BEVOpen3D: Towards Open-World 3D Object Detection in Bird's-Eye-View

Huyue Zeng[1], Jiaqi Yang[2], and Xian-Feng Han[1],✉

[1] College of Computer and Information Science, Southwest University, Chongqing 400715, China xianfenghan@swu.edu.cn
[2] National Engineering Laboratory for Integrated Aero-SpaceGround-Ocean Big Data Application Technology, Northwestern Polytechnical University, Xi'an 710072, China jqyang@nwpu.edu.cn

Abstract. The reliance on extensively annotated 3D data confines current autonomous driving systems to a closed-set detection paradigm, limiting their robustness in real-world open-world scenarios where novel objects frequently emerge. To bridge this gap, we propose **BEVOpen3D**, a novel distillation framework that facilitates open-world 3D object detection by effectively transferring knowledge from 2D vision-language models into the 3D domain via the Bird's-Eye-View (BEV) space under a partial-label setting. Our approach first generates initial 3D proposals for both seen and unseen categories using an open-vocabulary 2D detector combined with a greedy spatial search strategy. We then introduce a *Triple-Source Label Refinement* mechanism, which fuses original, locally rectified (via a *Seen-Guided Local Query*), and globally re-examined proposals to produce high-quality pseudo labels for unseen-category. Finally, we propose a *Heatmap Proposal Distillation* strategy, where an image-based BEV teacher transfers its vision-grounded semantic priors to a LiDAR-based student through aligned BEV heatmaps, enabling the student to detect both seen and unseen objects without requiring additional 3D annotations. Extensive experiments validate the effectiveness of BEVOpen3D in open-world 3D detection.

Keywords: 3D Detection · Vision-Language Model · Open-world

1 Introduction

Recent advances in autonomous driving have necessitated perception systems capable of operating in open-world environments. However, traditional *closed-set* training paradigms rely heavily on pre-defined category lists, rendering models ineffective when encountering object classes absent from the training set. This limitation severely constrains generalization and safety in complex real-world scenarios.

Concurrently, *open-vocabulary* recognition has achieved remarkable progress in the 2D image domain. Leveraging large-scale image-text pairs, vision-language pre-trained models such as CLIP [19] have demonstrated powerful zero-shot classification abilities, enabling the identification of objects never explicitly seen during training. Nevertheless, transferring this successful framework to 3D point

cloud detection faces substantial obstacles. The high cost of annotating 3D bounding boxes results in limited scale and sparse category diversity in existing open-source 3D datasets. Consequently, training a 3D-CLIP model remains a challenging issue.

To address this problem, prior works have explored indoor scenes. These works [16,22] typically adopt a two-stage paradigm: first training a class-agnostic generic object detector to capture foreground regions, and subsequently employing contrastive learning to align point cloud region features, back-projected image patch features, and text prompt features. This tri-modal alignment endows the detector with open-vocabulary classification capabilities.

However, directly transferring this indoor paradigm to outdoor autonomous driving scenarios presents serious challenges. First, CLIP is pre-trained on dense image-text pairs, making its feature space highly reliant on rich texture information. In contrast, outdoor point clouds are characterized by substantial geometric and textural discrepancies compared to 2D images. Enforcing end-to-end contrastive alignment across such a vast modality gap not only complicates optimization but also forces the 3D detector to sacrifice geometric localization accuracy to accommodate text semantics. Second, the paradigm of first training a class-agnostic detector and then injecting classification capabilities via contrastive learning appears inefficient for outdoor settings. Most outdoor 3D detectors have already acquired robust foreground capture and box regression capabilities through large-scale supervised learning, with decoding processes heavily reliant on category priors (e.g., heatmaps). Training a cross-modal projection head to re-learn classification capabilities introduces additional computational overhead and instability. Rather than struggling to align text semantics in noisy 3D feature spaces, a more direct approach is needed to transform the open-vocabulary capabilities of the 2D domain into effective supervision signals for 3D detection.

Addressing these challenges, the baseline method [5] proposes an offline strategy. A 2D open-vocabulary detector [12] identifies unknowns and generates 3D pseudo labels via frustum search. These labels are merged with ground-truth annotations for seen categories to supervise the 3D detector, employing multi-round self-training where inference results serve as updated pseudo labels for iterative refinement. This design indirectly transfers CLIP's classification ability to 3D detection. Despite demonstrating the viability of this decoupled strategy, the baseline scheme suffers from two critical limitations:

1. **Error Accumulation:** During self-training, the model relies on pseudo-labels generated by its own previous inference rounds. Although data augmentation is employed, erroneous annotations from initial stages are amplified during iteration, leading to performance bottlenecks or even degradation.
2. **Waste of Modal Information:** To avoid the domain gap, the baseline method completely discards the image modality. However, rich texture details in images are crucial for distinguishing objects with similar appearances but different categories (*e.g.*, cars vs. pickups). Existing research indicates

that cross-modal fusion often yields superior detection performance; thus, completely discarding this information represents a significant loss.

To address these problems, this paper proposes a novel framework for open-world 3D object detection. The framework adopts a teacher-student co-evolution architecture, where the teacher model (an image-based teacher generating 3D proposals via open-vocabulary detection) and the student model (a 3D point cloud detector) are trained simultaneously from scratch. This online distillation mechanism provides the student with smoother soft-label distributions compared to traditional hard labels, effectively mitigating overfitting risks caused by label sharpening and enhancing generalization robustness.

Specifically, we design a *Triple-Source Label Refinement* module to tackle error accumulation. This module leverages reliable predictions generated by the teacher model in each iteration to online update and correct the student's pseudo-labels, effectively blocking the propagation chain of error signals. Regarding the utilization of image modalities, we propose *Heatmap Proposal Distillation* strategy. Instead of performing expensive cross-modal alignment in the feature space, we employ category response heatmaps generated by the 2D teacher model as soft supervision signals, guiding the 3D student model to learn semantic distributions in the Bird's-Eye-View (BEV) space. This strategy avoids the inter-modal domain gap while fully integrating the fine-grained semantic information provided by image textures.

- We propose BEVOpen3D, a novel open-world 3D detection framework that resolves the error accumulation problem inherent in self-training processes and designs an efficient cross-modal information fusion strategy. It fully leverages image texture and 2D open-vocabulary model priors while bypassing the difficulties of direct domain alignment.
- We achieve significant performance improvements on outdoor dataset, where extensive experiments on the nuScenes dataset demonstrate that our method outperforms existing baselines under open-world settings, achieving state-of-the-art detection accuracy. In particular, it exhibits superior recall and classification accuracy for unseen categories, validating its potential for real-world deployment.

2 Related Work

2.1 BEV-based 3D Object Detection

BEV representation has emerged as a unified paradigm for 3D detection, enabling efficient convolution and multi-modal fusion. While LiDAR-based methods such as PointPillars [11] and CenterPoint [24] naturally exploit this space via pseudo-images, camera-based approaches address the ill-posed view transformation through two distinct strategies: explicit geometric lifting and implicit attention mechanisms. The former, pioneered by LSS [18] and advanced by

BEVDet4D [6], relies on depth estimation to splat features into BEV. Conversely, transformer-based methods like DETR3D [21] and BEVFormer [13] bypass explicit depth estimation by learning implicit spatial-temporal correlations via query-based attention, currently defining the state-of-the-art in end-to-end perception. Notably, Ji et al. [8] further improve this paradigm by augmenting queries with 2D detection-guided anchors.

2.2 Open-vocabulary 3D Object Detection

Open-vocabulary 3D object detection aims to localize and recognize objects beyond predefined categories, leveraging vision-language models (VLMs) such as CLIP [19] and GLIP [12]. Early approaches primarily focus on transferring 2D VLM knowledge to 3D domains. OV-3DET [16] pioneers this direction by distilling knowledge from 2D pre-trained detectors and CLIP into point clouds without requiring 3D annotations for unseen classes. CoDA [3] further eliminates the dependency on auxiliary 2D detectors through collaborative novel box discovery and cross-modal alignment.

Recent advances shift towards learning cross-modal representations via hierarchical alignment and unified frameworks. [10, 26] capture multi-granularity scene understanding by aligning local object features with global context from instance to scene levels. OV-Uni3DETR [22] proposes a unified framework via cycle-modality propagation, which enables seamless modality switching during inference by propagating 2D semantic knowledge to guide 3D novel class discovery while leveraging 3D geometric knowledge to supervise 2D detection images. This strategy achieves modality unification, scene unification, and open-vocabulary learning simultaneously.

To address the scarcity of 3D annotations, recent methods explore weakly supervised, unsupervised, and image-only training strategies. Weakly supervised methods [9, 25] demonstrate that 2D image-level annotations are sufficient for accurate 3D localization. Unsupervised approaches [17] distill CLIP-based semantic knowledge into point cloud clustering without any manually annotated labels. The latest works [7, 23] push the boundary further by training open-vocabulary 3D detectors solely from 2D images. ImOV3D [23] synthesizes pseudo point clouds from internet images through monocular depth estimation and point cloud rendering, effectively removing the need for exhaustive 3D supervision.

3 Method

We present BEVOpen3D, a distillation-based framework for open-world 3D object detection under the partial-label setting. As illustrated in Fig. 1, our pipeline leverages a vision-based teacher to refine noisy pseudo labels and distill knowledge into a LiDAR-only student. The core of our approach lies in three key components: (1) *Seen-Guided Local Query* mechanism (Sec. 3.3) that learns geometric corrections from seen-category ground truth by attending to local BEV regions. (2) the *Triple-Source Label Refinement* strategy (Sec. 3.4) that fuses

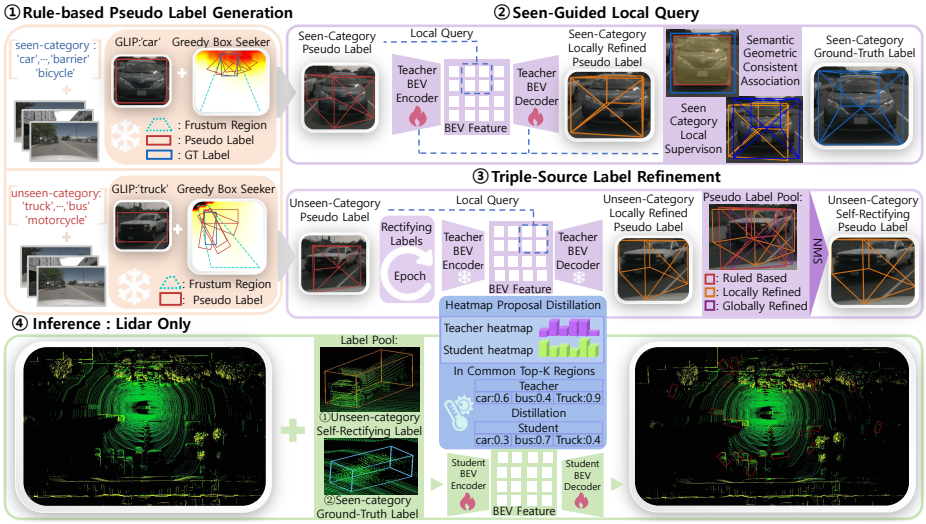

Fig. 1: Overall pipeline of our proposed BEVOpen3D framework, including four main modules: Rule-based Pseudo Label Generation is detailed in Section 3.2; Seen-Guided Local Query is described in Section 3.3; Triple-Source Label Refinement is explained in Section 3.4; and Heatmap Proposal Distillation is introduced in Section 3.5.

original, locally refined, and globally re-examined proposals via 3D-NMS to produce high-quality pseudo labels for unseen categories. (3) the *Heatmap Proposal Distillation* scheme (Sec. 3.5) that transfers vision-grounded semantics to the student through heatmap-level alignment, avoiding unstable cross-modal feature matching.

3.1 Overview

We address open-world 3D object detection problem under partial-label setting. **Training Input:** The model accepts a point cloud $\mathbf{P}$ and synchronized RGB images $\mathbf{I}$, supervised by ground-truth boxes $\mathbf{B}^{\mathrm{gt,s}}$ for *seen* categories $\mathcal{S}$ and textual names only for *unseen* categories $\mathcal{U}$. **Inference Output:** The goal is to localize 3D boxes $\hat{B}$ with labels $c_i \in \mathcal{S} \cup \mathcal{U}$ using only LiDAR input, despite the absence of 3D annotations for $\mathcal{U}$ during training.

Label Pool & Distillation: we employ an online distillation framework with distinct roles for the teacher and student. The *teacher* continuously refines rule-based proposals into high-quality pseudo labels for unseen categories and transfers semantic knowledge via heatmap-level distillation. Conversely, the lightweight LiDAR *student* is optimized on an aggregated label pool that it consumes, but never updates. This pool comprises: (1) *fixed ground-truth annotations* $\mathbf{B}^{\mathrm{gt,s}}$ for seen categories, and (2) *self-rectifying pseudo labels* $\mathbf{B}^{\mathrm{rp,u}}$ for unseen categories, which are dynamically generated by the teacher. By decoupling label genera-

tion from model optimization, this strategy prevents error accumulation while enabling robust open-world detection.

3.2 Rule-based Pseudo Label Generation

We bootstrap rule-based pseudo labels $\mathbf{B}^{\mathrm{p}}$ via a pipeline inspired by Greedy Box Seeker [5]. Given LiDAR point clouds and synchronized multi-view images, an open-vocabulary 2D detector first generates 2D boxes with semantic confidences. Each 2D detection is lifted to a 3D frustum, within which a spatial search over depth, scale, and orientation candidates yields the highest-scoring 3D proposal via density-IoU scoring–entirely without learnable parameters.

3.3 Rectify Labels via Vision

Built upon the canonical 3DETR architecture, the teacher network processes BEV feature maps to simultaneously support global heatmap detection and local instance refinement. Rather than treating rule-based proposals as fixed targets, we explicitly model the systematic bias inherent in hand-crafted features by leveraging the discrepancy between seen-category ground truth and their corresponding offline pseudo labels. By learning this corrective mapping on seen categories, the network establishes a robust association between geometric priors and visual evidence, enabling it to progressively rectify coarse proposals into high-quality position and category pseudo labels. This vision-grounded correction capability generalizes directly to unseen categories, upgrading the supervision signal without requiring additional parameters or auxiliary networks.

Seen-Guided Local Query To refine rule-based pseudo labels for seen categories, we construct a localized query under the supervision of ground-truth annotations. For each rule-based pseudo label $\mathbf{B}_m^{\mathrm{p,s}} \in \mathbb{R}^8$, we generate a binary mask $M_m \in \{0,1\}^{X\times Y}$ by rasterizing the bounding box onto the BEV feature map and collect all grid cells inside it:

$$\mathcal{S}^{\mathrm{loc}} = \big\{(x,y) \in \mathbb{Z}^2 \mid M_m^s[x,y] = 1\big\}. \tag{1}$$

A single local token is obtained by max-pooling over the indexed BEV features $\mathrm{BEV}_t[\mathcal{S}^{\mathrm{loc}}]$. To maintain coordinate consistency with the global decoder, we map the 3D center (x_m, y_m) of $\mathbf{B}_m^{\mathrm{p;s}}$ back to the BEV lattice (u_m, v_m) and augment the pooled feature with the identical sinusoidal positional encoding $\mathbf{p}_m \in \mathbb{R}^C$ used in the transformer backbone. The resulting query is passed through the shared decoder DEC_ϕ to yield the locally refined pseudo label:

$$\hat{\mathbf{b}}_m^{\mathrm{loc},s} = \mathrm{DEC}_\phi\Big(\mathrm{MaxPool}\big(\mathrm{BEV}_t[\mathcal{S}^{\mathrm{loc}}]\big) + \mathbf{p}_m\Big). \tag{2}$$

This output is supervised by ground-truth boxes and replaces the original noisy proposal.

To associate pseudo labels with ground truth, we propose a *semantic–geometric consistent matching* strategy. Given a 3D ground-truth box $\mathbf{B}_n^{\mathrm{gt,s}}$ and 2D open-vocabulary detector $\mathcal{D}_{\mathrm{2D}}^{\mathrm{OV}}$ from multi-view cameras, we first project the 3D box into each camera's image plane:

$$\mathcal{E}_{n,k} = \mathit{\Pi}_k\big(\mathbf{B}_n^{\mathrm{gt,s}}; \mathbf{R}_k, \mathbf{t}_k, \mathbf{K}_k\big), \quad \text{s.t.} \quad \mathbf{e}_3^\top(\mathbf{R}_k\mathbf{c}_n + \mathbf{t}_k) > 0, \tag{3}$$

where $\mathit{\Pi}_k(\cdot)$ denotes visibility-constrained projection that yields the axis-aligned enclosing rectangle $\mathcal{E}_{n,k}$ only if the object lies in front of camera k. We then perform bipartite matching between all projected boxes $\{\mathcal{E}_{n,k}\}$ and 2D detector $\mathcal{D}_{\mathrm{2D}}^{\mathrm{OV}}$ using the Hungarian algorithm, resulting in the assignment set:

$$\mathcal{M}_t^{\mathrm{s}} = \Big\{(n,m) \;\Big|\; \mathbf{B}_n^{\mathrm{gt,s}} \text{ is matched to } \mathbf{B}_m^{\mathrm{p,s}} \text{ at time } t\Big\}, \tag{4}$$

with the optimal assignment minimizing the category-aware cost:

$$\min_{\pi} \sum_{(i,j)\in\pi} \mathit{\Delta}_{ij}, \;\; \mathit{\Delta}_{ij} = \begin{cases} 1 - \mathrm{IoU}(\mathcal{E}_{i,k}, \mathcal{D}_{\mathrm{2D},j}^{\mathrm{OV}}), & c_i = c_j \\ +\infty, & \text{otherwise.} \end{cases} \tag{5}$$

This matching is performed offline and solely serves to link pseudo labels with ground truth for loss computation, restricted to seen classes to avoid leakage from unseen categories.

Heatmap-Driven Global Query Simultaneously, the global stream implements a conventional heatmap based 3D detection baseline. Operating on the entire BEV representation, this model generates object hypotheses for both seen and unseen categories without region constraints.

Specifically, the final BEV representation BEV_t is projected into a per-class center heatmap $H_t^{\mathrm{T}} \in \mathbb{R}^{X\times Y\times N_{\mathrm{cls}}}$. Top-$K$ peaks extracted across all N_{cls} channels serve as global object queries,

$$\mathcal{S}^{\mathrm{glb}} = \mathrm{TopK}\big(H_t^{\mathrm{T}}\big) \triangleq \big\{(x_k, y_k)\big\}_{k=1}^{K} \subset \mathbb{Z}^2. \tag{6}$$

Their C-dimensional feature vectors are sampled from BEV_t and refined by the transformer decoder to produce globally refined pseudo labels,

$$\hat{\mathbf{b}}^{\mathrm{glb}} = \mathrm{DEC}_\phi\Big(\mathrm{BEV}_t\big[\mathcal{S}^{\mathrm{glb}}\big]\Big). \tag{7}$$

Global supervision strategy The global stream is trained against a mixed label pool comprising ground truth labels $\mathbf{B}^{\mathrm{gt,s}}$ for seen categories and rule-based pseudo labels $\mathbf{B}^{\mathrm{p,u}}$ for unseen categories. Predictions $\hat{\mathbf{b}}^{\mathrm{glb}}$ are matched to this mixed pool via class-aware Hungarian assignment. Matched seen-category proposals are supervised by the standard L1 regression and focal classification losses. For matched unseen-category proposals, we attenuate supervision via dedicated loss weights, bounding-box regression keeps full gradients for center, size, and orientation, but the classification loss is down-weighted by 0.45. This design preserves high precision on seen-category proposals while providing continuous regression gradients for unseen regions, establishing a semantically smooth foundation that complements the fine-grained local refinement stream.

Optimization Objective We now present the complete optimization framework for the teacher network, which integrates both local and global streams into a unified training objective.

Local rectification loss. With established correspondences $\mathcal{M}_t^{\mathrm{s}}$, the local stream is supervised by ground truth labels through a regression loss,

$$\mathcal{L}_{\mathrm{rect}} = \sum_{(n,m)\in\mathcal{M}_t^{\mathrm{s}}} \mathcal{L}_{\mathrm{match}}\big(\hat{\mathbf{b}}_m^{\mathrm{loc},s},\, \mathbf{B}_n^{\mathrm{gt,s}}\big), \tag{8}$$

where $\mathcal{L}_{\mathrm{match}}$ combines smooth-L1 regression for box parameters and focal classification for category prediction. Minimizing this loss forces the shared decoder weights to perform a corrective mapping that pulls coarse rule-based proposals toward the ground-truth. Since these proposals are obtained offline by the Greedy Box Seeker, the residual between $\hat{\mathbf{b}}_m^{\mathrm{loc},s}$ and $\mathbf{B}_n^{\mathrm{gt,s}}$ directly encodes the systematic bias introduced by the geometry-driven pipeline. Learning this bias model on seen categories enables zero-shot transfer to unseen categories. Specifically, during the pseudo-label generation phase with the frozen teacher, we apply this mechanism to unseen-class proposals. The identical local refinement procedure processes coarse proposals $\mathbf{B}^{\mathrm{p,u}}$ without requiring ground truth supervision. Consequently, it produces locally refined pseudo labels $\hat{\mathbf{b}}_m^{\mathrm{loc},u}$. These refined labels inherit the corrective capability acquired from seen-class training, effectively rectifying geometric biases for novel objects.

Joint optimization.The total teacher objective is defined as,

$$\mathcal{L}_{\mathrm{teacher}} = \mathcal{L}_{\mathrm{global}}(\hat{\mathbf{b}}^{\mathrm{glb}}, \mathbf{B}^{\mathrm{gt,s}} \cup \mathbf{B}^{\mathrm{p,u}}) + \lambda_{\mathrm{rect}}\mathcal{L}_{\mathrm{rect}}. \tag{9}$$

3.4 Triple-Source Label Refinement

The Greedy Box Seeker (GBS) generates 3D pseudo labels offline without learnable parameters, but suffers from geometric bias that object centers are naively snapped to rays cast from 2D detection centroids, causing boxes to hover above the ground surface. These floating proposals are often suppressed during NMS.

We regard GBS proposals $\mathbf{B}^{\mathrm{p}}$ as noisy samples from an unknown conditional distribution $p(\mathbf{B}^{\mathrm{gt,s}} \mid \mathbf{B}^{\mathrm{p,s}})$ encoding the offset between geometry-driven guesses and real object surfaces. For seen categories, the teacher learns this residual distribution through ground-truth supervision. We then freeze the teacher and apply the same correction to unseen categories via two complementary streams, yielding three proposal sources fused by NMS.

Local Refinement For each GBS proposal $\mathbf{B}^{\mathrm{p,u}}$ of unseen category, we crop the corresponding BEV region, max-pool its features into a 256-dimensional token, augment with sinusoidal positional encoding $\mathbf{p}_m$, and forward through the frozen decoder,

$$\hat{\mathbf{B}}_m^{\mathrm{loc,u}} = \mathrm{DEC}_\phi\big(\mathrm{MaxPool}\big(\mathrm{BEV}_t\big[\,\mathcal{S}^{\mathrm{loc,u}}\big]\big) + \mathbf{p}_m\big)\,, \tag{10}$$

where $\mathcal{S}^{\mathrm{loc,u}}$ represents BEV cells enclosed by the bird's-eye rectangle of $\mathbf{B}^{\mathrm{p,u}}$. This process transforms geometric discrepancies into a structured output. While

traditional 3D detection models typically maintain an end-to-end box-to-box mapping, our approach leverages max pooling to unify multi-scale candidate regions into fixed-dimensional feature tokens. These tokens serve as queries for the decoder, which is designed to map such query embeddings directly to box parameter residuals. Consequently, the decoder outputs an 8D correction vector that encodes the necessary adjustments (e.g., center offsets, dimension scaling, and orientation corrections) to refine the initial GBS proposal, effectively compressing the complex geometric error distribution into a concise, learnable representation.

Global Refinement The frozen teacher re-evaluates the full BEV feature map BEV_t via 1×1 convolution producing heatmap H_t^{T}. Top-K peaks across all N_{cls} channels serve as queries,

$$\mathcal{S}^{\mathrm{glb,u}} = \mathrm{TopK}\big(H_t^{\mathrm{T}}\big) \triangleq \big\{(x_k, y_k)\big\}_{k=1}^{K}, \tag{11}$$

yielding decoded predictions $\{\hat{\mathbf{B}}_k^{\mathrm{glb,u}}, s_k^{\mathrm{glb}}\}_{k=1}^{K}$. This provides an unbiased opinion that recovers objects missed by GBS.

Triple-Source Fusion. The three proposal sets, namely original GBS boxes $\{\mathbf{B}^{\mathrm{p,u}}\}$, locally refined boxes $\{\hat{\mathbf{b}}^{\mathrm{loc,u}}\}$, and globally re-scored candidates $\{\hat{\mathbf{b}}^{\mathrm{glb,u}}\}$, are concatenated and fed into class-agnostic 3D NMS (IoU threshold 0.5). Surviving boxes constitute the upgraded unseen-category pseudo labels,

$$\{\mathbf{B}_n^{\mathrm{rectified,u}}\}_{n=1}^{N} = \text{3D-NMS}\Big(\{\mathbf{B}^{\mathrm{p,u}}\} \cup \{\hat{\mathbf{b}}^{\mathrm{loc,u}}\} \cup \{\hat{\mathbf{b}}^{\mathrm{glb,u}}\}\Big). \tag{12}$$

3.5 Heatmap Proposal Distillation

Image and LiDAR BEV features exhibit structural discrepancy, dense texture-rich grids versus sparse measurement-bound representations, making direct feature alignment suboptimal. Our method attempt to perform distillation at the heatmap proposal level, allowing the LiDAR student to indirectly acquire vision's texture-aware object discovery capability without cross-modal feature emulation.

To avoid overfitting to poorly calibrated teacher outputs, we adopt online distillation activated after a brief warm-up and at fixed intervals. The teacher extracts top-K high-response proposals from its dense heatmap, each with BEV center c_i^{T}, class probability vector p_i^{T}, and peak confidence w_i. The student gathers its own top-K heatmap responses. Distillation is restricted to mutually matched locations under class-consistent local constraints. A teacher proposal is matched to a student response only if they share the same class and lie within a spatial radius ρ, using greedy nearest-neighbour assignment. Let $\mathcal{M}$ denote the matched set.

The distillation loss at each matched location combines a prior term on the full class distribution and a score term on the teacher-selected class,

$$\mathcal{L}_{\mathrm{distill}} = \frac{1}{\sum_{i\in\mathcal{M}} w_i} \sum_{i\in\mathcal{M}} w_i \Big[\lambda_{\mathrm{prior}} \big\| p_i^{\mathrm{S}} - p_i^{\mathrm{T}} \big\|_2^2 + \lambda_{\mathrm{score}} \big| p_{i,c_i}^{\mathrm{S}} - p_{i,c_i}^{\mathrm{T}} \big| \Big], \tag{13}$$

Table 1: Performance Comparison of BEVOpen3D, OV-Uni3DETR, and Find n' Propagate Across Different Novel Class Settings on the Validation Set Here, 'StuDH.' denotes Student Dense Head, 'TeaDH.' refers to Teacher Dense Head, 'Trans.' indicates TransFusion, and 'Focal' represents FocalFormer3D.

(a) Four novel classes. Results with 4 novel categories (Truck, Bus, Motorcycle, Traffic Cone)

			Overall				Seen Categories						Unseen Categories			
Method	StuDH.	TeaDH.	mAP	NDS	AP_S	AP_U	Car	Const.	Trai.	Barr.	Bic.	Ped.	Truck	Bus	Motor.	Cone.
BEVOpen3D	Seed	Seed	44.10	46.25	51.42	33.11	83.32	14.74	32.38	68.66	28.81	80.64	25.98	26.15	34.97	45.36
	Seed	Trans.	42.87	45.38	50.32	31.70	81.12	14.24	33.02	67.11	27.02	79.42	26.32	21.79	33.56	45.13
	Seed	Focal.	40.98	41.19	50.07	27.34	80.91	12.10	34.12	59.66	28.02	85.64	24.78	24.80	34.97	24.81
	Trans.	Trans.	41.80	43.88	49.38	30.44	79.67	15.70	30.13	61.26	30.48	79.01	25.34	25.22	27.06	44.14
	Seed	Trans.	40.87	40.68	47.51	30.93	80.25	12.83	29.06	60.94	25.03	76.93	26.48	23.90	30.98	42.36
	Focal.	Trans.	43.88	46.74	49.94	34.80	81.26	14.74	28.71	67.22	28.44	79.24	27.17	27.12	36.48	48.44
	Focal.	Focal.	39.94	40.17	46.86	29.56	74.36	10.64	25.36	61.34	28.81	80.64	21.96	15.94	34.97	45.36
	Trans.	Focal.	41.36	42.72	49.02	29.88	77.49	10.14	27.37	69.66	26.81	82.62	25.99	20.15	35.03	38.36
	Seed	Focal.	40.08	42.52	45.71	31.62	73.73	12.57	23.71	65.08	26.81	72.36	25.11	20.15	38.79	42.46
OV-Uni3DETR	-	-	44.29	30.82	61.95	17.79	83.42	29.25	40.99	73.16	62.48	82.43	12.81	14.30	20.03	24.03
Find n' Propagate	-	-	43.79	46.08	51.20	32.67	83.30	14.86	31.99	69.12	28.48	79.43	25.77	26.78	33.92	44.23

(b) Seven novel classes. Results with 7 novel categories (all except Car, Bicycle, Pedestrian)

			Overall				Seen Categories			Unseen Categories						
Method	StuDH.	TeaDH.	mAP	NDS	AP_S	AP_U	Car	Bic.	Ped.	Bus	Trai.	Barr.	Motor.	Truck	Const.	Cone.
BEVOpen3D	Seed	Seed	27.39	29.12	62.04	12.53	82.40	25.30	78.43	28.29	0.34	0.12	22.57	14.23	0.95	21.23
	Seed	Trans.	27.65	28.93	59.90	13.83	75.49	27.94	76.29	22.77	0.19	0.22	19.33	23.17	0.91	30.23
	Seed	Focal.	29.71	31.08	63.56	15.20	79.45	30.19	81.04	29.18	0.12	0.18	26.24	21.20	0.62	28.83
	Trans.	Trans.	27.23	28.14	62.82	11.98	77.91	31.62	78.92	19.14	0.22	0.19	25.37	14.06	0.66	24.23
	Seed	Trans.	28.65	29.38	61.07	14.76	74.30	28.48	80.43	28.77	0.27	0.24	28.37	15.89	0.71	29.07
	Focal.	Trans.	26.25	28.03	60.31	11.65	71.11	27.41	82.41	2.98	0.21	0.24	29.10	18.23	0.56	30.23
	Focal.	Focal.	28.42	30.64	60.15	14.82	82.14	26.09	72.22	20.51	0.21	0.33	26.16	24.63	0.64	31.23
	Trans.	Focal.	30.10	31.26	62.40	16.26	76.25	28.53	82.43	26.93	0.22	0.28	32.55	22.73	0.84	30.23
	Seed	Focal.	28.19	29.43	63.01	13.27	78.30	26.91	83.82	19.83	0.24	0.23	26.12	18.58	0.64	27.22
OV-Uni3DETR	-	-	24.47	10.87	73.98	3.25	82.52	59.25	80.18	9.03	0.11	0.07	0.95	11.17	0.54	0.86
Find n' Propagate	-	-	29.37	31.62	63.07	14.93	80.30	28.48	80.43	22.78	0.27	0.24	29.92	20.23	0.86	30.23

where c_i is the teacher-assigned class for proposal i.

The student is trained on seen-category ground-truth $\mathbf{B}^{\mathrm{gt,s}}$ together with merged pseudo-labels from both offline frustum proposals and online self-training proposals refreshed by the teacher at scheduled epochs. The total objective is

$$\mathcal{L}_{\mathrm{student}} = \mathcal{L}_{\mathrm{det}}\big(\hat{\mathbf{b}}^{\mathrm{S}}, \mathbf{B}^{\mathrm{gt,s}} \cup \mathbf{B}^{\mathrm{pseudo}}\big) + \mathcal{L}_{\mathrm{distill}}. \tag{14}$$

At inference, only the lightweight student remains, processing raw point clouds without vision-side overhead.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate BEVOpen3D on the nuScenes dataset [2], a large-scale autonomous driving benchmark containing 1,000 driving scenes. Each scene provides 20-second video sequences captured by 6 surrounding cameras and a 32-beam LiDAR, with 3D bounding box annotations for 10 object categories. Following the standard open-vocabulary protocol [5,16], we partition the categories

into seen and unseen sets. For the 4-novel-class setting, we treat *Car*, *Construction Vehicle*, *Train*, *Barrier*, *Bicycle*, and *Pedestrian* as seen categories, while *Truck*, *Bus*, *Motorcycle*, and *Traffic Cone* serve as unseen categories. For the more challenging 7-novel-class setting, only *Car*, *Bicycle*, and *Pedestrian* are seen, with the remaining 7 categories treated as unseen classes.

Evaluation Metrics. We adopt the standard nuScenes detection metrics, including mean Average Precision (mAP) and nuScenes Detection Score (NDS). mAP is calculated by matching predictions to ground truth based on 2D BEV center distance (thresholds: 0.5m, 1m, 2m, 4m) and averaging over all categories. NDS combines mAP with additional quality measures including average translation error, scale error, orientation error, velocity error, and attribute error. For open-world evaluation, we specifically report AP_S (mAP on seen categories) and AP_U (mAP on unseen categories) to separately assess the model's ability to maintain performance on known classes while generalizing to novel ones.

Implementation Details. Our experiments are conducted within the OpenPCDet framework [20], a widely adopted codebase for 3D object detection. The teacher network employs Swin-Transformer [14] as the image backbone to extract multi-scale features from 6 surrounding-view images, which are then lifted into BEV space via view transformation. The student network utilizes VoxelNet [27] as the point cloud encoder, converting raw LiDAR points into voxel features for efficient processing.

4.2 Comparison with State-of-the-Art

We compare BEVOpen3D against two recent open-world 3D detection methods, OV-Uni3DETR [22] and Find n' Propagate [5]. OV-Uni3DETR is a unified detection framework designed for both indoor and outdoor scenarios, employing hierarchical cross-modal alignment to enable flexible modality switching. While primarily optimized for indoor environments, it demonstrates competitive performance on outdoor benchmarks through its scene-agnostic architecture. Find n' leverages an offline strategy that combines greedy spatial search with multi-round self-training to iteratively refine pseudo labels for transferring 2D vision-language knowledge to 3D point clouds.

Additionally, to validate the generalization capability of our proposed modules, we instantiate BEVOpen3D with two different dense detection heads built upon the TransFusion [1] baseline, Seed [15] and FocalFormer3D [4]. Both heads inherit TransFusion's query initialization mechanism based on BEV heatmap peak detection, making them fully compatible with our Truth-Rectified Pseudo Query (TR-PQuery) and Heatmap Proposal Distillation (HP-Distill) modules. Seed introduces a self-ensembling mechanism that aggregates multi-scale features for robust query refinement, while FocalFormer3D incorporates focal attention to concentrate on informative regions and suppress background noise. By testing multiple head configurations, we demonstrate that BEVOpen3D's improvements are orthogonal to specific detection head designs.

4.3 Main Results

Four Novel Classes. As shown in Table 1a, BEVOpen3D achieves a superior balance between seen and unseen category performance. The Seed-Seed configuration attains 44.10% mAP and 46.25% NDS, with a notable 33.11% AP_U on unseen classes. This significantly outperforms OV-Uni3DETR (17.79% AP_U). Furthermore, the Focal-Trans variant achieves the highest unseen performance (34.80% AP_U), demonstrating the robustness of our rectification mechanism across different detection paradigms. Compared to the baseline Find n' Propagate, BEVOpen3D maintains comparable overall accuracy while offering better generalization to novel objects.

Seven Novel Classes. In the more challenging 7-novel-class setting (Table 1b), BEVOpen3D's advantages become more pronounced. The Trans-Focal configuration yields 30.10% mAP and 16.26% AP_U, substantially surpassing both OV-Uni3DETR (3.25% AP_U) and Find n' Propagate (14.93% AP_U). Two key trends are reported. First, the performance gap widens as the number of unseen categories increases, validating the efficacy of our Triple-Source Label Refinement in high-novelty scenarios. Second, while OV-Uni3DETR suffers a drastic NDS drop (from 30.82% to 10.87%), BEVOpen3D maintains stable scores (28–31%), indicating robust velocity and attribute prediction even when most categories are unseen.

Analysis and Comparison. The improvements across various student-teacher head combinations confirm that BEVOpen3D's core modules (TR-PQuery and HP-Distill) are architecture-agnostic. While matched head pairs (e.g., Seed-Seed) generally excel in the 4-novel-class setting, mismatched configurations (e.g., Trans-Focal) show superior unseen performance in the 7-novel-class setting, suggesting that diverse architectural perspectives provide complementary signals when pseudo-label quality is critical. Finally, against Find n' Propagate, BEVOpen3D's progressive rectification strategy yields a clear margin (+0.73% mAP, +1.33% AP_U in the 7-novel-class setting). This confirms that actively correcting geometric biases via our Triple-Source Label Refinement method is more effective than relying solely on offline heuristics, preventing error accumulation during distillation.

5 Ablation Study and Visualisation

We conduct ablation studies under the standard partial-label protocol (6 seen, 4 unseen categories). All variants employ the Seed-Seed architecture. We evaluate four configurations to isolate the effects of **Local Query**, **Heatmap Proposal Distillation**, and **Triple-Source Label Refinement**.

1. **Full BEVOpen3D.** The complete framework. The teacher utilizes Local Query and Triple-Source fusion to generate rectified pseudo labels, while Heatmap Proposal Distillation transfers vision-grounded semantic priors to the LiDAR student.

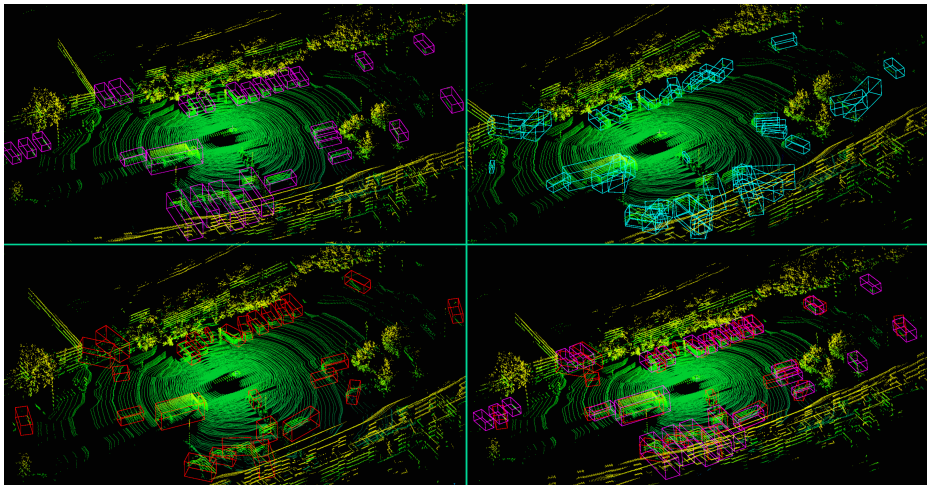

Fig. 2: Visualization of a scene from the nuScenes dataset. The composite image displays (*top-left*) Ground Truth labels with magenta bounding boxes; (*top-right*) Triple-Source Pseudo Labels (covering both known and unknown classes) with red bounding boxes; (*bottom-left*) Student model inference results with red bounding boxes; and (*bottom-right*) The combined visualization of inference results overlaid on Ground Truth labels.

2. **Local Query Only.** Using the *Seen-Guided Local Query* to refine seen-class proposals but relies on static, offline GBS pseudo labels for unseen classes without iterative rectification or fusion.
3. **Local + Triple-Source.** Incorporating the teacher-student paradigm with *Triple-Source* fusion to dynamically update unseen pseudo labels. However, *Heatmap Distillation* is disabled, removing direct vision-to-LiDAR semantic guidance.
4. **Distillation Only.** Employing only *Heatmap Proposal Distillation*, where the teacher transfers semantic priors to guide the student's self-training and pseudo-label updates. It excludes both *Local Query* refinement and *Triple-Source* fusion.

Results Analysis. As shown in Table 2, the *Distillation Only* variant (Config 4) establishes a strong foundation by leveraging vision-guided self-training, yet it remains limited by unrefined geometric proposals. Adding *Local Query* (Config 2) yields marginal gains by improving alignment for seen classes. Significant improvements arise in Config 3 (*Local + Triple-Source*), where the teacher's ability to iteratively rectify pseudo labels for unseen categories via fusion drastically reduces supervision noise. Finally, the full model (Config 1) achieves the best performance by incorporating all modules, ensuring both precise geometric alignment and robust semantic regularization. Visualizations in Fig. 2 confirm that our full framework effectively corrects the floating box issues prevalent in raw GBS proposals. In Table 3, we remove Top-K filtering and test two variants:

Table 2: Comparison of different query strategies and distillation methods on validation set.

Local query	Heatmap Proposal Distillation	Triple-Source Pseudo Label	mAP	NDS
✓	✓	✓	**44.10**	**46.25**
✓	✗	✗	38.72	43.16
✓	✗	✓	43.06	46.68
✗	✓	✗	35.42	34.12

Table 3: Comparison of pseudo-label generation strategies on validation set.

Method	mAP	NDS
without Top-K Filtering, Teacher-generated Pseudo-Labels	41.29	43.61
without Top-K Filtering, Student-generated Pseudo-Labels	34.20	38.18

teacher-generated pseudo-labels with distillation, and student-generated pseudo-labels with teacher-only distillation. The latter drops by 7.09 mAP and 5.43 NDS, showing clear degradation. This confirms that teacher-provided pseudo-labels are essential to block error accumulation, as student-generated labels suffer from confirmation bias that progressively amplifies initial prediction errors. Moreover, the substantial gap highlights that the teacher's role extends beyond distillation supervision to actively providing reliable pseudo-labels that anchor the student's learning.

6 Conclusion

We presented **BEVOpen3D**, a novel framework for open-world 3D detection under partial-label supervision. To address the limitations of noisy rule-based pseudo labels, our method introduces a *Seen-Guided Local Query* mechanism to learn geometric corrections from seen categories, which are then generalized to refine unseen proposals. Then we proposed a *Triple-Source Label Refinement* strategy that fuses original, locally rectified, and globally re-examined proposals to generate high-quality supervision signals. Furthermore, we designed a *Heatmap Proposal Distillation* scheme to effectively transfer vision-grounded semantic priors from an image-based teacher to a LiDAR-based student without suffering from cross-modal feature misalignment. Experiments on nuScenes demonstrate state-of-the-art performance in detecting unseen categories. By decoupling heavy vision reasoning during training from efficient LiDAR inference, BEVOpen3D offers a practical and robust solution for real-world autonomous driving in open-world scenarios.

References

1. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1090–1099 (2022)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
3. Cao, Y., Yihan, Z., Xu, H., Xu, D.: Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. Advances in Neural Information Processing Systems **36**, 71862–71873 (2023)
4. Chen, Y., Yu, Z., Chen, Y., Lan, S., Anandkumar, A., Jia, J., Alvarez, J.M.: Focalformer3d: focusing on hard instance for 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8394–8405 (2023)
5. Etchegaray, D., Huang, Z., Harada, T., Luo, Y.: Find n'propagate: Open-vocabulary 3d object detection in urban environments. In: European Conference on Computer Vision. pp. 133–151. Springer (2024)
6. Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054 (2022)
7. Huang, R., Zheng, H., Wang, Y., Xia, Z., Pavone, M., Huang, G.: Training an open-vocabulary monocular 3d detection model without 3d data. Advances in Neural Information Processing Systems **37**, 72145–72169 (2024)
8. Ji, H., Liang, P., Cheng, E.: Enhancing 3d object detection with 2d detection-guided query anchors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21178–21187 (2024)
9. Jiang, X., Jin, S., Lu, L., Zhang, X., Lu, S.: Weakly supervised monocular 3d detection with a single-view image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10508–10518 (2024)
10. Jiao, P., Zhao, N., Chen, J., Jiang, Y.G.: Unlocking textual and visual wisdom: Open-vocabulary 3d object detection enhanced by comprehensive guidance from text and image. In: European Conference on Computer Vision. pp. 376–392. Springer (2024)
11. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
12. 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., Gao, J.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10955–10965 (2022)
13. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 2020–2036 (2024)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

15. Liu, Z., Hou, J., Ye, X., Wang, T., Wang, J., Bai, X.: Seed: A simple and effective 3d detr in point clouds. In: European Conference on Computer Vision. pp. 110–126. Springer (2024)
16. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1190–1199 (2023)
17. Najibi, M., Ji, J., Zhou, Y., Qi, C.R., Yan, X., Ettinger, S., Anguelov, D.: Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8602–8612 (2023)
18. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: European conference on computer vision. pp. 194–210. Springer (2020)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
20. Team, O.D.: Openpcdet: An open-source toolbox for 3d object detection from point clouds. `https://github.com/open-mmlab/OpenPCDet` (2020)
21. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on robot learning. pp. 180–191. PMLR (2022)
22. Wang, Z., Li, Y., Liu, T., Zhao, H., Wang, S.: Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. In: European Conference on Computer Vision. pp. 73–89. Springer (2024)
23. Yang, T., Ju, Y., Yi, L.: Imov3d: Learning open vocabulary point clouds 3d object detection from only 2d images. Advances in Neural Information Processing Systems **37**, 141261–141291 (2024)
24. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021)
25. Zhang, G., Fan, J., Chen, L., Zhang, Z., Lei, Z., Zhang, L.: General geometry-aware weakly supervised 3d object detection. In: European Conference on Computer Vision. pp. 290–309. Springer (2024)
26. Zhao, Y., Lin, J., Lau, R.W.: Hierarchical cross-modal alignment for open-vocabulary 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10501–10509 (2025)
27. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)