Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062161

ARTICLE

A Category-Agnostic Hybrid Contrastive Learning Method for Few-Shot Point Cloud Object Detection

Xuejing Li*

The MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, 200240, China

*Corresponding Author: Xuejing Li. Email: lixuejing@sjtu.edu.cn

Received: 11 December 2024; Accepted: 25 February 2025; Published: 16 April 2025

ABSTRACT: Few-shot point cloud 3D object detection (FS3D) aims to identify and locate objects of novel classes within point clouds using knowledge acquired from annotated base classes and a minimal number of samples from the novel classes. Due to imbalanced training data, existing FS3D methods based on fully supervised learning can lead to overfitting toward base classes, which impairs the network's ability to generalize knowledge learned from base classes to novel classes and also prevents the network from extracting distinctive foreground and background representations for novel class objects. To address these issues, this thesis proposes a category-agnostic contrastive learning approach, enhancing the generalization and identification abilities for almost unseen categories through the construction of pseudo-labels and positive-negative sample pairs unrelated to specific classes. Firstly, this thesis designs a proposal-wise context contrastive module (CCM). By reducing the distance between foreground point features and increasing the distance between foreground and background point features within a region proposal, CCM aids the network in extracting more discriminative foreground and background feature representations without reliance on categorical annotations. Secondly, this thesis utilizes a geometric contrastive module (GCM), which enhances the network's geometric perception capability by employing contrastive learning on the foreground point features associated with various basic geometric components, such as edges, corners, and surfaces, thereby enabling these geometric components to exhibit more distinguishable representations. This thesis also combines category-aware contrastive learning with former modules to maintain categorical distinctiveness. Extensive experimental results on FS-SUNRGBD and FS-ScanNet datasets demonstrate the effectiveness of this method with average precision exceeding the baseline by up to 8%.

KEYWORDS: Contrastive learning; few-shot learning; point cloud object detection

1 Introduction

Deep learning based point cloud object detection has seen tremendous advancements over the past decade. Nonetheless, most existing methods [1–3] require a vast amount of labeled data for fully-supervised training. Such detectors can effectively identify objects belonging to categories present in the training set. However, they often struggle with rare categories that have only a few annotated samples. When confronted with a new task scenario, it is impractical to quickly obtain a large, annotated training dataset. Therefore, few-shot learning (FSL) has been proposed to improve the generalization capability over categories of neural networks, enabling them to quickly adapt to new task scenarios with almost unseen categories using only a few annotated samples. FSL aims to generalize the knowledge acquired from extensively labeled base classes to the understanding of novel classes with very limited labeled samples. It has proven effective across a

range of visual understanding tasks from classification [4–6] and semantic segmentation [7–9] to object detection [10–13].

More recently, preliminary efforts have also been made to introduce FSL to point cloud object detection [14–17]. Nonetheless, due to the sparsity of point cloud data and the coarse granularity of detection features, it is non-trivial to deploy off-the-shelf point cloud object detectors within an existing FSL framework. A main challenge of few-shot point cloud object detection (FS3D) lies in generalizing the detection ability of 3D detectors to almost unseen classes using only a few training samples [13]. Previous FS3D algorithms primarily employ a prototype-learning strategy. These methods utilize annotated samples as support point clouds to extract information called prototypes, which are a set of vectors that represent the typical characteristics of certain categories or data patterns and are later used to guide the detection of the input query point cloud. Prototypical VoteNet [14] and P-VAE [17] both extract primitive geometric prototypes and category-aware semantic prototypes from annotated training samples to enhance the network's perception of unlabeled query scenes. Yet, these methods utilize the provided labeled data for supervised training. Due to the overwhelming number of samples in base classes, lacking category-agnostic supervision for these prototypes can render the network biased and overfitted toward base classes, consequently impairing its generalization ability over novel classes. Moreover, Due to the lack of sufficient understanding of novel classes, they cannot be distinctly separated from the background.

Driven by the insights above, this thesis first proposes a novel category-agnostic contrastive learning approach for FS3D. This approach aims to enhance the network's ability to generalize to novel classes by applying class-agnostic supervision to point features extracted by the network and reducing its reliance on categorical labels. The category-agnostic contrastive learning comprises a geometric contrastive module (GCM) and a context contrastive module (CCM). Through similarity assignment, each foreground point feature can be categorized into a corresponding three-dimensional geometric component, such as edges, corners, or surfaces. By drawing point features that belong to the same geometric component closer and pushing those that belong to different geometric components apart, GCM ensures that the geometric components extracted by the network exhibit more discriminative representations and enhances the network's perception of 3D spatial relationships. CCM is proposal-wise, reduces the distance among foreground features within a proposal, which is a region identified by the network as potentially containing an object, and increases the distance between background and foreground features. This module aims to enrich region proposals with comprehensive contextual information while distinguishing the feature representations between foreground and background points, thereby preventing potential confusion in the surrounding environments of foreground objects. From 1-shot to 5-shot settings, the network has access to an increasing number of labeled training samples of novel classes, which means the network can acquire more category-aware knowledge from novel classes. Therefore, this thesis further employs category-aware contrastive learning to incorporate enhanced category-aware semantic information. This design aids in enhancing the network's ability to perceive novel classes by bringing features of the same class closer while pushing features of different classes apart.

In the following content, this paper introduces recent related works in Section 2; in Section 3, this paper provides descriptions of my methodology; Section 4 presents the experimental settings, a brief introduction to datasets and analysis of experimental results; in Section 5, this paper draws conclusions and discuss the possible future work.

2 Related Works

Few-shot 2D object detection (FS2D) works can be categorized into two primary approaches. The first approach [10,13] utilizes labeled samples to form a support branch, extracting guiding information to

refine the features of the query input, thereby aiding in its detection. The second approach [11,12] employs a pre-training and fine-tuning strategy. It involves pre-training the network on fully labeled base classes to develop its feature extraction capabilities, followed by fine-tuning using novel classes to adapt the prediction head to detect novel objects. Correspondingly, in FS3D works, approaches such as MetaDet3D [15] utilize prototypes extracted from labeled samples to assist in the detection of query point cloud scenes. On the other hand, Generalized FS3D [16] employs the strategy of fine-tuning additional prediction heads to adapt to the detection of novel classes.

As one of the earlier indoor FS3D works, MetaDet3D [15] extracts class-specific reweighting vectors from labeled samples. It subsequently uses these vectors to refine the features of the point clouds to be detected and the proposal features through the channel-wise product. This method effectively integrates the information from support samples into the query point cloud, enabling the network to perform object detection of respective categories under the guidance of category-specific information, which refers to the semantic categories of objects manually annotated, such as tables, chairs, etc. However, focusing solely on category-specific information can lead to overfitting the base classes, thereby weakening the network's ability to generalize to novel classes. In addition to category prototypes, Prototypical VoteNet [14] utilizes the similarity clustering of foreground point features to extract geometric prototypes shared among different categories. These category-agnostic geometric prototypes enhance the network's spatial perception ability of point cloud scenes. However, these geometric prototypes formed through similarity clustering are too rudimentary and cannot clearly represent the features of different geometric components. P-VAE [17] identifies a lack of fine-grained supervision in existing FS3D methods. Therefore, it introduces an additional point cloud reconstruction task as auxiliary supervision beyond the detection task. Through upsampling, it generates fine-grained point features to prevent the simple average of features at a coarse-grained level. However, this approach significantly increases the training burden of the network. Moreover, relying on category-specific supervision still tends to cause the network to overfit to base classes, making the foreground features of novel classes still difficult to distinguish from their surrounding environments.

Contrastive learning [18] has emerged as a powerful paradigm for learning effective representations from unlabeled data. Recently, it has also been introduced into FSL tasks [19–21]. ContrastBoundary [22] and Contextrast [23] utilize contrastive learning on the edges of different objects to enable the segmentation network to delineate clearer object boundaries. However, the application of contrastive learning to FSL still has not escaped the constraints of category annotations. Moreover, due to the coarse granularity of detection features, contrastive learning remains unexplored in FS3D tasks.

3 Methodology

Taking Prototypical VoteNet [14] as the baseline and PointNet++ [24] as the 3D point cloud backbone, the framework of this method is illustrated in Fig. 1. To reduce the dependency of network training on class annotations and enhance its generalization capability toward novel classes, this paper proposes a category-agnostic learning approach comprised of CCM and GCM. By constructing category-agnostic positive and negative sample pairs using inputs from one minibatch, this approach provides a supervision paradigm that does not rely on category annotations and data augmentation for FS3D tasks. Specifically, CCM utilizes context contrastive learning to obtain more distinctive feature representations for foreground and background points. GCM employs contrastive learning of geometric prototypes to enable the network to extract more discriminative representations of geometric components. To ensure the network's ability to differentiate between classes, this paper also approximately incorporates category-aware contrastive learning to refine category prototypes of different classes. As illustrated in Fig. 1, each contrastive learning module takes as input the point cloud features from the corresponding hierarchical layer. Based on the contrastive

learning strategy of each module, pseudo-labels are assigned to the input features to construct positive and negative sample pairs. The features are then processed by a projection layer, followed by the computation of contrastive loss. The detailed design of each contrastive learning module will be elaborated in the subsequent sections.

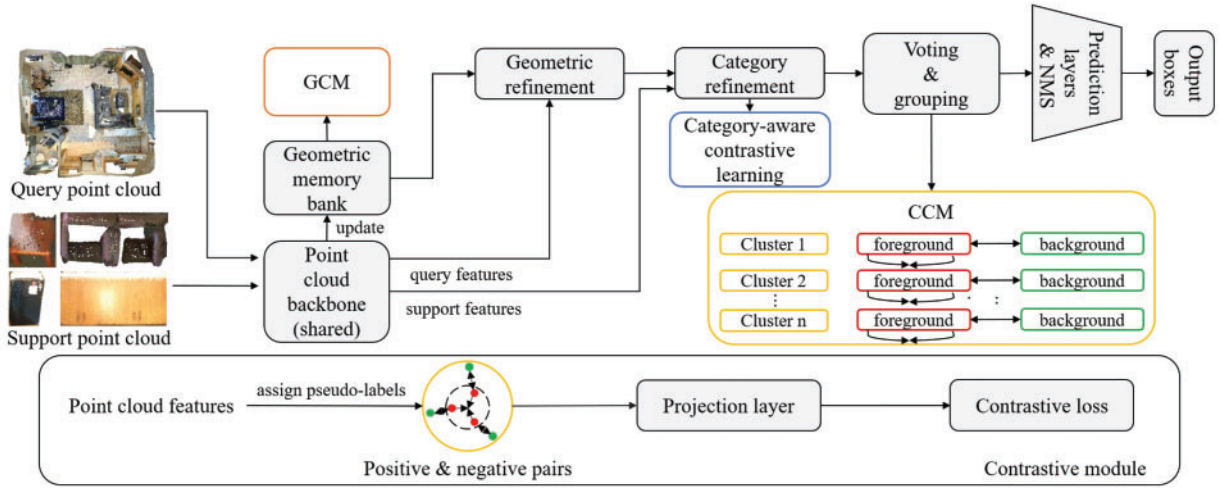

Figure 1: The proposed contrastive learning framework and the basic structure of each contrastive module

3.1 Overall Framework

3.1.1 Problem Definition

Following [14,15,17], the FS3D dataset D is divided into D_{base} and D_{novel}. Here, D_{base} represents the base classes with abundant labeled samples, while D_{novel} represents the novel classes with only a few labeled samples. $D_{base} \bigcup D_{novel} = D$, and $D_{base} \bigcap D_{novel} = \varnothing$. Typically, FS3D tasks organize data input using an episodic learning style. For an N-way K-shot FS3D task, there are N novel classes, with K labeled samples available for each novel class, and the objective is to identify and localize objects of these N categories within the input point cloud, with K labeled samples for each category serving as support instances. Each input is structured as a sub-task in the form of a $\{query, support\}$ set, consisting of a point cloud to be detected, represented by *query*, and its corresponding $N * K$ labeled samples, represented by *support*. Such sub-tasks are called episodes. A batch size of B contains B episodes. To implement episodic learning, inputs from both base classes and novel classes, as well as training and testing sets, are structured in this manner.

3.1.2 Framework

The process of this contrastive learning approach is outlined in Algorithm 1.

Algorithm 1: Framework of category-aware contrastive learning and category-agnostic contrastive learning

Input: Query point cloud *query* to be detected; Support point cloud *support* composed of labeled samples;
Output: Contrastive loss; Predicted boxes;

1: Extract point cloud features of *query* and *support* using point cloud backbone;
2: Perform **category-aware contrastive learning** using instance features extracted from *support*;
3: **Positive:** Instance features of the same category;

(Continued)

Algorithm 1 (continued)

Negative: Instance features of different categories;
Calculate **category-aware contrastive loss** L_{caw};
Allocate features of foreground points extracted from *query* to geometric memory bank by similarity;
Assign geometric pseudo-labels based on memory bank indices;
Use **GCM** for **geometric contrastive learning** and calculate **geometric contrastive loss** L_{geom};
Positive: Each foreground point feature & its corresponding geometric prototype;
Negative: Each foreground point feature & other geometric prototypes;
Refine point cloud features through geometric refinement and category refinement;
Form region proposals via voting & grouping layer;
Assign pseudo-labels based on foreground/background;
Use **CCM** for **context contrastive learning** and calculate **context contrastive loss** $L_{context}$;
Positive: Foreground points in each proposal;
Negative: Foreground & background points in each proposal;
return Weighted sum of L_{caw} and L_{cag}, **category-agnostic contrastive loss** $L_{cag} = L_{geom} + L_{context}$; Predicted boxes;

As shown in Fig. 2, after the input point cloud undergoes feature extraction via the 3D backbone, the network extracts a set of vectors from the coordinates of points, which encapsulate the positional information of the points as well as the surrounding structural information, referred to as features. The features of foreground points are allocated to a randomly initialized geometric memory bank in the order of similarity. The memory bank is used to record the prototypes of geometric components. To extract more discriminative representations of geometric components, this work employs GCM to conduct contrastive learning on these geometric prototypes. To be specific, based on the above allocation process, GCM assigns geometric pseudo-labels to each foreground point according to the index of its geometric prototype in the memory bank. Using these pseudo-labels as the basis for positive and negative sample matching, GCM treats each foreground point and its corresponding geometric prototype as a positive sample pair. Conversely, GCM considers the point and other geometric prototypes as negative sample pairs. Subsequently, the calculation of the geometric contrastive loss L_{geom} is performed.

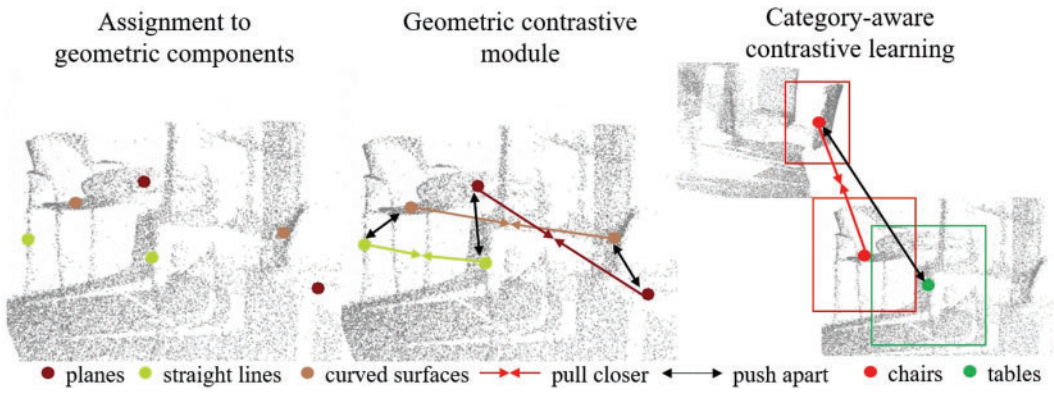

Figure 2: Illustrations of geometric contrastive module and category-aware contrastive learning

As shown in Fig. 3, after the refined point cloud features undergo the voting and grouping layer to form region proposals [25], CCM assigns pseudo-labels to points based on whether they originally belong to the

foreground or background of the point cloud scene. Within a proposal, foreground points are considered positive sample pairs, while background points are treated as negative sample pairs with foreground ones. Points in different proposals are considered independent of each other. Afterward, the context contrastive loss L_{context} is calculated.

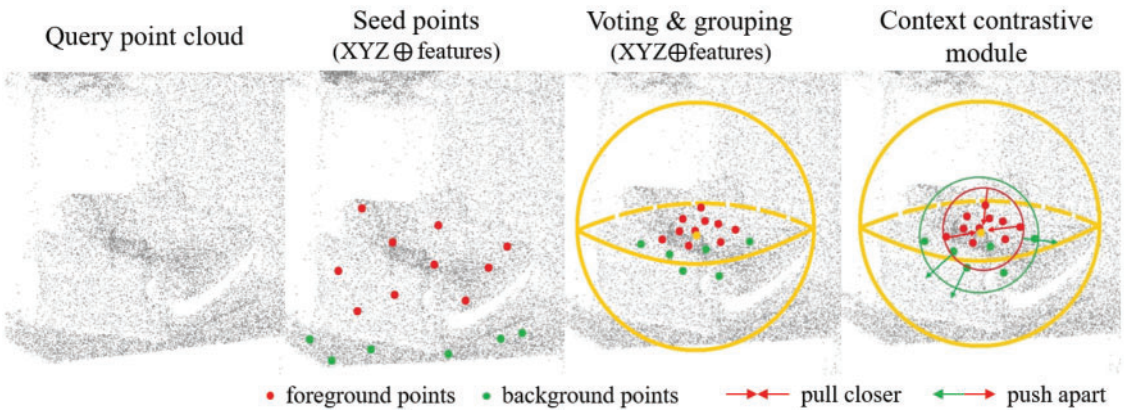

Figure 3: The process of context contrastive module

L_{geom} and L_{context} together form the category-agnostic contrastive loss L_{cag}, which is explained in detail in Section 3.2.1.

Category-aware contrastive learning is performed between the category prototypes extracted from support instances. To increase the number of sample pairs available for loss computation and to avoid the training burden that data augmentation can impose on the network, this work uses the entire minibatch of inputs to construct positive and negative sample pairs. As shown in Fig. 2, instance features belonging to the same category are considered as positive pairs, while those from different categories are treated as negative pairs for the calculation of the category-aware contrastive loss L_{caw}, as further explained in Section 3.2.2.

The training objective of this network is to optimize the loss

$$L = L_{\text{det}} + \lambda_1 L_{\text{cag}} + \lambda_2 L_{\text{caw}} \tag{1}$$

where L_{det} is the original training loss of Prototypical VoteNet [14]. λ_1 and λ_2 balance the effects of category-agnostic and category-aware contrastive learnings. For fewer shots with very limited categorical information for novel classes, due to the overwhelming number of training samples of base classes, category-aware supervision can lead to the network overfitting toward these base classes. Thus, this work sets λ_1 to be higher than λ_2 to emphasize category-agnostic contrastive learning. For more-shot settings, the network has access to richer category-aware knowledge from novel classes. Hence, this work slightly lowers λ_1 and increases λ_2 to calibrate their contributions.

3.2 Contrastive Learning Methods

3.2.1 Category-Agnostic Contrastive Learning

Context contrastive module. A classic point cloud object detector, VoteNet [25] encourages non-object seed points located near the object to vote for the object's center, thereby enhancing the contextual knowledge extracted for object detection. Such a fully-supervised 3D detector demands a large number of labeled training samples to train in order to learn clear boundaries between foreground and background.

However, in the scenario of FS3D, novel classes lack sufficient samples for the network to learn from in order to differentiate foreground from background. As a result, voting without additional supervision may lead to contextual confusion of novel classes, causing the boundaries of object features to become blurred with surrounding points, thereby impairing the model's detection capabilities. To enhance the distinction between features of objects and their surroundings and to reduce the dependence on labels, this paper proposes the proposal-wise category-agnostic CCM motivated by [22] and [23], but this method directly employs coarse-grained features tailored for the object detection task. For example, ContrastBoundary [22] introduces boundary contrastive learning in the point cloud semantic segmentation task. This approach increases the feature differentiation between points located at the boundaries of different objects. During its upsampling process, the boundary point features extracted from each layer of the network are subjected to such processing to enhance the distinction. However, in the FS3D task, redundant operations such as upsampling are typically unnecessary. Moreover, the positions of points processed through the voting layer do not precisely correspond to the actual physical boundaries. Furthermore, considering the goal of enhancing the network's generalization ability using a category-agnostic learning strategy, identifying object boundaries relies more heavily on annotations compared to distinguishing between foreground and background. The same applies to Contextrast [23] as well. Based on the above analysis, CCM directly utilizes coarse-grained point cloud features without performing upsampling and focuses on distinguishing foreground and background features within point clusters, which are later used to generate region proposals, thus enhancing the network's general perception capability for foreground objects.

As shown in Fig. 3, given a query scene, after the point cloud features are extracted by the 3D backbone and clustered through the voting and grouping module, the network obtains a series of proposals $\mathbb{P} = \{p_r\}_{r=1}^{P}$ along with the seed point features $\mathbb{S} = \{s_i\}_{i=1}^{S}$ that constitute each proposal. To be specific, $\mathbb{P}$ actually represents a series of clusters formed after the grouping layer applies farthest point sampling and clustering on the points obtained after the voting layer. The points $\mathbb{S}$ within $\mathbb{P}$ are considered highly likely to contain foreground information of the objects to be detected, as well as the surrounding environmental information. P and S represent the predefined number of proposals and the number of seed points in each proposal, respectively. Each s further integrates its coordinates and features through a multilayer perceptron (MLP), enabling contrastive learning to impose a unified constraint on the voting of both coordinates and features.

Following SimCLR [18], this work employs a projection layer $proj(\cdot)$ between $\mathbb{S}$ and the contrastive objective. $proj(\cdot)$ consists of a linear layer followed by a normalization layer, which projects the input features into a new feature space. This design aims to prevent information loss caused by backpropagation directly acting on the original input features. Concretely, the features used for CCM are generated as

$$s_i \leftarrow proj(\mathrm{MLP}(s_i)), i = 1, 2, ..., S \tag{2}$$

This thesis uses $\{s_i^f\}_{i=1}^{F}$ and $\{s_i^b\}_{i=1}^{B}$ to represent the foreground and background point features of a proposal, respectively, where F and B denote the number of foreground and background points, respectively.

The similarity of a positive pair of points in proposal p_r is calculated as

$$\mathrm{sim}_r^+ = \frac{1}{F * (F-1)} \sum_{i=1}^{F} \sum_{j=1, j\neq i}^{F} {s_i^f}^T \cdot s_j^f \tag{3}$$

and that of a negative pair is calculated as

$$\mathrm{sim}_r^- = \frac{1}{F * B} \sum_{i=1}^{F} \sum_{j=1}^{B} {s_i^f}^T \cdot s_j^b \tag{4}$$

The proposal-wise category-agnostic CCM takes the form of the InfoNCE loss [26]:

$$L_{\text{context}} = -\frac{1}{P}\sum_{r=1}^{P} log\frac{exp(\text{sim}_r^+/\tau)}{exp(\text{sim}_r^+/\tau) + exp(\text{sim}_r^-/\tau)} \tag{5}$$

where τ is the temperature parameter.

Geometric contrastive module. As illustrated in Fig. 2, the proposed GCM resembles the design of CCM. First, foreground points are assigned to the geometric component that best matches their features based on their similarity to geometric prototypes. This geometry prototype assignment method based on similarity matching can be regarded as a clustering process. Since contrastive learning inherently aims to make clusters of samples belonging to the same class more compact and those of different classes more distinct, the geometric prototypes generated through such a clustering approach can also be optimized via contrastive learning. Thus, GCM labels these points with pseudo-labels corresponding to their respective geometric components. Subsequently, point features and geometric prototypes are projected through a projection layer similar to the one in Eq. (2). Positive pairs are constructed from the geometric prototypes and the foreground point features assigned to them, while negative pairs are formed from the prototypes and the foreground point features belonging to different geometric components. The GCM loss is computed as

$$L_{\text{geom}} = -\frac{1}{G}\sum_{i=1}^{G} log\frac{exp(\text{sim}(f_i, g_i)/\tau)}{\sum_{j=1}^{G} exp(\text{sim}(f_j, g_i)/\tau)} \tag{6}$$

where f_i represents the foreground point features annotated to geometric prototype g_i, and G is the predefined number of geometric prototypes. The similarity $\text{sim}(\cdot,\cdot)$ is calculated using the dot product between feature vectors. Finally, the category-agnostic contrastive loss is simply a sum of the above two losses, i.e., $L_{\text{cag}} = L_{\text{context}} + L_{\text{geom}}$.

3.2.2 Category-Aware Contrastive Learning

From 1-shot to 5-shot settings, as the number of labeled samples in novel classes increases, the training set can provide more category-specific semantic information for these novel classes. In this context, appropriately incorporating category-aware contrastive learning helps enhance the network's reasoning about the novel classes. For the point features extracted from one support instance, this work takes the average of them as the feature representation of this entire instance. This approach helps to prevent the network from overfitting to category-specific information, given that the number of training samples for base classes still far exceeds that for novel classes. For an N-way, K-shot FS3D task, each input support branch contains objects of N categories, with K support samples per category. For an input batch of size B, since the N categories of different sub-tasks within a batch may not be identical, this work computes the similarity for each sub-task separately. Denoting one instance feature of category n as P_i^n, $i \in [1, K]$, $n \in [1, N]$, the similarity of positive pairs among the same category n in batch b, $b \in [1, B]$ can be expressed as

$$\text{sim}_b(n, n) = \frac{1}{K * (K-1)}\sum_{i=1}^{K}\sum_{j=1, j\neq i}^{K} {P_i^n}^T \cdot P_j^n \tag{7}$$

While the similarity of negative pairs between category n and category m is

$$\text{sim}_b(n, m) = \frac{1}{K * K}\sum_{i=1}^{K}\sum_{j=1}^{K} {P_i^n}^T \cdot P_j^m \tag{8}$$

Category-aware contrastive loss is computed according to the following expression:

$$L_{\text{caw}} = -\frac{1}{B}\sum_{b=1}^{B}\frac{1}{N}\sum_{n=1}^{N} log\frac{exp(\text{sim}_b(n,n)/\tau)}{\sum_{m=1}^{N} exp(\text{sim}_b(n,m)/\tau)} \tag{9}$$

4 Experiments

4.1 Experiment Setup

The experiments are conducted on one NVIDIA GeForce RTX 4090 GPU. The training and evaluation setups align with Prototypical VoteNet [14]. In the 1–3 shot task settings, λ_1 and λ_2 are set to 0.1 and 0.01, respectively; for the 4–5 shot tasks, they are set to 0.05 and 0.1. The projection layer projects the features from 256 channels to 128 channels. τ is set to 0.2 in the contrastive losses. For evaluation, this paper follows the standard 3D object detection evaluation protocol and reports mean average precision at IoU thresholds of 0.25 and 0.50 (i.e., AP_{25} and AP_{50}), respectively.

To validate the effectiveness and generalizability of this method, this paper evaluates its performance on FS-SUNRGBD and FS-ScanNet [14] datasets. To validate the effectiveness of each contrastive learning module in this study, ablation experiments on L_{context}, L_{geom}, and L_{caw} are conducted using the FS-SUNRGBD dataset with $K = 5$. To demonstrate the rationale behind the combination of category-agnostic contrastive learning and category-aware contrastive learning, I conduct comparative experiments using the FS-SUNRGBD dataset with K = 1–5. To prove the necessity of the projection layers, this paper performs ablation studies using the FS-SUNRGBD dataset with $K = 1, 3, 5$.

4.2 Datasets

Following previous studies [14,15,17], this approach is developed and assessed using FS-ScanNet and FS-SUNRGBD [14] datasets. These datasets are derived from ScanNet [27] and SUNRGBD [28] and are designed for few-shot 3D indoor scene comprehension.

FS-ScanNet is made up of 1513 point clouds categorized into 18 semantic groups. Out of these, six categories are randomly chosen as novel classes, with the remainder serving as base classes. The dataset employs two random splits, Split-1 and Split-2, for the division of base and novel classes. Each novel class is given K labeled samples, with K being 1, 3, or 5. **FS-SUNRGBD** includes 5000 RGB-D samples annotated across 10 categories. From these, four categories are randomly picked as novel classes, leaving the rest as base classes. Each novel class has K annotated examples, where K ranges from 1 to 5.

4.3 Results

Table 1 displays the outcomes of my approach on the **FS-ScanNet** dataset. My method demonstrates an AP_{25} increase of up to 6% over the baseline [14], while AP_{50} demonstrates an improvement reaching 7%. In comparison to the existing state-of-the-art [17], my method also offers an enhancement of up to 5%.

Table 1: Results on **FS-ScanNet**

Method	Novel Split 1							Novel Split 2						
	Novel num.[2]	1-shot		3-shot		5-shot		Novel num.[2]	1-shot		3-shot		5-shot	
		AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}		AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}
Prototypical VoteNet [14]	6	15.34	8.25	31.25	16.01	32.25	19.52	6	11.01	2.21	21.14	8.39	28.52	12.35
VoteNet-JT[1] [25]	2	1.35	–	1.50	–	1.95	–	4	1.18	–	4.83	–	5.83	–
VoteNet-FT[1] [25]	2	3.85	–	3.90	–	4.85	–	4	2.65	–	4.20	–	5.23	–
VoteNet-2[1] [25]	2	4.85	–	4.90	–	6.40	–	4	3.20	–	5.27	–	6.57	–
MetaDet3D [15]	6	10.28	4.03	23.42	10.64	25.65	13.88	6	5.21	1.32	15.44	4.37	22.13	7.09
Generalized FS3D [16]	6	12.03	8.19	24.90	10.26	29.29	16.67	6	9.19	1.87	19.41	6.80	25.18	12.74
P-VAE [17]	6	16.00	10.22	31.60	19.37	32.84	22.39	6	12.66	4.15	21.27	10.09	31.70	14.43
Mine	6	**17.50**	**10.34**	**34.41**	**19.75**	**38.25**	**25.85**	6	**15.41**	**5.03**	**25.67**	**11.09**	**35.59**	**16.20**

Note: [1] These methods actually employ 10-shot, 30-shot, and 50-shot configurations and only report AP_{25}, which are respectively listed under the 1-shot, 3-shot, and 5-shot columns in the table. [2] "Novel num." refers to the number of novel categories.

As shown in Table 2, my method achieves an improvement of over 7% in AP_{25} compared to the baseline [14], and up to 6% in AP_{50} on the **FS-SUNRGBD** dataset. Moreover, my method outperforms state-of-the-art P-VAE [17] in the 1, 2, 3, and 5-shot settings by a margin up to 2% in AP_{25} and 3% in AP_{50}, while performing on par with P-VAE in the 4-shot setting.

Table 2: Results on **FS-SUNRGBD**

Method	Novel num.[2]	1-shot		2-shot		3-shot		4-shot		5-shot	
		AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}
Prototypical VoteNet [14]	4	12.39	1.52	14.54	3.05	21.51	6.13	24.78	7.17	29.95	8.16
VoteNet-JT[1] [25]	2	3.83	–	–	–	4.83	–	–	–	5.83	–
VoteNet-FT[1] [25]	2	7.00	–	–	–	8.10	–	–	–	9.80	–
VoteNet-2[1] [25]	2	8.73	–	–	–	9.00	–	–	–	11.30	–
MetaDet3D [15]	4	6.77	0.73	8.29	1.21	15.37	2.99	19.60	4.67	24.22	5.68
Generalised FS3D [16]	4	6.81	1.58	12.21	2.02	17.52	4.69	22.12	5.97	22.84	6.76
P-VAE [17]	4	14.36	2.42	22.28	4.30	27.70	8.73	31.55	13.84	33.21	13.98
Mine	4	17.11	4.76	23.04	6.71	29.30	12.58	32.04	13.73	34.07	14.38

Note: [1] These methods actually employ 10-shot, 30-shot, and 50-shot configurations and only report AP_{25}, which are respectively listed under the 1-shot, 3-shot, and 5-shot columns in the table. [2] "Novel num." refers to the number of novel categories.

Quantitative results indicate that my method can more accurately detect novel class objects from point cloud scenes. It not only succeeds in isolating more novel class objects from the background but also more precisely identifies the category to which the object belongs, as illustrated in Fig. 4.

Table 3 validates the effectiveness of each contrastive learning module and their combinations. As shown in the table, using each module individually results in performance improvements compared to the baseline. Combining any two modules yields further performance improvements. By jointly considering foreground-background information, geometric information, and semantic information, utilizing all three modules achieves the best performance. Table 4 demonstrates the effectiveness of the weight settings of the combination of category-agnostic and category-aware contrastive learning. The results indicate that in the 1–3 shot settings, overly focusing on categorical information can harm the network's identification

capability for novel classes. In the 4–5 shot settings, incorporating category-aware contrastive learning can enhance the network's reasoning about novel classes. Table 5 examines how projection layers influence model performance. Through the addition and removal of projection layers, this paper illustrates their importance. It is evident that the exclusion of these projection layers leads to a notable decrease in network performance.

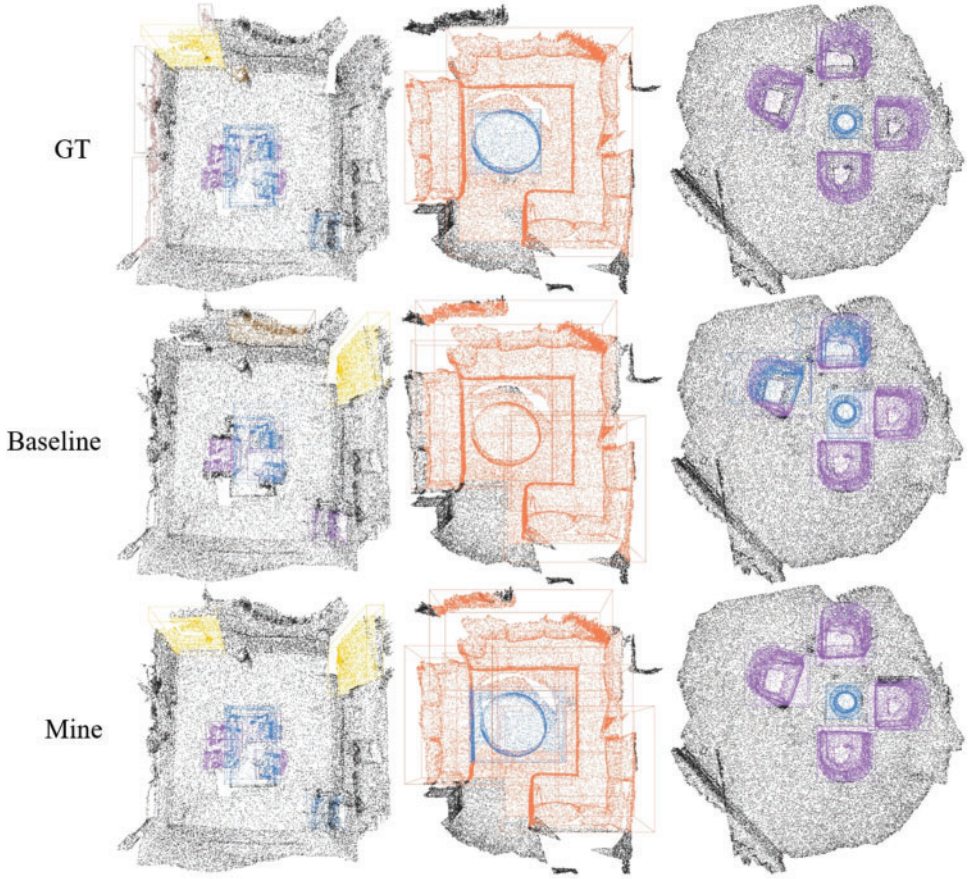

Figure 4: Qualitative results on novel split-1 of FS-ScanNet with $K = 5$

Table 3: Ablation studies on L_{context}, L_{geom} and L_{caw} using **FS-SUNRGBD** with $K = 5$

Method			5-shot	
L_{context}	L_{geom}	L_{caw}	AP_{25}	AP_{50}
–	–	–	29.95	8.16
+	–	–	32.23	9.00
–	+	–	32.08	10.34
–	–	+	31.71	11.32
+	+	–	33.33	12.69
+	–	+	32.77	11.90
–	+	+	33.55	13.56
+	+	+	34.07	14.38

Table 4: Comparative experiments for different weight settings on **FS-SUNRGBD**

Method		1-shot		2-shot		3-shot		4-shot		5-shot	
λ_1	λ_2	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}
0.1	0.01	17.11	4.76	23.04	6.71	29.30	12.58	29.65	11.61	32.61	13.31
0.05	0.1	16.07	4.05	22.95	6.42	28.85	11.18	32.04	13.73	34.07	14.38

Table 5: Ablation study on projection layers using **FS-SUNRGBD** with $K = 1$, 3 and 5

Method	1-shot		3-shot		5-shot	
projection layers	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}
–	16.02	3.47	27.39	9.78	31.10	13.35
+	17.11	4.76	29.30	12.58	34.07	14.38

4.4 Discussion

VoteNet-JT, VoteNet-FT, and VoteNet-2 [25] represent models trained on FS3D datasets using the VoteNet [25] framework. Specifically, VoteNet-JT refers to the model trained through joint training on both base and novel classes without applying fine-tuning. VoteNet-FT denotes the model pre-trained on base classes and subsequently fine-tuned using novel classes. Finally, VoteNet-2 corresponds to the model trained using the episodic learning strategy described in Section 3.1.1. The results demonstrate that simply applying conventional 3D detectors to the FS3D task does not yield satisfactory performance. Therefore, building upon the baseline, my method introduces three contrastive learning modules, incorporating prototype extraction and feature optimization. These adjustments allow the network to leverage information from annotated samples more effectively, providing robust guidance for novel class recognition and significantly improving FS3D performance. Moreover, my method significantly outperforms these three approaches in scenarios involving the detection of a larger number of novel classes (i.e., two/four novel classes vs. six novel ones on FS-ScanNet and two vs. four on FS-SUNRGBD), even with fewer annotations for these novel classes, as shown in Tables 1 and 2. This demonstrates the superior inter-class generalization capability of my method. Compared to MetaDet3D [15], which utilizes category prototypes to guide query point cloud detection, my method employs category-aware contrastive learning to extract more distinctive category prototypes. This enhancement improves the network's ability to distinguish between categories, thereby increasing its capability to identify novel class objects among numerous point features. In contrast to Generalized FS3D [16], which fine-tunes additional detection heads, my contrastive learning approach can impose constraints on the network's feature extractor. This results in an overall enhancement of the network's feature perception capabilities. Compared to Prototypical VoteNet [14], which utilizes category prototypes and geometric prototypes, my method constrains the feature spaces of both, making their feature representations more discriminative. Additionally, the introduction of context contrastive learning enhances the network's ability to distinguish between foreground and background features, thereby comprehensively improving its performance on FS3D tasks. Compared to P-VAE [17], which incorporates a scene reconstruction task as fine-grained supervision, the class-agnostic contrastive learning reduces reliance on fine-grained class annotations, thereby decreasing the network's overfitting to base classes and enhancing its generalization capability toward novel classes. Furthermore, my method directly employs coarse-grained detection features

for contrastive learning and constructs positive and negative sample pairs within a minibatch, avoiding the use of additional data that could increase the network's training burden.

Regarding the weights of category-agnostic and category-aware contrastive loss, this paper believes that as the categorical information for novel classes varies across different task settings, it is necessary to apply different weight configurations to these contrastive learning components. In the 1–3 shot settings, due to the scarcity of labeled samples in novel classes, this paper mainly employs category-agnostic contrastive learning with minimal reliance on category-aware contrastive learning. While for the 4 and 5-shot settings, where labeled samples in novel classes are slightly increased, this paper increases the weight of the category-aware contrastive loss and slightly reduces the emphasis on category-agnostic contrastive loss for better performance. Furthermore, the usage of projection layers in the contrastive learning components is necessary because projection layers help preserve feature information by cushioning the direct effects of contrastive loss on the original features.

In practical application scenarios, the proposed method can be applied to autonomous driving, drone navigation, household cleaning robots, and other contexts where models need to quickly adapt to new environments, particularly when these environments contain instances that are rarely seen in the training dataset. However, deep learning methods are inherently data-driven approaches, and the limitation of few-shot learning lies in the scarcity of data for novel categories, as well as the lack of more comprehensive annotations, such as multimodal information. Therefore, obtaining more annotated data, especially multimodal data, such as the combination of 2D images and 3D point clouds, and performing multimodal training may further enhance the performance of few-shot point cloud object detection networks.

5 Conclusion

In this thesis, I have presented a novel framework for the FS3D task, which encompasses both category-agnostic and category-aware contrastive learning. Such a hybrid contrastive learning method not only provides a supervision paradigm for FS3D tasks without reliance on category labels but also maintains the network's ability to distinguish different categories. The category-agnostic contrastive learning approach comprises the geometric contrastive module and the context contrastive module. By enhancing geometric prototype representations and foreground-background distinction within proposals, the category-agnostic contrastive learning significantly boosts the network's generalization ability from base classes to novel classes. The category-aware contrastive learning enables the network to extract more distinguishable prototypes for different classes. By integrating category-aware contrastive learning, this thesis also ensures the network's ability to discriminate between classes. The experimental results prove the effectiveness of this method and the soundness of my design. This research demonstrates that reducing reliance on class labels during training is crucial for improving the detection performance of FS3D models on novel classes. However, due to computational power constraints, this research is conducted on lightweight network architectures. As computational power advances, exploring the integration of FS3D with multimodal large models holds promising potential for future developments.

Acknowledgement: I would like to thank the staff at the MoE Key Laboratory for their technical assistance. I also appreciate the constructive suggestions provided by the reviewers and editors, which have significantly improved the quality of this thesis.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: The data that support the findings of this study are openly available on GitHub at https://github.com/cvmi-lab/fs3d (accessed on 24 February 2025). The code for this work is available on GitHub at https://github.com/offscuminSJTU/HC (accessed on 24 February 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest to report regarding the present study.

References

1. Rukhovich D, Vorontsova A, Konushin A. TR3D: towards real-time indoor 3D object detection. arXiv:230202858. 2023.
2. Shen Y, Geng Z, Yuan Y, Lin Y, Liu Z, Wang C, et al. V-DETR: DETR with vertex relative position encoding for 3D object detection. arXiv:230804409. 2023.
3. Yang H, Shi C, Chen Y, Wang L. Boosting 3D object detection via object-focused image fusion. arXiv:220710589. 2022.
4. Zhang B, Ling H, Li P, Wang Q, Shi Y, Wu L, et al. Multi-head attention graph network for few shot learning. Comput Mater Contin. 2021;68(2):1505–17. doi:10.32604/cmc.2021.016851.
5. Hegselmann S, Buendia A, Lang H, Agrawal M, Jiang X, Sontag D. TabLLM: few-shot classification of tabular data with large language models. In: Proceedings of the 26th International Conference on Artificial Intelligence and Statistics; Valencia, Spain. 2023.
6. Ke T, Cao H, Ling Z, Zhou F. Revisiting logistic-softmax likelihood in bayesian meta-learning for few-shot classification. In: Proceedings of Neural Information Processing Systems; New Orleans, LA, USA. 2023.
7. Yang Y, Chen Q, Feng Y, Huang T. MIANet: aggregating unbiased instance and general information for few-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Vancouver, BC, Canada. 2023.
8. Xu S, Zhang L, Jiang G, Hua Y, Liu Y. Part-whole relational few-shot 3D point cloud semantic segmentation. Comput Mater Contin. 2024;78(3):3021–39. doi:10.32604/cmc.2023.045853.
9. Liu SA, Zhang Y, Qiu Z, Xie H, Zhang Y, Yao T. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Vancouver, BC, Canada. 2023.
10. Fan Q, Zhuo W, Tang CK, Tai YW. Few-shot object detection with attention-RPN and multi-relation detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Seattle, WA, USA. 2020.
11. Zhang D, Pu H, Li F, Ding X, Sheng VS. Few-shot object detection based on the transformer and high-resolution network. Comput Mater Contin. 2023;74(2):3439–54. doi:10.32604/cmc.2023.027267.
12. Wu S, Pei W, Mei D, Chen F, Tian J, Lu G. Multi-faceted distillation of base-novel commonality for few-shot object detection. In: Proceedings of the European Conference on Computer Vision; Tel Aviv, Israel. 2022.
13. Han G, Ma J, Huang S, Chen L, Chang SF. Few-shot object detection with fully cross-transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; New Orleans, LA, USA. 2022.
14. Zhao S, Qi X. Prototypical votenet for few-shot 3D point cloud object detection. In: Proceedings of Advances in Neural Information Processing Systems; New Orleans, LA, USA. 2022.
15. Yuan S, Li X, Huang H, Fang Y. Meta-Det3D: learn to learn few-shot 3D object detection. In: Proceedings of the Asian Conference on Computer Vision; Macau, China. 2022.
16. Liu J, Dong X, Zhao S, Shen J. Generalized few-shot 3D object detection of LiDAR point cloud for autonomous driving. arXiv:230203914. 2023.
17. Tang W, Yang B, Li X, Liu YH, Heng PA, Fu CW. Prototypical variational autoencoder for 3D few-shot object detection. In: Proceedings of Advances in Neural Information Processing Systems; Vancouver, BC, Canada. 2024.
18. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning; Vienna, Austria. 2020.
19. Yang Z, Wang J, Zhu Y. Few-shot classification with contrastive learning. In: Proceedings of the European Conference on Computer Vision; Tel Aviv, Israel. 2022.
20. Ouali Y, Hudelot C, Tami M. Spatial contrastive learning for few-shot classification. In: Proceedings of the Machine Learning and Knowledge Discovery in Databases; Bilbao, Spain. 2021.

21. Liu C, Fu Y, Xu C, Yang S, Li J, Wang C, et al. Learning a few-shot embedding model with contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence; Vancouver, BC, Canada. 2021.
22. Tang L, Zhan Y, Chen Z, Yu B, Tao D. Contrastive boundary learning for point cloud segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; New Orleans, LA, USA. 2022.
23. Sung C, Kim W, An J, Lee W, Lim H, Myung H. Contextrast: contextual contrastive learning for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Seattle, WA, USA. 2024.
24. Qi CR, Yi L, Su H, Guibas LJ. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of Advances in Neural Information Processing Systems; Los Angeles, CA, USA. 2017.
25. Qi CR, Litany O, He K, Guibas LJ. Deep hough voting for 3D object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; Seoul, Republic of Korea. 2019.
26. Chen D, Chen Y, Li Y, Mao F, He Y, Xue H. Self-supervised learning for few-shot image classification. In: Proceedings of the IEEE International Conference on Acoustics; Toronto, ON, Canada. 2021.
27. Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M. Scannet: richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA. 2017.
28. Song S, Lichtenberg SP, Xiao J. Sun RGB-D: a RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Boston, MA, USA. 2015.