CGNL+. The proposed method effectively improves the performance of Votenet, realizing object detection in indoor scenes. Experiments show that on the ScanNet, the proposed method is 2.4% and 2.1% higher than Votenet at mAP@0.25 and mAP@0.5, respectively. As future work, we need to verify the effectiveness of our method on more datasets. This is essential for the development of indoor service robots.

## References

[1] Abeywickrama T., Cheema M., and Taniar D., “k-Nearest Neighbors on Road Networks: A Journey in Experimentation and In-Memory Implementation,” *Proceedings of the VLDB Endowment*, vol. 9, no. 6, pp. 1-12 , 2016. https://doi.org/10.14778/2904121.2904124

[2] Balcazar J., Dai Y., and Watanabe O., “Provably Fast Training Algorithms for Support Vector Machines,” *in Proceedings of the IEEE International Conference on Data Mining*, San Jose, pp. 43-50, 2001. doi: 10.1109/ICDM.2001.989499.

[3] Bruna J., Zaremba W., Szlam A., and LeCun Y., “Spectral Networks and Locally Connected Networks on Graphs,” *arXiv Preprint arXiv 1312.6203*, pp. 1-14, 2014. https://doi.org/10.48550/arXiv.1312.6203

[4] Dalal N. and Triggs B., “Histograms of Oriented Gradients for Human Detection,” *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, pp. 886-893, 2005. DOI: 10.1109/CVPR.2005.177

[5] Dai A., Chang A., Savva M., Halber M., Funkhouser T., and Nießner M., “Scannet: Richly-Annotated 3D Reconstructions,” *IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, pp. 5828-5839, 2017. https://doi.org/10.48550/arXiv.1702.04405

[6] Engelcke M., Rao D., Wang D., Tong C., and Posner I., “Vote3deep: Fast Object Detection in 3D Point Clouds Using Efficient Convolutional Neural Networks,” *in Proceedings of the IEEE International Conference on Robotics and Automation*, Singapore, pp. 1355-1361, 2017. doi: 10.1109/ICRA.2017.7989161.

[7] Felzenszwalb P., Girshick R., McAllester D., and Ramanan D., “Object Detection with Discriminatively Trained Part-Based Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, 2010. DOI: 10.1109/TPAMI.2009.167

[8] Girshick R., Donahue J., Darrell T., and Malik J., “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, pp. 580-587, 2014. doi: 10.1109/CVPR.2014.81.

[9] Griffin B., “Mobile Robot Manipulation Using Pure Object Detection,” *in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, pp. 561-571, 2023. DOI: 10.1109/WACV56688.2023.00063

[10] He K., Gkioxari G., Dollár P., and Girshick R., “Mask R-CNN,” *in Proceedings of the IEEE International Conference on Computer Vision*, Venice, pp. 2961-2969, 2017. doi: 10.1109/ICCV.2017.322.

[11] Hou J., Dai A., and Nießner M., “3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, California, pp. 4421-4430, 2019. DOI Bookmark: 10.1109/CVPR.2019.00455

[12] Li B., Zhang T., and Xia T. “Vehicle Detection from 3D Lidar Using Fully Convolutional

Network," arXiv Preprint, arXiv 1608.07916, pp.

[13] Li C., Li L., Jiang H., and Weng K., "YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications," *arXiv Preprint, arXiv: 2209.02976*, pp. 1-17, 2022. https://arxiv.org/pdf/2209.02976.pdf

[14] Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C., and Berg A., "SSD: Single Shot MultiBox Detector," *in Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, pp. 21-37, 2016. https://doi.org/10.1007/978-3-319-46448-0_2

[15] Lowe D., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004. DOI:10.1023/B:VISI.0000029664.99615.94

[16] Ojala T., Pietikainen M., and Maenpaa T., "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002. doi: 10.1109/TPAMI.2002.1017623.

[17] Qi C., Litany O., He K., and Guibas L., "Deep Hough Voting for 3D Object Detection in Point Clouds," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, pp. 9277-9286, 2019. doi: 10.1109/ICCV.2019.00937.

[18] Qi C., Liu W., Wu C., Su H., and Guibas L., "Frustum PointNets for 3D Object Detection from RGB-D Data," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Utah, pp. 918-927, 2018. doi: 10.1109/CVPR.2018.00102.

[19] Qi C., Su H., Mo K., and Guibas L., "Pointnet: Deep Learning on Point Sets for 3D Classification and Segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, pp. 652-660, 2017. DOI Bookmark: 10.1109/CVPR.2017.16

[20] Qi C., Yi L., Su H., and Guibas L., "Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," *in Proceedings of the 31st Conference on Neural Information Processing Systems*, California, pp. 1-10, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/d8bf84be3800d12f74d8b05e9b89836f-Paper.pdf

[21] Rafique A., Jalal A., and Kim K., "Statistical Multi-Objects Segmentation for Indoor/Outdoor Scene Detection and Classification via Depth Images," *in Proceedings of the 17th International Bhurban Conference on Applied Sciences and Technology*, Islamabad, pp. 271-276, 2020. doi: 10.1109/IBCAST47879.2020.9044576.

[22] Redmon J., Divvala S., Girshick R., and Farhadi A., "You Only Look Once: Unified, Real-Time Object Detection," *in Proceedigs of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 779-788, 2016. DOI Bookmark: 10.1109/CVPR.2016.91

[23] Ren S., He K., Girshick R., and Sun J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

[24] Simon M., Amende K., Kraus A., and Honer J., "Complexer-YOLO: Real-Time 3D Object Detection and Tracking on Semantic Point Clouds," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, California, pp. 1190-1199, 2019. DOI:10.1109/CVPRW.2019.00158

[25] Song S. and Xiao J., "Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 808-816, 2016. doi: 10.1109/CVPR.2016.94.

[26] Wang C., Bochkovskiy A., and Liao H., "YOLOv7: Trainable Bag-Of-Freebies Sets New State-Of-The-Art for Real-Time Object Detectors," *arXiv Preprint, arXiv 2207.02696*, pp. 1-15, 2022. https://arxiv.org/abs/2207.02696

[27] Wang Y., Sun Y., Liu Z., Sarma S., Bronstein M., and Solomon J., "Dynamic Graph CNN for Learning on Point Clouds," *ACM Transactions on Graphics*, vol. 1, no. 1, pp. 1-13, 2019. https://arxiv.org/pdf/1801.07829.pdf

[28] Wang Y. and Solomon J., "Object DGCNN: 3D Object Detection Using Dynamic Graphs," *in Proceedings of the 35th Conference on Neural Information Processing Systems*, Sydney, pp. 1-16, 2021. https://arxiv.org/pdf/2110.06923.pdf

[29] Xie Q., Lai Y., Wu J., Wang Z., Zhang Y., Xu K., and Wang J., "Mlcvnet: Multi-Level Context Votenet for 3D Object Detection," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp. 10447-10456, 2020. doi: 10.1109/CVPR42600.2020.01046

[30] Yi L., Zhao W., Wang H., Sung M., and Guibas L., "GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, California, pp. 3947-3956, 2019. doi: 10.1109/CVPR.2019.00407.

[31] Yue K., Sun M., Yuan Y., Zhou F., Ding E., and Xu F., "Compact Generalized Non-Local Network," *in Proceedings of the 32nd Conference on Neural Information Processing Systems*,

Montréal, pp. 1-10, 2018. https://dl.acm.org/doi/pdf/10.5555/3327757.3327758

[32] Zhou Y. and Tuzel O., "Voxelnet: End-To-End Learning For Point Cloud Based 3D Object Detection," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Utah, pp. 4490-4499, 2018. DOI Bookmark: 10.1109/CVPR.2018.00472

**Chunmei Chen** an Associate Professor at Southwest University of Science and Technology, China. She holds a Ph.D. degree and is a supervisor for master's students. Her main research and teaching areas include wireless network communication technology, Internet of Things engineering, intelligent terminal software development, and image processing.

**Zhiqiang Liang** master of Information Engineering School, Southwest University of Science and Technology, China. His research interests include computer vision and deep learning, especially in the field of 3D point cloud object detection.

**Haitao Liu** master of Information Engineering School, Southwest University of Science and Technology, China. His research interests include artificial intelligence and data mining, especially in the field of data mining.

**Xin Liu** master of Information Engineering School, Southwest University of Science and Technology, China. Her research interests include computer vision and deep learning, especially in the field of 3D reconstruction.