

SPGroup3D: Superpoint Grouping Network for Indoor 3D Object Detection

Yun Zhu¹, Le Hui², Yaqi Shen¹, Jin Xie^{1*}

¹ PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology, China

² Shaanxi Key Laboratory of Information Acquisition and Processing, Northwestern Polytechnical University, China
zhu.yun@njust.edu.cn, huile@nwpu.edu.cn, syq@njust.edu.cn, csjie@njust.edu.cn

Abstract

Current 3D object detection methods for indoor scenes mainly follow the voting-and-grouping strategy to generate proposals. However, most methods utilize instance-agnostic groupings, such as ball query, leading to inconsistent semantic information and inaccurate regression of the proposals. To this end, we propose a novel superpoint grouping network for indoor anchor-free one-stage 3D object detection. Specifically, we first adopt an unsupervised manner to partition raw point clouds into superpoints, areas with semantic consistency and spatial similarity. Then, we design a geometry-aware voting module that adapts to the centerness in anchor-free detection by constraining the spatial relationship between superpoints and object centers. Next, we present a superpoint-based grouping module to explore the consistent representation within proposals. This module includes a superpoint attention layer to learn feature interaction between neighboring superpoints, and a superpoint-voxel fusion layer to propagate the superpoint-level information to the voxel level. Finally, we employ effective multiple matching to capitalize on the dynamic receptive fields of proposals based on superpoints during the training. Experimental results demonstrate our method achieves state-of-the-art performance on ScanNet V2, SUN RGB-D, and S3DIS datasets in the indoor one-stage 3D object detection. Source code is available at <https://github.com/zyrant/SPGroup3D>.

Introduction

As one of the basic tasks of 3D scene understanding, the goal of 3D object detection is to estimate the oriented 3D bounding boxes and semantic labels of objects in point clouds. It has been used in many application scenarios, such as autonomous driving, augmented reality, and robotics. Since indoor scenarios have more complex geometry and occlusion than outdoor scenarios (Chen et al. 2022; He et al. 2022; Wu et al. 2023; Feng et al. 2023), it remains challenging to cluster the scenes and get the proposals accurately.

Previous state-of-the-art 3D object detection methods mainly follow the bottom-up paradigm, involving two key components: voting and grouping. These components are

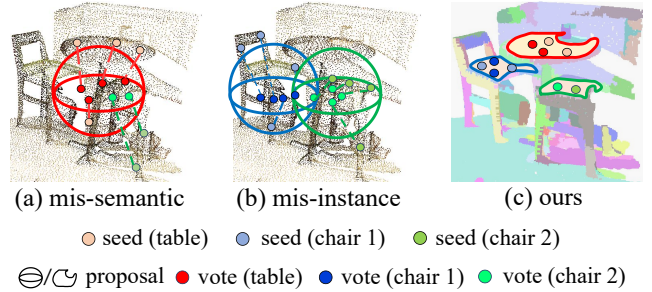


Figure 1: Different types of grouping strategy: (a) semantic-agnostic grouping; (b) semantic-aware grouping (Wang et al. 2022a); (c) our superpoint-based grouping. The proposal corresponding to the table is shown in red, and the two chairs are blue and green, respectively. For superpoint-based grouping, we represent different superpoints with different colors, and the features with the same superpoint will be grouped.

used to learn the point-wise center offsets and aggregate the points that vote to instance-agnostic local region, respectively. The conventional voting tends to push all proposals close to the object centers. It may not be suitable for FCOS-like (Tian et al. 2019; Rukhovich, Vorontsova, and Konushin 2022) detection methods since the distances from the proposals to the object centers play a crucial role in determining the quality of proposals. Besides, the commonly used grouping module can be categorized into two types: semantic-agnostic and semantic-aware. The former, such as VoteNet and its subsequent works (Qi et al. 2019; Cheng et al. 2021; Liang, An, and Ma 2022; Sun et al. 2022), groups the scenes based on the distance between coordinates. However, this approach fails in clustering indoor scenes when various instances are close but belong to different categories, as shown in Fig. 1(a). The latter one, such as CAGroup3D (Wang et al. 2022a), which additionally considers semantics, still suffers from a similar issue when instances of the same category are close to each other, as shown in Fig. 1(b). In summary, the conventional voting and the instance-agnostic grouping introduce more noise and outliers to the proposals, leading to performance bottlenecks in bottom-up methods.

In this paper, we propose a superpoint grouping network for indoor anchor-free one-stage 3D object detection (SP-

*Corresponding author.

Group3D) to solve these problems. Specifically, we first use a sparse convolution-based backbone to obtain the voxels, which include coordinates and features. Then, we introduce a geometry-aware voting module to adapt to the concept of centerness in the anchor-free detection methods (Tian et al. 2019; Rukhovich, Vorontsova, and Konushin 2022). For each voxel, by merging the seed parts and the vote parts, we can ensure the relative geometric relationships between proposals and object centers, making the model easier to filter out low-quality proposals. Consequently, we address grouping problem by introducing superpoint (Landrieu and Simonovsky 2018). As shown in Fig. 1(c), superpoint is a natural instance-aware local unit, consisting of adjacent points sharing similar semantics and spatiality. Based on superpoint, we construct a superpoint-based grouping module. Superpoint-based grouping includes superpoint attention and superpoint-voxel fusion to facilitate superpoint-to-superpoint and superpoint-to-voxel information interactions, respectively. Superpoint attention is proposed to enhance the interaction between non-overlapping superpoints, adaptively exploring relevant superpoint features within their neighborhoods. The following superpoint-voxel fusion based on the sparse convolution is used for the interaction of voxels and superpoints. Finally, we employ multiple matching to assess the discrepancy between each superpoint-based proposal and the ground truth, and select positive samples during training. Extensive experiments indicate our method achieves state-of-the-art performance in the indoor one-stage 3D object detection task on three datasets, in terms of mAP@0.25 on ScanNet V2 (+1.1), SUN RGB-D (+1.2), and S3DIS (+2.5).

The main contributions are listed as follows:

- We introduce a geometry-aware voting module, which preserves the relative geometry of superpoints in coordinate space, to adapt to anchor-free detection.
- We design a superpoint-based grouping module including superpoint attention and superpoint-voxel fusion to enable superpoint-to-superpoint and superpoint-to-voxel feature interactions, respectively.
- We present a multiple matching strategy, which can discriminate between positive and negative samples of superpoint-based proposals during training.

Related Work

Top-down 3D object detection. Top-down models are primarily applied in outdoor autonomous driving scenarios. VoxelNet (Zhou and Tuzel 2018) is a pioneering work that partitions the raw point clouds into voxels and realizes end-to-end 3D object detection in this field. PV-RCNN (Shi et al. 2020) argues that voxel quantization may lead to information loss and proposes a second stage for incorporating fine-grained features at the point level. For indoor 3D object detection, FCAF3D (Rukhovich, Vorontsova, and Konushin 2022) introduces the first anchor-free model with fully sparse convolution and further proposes an improved version, TR3D (Rukhovich, Vorontsova, and Konushin 2023).

Bottom-up 3D object detection. Bottom-up methods, inspired by PointNet (Qi et al. 2017a) and its variation (Qi et al. 2017b), have gained widely attention for their ability to predict 3D bounding boxes from point clouds. VoteNet (Qi et al. 2019) adopts deep hough voting to group features and generate proposals. Building upon this framework, ML-CVNet (Xie et al. 2020) adopts different level context information to explore relationships between objects. H3DNet (Zhang et al. 2020) extends the key points prediction of centers in VoteNet, including the centers of bounding boxes, surfaces, and edges. VENet (Xie et al. 2021) designs a vote-weighting module to improve the voting process. BRNet (Cheng et al. 2021) proposes using clustered centers to predict the angles and spatial positions of entire objects. RBGNet (Wang et al. 2022b) introduces a ray-based feature grouping module to capture points on the object surface. CAGroup3D (Wang et al. 2022a) leverages the powerful expressive capability of sparse convolution to enhance feature extraction and proposes a two-stage method. However, these methods essentially construct instance-agnostic grouping to generate proposals, which inevitably leads to the presence of multiple instance features within a proposal.

Superpoints. Recently, superpoints have been introduced into other 3D point cloud tasks (Hui et al. 2023; Shen et al. 2023; Tang, Hui, and Xie 2022). The concept of superpoints is initially proposed by SPG (Landrieu and Simonovsky 2018) and applied to 3D semantic segmentation. SSP (Landrieu and Boussaha 2019) proposes a learnable strategy for over-segmenting superpoints, while SPNet (Hui et al. 2021) further extends the concept of superpoint over-segmentation to an end-to-end approach. In the domain of instance segmentation, SSTNet (Liang et al. 2021) constructs a superpoint tree network to aggregate superpoints with the same semantic information. GraphCut (Hui et al. 2022) proposes a bilateral graph attention mechanism to generate precise instances using the superpoint graph cutting network. SPFormer (Sun et al. 2023) utilizes learnable query to predict instance segmentation from superpoint features as a top-down pipeline. Although these methods have explored a variety of uses for superpoints, they have not directly enhanced the representation of superpoints nor made them more suitable for anchor-free detection. We employ a superpoint-based grouping to enhance the expression of superpoints and design a geometry-aware voting specifically for anchor-free detection to make the model easier to filter out low-quality proposals.

Method

Overview

The overall architecture of our method is depicted in Fig. 2. The input point cloud typically consists of N points, which can be represented as $\{\mathbf{p}_i\}_{i=1}^N$, with $\mathbf{p}_i \in \mathbb{R}^6$. Each point has coordinates x, y, z and colors r, g, b . Following previous method (Wang et al. 2022a), we first extract M high-resolution non-empty seed voxels from a sparse 3D backbone network (Choy, Gwak, and Savarese 2019). Next, these voxels are passed through a geometry-aware voting module and several superpoint-based grouping

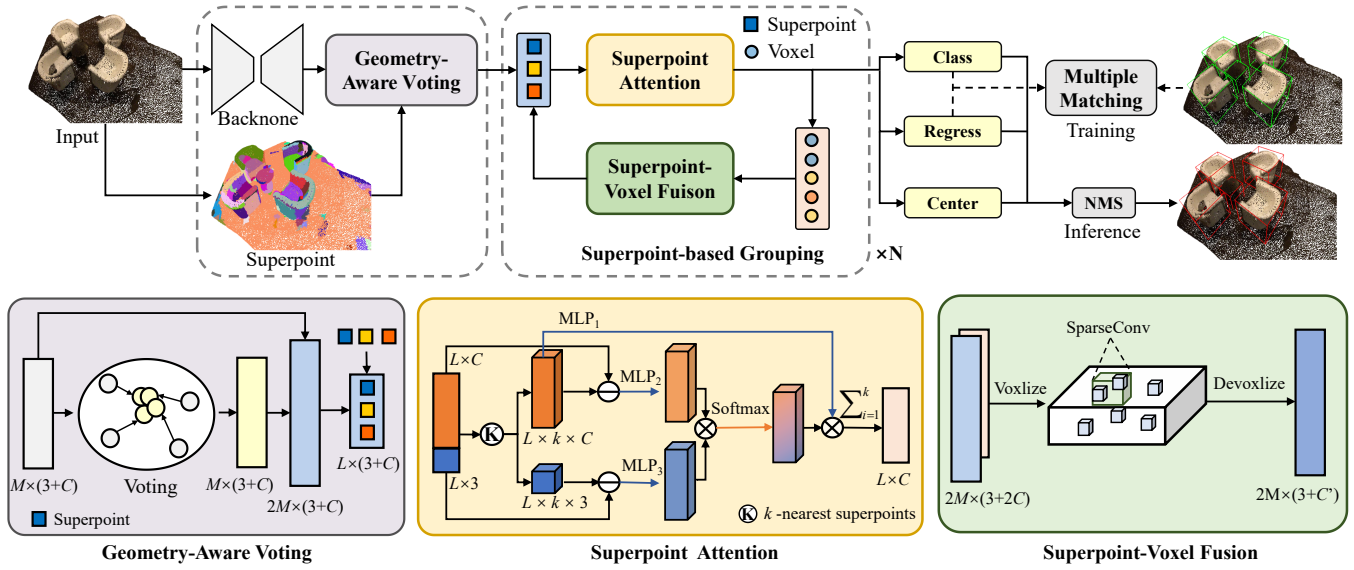


Figure 2: Framework of the superpoint grouping network for indoor 3D object detection (SPGroup3D). Given the input, we first extract seed voxels through the backbone. Subsequently, we construct geometry-aware voting to preserve the relative positions of the proposals and the object centers. Following this, based on superpoint-based grouping, we iteratively optimize the feature representations of the superpoints. Finally, multiple matching is employed to select positive samples during training and 3D NMS is applied to eliminate redundant proposals during the inference time.

modules to generate proposals. Finally, multiple matching is employed to choose positive proposals during training and 3D non-maximum suppression (NMS) is applied to remove redundant proposals during inference.

Geometry-Aware Voting

In anchor-free detection, we usually assign higher scores/centernesses to the proposals/superpoints that are close to object centers and lower scores to proposals that are far away from object centers, and discard the proposals with low scores. However, traditional voting (Qi et al. 2019) tends to push proposals close to the object centers, resulting in high scores, which makes it difficult to filter out low-quality proposals. In this paper, we present a geometry-aware voting to solve this problem. Geometry-aware voting preserves the relative position information from the proposals to the object centers in geometric space, enabling the model to allocate low scores for relatively distant proposals, thus making it easier to filter out low-quality proposals in post-processing.

Specifically, given M high-resolution non-empty seed voxels $\{v_i\}_{i=1}^M$ from the backbone, where $v_i = [v_i^c; v_i^f]$ with $v_i^c \in \mathbb{R}^3$ and $v_i^f \in \mathbb{R}^C$. The next key step is to generate proposals/superpoints. The most straightforward way to group voxels into superpoints is to cluster $\{v_i\}_{i=1}^M$ using offline pre-computed superpoints. However, as shown in Fig. 3(b) and (f), only using $\{v_i\}_{i=1}^M$ will lead to the generated superpoints mainly existing on the surface of the objects, which are far away from the center of the instances, making it difficult to accurately regress (Qi et al. 2019).

Therefore, similar to VoteNet (Qi et al. 2019), we set a voting branch to learn the offsets of the voxels to the cor-

responding bounding box centers. Each seed voxel includes coordinate offset $\Delta v_i^c \in \mathbb{R}^3$ and feature offset $\Delta v_i^f \in \mathbb{R}^C$. The vote voxel o_i is generated by adding the offsets as follows:

$$\{o_i \mid o_i = [v_i^c + \Delta v_i^c; v_i^f + \Delta v_i^f]\}_{i=1}^M \quad (1)$$

where the corresponding coordinate and feature of o_i are denoted by $o_i^c \in \mathbb{R}^3$ and $o_i^f \in \mathbb{R}^C$, respectively. The predicted offset Δv_i^c is explicitly supervised by a smooth- ℓ_1 loss with the ground-truth displacement from the coordinate of seed voxel v_i to its corresponding bounding box center. Depicted in Fig. 3(c) and (g), the generated superpoints primarily emphasize the object centers with this traditional voting, weakening the geometry representation and relative position to object centers (Cheng et al. 2021).

This interference poses challenges to the post-processing in the anchor-free detection methods (Tian et al. 2019; Rukhovich, Vorontsova, and Konushin 2022) since the superpoints belonging to the current object tend to gather around the object center. Hence, unlike VoteNet, which directly utilizes the vote voxels, we further merge the seed voxels. This merge processing can be formulated as follows:

$$\{f_i \mid f_i = [(v_i^c, o_i^c); (v_i^f, o_i^f)]\}_{i=1}^{2M} \quad (2)$$

where f_i denotes the i -th voxel of $\{f_i\}_{i=1}^{2M}$. This processing ensures that the generated superpoints remain within their corresponding objects while preserving their original relative position to the object centers, as illustrated in Fig. 3(d) and (h).

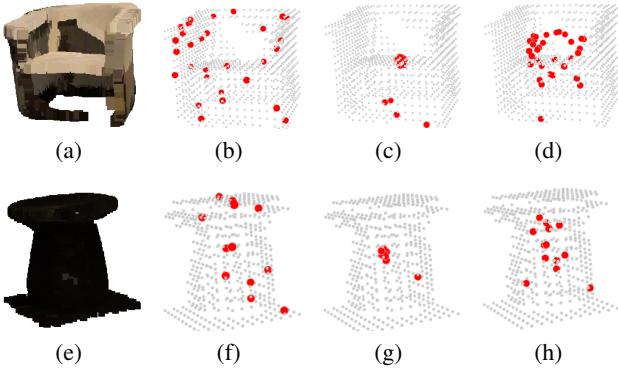


Figure 3: The visualization of the locations of generated superpoints. The red points represent superpoints. To better represent the geometry of the object, we use gray points to indicate the voxels. (a) and (e) represent the input point clouds. (b) and (f) represent superpoints without voting. (c) and (g) represent superpoints with traditional voting. (d) and (h) represent superpoints with geometry-aware voting.

Superpoint-based Grouping

For the purpose to realize the consistent instance representation within the same proposals, we propose superpoint-based grouping to aggregate voxels into superpoints, consisting of superpoint attention and superpoint-voxel fusion. In our setting, we iterate the superpoint-based grouping multiple times, *i.e.*, 3. For each iteration, the output of superpoint attention is regarded as the corresponding output. Conversely, the output of superpoint-voxel fusion serves as the input for the next iteration of superpoint-based grouping.

Superpoint Attention. Superpoints are non-overlapping local regions of the input voxels, which become the bottleneck of the model perception, preventing dense prediction tasks like object detection from understanding the instances. In this paper, inspired by attention mechanism (Vaswani et al. 2017), we propose a superpoint attention mechanism to enable feature interaction within neighboring local regions. The superpoint attention mechanism facilitates information propagation and integration, leading to a better understanding of the semantic and structural aspects of the instances.

Specifically, given the $\{\mathbf{f}_i\}_{i=1}^{2M}$ obtained from the former module, we first group them into initial L superpoints $\{\mathbf{s}_i\}_{i=1}^L$, where $\mathbf{s}_i = [\mathbf{s}_i^c; \mathbf{s}_i^f]$ with $\mathbf{s}_i^c \in \mathbb{R}^3$ and $\mathbf{s}_i^f \in \mathbb{R}^C$, using the superpoints. Subsequently, to accelerate the algorithm convergence and simplify its complexity, we apply the k -nearest neighbours (k -NN) algorithm to obtain the k nearest neighbours based on coordinate space for each superpoint. The attention operation is then performed only within the k nearest neighbours of each superpoint. These processes can be formulated as follows:

$$\{\mathbf{s}_i\}_{i=1}^L = \text{Scatter} \{\mathbf{f}_i\}_{i=1}^{2M}, \quad \{\mathbf{n}_{ij}\}_{j=1}^k = \text{knn}(\mathbf{s}_i) \quad (3)$$

where Scatter and $\{\mathbf{n}_{ij}\}_{j=1}^k$ represent Scatter¹ function

¹https://github.com/rusty1s/pytorch_scatter

and the k nearest superpoints of $\mathbf{s}_i$, respectively. Each superpoint in $\{\mathbf{n}_{ij}\}_{j=1}^k$ can be represented by coordinate $\mathbf{n}_{ij}^c \in \mathbb{R}^3$ and feature $\mathbf{n}_{ij}^f \in \mathbb{R}^C$.

Inspired by (Hui et al. 2022), we also explicitly calculate the similarity in both coordinate and feature space instead of embedding the coordinate into feature space. Different from their method that computes similarities between graphs for graph cutting in that realizes instance segmentation, our algorithm aims to facilitate interaction between superpoints to capture more contextual information. The specific steps are described as follows:

$$\{\mathbf{w}_{ij}^c\}_{j=1}^k = \text{MLP}(\mathbf{s}_i^c - \{\mathbf{n}_{ij}^c\}_{j=1}^k) \quad (4)$$

$$\{\mathbf{w}_{ij}^f\}_{j=1}^k = \text{MLP}(\mathbf{s}_i^f - \{\mathbf{n}_{ij}^f\}_{j=1}^k) \quad (5)$$

where $\{\mathbf{w}_{ij}^c\}_{j=1}^k$ and $\{\mathbf{w}_{ij}^f\}_{j=1}^k$ represent the wights in coordinate space and feature space of k nearest superpoints, respectively. MLP indicates a fully-connected layer. It is noted that to guarantee the above process, we copy $\mathbf{s}_i^c$ and $\mathbf{s}_i^f$ k times.

The fusion weights can be obtained by multiplying the coordinate space weights and feature space weights, followed by applying a softmax function to obtain normalized fusion weights. Therefore, the fusion wight w_{ij} for i -th superpoint and its j -th nearest superpoint can be formulated as:

$$w_{ij} = \frac{\exp(\mathbf{w}_{ij}^c \times \mathbf{w}_{ij}^f)}{\sum_{j=1}^k \exp(\mathbf{w}_{ij}^c \times \mathbf{w}_{ij}^f)} \quad (6)$$

Finally, these fusion weights are multiplied by the superpoint features of the k -nearest neighbours to obtain updated superpoint feature, while the coordinate of the superpoint remains unchanged. The step can be formulated as:

$$\mathbf{a}_i^f = \sum_{j=1}^k (w_{ij} \times \text{MLP}(\mathbf{n}_{ij}^f)), \quad \mathbf{a}_i^c = \mathbf{s}_i^c \quad (7)$$

where $\mathbf{a}_i^f$ and $\mathbf{a}_i^c$ denote one of the features and coordinates of the updated superpoints in $\{\mathbf{a}_i\}_{i=1}^L$, respectively. In addition to above steps, similar to the original attention algorithm, we also incorporate residual and normalization operations at last.

Superpoint-Voxel Fusion. From voxels to superpoints is a quantification process, inevitably leading to a loss of fine-grained features. Therefore, we propose superpoint-voxel fusion based on sparse convolution to achieve the interaction of coarse-grained superpoints and fine-grained voxels.

First, we need to broadcast the superpoint features to match the size of the merge voxels $\{\mathbf{f}_i\}_{i=1}^{2M}$ using the superpoints. The superpoint features after the broadcast are denoted as $\{\mathbf{h}_i^f\}_{i=1}^{2M}$. Then, we concatenate the features from $\{\mathbf{f}_i\}_{i=1}^{2M}$ and $\{\mathbf{h}_i^f\}_{i=1}^{2M}$ to obtain the initial fusion voxel features $\{\mathbf{g}_i^f\}_{i=1}^{2M}$. The step can be formulated as:

$$\{\mathbf{h}_i^f\}_{i=1}^{2M} = \text{Broadcast} \left(\{\mathbf{a}_i^f\}_{i=1}^L \right) \quad (8)$$

$$\{\mathbf{g}_i^f\}_{i=1}^{2M} = \text{Concat} \left(\{\mathbf{f}_i^f\}_{i=1}^{2M}; \{\mathbf{h}_i^f\}_{i=1}^{2M} \right) \quad (9)$$

where $\mathbf{g}_i^f \in \mathbb{R}^{2C}$ is one the feature of fusion voxels $\{\mathbf{g}_i\}_{i=1}^{2M}$. Besides, the coordinates of the fusion voxels are the same as those of the merge voxels. Next, based on the coordinates, we will voxelize fusion voxels again with a voxel size equal to the resolution of output voxels from the backbone. The revoxelized fusion voxels pass through an SPFFN, consisting of a sparse convolution layer, a normalization layer, and an activation layer to obtain the final output. This process can be formulated as follows:

$$\{\mathbf{g}'_i\}_{i=1}^M = \text{SPFFN}(\text{Voxelize}(\{\mathbf{g}_i\}_{i=1}^{2M})) \quad (10)$$

where $\mathbf{g}'_i$ denotes one of updated fusion voxels. Finally, $\{\mathbf{g}'_i\}_{i=1}^M$ will be converted back to input-level by the sparse tensor matrix mapping and then get the refined voxels $\{\mathbf{f}_{i+1}\}_{i=1}^{2M}$ for next iteration.

Multiple Matching and Loss Function

Previous anchor-free indoor 3D detection methods (Rukhovich, Vorontsova, and Konushin 2022; Wang et al. 2022a) rely on constructing proposals using regular neighbourhoods, *i.e.*, voxels. Assume we consider the receptive field as a criterion for determining the importance of the current proposal, proposals within the same features level would be treated equally. Based on this pattern, they could directly set positive and negative samples by judging the distance between the location of the proposals and the center of the bounding boxes.

However, this approach is not suitable for proposals generated by superpoints due to the fact that the proposals in our method have dynamic receptive fields. We must improve the matching strategy to select multiple positive samples for each ground truth. Here, to ensure consistency between training and registration, inspired by DETR (Carion et al. 2020), we directly adopt the loss during the training process as the cost function for matching and consider the cost of both classification and regression simultaneously. It is worth mentioning that our approach does not require Hungarian Matching and assigns multiple samples for each ground truth, which is different from Bipartite Matching in DETR. Cost_{ik} is to evaluate the similarity of the i -th proposal and the k -th ground truth. As defined by the following formula:

$$\text{Cost}_{ik} = -\lambda_{cls}\text{cost}_{ik}^{cls} - \lambda_{reg}\text{cost}_{ik}^{reg} \quad (11)$$

where cost_{ik}^{cls} and cost_{ik}^{reg} represent the focal cost function (Carion et al. 2020) and DIOU cost function (Zheng et al. 2020), respectively. λ_{cls} and λ_{reg} are the corresponding coefficients for each term. Finally, we directly select the top- r (*i.e.*, $r=18$) proposals with the minimum cost as positive samples for each ground truth, while the rest are considered as negative. In our experiments, only the proposals within the bounding box are considered, and both λ_{cls} , λ_{reg} are set to 1.

After assignment, our method is trained from scratch with voting loss $\mathcal{L}_{vote}$, centerness loss $\mathcal{L}_{cntr}$, bounding box estimation loss $\mathcal{L}_{box}$, and classification loss $\mathcal{L}_{cls}$ for object detection, which are formulated as follows:

$$L = \beta_{vote}\mathcal{L}_{vote} + \beta_{cntr}\mathcal{L}_{cntr} + \beta_{box}\mathcal{L}_{box} + \beta_{cls}\mathcal{L}_{cls} \quad (12)$$

$\mathcal{L}_{vote}$ is a smooth- ℓ_1 loss for predicting the center offset of each voxel. In terms of proposal generation, $\mathcal{L}_{cntr}$, $\mathcal{L}_{box}$, and $\mathcal{L}_{cls}$ utilize Cross-Entropy loss, DIOU loss (Zheng et al. 2020), and Focal loss (Lin et al. 2017) to optimize object centerness, bounding box prediction, and classification, respectively. β_{vote} , β_{cntr} , β_{box} , and β_{cls} represent the corresponding coefficients and are set to 1.

Experiments

Datasets and Evaluation Metric

Our SPGroup3D is evaluated on three indoor challenging 3D scene datasets, *i.e.*, ScanNet V2 (Dai et al. 2017), SUN RGB-D (Song, Lichtenberg, and Xiao 2015), S3DIS (Armeni et al. 2016). For all datasets, we follow the standard data splits adopted in (Qi et al. 2019) and (Gwak, Choy, and Savarese 2020).

ScanNet V2. ScanNet is a richly annotated dataset that provides a comprehensive collection of 3D indoor scans, with annotated 3D rebuilt indoor scenes and bounding boxes for the 18 object categories. The dataset is divided into 1,201 training samples, with the remaining 312 used for validation.

SUN RGB-D. SUN RGB-D is a widely recognized dataset designed for 3D object detection in indoor environments. This dataset is divided into approximately 5,000 training and 5,000 validation samples. Each sample is annotated with oriented 3D bounding boxes and per-point semantic labels, covering 37 different object categories. Following the approach in (Qi et al. 2019), we select 10 categories.

S3DIS. S3DIS is a comprehensive 3D indoor dataset that comprises 3D scans from 272 rooms across 6 buildings, with annotations for both 3D instances and semantic categories. We follow the standard division, where Area 5 is used for validation, while the remains are the training subset.

We evaluate all the experiment results by a standard evaluation protocol (Qi et al. 2019; Rukhovich, Vorontsova, and Konushin 2022), which uses mean average precision (mAP) with different IoU thresholds, *i.e.*, 0.25 and 0.5.

Implementation Details

We set the voxel size as $0.02m$ for all datasets. For the backbone, we use the same backbone introduced in (Wang et al. 2022a) as the voxel feature extractor and the voxel size of output high-resolution is $0.04m$. In terms of the superpoint-based grouping, we set the iteration number to 3 and the neighbour number to 8. Moreover, following the setting in FCAF3D (Rukhovich, Vorontsova, and Konushin 2022), we set the number of positive samples to 18 in multiple matching.

In our experiments, we train the model in an end-to-end manner using the MMDetection3D framework (Contributors 2020). Following the approach in (Rukhovich, Vorontsova, and Konushin 2022, 2023), we employ the AdamW optimizer (Kingma and Ba 2014) with batch size, initial learning rate, and weight decay set to 4, 0.001, and 0.0001, respectively. Training is performed for 15 epochs on each dataset, with a learning rate decay by a factor of 10 at the 9-th and 12-th epochs. The experiments are conducted on four

Methods	Presented at	ScanNet V2		SUN RGB-D		S3DIS	
		mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
VoteNet (Qi et al. 2019)	ICCV'19	58.6	33.5	57.7	-	-	-
3D-MPA (Engelmann et al. 2020)	CVPR'20	64.2	49.2	-	-	-	-
HGNet (Chen et al. 2020)	CVPR'20	61.3	34.4	61.6	-	-	-
MLCVNet (Xie et al. 2020)	CVPR'20	64.5	41.4	59.8	-	-	-
GSDN (Gwak et al. 2020)	ECCV'20	62.8	34.8	-	-	47.8	25.1
H3DNet (Zhang et al. 2020)	ECCV'20	67.2	48.1	60.1	39.0	-	-
BRNet (Cheng et al. 2021)	CVPR'21	66.1	50.9	61.1	43.7	-	-
3DETR (Misra et al. 2021)	ICCV'21	65.0	47.0	59.1	32.7	-	-
VENet (Xie et al. 2021)	ICCV'21	67.7	-	62.5	39.2	-	-
GroupFree (Liu et al. 2021)	ICCV'21	69.1 (68.6)	52.8 (51.8)	63.0 (62.6)	45.2 (44.4)	-	-
RBGNet (Wang et al. 2022b)	CVPR'22	70.6 (69.6)	55.2 (54.7)	64.1 (63.6)	47.2 (46.3)	-	-
HyperDet3D (Zheng et al. 2022)	CVPR'22	70.9	57.2	63.5	47.3	-	-
FCAF3D (Rukhovich et al. 2022)	ECCV'22	71.5 (70.7)	57.3 (56.0)	64.2 (63.8)	48.9 (48.2)	66.7 (64.9)	45.9 (43.8)
CAGroup3D* (Wang et al. 2022a)	NeurIPS'22	73.2	57.1	-	-	-	-
SPGroup3D (ours)	-	74.3 (73.5)	59.6 (58.3)	65.4 (64.8)	47.1 (46.4)	69.2 (67.7)	47.2 (43.6)

Table 1: 3D detection results on validation of ScanNet V2, SUN RGB-D, and S3DIS datasets. The main comparison is based on the best results of multiple experiments between different methods, and the average value of 25 trials is given in brackets. Since we only focus on one-stage methods in this paper, we report the one-stage result from the ablation study of the origin paper (Wang et al. 2022a) for fair comparison, dubbed as “CAGroup3D*”.

Geometry	Group	Match	mAP@0.25	mAP@0.5
			68.2	53.2
✓			71.3	57.0
✓	✓		72.7	57.9
✓	✓	✓	73.5	58.3

Table 2: Ablation study of key components including geometry-aware voting (Geometry), superpoint-based grouping (Group), and multiple matching (Match).

NVIDIA RTX 3090 GPUs. We follow the evaluation scheme from (Liu et al. 2021), which involved training the models five times and testing each trained model five times. The reported results include the best and average performance across all results.

Benchmarking Results

We compare our method with the recent state-of-the-art 3D detection methods on ScanNet V2 (Dai et al. 2017), SUN RGB-D (Song, Lichtenberg, and Xiao 2015) and S3DIS (Armeni et al. 2016) benchmarks. As indicated in Tab. 1, SPGroup3D outperforms the previous state-of-the-art methods in almost all metrics. In terms of mAP@0.25, our method achieves 1.1, 1.2, and 2.5 improvements over the previous state-of-the-art methods on ScanNet, SUN RGB-D and S3DIS, respectively. Regarding mAP@0.5, our method shows 2.3 and 1.3 improvements on ScanNet and S3DIS. The visualization of 3D Object detection with predicted bounding boxes on Scannet V2 is shown in Fig. 5.

Ablation Study

We conduct extensive ablation studies on the validation set of ScanNet V2 to analyze individual components of our

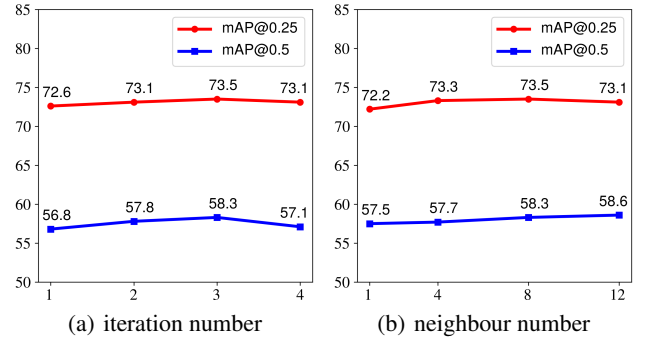


Figure 4: Comparison of performance results for different iteration numbers and neighbour numbers.

proposed method.

Effect of different components of SPGroup3D. We first ablate the effects of different components of SPGroup3D. As seen in Tab. 2, the base model (1st row) is the fully sparse convolutional VoteNet-style (Qi et al. 2019; Wang et al. 2022a) model. Comparing the 1st and 2nd row, we introduce a geometry-aware voting (dubbed as “Geometry”), eliminate superpoint-based grouping and multiple matching, and directly group voxels into superpoints. The results of this variant model are significantly improved from 68.2 to 71.3, and mAP@0.5 from 53.2 to 57.0. Comparing the 2nd and 3rd row, by adding several superpoint based groupings (dubbed as “Group”), the performance of this variant further improved, i.e., mAP@0.25 and mAP@0.5, 71.3 to 72.7 and 57.0 to 57.9, respectively. In conjunction with multiple matching (dubbed as “Match”), from the 4th row, the result of mAP@0.25 increased from 72.7 to 73.5, and mAP@0.5 from 57.9 to 58.3. These experiments

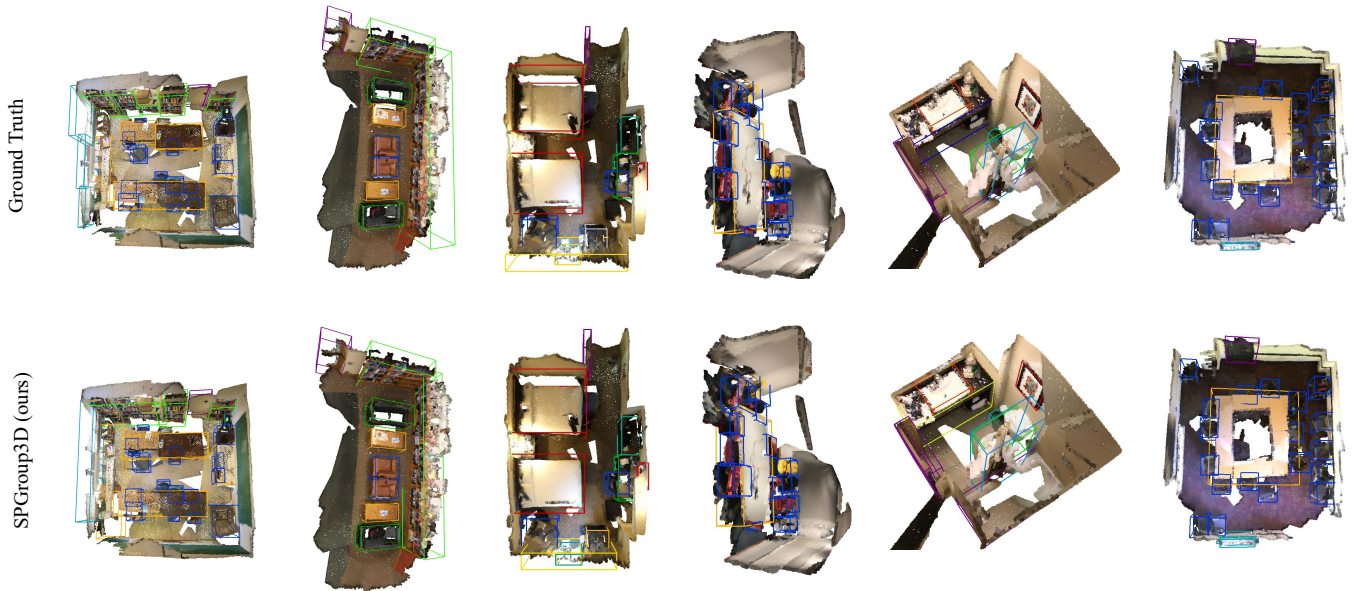


Figure 5: Qualitative results on validation of ScanNet V2. Different classes are indicated by bounding boxes in different colors.

Settings	mAP@0.25	mAP@0.5
Semantic-agnostic	68.2	53.2
Semantic-aware	73.2	57.1
SPGroup3D (ours)	73.5	58.3

Table 3: Comparison with other grouping-based strategies.

demonstrate the effectiveness of our geometry-aware voting, superpoint-based grouping and multiple matching.

Effect of superpoint based grouping. In the superpoint-based grouping, there are two hyperparameters: iteration number ($iter$) and neighbour number (k). According to Fig. 4, we can see that the model is insensitive to changes in hyperparameters. Fig. 4(a) shows that the best results are obtained at an iteration number of 3 (73.5 on mAP@0.25 and 58.3 on mAP@0.5). Therefore, in our experiments, we set the number of iterations to 3. We choose several values for the number of neighbours: 1, 4, 8, and 12, as shown in Fig. 4(b). The results of mAP@0.25 and mAP@0.5 are 72.2, 57.5 ($k=1$), 73.3, 57.7 ($k=4$), 73.5, 58.3 ($k=8$) and 73.1, 58.6 ($k=12$), respectively. In order to strike a balance between performance and efficiency, this paper chooses 8 as the number of neighbours.

To further demonstrate the effect of superpoint-based grouping, we compare it with other grouping strategies. In Tab. 3, “semantic-agnostic” and “semantic-aware” correspond to VoteNet-style model and the one-stage CA-Group3D (Wang et al. 2022a), respectively. Our method achieve more reliable detection results (73.5 for mAP@0.25, 58.3 for mAP@0.5) compared to these methods. This proves that instance-aware proposals produced by superpoint-based

grouping lead to better detection.

Effect of geometry-aware voting. In our geometry-aware voting, we need to preserve both features before and after voting. Here we study its impact. Only using the features before voting achieves 72.0 for mAP@0.25 and 55.1 for mAP@0.5. The variant model only with the features after voting obtains 70.9 for mAP@0.25 and 56.2 for mAP@0.5, and with both features can achieve 73.5 for mAP@0.25 and 58.3 for mAP@0.5. It can be observed that by utilizing both features, our model achieves improvement. This shows that preserving the geometric distribution of superpoints in coordinate space is beneficial to the anchor-free method. Additionally, we claim that our model is robust to incorrectly partitioned superpoints, which usually appear at the edges of objects and are more likely to generate low-quality proposals. With geometry-aware voting, the position of superpoints relative to the center of objects remains unchanged, allowing us to easily use centerness scores to filter out low-quality proposals during post-processing.

Conclusion

In this paper, we proposed a novel end-to-end one-stage method, SPGroup3D, for indoor 3D object detection. SPGroup3D first utilizes geometry-aware voting to refine the positions of superpoints and then employs superpoint-based grouping to group the bottom-up latent features of voxels into superpoints. During the training phase, multiple matching is used to select positive superpoint-based proposals. Extensive experiments on ScanNet V2, SUN RGB-D, and S3DIS benchmarks demonstrate that the proposed method achieves state-of-the-art performance on the indoor one-stage 3D object detection task.

Acknowledgments

This work was supported by the National Science Fund of China (Grant Nos. 62276144, 62306238) and the Fundamental Research Funds for the Central Universities.

References

- Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1534–1543.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision*, 213–229.
- Chen, C.; Chen, Z.; Zhang, J.; and Tao, D. 2022. SASA: Semantics-Augmented Set Abstraction for Point-Based 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 221–229.
- Chen, J.; Lei, B.; Song, Q.; Ying, H.; Chen, D. Z.; and Wu, J. 2020. A Hierarchical Graph Network for 3D Object Detection on Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 392–401.
- Cheng, B.; Sheng, L.; Shi, S.; Yang, M.; and Xu, D. 2021. Back-tracing Representative Points for Voting-based 3D Object Detection in Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8963–8972.
- Choy, C.; Gwak, J.; and Savarese, S. 2019. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3075–3084.
- Contributors, M. 2020. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d>.
- Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5828–5839.
- Engelmann, F.; Bokeloh, M.; Fathi, A.; Leibe, B.; and Nießner, M. 2020. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9031–9040.
- Feng, X.; Du, H.; Fan, H.; Duan, Y.; and Liu, Y. 2023. SEFormer: Structure Embedding Transformer for 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 632–640.
- Gwak, J.; Choy, C.; and Savarese, S. 2020. Generative Sparse Detection Networks for 3D Single-Shot Object Detection. In *Proceedings of the European Conference on Computer Vision*, 297–313.
- He, Q.; Wang, Z.; Zeng, H.; Zeng, Y.; and Liu, Y. 2022. SVGA-Net: Sparse Voxel-Graph Attention Network for 3D Object Detection from Point Clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 870–878.
- Hui, L.; Tang, L.; Dai, Y.; Xie, J.; and Yang, J. 2023. Efficient LiDAR Point Cloud Oversegmentation Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18003–18012.
- Hui, L.; Tang, L.; Shen, Y.; Xie, J.; and Yang, J. 2022. Learning Superpoint Graph Cut for 3D Instance Segmentation. *Advances in Neural Information Processing Systems*, 36804–36817.
- Hui, L.; Yuan, J.; Cheng, M.; Xie, J.; Zhang, X.; and Yang, J. 2021. Superpoint Network for Point Cloud Oversegmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5510–5519.
- Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.
- Landrieu, L.; and Boussaha, M. 2019. Point Cloud Oversegmentation with Graph-Structured Deep Metric Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7440–7449.
- Landrieu, L.; and Simonovsky, M. 2018. Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4558–4567.
- Liang, J.; An, P.; and Ma, J. 2022. Distribution Aware VoteNet for 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1583–1591.
- Liang, Z.; Li, Z.; Xu, S.; Tan, M.; and Jia, K. 2021. Instance Segmentation in 3D Scenes using Semantic Superpoint Tree Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2783–2792.
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2980–2988.
- Liu, Z.; Zhang, Z.; Cao, Y.; Hu, H.; and Tong, X. 2021. Group-Free 3D Object Detection via Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2949–2958.
- Misra, I.; Girdhar, R.; and Joulin, A. 2021. An End-to-End Transformer Model for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2906–2917.
- Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep Hough Voting for 3D Object Detection in Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9277–9286.
- Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet: Deep learning on Point sets for 3D Classification and Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 652–660.
- Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Advances in Neural Information Processing Systems*, 4490–4499.

- Rukhovich, A.; Vorontsova, A.; and Konushin, A. 2022. FCAF3D: Fully Convolutional Anchor-Free 3D Object Detection. In *Proceedings of the European Conference on Computer Vision*, 477–493.
- Rukhovich, D.; Vorontsova, A.; and Konushin, A. 2023. TR3D: Towards Real-Time Indoor 3D Object Detection. *arXiv preprint arXiv:2302.02858*.
- Shen, Y.; Hui, L.; Xie, J.; and Yang, J. 2023. Self-Supervised 3D Scene Flow Estimation Guided by Superpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5271–5280.
- Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10529–10538.
- Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 567–576.
- Sun, J.; Fang, H.-S.; Zhu, X.; Li, J.; and Lu, C. 2022. Correlation Field for Boosting 3D Object Detection in Structured Scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2298–2306.
- Sun, J.; Qing, C.; Tan, J.; and Xu, X. 2023. Superpoint Transformer for 3D Scene Instance Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2393–2401.
- Tang, L.; Hui, L.; and Xie, J. 2022. Learning Inter-Superpoint Affinity for Weakly Supervised 3D Instance Segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 1282–1297.
- Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9627–9636.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is all You Need. *Advances in Neural Information Processing Systems*, 6000–6010.
- Wang, H.; Dong, S.; Shi, S.; Li, A.; Li, J.; Li, Z.; Wang, L.; et al. 2022a. CAGroup3D: Class-Aware Grouping for 3D Object Detection on Point Clouds. *Advances in Neural Information Processing Systems*, 29975–29988.
- Wang, H.; Shi, S.; Yang, Z.; Fang, R.; Qian, Q.; Li, H.; Schiele, B.; and Wang, L. 2022b. RBGNet: Ray-Based Grouping for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1110–1119.
- Wu, H.; Wen, C.; Li, W.; Li, X.; Yang, R.; and Wang, C. 2023. Transformation-Equivariant 3D Object Detection for Autonomous Driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2795–2802.
- Xie, Q.; Lai, Y.-K.; Wu, J.; Wang, Z.; Lu, D.; Wei, M.; and Wang, J. 2021. VENet: Voting Enhancement Network for 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3712–3721.
- Xie, Q.; Lai, Y.-K.; Wu, J.; Wang, Z.; Zhang, Y.; Xu, K.; and Wang, J. 2020. MICVNet: Multi-Level Context VoteNet for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10447–10456.
- Zhang, Z.; Sun, B.; Yang, H.; and Huang, Q. 2020. H3DNet: 3D Object Detection Using Hybrid Geometric Primitives. In *Proceedings of the European Conference on Computer Vision*, 311–329.
- Zheng, Y.; Duan, Y.; Lu, J.; Zhou, J.; and Tian, Q. 2022. HyperDet3D: Learning a Scene-conditioned 3D Object Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5585–5594.
- Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and Better Learning for Bounding Box Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12993–13000.
- Zhou, Y.; and Tuzel, O. 2018. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4490–4499.