

# From Dataset to Real-world: General 3D Object Detection via Generalized Cross-domain Few-shot Learning

Shuangzhi Li<sup>1</sup>, Junlong Shen<sup>1</sup>, Lei Ma<sup>2,1</sup>, and Xingyu Li<sup>1</sup>

<sup>1</sup>University of Alberta, Canada

<sup>2</sup>The University of Tokyo, Japan

{shuangzh, junlong6, xingyu}@ualberta.ca, ma.lei@acm.org

## Abstract

LiDAR-based 3D object detection models often struggle to generalize to real-world environments due to limited object diversity in existing datasets. To tackle it, we introduce the first generalized cross-domain few-shot (GCFS) task in 3D object detection, aiming to adapt a source-pretrained model to both common and novel classes in a new domain with only few-shot annotations. We propose a unified framework that learns stable target semantics under limited supervision by bridging 2D open-set semantics with 3D spatial reasoning. Specifically, an image-guided multi-modal fusion injects transferable 2D semantic cues into the 3D pipeline via vision-language models, while a physically-aware box search enhances 2D-to-3D alignment via LiDAR priors. To capture class-specific semantics from sparse data, we further introduce contrastive-enhanced prototype learning, which encodes few-shot instances into discriminative semantic anchors and stabilizes representation learning. Extensive experiments on GCFS benchmarks demonstrate the effectiveness and generality of our approach in realistic deployment settings.

**Code** — <https://github.com/Castiel-Lee/GCFS-3Det>

## 1 Introduction

LiDAR-based 3D object detection (Zhang et al. 2025c; Baur, Moosmann, and Geiger 2024; Mao et al. 2023) has significantly advanced autonomous driving by leveraging annotated datasets collected across diverse global locations (Geiger, Lenz, and Urtasun 2012; Caesar et al. 2020; Sun et al. 2020; Geyer et al. 2020). However, as summarized in Table 1, existing datasets primarily focus on a limited set of common object categories (such as cars, pedestrians, and bicycles) within selected urban areas (e.g., USA, Singapore, and German cities). In contrast, real-world deployment introduces new geographic regions and novel object categories, such as electric scooters in Chinese cities or tuk-tuks in Thailand. Collecting and annotating large-scale LiDAR datasets for each new environment is both time-consuming and resource-prohibitive, which makes it unsuitable for rapid adaptation. This practical limitation highlights the need for methods that can generalize beyond the constraints of existing datasets: adapting to new domains and emerging object categories with minimal supervision.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

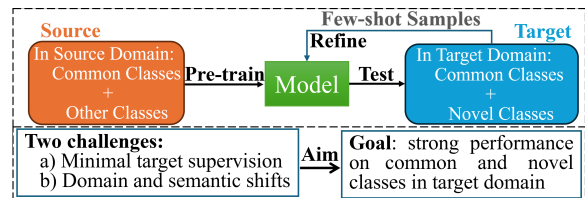


Figure 1: GCFS in 3D object detection aims to adapt source-pretrained models for strong performance on common and novel classes in the target domain via limited target samples.

Despite growing interest in these challenges, existing LiDAR-based 3D detection methods still face key limitations in effectively generalizing to novel categories with limited target-domain data. Among existing approaches, semi-supervised learning (Wang et al. 2023) and 3D open-vocabulary detection (OVD) (Etchegaray et al. 2024; Zhang et al. 2025a; Cao et al. 2024) often assume the availability of large amounts of unlabeled target data, which isn’t always feasible in model deployment. While 3D domain adaptation (DA) (Wang et al. 2020b; Yang et al. 2022; Hegde and Patel 2024) focuses on addressing domain shifts, it does not explicitly account for novel object categories unseen during training. Simply labeling novel objects as “others” is often insufficient in safety-critical scenarios where object-specific recognition is necessary for decision making.

To bridge the gap from dataset-based training to real-world deployment, we tackle a new task, *generalized cross-domain few-shot (GCFS)* learning, for LiDAR-based 3D object detection. As conceptualized in Fig. 1, the GCFS task comprehensively considers efficient adaptation to the target domain and stable semantic learning for novel and common categories via minimal target supervision, offering a cost-effective solution for rapid deployment in diverse environments. Unlike existing 3D few-shot learning (FSL) (Zhao and Qi 2022; Tang et al. 2024; Li, Zhang, and Ma 2024), or its extension, 3D generalized few-shot learning (GFSL) (Liu et al. 2023), which assumes the same distribution between training and deployment environments, GCFS accommodates both domain discrepancies and semantic adaptation target under limited target supervision.

Specifically, in GCFS tasks, a 3D object detection model is initially trained on a source dataset including common ob-

| Datasets           | Locations               | Classes | Categories-of-interest   |
|--------------------|-------------------------|---------|--|
| KITTI (2012)       | Karlsruhe (Germany)     | 7       | <u>Car</u> , <u>Pedestrian</u> , <u>Truck</u> , <u>Van</u> , <u>Person_sitting</u> , <u>Cyclist</u> , <u>Tram</u>      |
| NuScenes (2020)    | Boston (USA), Singapore | 23      | <u>Car</u> , <u>Pedestrian</u> , <u>Truck</u> , <u>Barrier</u> , <u>Construction_vehicle</u> , etc.                    |
| Waymo (2020)       | 3 cities in USA         | 4       | <u>Vehicle(car, truck, and bus)</u> , <u>Pedestrian</u> , <u>Cyclist</u> , <u>Sign</u>                                 |
| Argoverse 2 (2023) | 6 cities in USA         | 30      | <u>Car</u> , <u>Pedestrian</u> , <u>Truck</u> , <u>Bicycle</u> , <u>Motorcycle</u> , <u>Bus</u> , <u>Barrel</u> , etc. |
| A2D2 (2020)        | 50 cities in Germany    | 14      | <u>Car</u> , <u>Pedestrian</u> , <u>Truck</u> , <u>Bicycle</u> , <u>Bus</u> , <u>UtilityVehicle</u> , etc.             |

Table 1: Summary of common 3D Object Detection Datasets, where the most common detection categories are underlined.

ject classes along with other possible source-specific classes. In the target environment, which may have a domain gap from the source data due to environmental factors, sensor configurations, and object appearances (Yang et al. 2022; Hegde and Patel 2024; Li, Ma, and Li 2025), we assume the presence of additional target-specific classes (i.e., novel classes) alongside the common classes. Given the practical feasibility and high cost of LiDAR data collection and annotation, we further assume that access to annotated data in the target environment is restricted to only a minimal amount (e.g., few-shot samples). The GCFS task, therefore, aims to enable the pre-trained model to adapt with minimal supervision in the target environment, ensuring strong performance on both common and target-specific novel categories. Although certain tasks in 2D object detection, such as few-shot domain adaptation (Gao et al. 2023; Nakamura et al. 2022) and generalized few-shot learning (Fan et al. 2021; Zhang et al. 2023b), offer methodological insights into combining limited data adaptation with domain gap bridging, extending these 2D solutions effectively to the 3D domain remains challenging due to the higher-dimensional complexity and unique spatial characteristics of 3D data.

In this work, we introduce the first effective solution to comprehensively address the challenge of stable semantic representation learning under minimal target supervision in GCFS tasks. Our key insight is that generalization across domains and object categories is possible by bridging 2D open-set semantics and 3D spatial reasoning. By aligning sparse 3D observations with rich 2D vision-language priors and refining object understanding through prototype-based semantic anchoring, models can adapt robustly to both domain shifts and novel object classes from a few labeled examples. To realize this, we propose a unified GCFS framework built on two synergistic components: (1) an image-guided multi-modal fusion module that injects transferable 2D semantic cues into the 3D detection pipeline, improving proposal quality even in sparse point clouds; and (2) a contrastive-enhanced prototype learning mechanism that encodes few-shot target samples into discriminative, class-specific semantic anchors. Notably, we introduce a physically-aware box search strategy to improve 2D-to-3D alignment, and use contrastive learning to stabilize semantic prototypes under limited data. Together, these components enable robust adaptation with minimal supervision, offering a practical and generalizable solution for real-world 3D object detection. In evaluation, we design four GCFS benchmark settings and conduct extensive experiments to illustrate the effectiveness of our solution. In sum, our contributions are:

- We formulate the generalized cross-domain few-shot task for 3D object detection and propose the first GCFS solution, holistically addressing domain shifts and novel object categories under limited supervision.
- We propose a unified framework that leverages image-guided semantic grounding and contrastive prototype refinement to learn transferable object-level representations from sparse 3D data. Our framework illustrates that combining 2D vision-language priors with 3D geometry and few-shot semantic anchoring enables robust generalization across diverse environments and categories.
- We establish four GCFS benchmark settings and show that our approach outperforms existing methods, providing a standardized framework for future research on 3D detection under domain and data constraints.

## 2 Related Works

### 2.1 LiDAR-based 3D Object Detection

LiDAR-based 3D object detection (Zhang et al. 2025c; Gambashidze et al. 2024; Mao et al. 2023) aims to locate and classify objects of interest from input point clouds. Its models are primarily categorized into point-based, voxel-based, and point-voxel-based methods. Point-based models (Pan et al. 2021; Shi, Wang, and Li 2019; Shi and Rajkumar 2020) incorporate raw points and the PointNet-based backbones for fine-grained representation at the point level, albeit with high computational demands. Voxel-based methods (Yan, Mao, and Li 2018; Mao et al. 2021; Deng et al. 2021; Zhou and Tuzel 2018) represent the point cloud within a structured voxel grid and utilize sparse convolution for feature extraction, offering a trade-off between computational efficiency and spatial resolution. Point-voxel-based methods (Shi et al. 2023, 2020) combine both, achieving a balance between efficiency and representation resolution, but often coming with increased model complexity and computation.

### 2.2 Few-shot Learning in Object Detection

In object detection, FSL aims to enable models to detect objects with limited labeled samples. In 2D, extensive studies (Zhang et al. 2025b; Xin et al. 2024) tackle data scarcity by exploiting techniques like meta-learning (Yan et al. 2019; Ren et al. 2022), transfer learning (Wang et al. 2020a; Chen et al. 2018), and data augmentation (Wu et al. 2020). In 3D object detection, most works focus on indoor scenarios. Based on VoteNet (Qi et al. 2019), Proto-Vote (Zhao and Qi 2022) introduces a prototypical vote module for local features refinement and a prototypical head module for global

feature enhancement. On top of it, a VAE-based prototype learning (Tang et al. 2024) is designed, and contrastive learning (Li, Zhang, and Ma 2024) is further exploited to learn more refined prototypical representations. However, extending 3D indoor object detection methods to outdoor scenarios is challenging due to sparse point clouds at greater distances, dynamic objects, and varying lighting and weather conditions. A recent work (Liu et al. 2023) proposes the first outdoor generalized FSL solution for novel class learning. Yet, without dealing with domain gaps in cross-domain scenarios, it leads to limited performance on GCFS settings.

### 2.3 Domain Adaptation in 3D Object Detection

The study of domain adaptation in 3D object detection mainly focuses on unsupervised or semi-supervised settings. Works (Yang et al. 2022; Chen et al. 2018) employ a hybrid quality-aware triplet memory to generate pseudo-labels for unlabeled target-domain data. A source-free unsupervised DA approach (Hegde and Patel 2024) utilizes class prototypes to suppress noisy pseudo-labels on target data. Density-resampling-based augmentation and test-time adaptation (Li, Ma, and Li 2025) are proposed to bridge density-related domain gaps. Yet, dependence on large target datasets and the inability to handle novel classes make these methods inapplicable to GCFS tasks

### 2.4 Open-vocabulary 3D Object Detection

Recently, open-vocabulary object detection (Wu et al. 2024; Zareian et al. 2021; Gu et al. 2021; Zhang et al. 2023a; Li\* et al. 2022) has garnered significant attention. In 3D object detection, these methods usually take advantage of 2D VLMs to acquire novel open-set semantics and enable detection on novel objects without annotations. For instance, Lu et al. 2023 proposes to utilize CLIP-based VLMs to connect open-set textual knowledge and point-cloud representations for novel object identification. Auto-label methods (Najibi et al. 2023; Etchegaray et al. 2024) are applied to point cloud sequences via a pretrained 2D VLM and enable novel semantic discovery for self-training. A 2D-3D co-modeling approach (Zhang et al. 2025a) estimates corresponding 3D boxes from 2D insights with temporal and spatial constraints. Since these 3D-OVD methods rely on large volumes of target data (including novel objects), their performance on the GCFS task remains to be validated.

## 3 Methodology

**Problem Statement:** To formulate the GCFS of 3D object detection, we distinguish LiDAR data from the source dataset and target environment (dataset) with superscripts  $s$  and  $t$ , respectively. In the source dataset used to pre-train the model  $M_{\text{pretrained}}$ , we assume access to sufficient annotated data  $\mathbb{D}^s = \{\mathbf{B}_i^s, \mathbf{C}_i^s, \mathbf{P}_i^s\}_{i=1}^{N^s}$ , where  $\mathbf{P}_i^s \in \mathbb{R}^{N_{\text{pts}} \times 3}$  denotes the point cloud,  $\mathbf{B}_i^s = \{\mathbf{b}^s \mid \mathbf{b}^s = [x, y, z, h, w, l, \theta]\}_{l=1}^{N_{\text{obj}}}$  the 3D bounding boxes, and  $\mathbf{C}_i^s$  the corresponding object category belonging to the source category space  $\mathbb{C}^s$ . For the target dataset  $\mathbb{D}^t = (\mathbf{B}_i^t, \mathbf{C}_i^t, \mathbf{P}_i^t)_{i=1}^{N^t}$ , only limited (few-shot) samples are available for each target object category

in the target category set  $\mathbb{C}^t$ . Here, we assume some categories are shared in  $\mathbb{C}^t$  and  $\mathbb{C}^s$ , so certain knowledge in  $M_{\text{pretrained}}$  is valuable to the target task. Formally, these common classes are defined by  $\mathbb{C}_{\text{com}} = \mathbb{C}^t \cap \mathbb{C}^s \neq \emptyset$ . We use  $\mathbb{C}_{\text{nov}}^s = \mathbb{C}^s \setminus \mathbb{C}_{\text{com}}$  and  $\mathbb{C}_{\text{nov}}^t = \mathbb{C}^t \setminus \mathbb{C}_{\text{com}}$  to denote the domain-specific novel classes. That is, objects belonging to  $\mathbb{C}_{\text{nov}}^t$  are unseen in the source dataset. The goal of GCFS tasks is to obtain a strong detection model  $M_{\text{finetuned}}$  through refining  $M_{\text{pretrained}}$  with the  $K$ -shot examples in  $\mathbb{D}^t$ .

Fig. 2 presents an overview of our framework. To learn stable target semantics under limited supervision, we integrate two key components: an *image-guided multi-modal fusion* module and a *class-specific contrastive prototype learning* module. The fusion module exploits vision-language models (VLMs) to extract open-set semantic cues from point-cloud-aligned images, guided by a physically-aware box searching strategy that models LiDAR scanning behavior in the 3D geometric space. Meanwhile, the prototype learning module encodes class-level semantics from few-shot target samples into discriminative prototype anchors, which refine and align object features during inference.

### 3.1 Image-guided Multi-modal Fusion (IMMF)

In GCFS tasks, detectors trained on source data must adapt to new domains and categories via minimal target supervision. Yet, LiDAR data, which is sparse and geometry-focused, offers limited semantic richness, especially for novel objects. In contrast, aligned RGB images offer dense, transferable visual features and access to open-set semantics via pre-trained VLMs. To bridge this semantic gap, we introduce an image-guided multi-modal fusion that enriches 3D point representations with 2D semantic cues extracted from Grounding DINO (GDino) (Liu et al. 2024) and SAM (Kirillov et al. 2023), improving detection robustness under domain and category shifts.

**Image-guided feature fusion.** Given the point cloud  $\mathbf{P}$ , we extract the non-empty voxel feature  $\mathbf{F}^{\text{voxel}} \in \mathbb{R}^{N_{\text{voxel}} \times C}$  via a 3D backbone, where  $C$  denotes the 3D feature dimension. For the paired image  $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ , we use object category names (i.e.,  $\mathbb{C}_{\text{com}}$  and  $\mathbb{C}_{\text{nov}}^t$ ) as text prompts to activate GDino, producing  $N_{\text{obj}}$  2D boxes  $\mathbf{B}^{2D} \in \mathbb{R}^{N_{\text{obj}} \times 4}$  with class labels as potential semantic clues. After non-maximum suppression and confidence filtering, SAM takes  $\mathbf{B}^{2D}$  as box prompts and generates dense object masks  $\mathbf{M}^{2D} \in \mathbb{R}^{H \times W \times |\mathbb{C}^t|}$ . We then project the coordinates  $\mathbf{P}^{\text{voxel}}$  of  $\mathbf{F}^{\text{voxel}}$  onto the image to identify the object masks and obtain the voxel-aligned object mask  $\mathbf{M}^{\text{vxl-obj}} \in \mathbb{R}^{N_{\text{voxel}} \times |\mathbb{C}^t|}$ :

$$\mathbf{M}^{\text{vxl-obj}} = f_{\text{proj}}(\mathbf{M}^{2D}, \mathbf{P}^{\text{voxel}}), \quad (1)$$

where  $f_{\text{proj}}(\cdot)$  denotes 2D-to-3D mapping based on known camera intrinsics and extrinsics. To integrate 2D semantic cues into the 3D representation, we apply an MLP to align the channel dimensions and fuse the features:

$$\mathbf{F}^{\text{fused}} = \mathbf{F}^{\text{voxel}} + \text{MLP}(\mathbf{M}^{\text{vxl-obj}}). \quad (2)$$

This fused feature  $\mathbf{F}^{\text{fused}}$  enhances the downstream region proposal network (RPN), improving object recall for both common and novel categories.

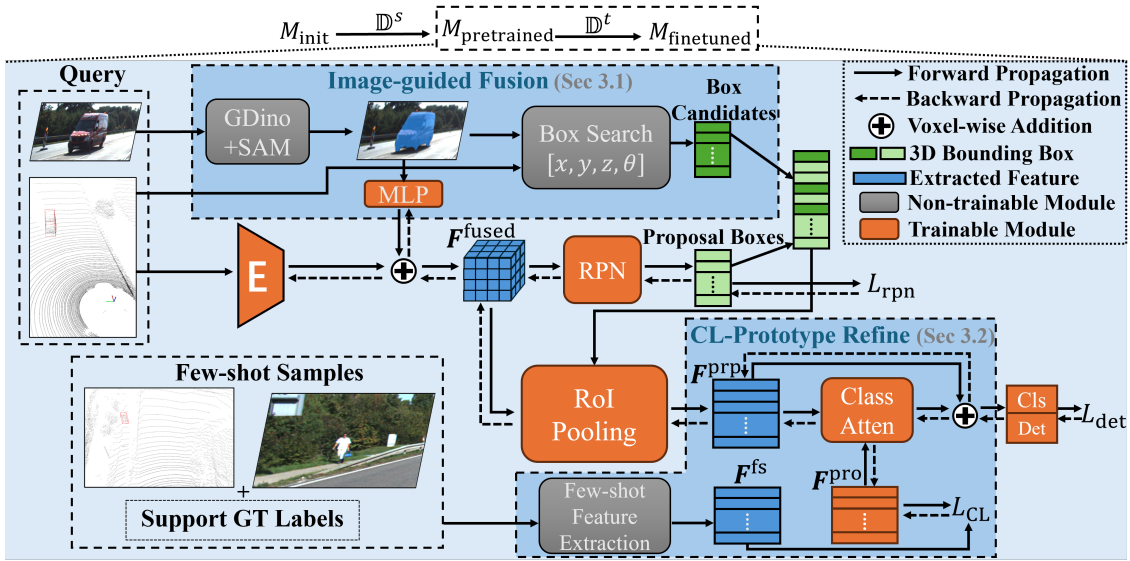


Figure 2: Proposed GCFS Framework. We first pretrain a detection model with source data. During model finetuning using target few-shot samples, each query—the image and point cloud pair—is processed by GDino+SAM and 3D backbone for 2D instance-level masks and 3D features (top block). Insights from 2D context contribute to 1) enriching 3D features  $F^{\text{fused}}$  with 2D semantic clues and 2) proposing high-quality “Box Candidates” via a novel 2D-to-3D box search. Proposal features  $F^{\text{PP}}$  are refined by learnable prototypes  $F^{\text{P}^0}$  with an attention mechanism, and then passed to the final prediction (bottom block).

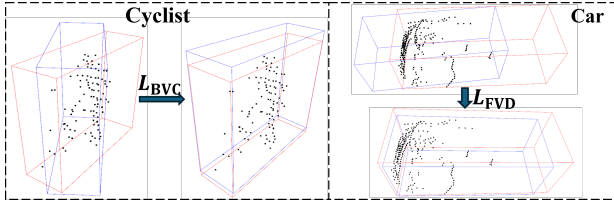


Figure 3: Physical-aware box searching. Red boxes are GT boxes, and blue ones are searched boxes. Regarding “Cyclist” (left) and “Car” (right), angle and center biases on searched boxes are corrected by  $L_{\text{BVC}}$  and  $L_{\text{FVD}}$ .

**Physical-aware 3D box searching from 2D masks.** While VLMs provide rich semantics, transferring these 2D cues into the 3D space is inherently noisy in sparse LiDAR settings. Calibration inaccuracies and vision misalignment can lead to imprecise 2D-to-3D mappings. To ensure 2D semantic cues are projected to geometrically plausible 3D box proposals, we introduce a physically-aware box search strategy that filters and aligns proposals based on spatial consistency.

Specifically, to estimate fine-grained box locations from the 2D object masks  $\mathbf{M}^{2\text{D}}$ , we first project the raw point cloud  $\mathbf{P}$  into the image and identify points within masks by  $\mathbf{P}^{\text{pts}} = f_{\text{proj}}(\mathbf{M}^{2\text{D}}, \mathbf{P})^\top \mathbf{P}$ , where  $\mathbf{P}^{\text{pts}}$  denotes the points of all object masks. For the  $i^{\text{th}}$  object, we extract its points  $\mathbf{P}_i^{\text{pts}} \in \mathbf{P}^{\text{pts}}$  and use the mean and  $2 \times$  standard deviation of point coordinates as the center and boundary of the valid range to eliminate background points. For each class  $c \in \mathcal{C}^t$ , we pre-define an anchor box with the size  $[h^c, w^c, l^c]$  via the mean size of target few-shot objects. The goal of box search-

ing is to find the optimal center  $[x, y, z]$  and heading angle  $\theta$  of the anchor box for each object. Specifically, for  $i$ -th object,  $[x, y, z, \theta]$  defines a rotation transformation  $\mathbf{T}$  (see the supplementary for details), and centered coordinates  $\mathbf{P}_i^{\text{local}}$  are obtained by  $\mathbf{P}_i^{\text{local}} = \mathbf{T}\mathbf{P}_i^{\text{pts}}$ . We first design an outside distance loss  $L_{\text{OD}}$  to constrain  $\mathbf{P}_i^{\text{local}}$  in the box,

$$L_{\text{OD}} = \sum_{\mathbf{p} \in \mathbf{P}_i^{\text{local}}} \min(\text{abs}(\mathbf{p}) - \mathbf{BD}^c, \mathbf{0}). \quad (3)$$

Here,  $\mathbf{BD}^c = [h^c/2, w^c/2, l^c/2]$  denotes the local box boundary for class  $c$ .

Furthermore, we notice that, due to central unidirectional scanning, LiDAR-scanned object points present significant differences in point distribution regarding different structural complexities. For simple structural objects with flat surfaces, like vehicles (e.g., cars and buses), most points are on smooth surfaces and front-viewed by LiDAR. For complex structural objects with irregular surfaces (e.g., pedestrians and bicycles), points are more to shape the whole objects in the bird’s eye view. Motivated by this observation, we categorize general objects into two types: simple structural (SS) objects and complex structural (CS) ones. For SS objects, we design the front-viewed distance (FVD) loss to make points closer to the front-viewed boundaries of the box,

$$L_{\text{FVD}} = \sum_{\mathbf{p} \in \mathbf{P}_i^{\text{local}}} \|\mathbf{p} - \mathbf{FB}^c\| \cdot \mathbf{1}(\mathbf{P}_i^{\text{local}} \in \text{SS}), \quad (4)$$

where the LiDAR front-viewed box boundaries  $\mathbf{FB}^c$  is defined by  $[x, y, z, \theta]$  (see the supplementary for details). For CS objects, we design the bird-viewed center (BVC) loss to

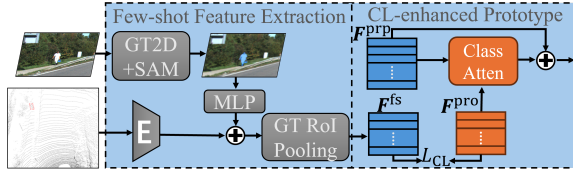


Figure 4: Few-shot feature extraction and CL-enhanced prototype learning. In few-shot feature extraction, 2D and 3D ground-truth labels replace GDino and RPN outputs to extract object features.

align the centers of points and boxes.

$$L_{BVC} = \sum_{\mathbf{p} \in \mathbf{P}_i^{local}} \|f_{2D}(\mathbf{p})\| \cdot \mathbf{1}(\mathbf{P}_i^{local} \in CS), \quad (5)$$

where  $f_{2D}(\cdot)$  simply obtains  $[x, y]$  of points. Applying  $L_{BVC}$  and  $L_{FVD}$  facilitates the discovery of the correct centers and heading angles for boxes, as shown in Fig. 3. In summary, the box-searching loss for optimizing  $[x, y, z, \theta]$  is:

$$L_{box} = L_{OD} + \lambda_1 L_{FVD} + \lambda_2 L_{BVC}. \quad (6)$$

Since the computational load of box searching is low (due to sparse object points), we use the Quasi-Newton BFGS optimization (Head and Zerner 1985) to efficiently optimize  $[x, y, z, \theta]$  for each object. In essence, our physically-aware box search acts as a semantic gatekeeper-ensuring that 2D-to-3D knowledge transfer remains spatially coherent.

### 3.2 Class-specific Contrastive-Enhanced Learnable Prototype and Feature Refinement

While our IMMf module improves proposal accuracy, domain shifts and limited annotations still hinder reliable feature learning via simple fine-tuning. To overcome this, we propose a contrastive prototype learning strategy that builds robust, class-specific semantic anchors from limited examples and enhances them using contrastive learning to increase generalization and inter-class separability. Unlike the work (Li, Zhang, and Ma 2024), which uses contrastive learning to enhance static prototypes, our approach uses few-shot-driven contrastive learning on learnable prototypes, making our prototypes more discriminative.

**Class-specific contrastive prototype learning.** We build a learnable target-specific feature bank  $\mathbf{F}^{pro} \in \mathbb{R}^{|\mathcal{C}^t| \times d}$  for all object classes, where  $d$  is the dimension of features. These prototypes are optimized together with the model fine-tuning update. To accelerate convergence under limited data, we introduce a contrastive loss for the learnable prototypes. As shown in Fig. 4, we group the features of the few shots  $\mathbf{F}^{fs}$  according to their box annotation as contrastive anchors. Then for each class  $c \in \mathcal{C}^t$ , we construct positive pairs with the corresponding prototype  $\mathbf{F}_c^{pro}$  and its anchor  $\mathbf{F}_c^{fs}$ . The remaining prototypes in the feature bank, denoted by  $\mathbf{F}_s^{pro}$ , are negative samples of the anchor.

$$L_{CL} = - \sum_{c \in \mathcal{C}^t} \log \frac{\exp(\text{Sim}(\mathbf{F}_c^{fs}, \mathbf{F}_c^{pro})/\tau)}{\sum_{s \in \mathcal{C}^t} \exp(\text{Sim}(\mathbf{F}_c^{fs}, \mathbf{F}_s^{pro})\tau)}, \quad (7)$$

where  $\text{Sim}(\cdot, \cdot)$  calculates the cosine similarity between two features in the InfoNCE loss (Oord, Li, and Vinyals 2018) with a temperature  $\tau$ . Since the anchors are directly obtained from target-domain examples, our contrastive-enhanced features help bridge the domain gap between source and target environments and speed up  $\mathbf{F}^{pro}$  acquiring semantic essences of various classes under limited training data.

**Feature refinement by prototypes.** After obtaining the  $\mathbf{F}^{pro}$  along with the model finetuning process, we use them to refine the proposal features  $\mathbf{F}^{prp}$  of the query input. In the multi-head cross-attention, we take  $\mathbf{F}^{pro}$  to form the key and value, and  $\mathbf{F}^{prp}$  as the query.

$$\hat{\mathbf{F}}^{prp} = \text{Softmax}\left(\frac{\mathbf{F}^{prp} \mathbf{W}_Q (\mathbf{F}^{pro} \mathbf{W}_K)^T}{\sqrt{d}}\right) \mathbf{F}^{pro} \mathbf{W}_V, \quad (8)$$

where  $[\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V]$  is the trainable transformation of the query, key, and value. Finally,

$$\tilde{\mathbf{F}}^{prp} = \hat{\mathbf{F}}^{prp} + \mathbf{F}^{prp}, \quad (9)$$

is passed to the object detection head for object detection.

### 3.3 Model Optimization and Inference

The model parameter update and our prototype learning are conducted together. The Overall loss to optimize them is:

$$L = L_{rpn} + L_{det} + \lambda L_{CL}, \quad (10)$$

where  $L_{rpn}$  and  $L_{det}$  are the standard losses of RPN and detection head, and  $\lambda$  is a weight hyper-parameter. To further enable the model’s adaptability to a new domain under limited data, we adopt an MAML-based (Finn, Abbeel, and Levine 2017) training scheme. Briefly, during meta-training, we leverage the source data to set up the  $K$ -shot meta-task. This meta-training facilitates finding a set of model parameters and  $\mathbf{F}^{pro}$  for the quick model adaptation in the unseen domain (see the supplementary for more details). During deployment, aligned point clouds and images undergo the proposed image-guided fusion to enhance semantic discovery in proposals. After ROI pooling, object features are further refined with class prototypes to improve discrimination.

## 4 Experimentation

### 4.1 Experimental Settings<sup>1</sup>

**Benchmarks.** Since no prior study on GCFS tasks in 3D object detection, we leverage Nuscenes (2020), Waymo (2020), KITTI (2012), A2D2 (2020), and Argoverse 2 (2023) to construct 4 GCFS benchmarks: NuScenes→FS-KITTI, Waymo→FS-KITTI, KITTI→FS-A2D2, and KITTI→FS-Argo2. Specifically, we construct few-shot datasets by sampling  $K$ -shot objects per class from the *train* set of KITTI, A2D2, and Argoverse 2, forming FS-KITTI, FS-A2D2, FS-Argo2. We set  $K = 5$  for main experiments, while our ablation study explores  $K \in \{1, 3, 5, 10, 20, 40\}$  for a comprehensive evaluation. The *val* sets of KITTI and Argoverse 2 and the *test* set of A2D2

<sup>1</sup>Details on the benchmark setup and implementation are provided in the **supplementary** linked in the **GitHub repository**.

| Methods     | NuScenes→FS-KITTI             |                               |                               | Waymo→FS-KITTI         |                               |                               | KITTI→FS-A2D2                |                              |                              | KITTI→FS-Argo2               |                              |                              |
|-------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|-------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
|             | common                        | novel                         | overall                       | common                 | novel                         | overall                       | common                       | novel                        | overall                      | common                       | novel                        | overall                      |
| Source-only | 14.24                         | -                             | -                             | <b>26.85</b>           | -                             | -                             | 3.81                         | -                            | -                            | 6.65                         | -                            | -                            |
| Target-FT   | 12.77 <sub>(1.9)</sub>        | 5.48 <sub>(1.2)</sub>         | 8.61 <sub>(1.5)</sub>         | 23.06 <sub>(1.6)</sub> | 12.47 <sub>(1.9)</sub>        | 17.01 <sub>(1.8)</sub>        | 5.09 <sub>(1.1)</sub>        | 0.70 <sub>(0.2)</sub>        | 2.90 <sub>(0.6)</sub>        | 3.18 <sub>(0.2)</sub>        | 0.62 <sub>(0.1)</sub>        | 1.39 <sub>(0.1)</sub>        |
| Proto-Vote  | 7.56 <sub>(2.4)</sub>         | 5.74 <sub>(1.5)</sub>         | 6.52 <sub>(1.9)</sub>         | 17.36 <sub>(2.7)</sub> | 12.08 <sub>(1.8)</sub>        | 14.34 <sub>(2.2)</sub>        | 3.61 <sub>(0.8)</sub>        | 1.86 <sub>(0.5)</sub>        | 2.74 <sub>(0.7)</sub>        | 3.33 <sub>(0.9)</sub>        | 0.90 <sub>(0.4)</sub>        | 1.63 <sub>(0.5)</sub>        |
| PVAE-Vote   | 8.01 <sub>(2.8)</sub>         | 6.38 <sub>(2.2)</sub>         | 7.08 <sub>(2.5)</sub>         | 18.19 <sub>(2.9)</sub> | 12.79 <sub>(2.2)</sub>        | 15.10 <sub>(2.5)</sub>        | 3.43 <sub>(0.9)</sub>        | 1.97 <sub>(0.5)</sub>        | 2.70 <sub>(0.7)</sub>        | 3.10 <sub>(1.0)</sub>        | 0.92 <sub>(0.3)</sub>        | 1.58 <sub>(0.5)</sub>        |
| CP-Vote     | 10.69 <sub>(2.3)</sub>        | 7.84 <sub>(1.9)</sub>         | 9.06 <sub>(2.0)</sub>         | 17.66 <sub>(2.4)</sub> | 12.17 <sub>(1.9)</sub>        | 14.52 <sub>(2.1)</sub>        | 4.28 <sub>(0.9)</sub>        | 2.72 <sub>(0.9)</sub>        | 3.50 <sub>(0.9)</sub>        | 2.72 <sub>(0.9)</sub>        | 0.93 <sub>(0.4)</sub>        | 1.47 <sub>(0.5)</sub>        |
| GFS-Det     | 12.83 <sub>(2.4)</sub>        | 1.18 <sub>(0.4)</sub>         | 6.17 <sub>(1.2)</sub>         | 22.74 <sub>(2.8)</sub> | 1.26 <sub>(0.4)</sub>         | 10.47 <sub>(1.4)</sub>        | 4.39 <sub>(0.6)</sub>        | 0.22 <sub>(0.1)</sub>        | 2.30 <sub>(0.3)</sub>        | 6.11 <sub>(0.1)</sub>        | 0.03 <sub>(0.0)</sub>        | 1.86 <sub>(0.0)</sub>        |
| <b>Ours</b> | <b>15.99</b> <sub>(1.6)</sub> | <b>11.72</b> <sub>(1.4)</sub> | <b>13.55</b> <sub>(1.5)</sub> | 25.40 <sub>(2.0)</sub> | <b>17.75</b> <sub>(1.7)</sub> | <b>21.03</b> <sub>(1.8)</sub> | <b>7.78</b> <sub>(0.7)</sub> | <b>5.22</b> <sub>(0.6)</sub> | <b>6.50</b> <sub>(0.6)</sub> | <b>6.71</b> <sub>(0.2)</sub> | <b>2.07</b> <sub>(0.2)</sub> | <b>3.46</b> <sub>(0.2)</sub> |
| Full-Target | 41.34                         | 18.35                         | 28.21                         | 41.34                  | 18.35                         | 28.21                         | 36.61                        | 5.99                         | 21.30                        | 31.75                        | 18.48                        | 22.46                        |

Table 2: Performance in mAP(%) of VoxelRCNN for NuScenes → 5shot-KITTI, Waymo → 5shot-KITTI, KITTI → 5shot-A2D2, and KITTI → 5shot-Argo2. The bold values represent the best performance except Full-Target. Subscript values in parentheses are standard deviations. Please refer to the supplementary for specifics across various categories.

are used for model evaluation. According to Table 1, we select [Car, Pedestrian, Truck] as common classes for all datasets. For sufficient samples for model evaluation and avoiding class ambiguity, we target novel classes: [Van, Person\_sitting, Cyclist, Tram] in FS-KITTI, [Bicycle, Utility\_vehicle, Bus] in FS-A2D2, and [Construction\_barrel, Traffic\_cone, Large\_vehicle, Bicycle, Bus, Motorcycle, Sign] in FS-Argo2. We use Average Precision (AP) to measure precision-recall trade-offs for each class (Geiger, Lenz, and Urtasun 2012) and mean Average Precision (mAP) across multi-classes to assess overall performance. We conduct experiments 5 times and report the average mAP across trials, along with the standard deviation for stability evaluation.

**Implementation Details.** We use VoxelRCNN (Deng et al. 2021) (voxel-based) and PV-RCNN++ (Shi et al. 2023) (point-voxel-based) as base detectors. Pre-training applies standard augmentations: random world flipping, scaling, and rotation. In fine-tuning, we additionally use ground-truth object sampling to ensure all target classes are present in each iteration. For box searching, we define SS classes [Car, Truck, Van, Tram, Bus, Construction\_barrel, Large\_vehicle, Sign] and CS classes [Pedestrian, Person\_sitting, Cyclist, Bicycle, Utility\_vehicle, Traffic\_cone, Motorcycle]. The Adam-OneCycle optimizer (Team 2020; Song et al. 2024) is used with a 0.01 learning rate. All models are pre-trained for 30 epochs on NuScenes and Waymo, 80 epochs on KITTI, and fine-tuned for 100 epochs in FS-datasets. Batch sizes are 2 in pre-training and 1 in fine-tuning and testing.

**Compared Methods.** As no prior work has specifically tackled GCFS tasks for outdoor 3D object detection, we use a simple fine-tuning on few-shot target data (Target-FT) as the baseline. Source-only training and full target supervision (Source-only and Full-Target) serve as the performance with no and full adaptation. To benchmark our method, we compare against SOTA 3D-FSL methods, Proto-Vote (Zhao and Qi 2022), PVAE-Vote (Tang et al. 2024), and CP-Vote (Li, Zhang, and Ma 2024), as well as the 3D-GFSL method GFS-Det (Liu et al. 2023). Note that current outdoor OVD methods (i.e., Unsup3D (Najibi et al. 2023), FnP (Etchegaray et al. 2024), and OpenSight (Zhang et al. 2025a)) and 3D-DA methods (i.e., SN (Wang et al. 2020b), ST3D++ (Yang

|     | Target-FT | Image-Fusion | CL-Proto | Common       | Novel        | Overall      |
|-----|-----------|--------------|----------|--------------|--------------|--------------|
| (a) | ✓         |              |          | 12.77        | 5.48         | 8.61         |
| (b) | ✓         |              | ✓        | 14.80        | 8.10         | 10.97        |
| (c) | ✓         | ✓            |          | 14.69        | 11.17        | 12.68        |
| (d) | ✓         | ✓            | ✓        | <b>15.99</b> | <b>11.72</b> | <b>13.55</b> |

Table 3: Component ablations in mAP(%). Image-Fusion is our proposed IMM module and CL-Proto is our proposed contrastive-learning-enhanced prototype learning.

et al. 2022, 2021), and DenResamp (Li, Ma, and Li 2025)) are not directly applicable to our GCFS benchmark, as they rely on extensive unannotated data for unsupervised learning. To further assess the generalizability and potential of our approach, we extend our ablation study to a more complex unsupervised few-shot learning setting, where these 3D-OVD and 3D-DA methods can be evaluated under conditions more aligned with their original assumptions.

## 4.2 Experimental Results on GCFS Benchmark

As shown in Table 2, our method consistently achieves superior performance in all GCFS benchmarks, demonstrating strong generalization to both common and novel categories under limited supervision. It arises from two key strengths. First, our method exhibits robust cross-domain transferability under diverse density-domain shifts, including varying LiDAR configurations across NuScenes (32-beam), Waymo (64-beam), KITTI (64-beam), A2D2 (16-beam), and Argoverse 2 (32-beam). It effectively maintains detection quality despite drastic variations in point density and sensor characteristics. Second, our approach enables efficient few-shot adaptation to target semantic concepts, as evidenced by its performance in semantically challenging settings like KITTI → 5shot-Argo2, involving seven diverse novel classes. In contrast, 3D-FSL methods show limited robustness on common classes due to their reliance on dense, close-range point clouds. Meanwhile, GFSL-Det struggles

| Prototype | Common       | Novel        |
|-----------|--------------|--------------|
| w/o CL    | 15.23        | 10.33        |
| w/ CL     | <b>15.99</b> | <b>11.72</b> |

Table 4: Performance in mAP(%) of prototype learning with or without contrastive learning (CL).

| Box Search | CS          | SS           |
|------------|-------------|--------------|
| $L_{OD}$   | 4.65        | 12.56        |
| $L_{box}$  | <b>6.74</b> | <b>13.58</b> |

Table 5: Performance in mAP(%) with box searching by  $L_{OD}$  only or  $L_{box}$  (w/  $L_{OD}$ ,  $L_{FVD}$ , and  $L_{BVC}$ ).

| Methods | Target-FT | Proto-Vote | PVAE-Vote | CP-Vote | GFS   | Ours         |
|---------|-----------|------------|-----------|---------|-------|--------------|
| Common  | 15.28     | 6.97       | 7.43      | 8.73    | 17.37 | <b>18.06</b> |
| Novel   | 6.39      | 7.05       | 7.53      | 7.17    | 1.16  | <b>11.11</b> |
| Overall | 10.20     | 7.02       | 7.49      | 7.84    | 8.10  | <b>14.09</b> |

Table 6: Performance in mAP(%) of PV-RCNN for NuScenes  $\rightarrow$  5shot-KITTI. Please refer to the supplementary for specifics across other GCFS tasks.

to generalize to novel classes, as its simplistic incremental learning strategy lacks mechanisms for semantic transfer from common classes to novel ones.

**Limitations and Future Work.** Our method shows limited gains on certain hard classes (e.g. ‘‘Person\_sitting’’) due to ambiguous and diverse structures. In low-shift scenarios without semantic changes (Waymo  $\rightarrow$  FS-KITTI), improvements on common classes are marginal, due to the interruption of novel classes. Future work will focus on hard class learning, adaptability in shiftless settings, and code optimization for computation speed-up.

### 4.3 Ablation Studies

We conduct ablation experiments mainly on NuScenes  $\rightarrow$  5shot-KITTI with VoxelRCNN as detection model, to further analyze our method (see the supplementary for details). **Component Ablation.** Table 3 (a) $\rightarrow$ (b) indicates that our adaptive prototype learning enhances performance in common and novel classes, illustrating its swift adaptation to limited samples in the target domain. Applying our image-guided multi-modal fusion (a) $\rightarrow$ (c) yields marked improvement, especially on novel classes, showing its boost on object recall. By combining both, our GCFS method achieves the highest performance, demonstrating the complementarity of the two approaches. Notably, removing MAML lowers AP to 12.35, and replacing our box search with FnP gives AP of 12.58, showing our method’s effectiveness.

We also conduct ablations on our proposed prototype learning and box searching components. Table 4 shows that the contrastive loss boosts model performance, indicating its ability to swiftly adapt prototypes to few-shot data in the target domain. In Table 5, integrating  $L_{OD}$  with  $L_{FVD}$  and  $L_{BVC}$  yields improvements on both CS and SS objects, showing  $L_{FVD}$  and  $L_{BVC}$  enhancing recall rates for objects with diverse structural complexities, thereby further optimizing model effectiveness.

**Ablation on Detection Backbone.** We further evaluate our

| Shots   | K=1  | K=3   | K=5   | K=10  | K=20  | K=40  | Full-shot |
|---------|------|-------|-------|-------|-------|-------|-----------|
| Common  | 7.27 | 12.27 | 15.99 | 23.56 | 27.59 | 32.05 | 41.34     |
| Novel   | 0.57 | 7.76  | 11.72 | 12.21 | 17.49 | 21.55 | 18.35     |
| Overall | 3.44 | 9.70  | 13.55 | 17.08 | 21.82 | 26.05 | 28.21     |

Table 7: Performances in mAP(%) with different  $K$ . Full-shot denotes the training on the complete KITTI train set.

| Method         | DA    |        |           | OVD   |              |
|----------------|-------|--------|-----------|-------|--------------|
|                | SN    | ST3D++ | DenResamp | FnP   | Ours-OVD     |
| <b>Common</b>  | 12.09 | 21.00  | 14.89     | 10.59 | <b>22.25</b> |
| <b>Novel</b>   | -     | -      | -         | 2.66  | <b>8.26</b>  |
| <b>Overall</b> | -     | -      | -         | 6.06  | <b>14.26</b> |

Table 8: Comparison in mAP(%) for OVD and DA methods in the unsupervised few-shot setting.

GCFS framework using the point-voxel-hybrid detector PV-RCNN. As shown in Table 6, our approach consistently outperforms others across all three metrics, demonstrating the generalizability of our solution.

**Ablation on Numbers of Shots.** Table 7 shows that our method scales well with increasing  $K$ . At  $K = 40$ , the overall performance approaches the full-shot, narrowing the supervision gap. Despite a reasonable gap in common-class performance due to limited data, the novel-class performance surpasses the full-shot result, due to class imbalance in full-shot training and our image-guided design enhancing novel object discovery. These results confirm the scalability and generalization of our method under limited supervision. **Unsupervised few-shot ablation with OVD and DA methods.** We establish an unsupervised few-shot setting with no annotations for all classes. Our approach is benchmarked against the SOTA 3D-OVD solution and well-established 3D-DA methods in Table 8. To create an OVD version of our method, we incorporate a physical-aware box searcher to generate high-quality pseudo-labels for target-specific training. Compared to OVD and DA methods, our OVD method achieves the highest mAPs, showing strong domain gap bridging capability and high learning efficiency from unlabeled samples. Please refer to the supplementary for implementation and result details.

## 5 Conclusion

This paper tackled the generalized cross-domain few-shot task in 3D object detection and introduced the first GCFS solution. Beyond achieving state-of-the-art performance on four GCFS benchmarks, our work demonstrated a generalizable approach to few-shot 3D adaptation, grounded in the idea that semantic alignment across modalities and domains could be achieved by combining 2D open-set priors with 3D structural cues and few-shot supervision. We believed this framework opens new possibilities for 3D perception systems that must continually adapt to new environments and emerging object types, without relying on exhaustive data collection or domain-specific engineering.

## Acknowledgements

This work was supported in part by JST CRONOS Grant (No. JPMJCS24K8), JSPS KAKENHI Grant (No.JP21H04877, No.JP23H03372, and No.JP24K02920), Canada CIFAR AI Chairs Program, the Natural Sciences and Engineering Research Council of Canada, and the Autoware Foundation.

## References

- Baur, S. A.; Moosmann, F.; and Geiger, A. 2024. LISO: Lidar-only self-supervised 3d object detection. In *European Conference on Computer Vision*, 253–270.
- Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.
- Cao, Y.; Yihan, Z.; Xu, H.; and Xu, D. 2024. CoDA: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *Advances in Neural Information Processing Systems*, 36.
- Chen, H.; Wang, Y.; Wang, G.; and Qiao, Y. 2018. LSTD: A low-shot transfer detector for object detection. In *AAAI conference on artificial intelligence*, volume 32.
- Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; and Li, H. 2021. Voxel R-CNN: Towards high performance voxel-based 3d object detection. In *AAAI conference on artificial intelligence*, volume 35, 1201–1209.
- Etcheberry, D.; Huang, Z.; Harada, T.; and Luo, Y. 2024. Find n<sup>2</sup>Propagate: Open-Vocabulary 3D Object Detection in Urban Environments. In *European Conference on Computer Vision*, 133–151.
- Fan, Z.; Ma, Y.; Li, Z.; and Sun, J. 2021. Generalized few-shot object detection without forgetting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4527–4536.
- Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135.
- Gambashidze, A.; Dadukin, A.; Golyadkin, M.; Razzhivina, M.; and Makarov, I. 2024. Weak-to-strong 3d object detection with x-ray distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15055–15064.
- Gao, Y.; Lin, K.-Y.; Yan, J.; Wang, Y.; and Zheng, W.-S. 2023. AsyFOD: An asymmetric adaptation paradigm for few-shot domain adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3261–3271.
- Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361.
- Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A. S.; Hauswald, L.; Pham, V. H.; Mühlegg, M.; Dorn, S.; et al. 2020. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*.
- Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *International Conference on Learning Representations*.
- Head, J. D.; and Zerner, M. C. 1985. A Broyden—Fletcher—Goldfarb—Shanno optimization procedure for molecular geometries. *Chemical physics letters*, 122(3): 264–270.
- Hegde, D.; and Patel, V. M. 2024. Attentive prototypes for source-free unsupervised domain adaptive 3d object detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 3066–3076.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, 4015–4026.
- Li\*, L. H.; Zhang\*, P.; Zhang\*, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022. Grounded Language-Image Pre-training. In *CVPR*.
- Li, S.; Ma, L.; and Li, X. 2025. Domain Generalization of 3D Object Detection by Density-Resampling. In *European Conference on Computer Vision*, 456–473.
- Li, X.; Zhang, W.; and Ma, C. 2024. CP-VoteNet: Contrastive Prototypical VoteNet for Few-Shot Point Cloud Object Detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 461–475.
- Liu, J.; Dong, X.; Zhao, S.; and Shen, J. 2023. Generalized few-shot 3d object detection of lidar point cloud for autonomous driving. *arXiv preprint arXiv:2302.03914*.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *European Conference on Computer Vision*, 38–55. Springer.
- Lu, Y.; Xu, C.; Wei, X.; Xie, X.; Tomizuka, M.; Keutzer, K.; and Zhang, S. 2023. Open-vocabulary point-cloud object detection without 3d annotation. In *IEEE/CVF conference on computer vision and pattern recognition*, 1190–1199.
- Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8): 1909–1963.
- Mao, J.; Xue, Y.; Niu, M.; Bai, H.; Feng, J.; Liang, X.; Xu, H.; and Xu, C. 2021. Voxel transformer for 3d object detection. In *IEEE/CVF international conference on computer vision*, 3164–3173.
- Najibi, M.; Ji, J.; Zhou, Y.; Qi, C. R.; Yan, X.; Ettinger, S.; and Angelov, D. 2023. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *IEEE/CVF International Conference on Computer Vision*, 8602–8612.
- Nakamura, Y.; Ishii, Y.; Maruyama, Y.; and Yamashita, T. 2022. Few-shot adaptive object detection with cross-domain cutmix. In *Asian Conference on Computer Vision*, 1350–1367.
- Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.
- Pan, X.; Xia, Z.; Song, S.; Li, L. E.; and Huang, G. 2021. 3d object detection with pointformer. In *IEEE/CVF conference on computer vision and pattern recognition*, 7463–7472.
- Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *IEEE/CVF International Conference on Computer Vision*, 9277–9286.
- Ren, X.; Zhang, W.; Wu, M.; Li, C.; and Wang, X. 2022. Meta-YOLO: Meta-Learning for Few-Shot Traffic Sign Detection via Decoupling Dependencies. *Applied Sciences*, 12(11).
- 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 *IEEE/CVF conference on computer vision and pattern recognition*, 10529–10538.

- Shi, S.; Jiang, L.; Deng, J.; Wang, Z.; Guo, C.; Shi, J.; Wang, X.; and Li, H. 2023. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection. *International Journal of Computer Vision*, 131(2): 531–551.
- Shi, S.; Wang, X.; and Li, H. 2019. PointRCNN: 3d object proposal generation and detection from point cloud. In *IEEE/CVF conference on computer vision and pattern recognition*, 770–779.
- Shi, W.; and Rajkumar, R. 2020. Point-GNN: Graph neural network for 3d object detection in a point cloud. In *IEEE/CVF conference on computer vision and pattern recognition*, 1711–1719.
- Song, Z.; Zhang, G.; Liu, L.; Yang, L.; Xu, S.; Jia, C.; Jia, F.; and Wang, L. 2024. RoboFusion: Towards robust multi-modal 3d object detection via SAM. In *33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 141.
- Sun, P.; Kretschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.
- Tang, W.; Yang, B.; Li, X.; Liu, Y.-H.; Heng, P.-A.; and Fu, C.-W. 2024. Prototypical variational autoencoder for 3d few-shot object detection. *Advances in Neural Information Processing Systems*, 36.
- Team, O. D. 2020. OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds. <https://github.com/open-mmlab/OpenPCDet>. Accessed: 2025-12-04.
- Wang, X.; Huang, T. E.; Darrell, T.; Gonzalez, J. E.; and Yu, F. 2020a. Frustratingly simple few-shot object detection. In *37th International Conference on Machine Learning (ICML)*, 9919–9928.
- Wang, Y.; Chen, X.; You, Y.; Li, L. E.; Hariharan, B.; Campbell, M.; Weinberger, K. Q.; and Chao, W.-L. 2020b. Train in Germany, test in the USA: Making 3d object detectors generalize. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11713–11723.
- Wang, Y.; Yin, J.; Li, W.; Frossard, P.; Yang, R.; and Shen, J. 2023. SSDA3D: Semi-supervised domain adaptation for 3d object detection from point cloud. In *AAAI Conference on Artificial Intelligence*, volume 37, 2707–2715.
- Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J. K.; et al. 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *CoRR*, abs/2301.00493.
- Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; et al. 2024. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Wu, J.; Liu, S.; Huang, D.; and Wang, Y. 2020. Multi-scale positive sample refinement for few-shot object detection. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 456–472.
- Xin, Z.; Chen, S.; Wu, T.; Shao, Y.; Ding, W.; and You, X. 2024. Few-shot object detection: Research advances and challenges. *Information Fusion*, 102307.
- Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; and Lin, L. 2019. Meta R-CNN: Towards general solver for instance-level low-shot learning. In *IEEE/CVF International Conference on Computer Vision*, 9577–9586.
- Yan, Y.; Mao, Y.; and Li, B. 2018. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.
- Yang, J.; Shi, S.; Wang, Z.; Li, H.; and Qi, X. 2021. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *IEEE/CVF conference on computer vision and pattern recognition*, 10368–10378.
- Yang, J.; Shi, S.; Wang, Z.; Li, H.; and Qi, X. 2022. St3d++: De-noised self-training for unsupervised domain adaptation on 3d object detection. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 6354–6371.
- Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14393–14402.
- Zhang, H.; Li, F.; Zou, X.; Liu, S.; Li, C.; Yang, J.; and Zhang, L. 2023a. A simple framework for open-vocabulary segmentation and detection. In *IEEE/CVF International Conference on Computer Vision*, 1020–1031.
- Zhang, H.; Xu, J.; Tang, T.; Sun, H.; Yu, X.; Huang, Z.; and Yu, K. 2025a. Opensight: A simple open-vocabulary framework for lidar-based object detection. In *European Conference on Computer Vision*, 1–19.
- Zhang, J.; Liu, L.; Silven, O.; Pietikäinen, M.; and Hu, D. 2025b. Few-shot class-incremental learning for classification and object detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhang, P.; Li, X.; Lin, X.; and He, L. 2025c. A new literature review of 3D object detection on autonomous driving. *Journal of Artificial Intelligence Research*, 82: 973–1015.
- Zhang, T.; Zhang, X.; Zhu, P.; Jia, X.; Tang, X.; and Jiao, L. 2023b. Generalized few-shot object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195: 353–364.
- Zhao, S.; and Qi, X. 2022. Prototypical votenet for few-shot 3d point cloud object detection. *Advances in neural information processing systems*, 35: 13838–13851.
- Zhou, Y.; and Tuzel, O. 2018. VoxelNet: End-to-end learning for point cloud based 3d object detection. In *IEEE conference on computer vision and pattern recognition*, 4490–4499.