

OV-SCAN: Semantically Consistent Alignment for Novel Object Discovery in Open-Vocabulary 3D Object Detection

Adrian Chow Evelien Riddell Yimu Wang Sean Sedwards Krzysztof Czarnecki
University of Waterloo

{ahtchow, evelien.riddell, yimu.wang, sean.sedwards, krzysztof.czarnecki}@uwaterloo.ca

Abstract

Open-vocabulary 3D object detection for autonomous driving aims to detect novel objects beyond the predefined training label sets in point cloud scenes. Existing approaches achieve this by connecting traditional 3D object detectors with vision-language models (VLMs) to regress 3D bounding boxes for novel objects and perform open-vocabulary classification through cross-modal alignment between 3D and 2D features. However, achieving robust cross-modal alignment remains a challenge due to semantic inconsistencies when generating corresponding 3D and 2D feature pairs. To overcome this challenge, we present OV-SCAN, an Open-Vocabulary 3D framework that enforces Semantically Consistent Alignment for Novel object discovery. OV-SCAN employs two core strategies: discovering precise 3D annotations and filtering out low-quality or corrupted alignment pairs (arising from 3D annotation, occlusion-induced, or resolution-induced noise). Extensive experiments on the nuScenes dataset demonstrate that OV-SCAN achieves state-of-the-art performance. Our code is available at <https://github.com/ahtchow/OV-SCAN>.

1. Introduction

Recent advances in autonomous driving have intensified research in 3D object detection, with most methods operating in a closed-set setting, classifying objects into a small set of predefined categories (e.g. vehicle, cyclist, pedestrian). However, such high-level object identification is inadequate for navigating complex real-world environments. Autonomous systems need to recognize objects at a deeper semantic level. For instance, distinguishing between a fire truck and a school bus, rather than simply labeling them as trucks or buses, is critical for safe and context-aware decision-making.

To address the need for deeper semantic understanding in 3D vision systems, open-vocabulary (OV)-3D object detection has emerged. In OV-2D object detection, OV ca-

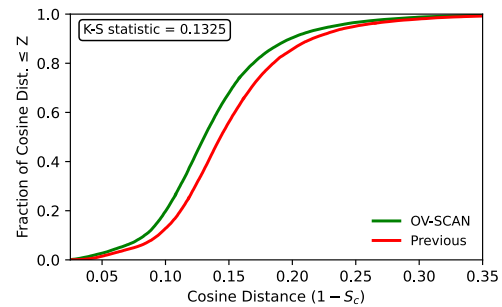


Figure 1. **Cross-modal Alignment Performance.** The red CDF shows the distance distribution between 3D embeddings from a baseline OV-3D detector and 2D CLIP embeddings on the nuScenes validation set. The green CDF, using OV-SCAN instead of the baseline, shifts left, indicating improved 3D-2D alignment enabled by our higher-quality proposal pairings and alignment head.

capabilities rely on large-scale datasets of image-text pairs, which are often unavailable for 3D data [8, 20, 21, 28, 43]. Without similar large-scale 3D-text datasets, methods must find alternative ways to provide accurate 3D annotations for novel objects and achieve alignment between 3D and text.

OV-3D object detection faces two main challenges: (1) novel object discovery (NOD), which involves generating 3D labels for novel objects in order to train an off-the-shelf LiDAR-based detector, and (2) semantic alignment between 3D features and textual embeddings. Existing NOD approaches generally fall into two categories: online and offline methods. The former starts with a set of base objects and progressively expands the training labels to include novel objects during training [4, 5]. The latter generates pseudo-labels for novel objects ahead of time [10, 25, 49]. Online methods often struggle to balance training with discovery, leading to biases in object shape and scale that result in overfitting to base categories. Due to the aforementioned challenges with online methods, recent work has increasingly shifted toward offline approaches. In these offline pipelines, OV-2D detectors such as Grounding DINO [21] and OWL-ViT [27] first generate 2D proposals from multi-view images, which are typically paired with a point cloud.

These 2D proposals are then projected into 3D space and reformulated as 3D pseudo-labels for training. However, because LiDAR point clouds are inherently sparse, naïve projection methods such as simple clustering struggle to accurately fit boxes to objects that are only partially observable and not visibly dense. Most recently, Find n’ Propagate [10] employs an exhaustive greedy search to optimally choose 3D bounding box parameters for each 2D proposal, yet this approach still suffers from several 3D annotation errors shown in Fig. 2. Such noise hinders the 3D object detector’s ability to extract discriminative 3D features.

To achieve semantic alignment between 3D data and text, existing methods use the 2D image space as an implicit intermediary [4, 25, 48, 49]. Each 3D bounding box label is associated with a corresponding 2D image region from which a visual embedding is extracted using a VLM such as CLIP [33]. During training, a base 3D object detector is supervised using the 3D annotations for box regression, while simultaneously aligning 3D object features with target 2D image embeddings. During test time, novel objects are classified via prompt-based classification, assigning labels based on the highest similarity between 3D features and text embeddings. Achieving strong alignment requires both semantically distinct 3D and 2D features. However, existing methods often overlook common autonomous driving scenarios where objects are partially occluded (Fig. 3a) or appear small at longer distances (Fig. 3b). In such cases, the 2D features become ambiguous or lack sufficient representation, leading to confusion during cross-modal alignment.

In this work, we introduce OV-SCAN, an OV-3D object detector which targets strong OV detection performance from enhanced semantic alignment between 3D objects and 2D image features. More specifically, we introduce the Semantically-Consistent Novel-Object Discovery (SC-NOD) module to handle the inherent challenges of noisy cross-modal alignment. To address *3D annotation noise*, SC-NOD reformulates 3D box search as a non-linear optimization problem and employs an adaptive search strategy to efficiently optimize 3D bounding box parameters from 2D proposals. To mitigate *occlusion-induced noise* and *resolution-induced noise*, SC-NOD enables a selective alignment mechanism by identifying the alignment pairs that are semantically consistent. During training, all 3D annotations contribute to box-related losses in order to maintain high recall, while only semantically consistent alignment pairs guide the cross-modal alignment. We further propose the Hierarchical Two-Stage Alignment (H2SA) head to enhance cross-modal alignment. Novel classes are organized hierarchically by first grouping them into broad, high-level categories (e.g., car) and then subdividing these categories into more detailed subclasses (e.g., SUV, sedan). This hierarchical structure allows H2SA to initially perform coarse-grained classification, before carefully aligning

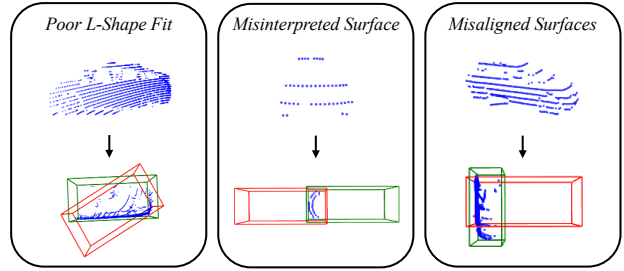


Figure 2. **3D Annotation Errors.** Common 3D annotation errors during box parametrization, including but not limited to, poor L-shape fitting, misinterpreted surfaces, and misaligned surfaces.



Figure 3. **Sources of Semantic Discrepancies.** (a) CLIP similarity scores for a truck reveal that occlusion cases result in an ambiguous 2D image feature. (b) CLIP similarity scores for a distant pedestrian demonstrate that insufficient resolution leads to degraded 2D image features.

class-informed object features for fine-grained discrimination among closely related subclasses. We summarize our main contributions as follows:

- We present OV-SCAN, an OV-3D object detector benefiting from improved cross-modal alignment, see Fig. 1.
- We propose the SC-NOD module, which extends existing 3D box search methods and introduces selective cross-modal alignment during training. In addition, the proposed H2SA head effectively aligns 3D-to-2D alignment pairs by introducing a two-stage alignment process.
- We validate OV-SCAN on the nuScenes [2] and KITTI [12] datasets, demonstrating that OV-SCAN achieves state-of-the-art performance in an OV-3D setting without supervising with any human-annotated data.

2. Related Work

2.1. Traditional 3D Object Detection

Traditional 3D object detection methods focus primarily on closed-set scenarios, training models on a limited number of common classes. LiDAR-based approaches are catego-

rized into single-representation, hybrid-representation, and transformer-based methods. Single-representation methods such as VoxelNet [50] utilize voxel grids processed by 3D CNNs but are computationally intensive. Improvements include sparse convolutions [7, 41], pillars [18], and BEV representations [45] for efficiency. Point-based methods like PointNet [32] and PointRCNN [34] directly process raw point clouds, achieving detailed geometry at the cost of speed. Hybrid models like PV-RCNN [35, 36] integrate voxel and point features for improved accuracy. Transformer-based architectures [24, 26, 29, 31, 37] deploy self-attention to capture global context. LiDAR-camera fusion approaches [1, 23, 44] combine LiDAR’s spatial precision with RGB images to enhance semantic understanding.

2.2. Open-Vocabulary 2D Object Detection

The emergence of large-scale pretrained models with strong text-vision alignment and impressive zero-shot transfer capabilities has driven significant interest in tackling the OV problem in 2D object detection. Approaches like OVR-CNN [47], ViLD [13], OWL-ViT [27], and OV-DETR [46] address this challenge through knowledge distillation, using pretrained VLMs such as CLIP [33]. These methods integrate semantic knowledge into traditional detectors by contrastively aligning region features or object embeddings. Subsequently, GLIP [20] reformulates object detection as phrase grounding, exploiting large-scale image-text pair pretraining to unify localization and classification while enabling generalization to unseen categories through semantic understanding. Grounding DINO [21] advances this idea further by applying tightly coupled fusion strategies on top of DINO [6], ensuring effective integration of language and vision. Extending OV-2D object detection to real-time applications, YOLO-World [8] uses vision-language modelling within the efficient YOLO architecture [15].

2.3. Open-Vocabulary 3D Object Detection

Building on the success of OV-2D object detectors, recent efforts have aimed to extend these capabilities to 3D. A key challenge in this adaptation is the scarcity of diverse labeled data required for large-scale 3D training. OV-3Det [25] introduced the 2D-to-3D paradigm in an indoor setting, where VLMs generate annotations for a diverse set of objects, which are then lifted to 3D using simple point cloud clustering. CoDA [4] pretrains a base detector with a small set of human-annotated labels and introduces a framework to incorporate novel labels during training. FM-OV3D [48] further improves detection performance by exploiting the strengths of multiple foundational models. OpenSight [49] extends open-vocabulary detection into the outdoor domain, incorporating temporal and spatial awareness in box annotations. In addition, their method introduces 2D-to-3D geometric priors to guide annotation under sparse point cloud

conditions. Find n’ Propagate [10] addresses the issue of insufficient recall of novel objects and propagates detection capabilities to more distant areas. ImOV3D [42] demonstrates an effective method to train an OV-3D object detector without any 3D point clouds, instead using a pseudo-multimodal representation. Despite advancements, existing methods often neglect the impact of noise sources that weaken cross-modal alignment.

3. Method

In this section, we present the details of OV-SCAN. An overview is provided in Fig. 4. SC-NOD (Sec. 3.2) generates 3D annotations from 2D proposals to form cross-modal alignment pairs. SC-NOD then determines the subset of semantically consistent pairs to guide cross-modal alignment. OV-SCAN comprises an off-the-shelf 3D object detector with a specialized alignment head (Sec. 3.3), trained using novel objects detected by SC-NOD (Sec. 3.4).

3.1. Notation and Preliminaries

3D Object Detection. In traditional LiDAR-based 3D object detection, the objective is to train a detector using input-target pairs $\mathbf{D} = \{\mathcal{P}, \Omega\}$. Given a point cloud $\mathcal{P} = \{p_i\}_{i=1}^M$ with M points, where each point $p_i = (x_i, y_i, z_i)$ represents spatial coordinates, the detector predicts targets $\Omega = \{(B_i, c_i)\}_{i=1}^N$ for N objects. Each target consists of a 3D bounding box $B_i = (x_i, y_i, z_i, l_i, w_i, h_i, (r_y)_i)$ with center (x_i, y_i, z_i) , dimensions (l_i, w_i, h_i) , orientation $(r_y)_i$, and a class label c_i . We assume that each LiDAR frame is accompanied by a set of K images from different perspectives (multi-view), represented as $\mathcal{I} = \{I_j\}_{j=1}^K$. LiDAR points can be mapped to image space by using the extrinsic and intrinsic calibrations T_{ext} and T_{int} .

Open-Vocabulary 3D Object Detection. In an OV setting, 3D object detectors do not have access to any target labels Ω and instead rely on VLMs to generate annotations for novel objects, creating a *novel object bank*. Our method extends the traditional target pair of 3D bounding box and class label, into a triplet target denoted by $\Omega' = \{(B_i, c_i, \mathcal{A}_{2D,i})\}_{i=1}^N$. $\mathcal{A}_{2D,i} \in \mathbb{R}^D$ represents a 2D alignment embedding that captures the semantic correspondence between the 3D object and its 2D image projection. Traditional LiDAR-based 3D object detection methods are designed to regress 3D object features $\mathcal{O}_{3D} \in \mathbb{R}^H$ given the input point-cloud \mathcal{P} . Our method aims to learn a function f that maps a set of 3D object features \mathcal{O}_{3D} to both the set of predicted 3D bounding boxes B and predicted alignment vectors \mathcal{A}_{2D} , thereby bridging the 3D and 2D domains in OV object detection. These alignment features are then used for prompt-based classification by comparing them with text embeddings generated from class prompts, enabling fine-grained recognition of novel objects.

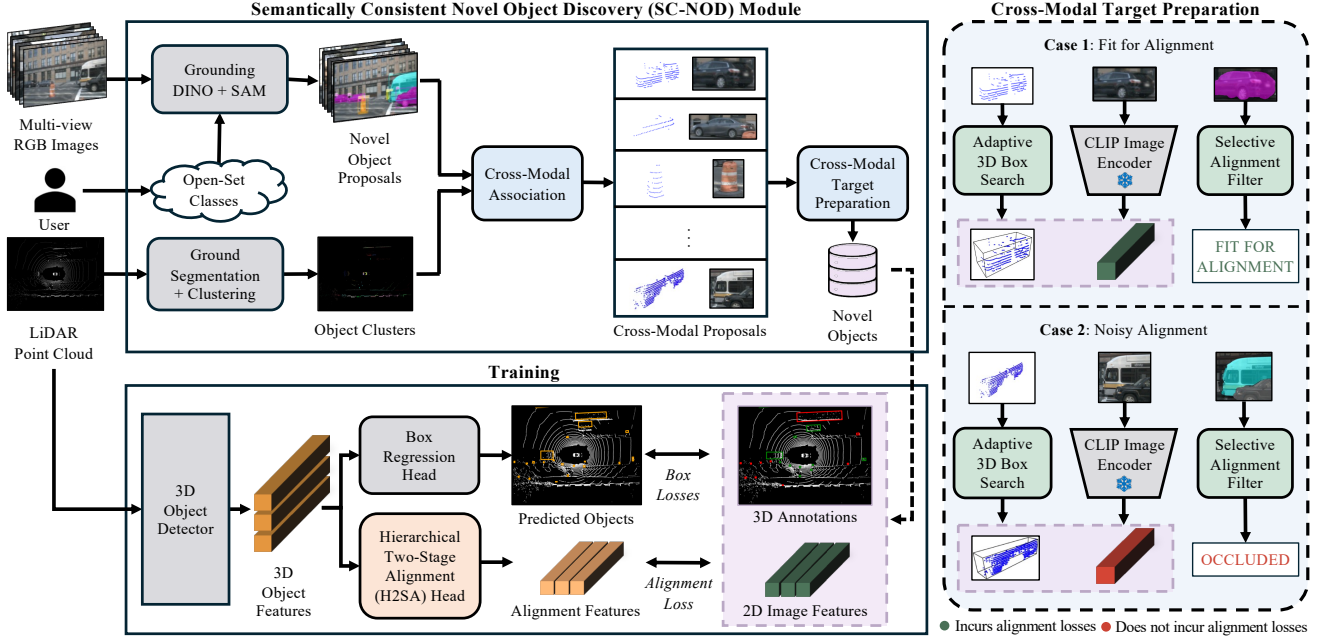


Figure 4. **Overall Framework for OV-SCAN.** During novel object discovery, SC-NOD associates novel object proposals with corresponding object clusters, creating cross-modal proposals. SC-NOD performs an adaptive search to fit 3D annotations and extracts 2D image features to prepare cross-modal targets for supervision. SC-NOD identifies which samples are fit for alignment based on cross-modal semantic consistency. During training, all 3D annotations are used, while only consistent novel objects guide cross-modal alignment.

3.2. Semantically Consistent NOD (SC-NOD)

Novel Object Proposals. A key component of open-vocabulary detection is identifying novel objects from an arbitrary set of classes. Following prior approaches [10, 48, 49], we employ Grounding DINO [21] to detect proposals from a given list of novel classes. For each LiDAR frame, novel object proposals \mathbf{P}_{2D} are generated on K multi-view images, capturing objects from multiple perspectives. These proposals, defined as 2D bounding boxes, are represented for each image I_k as $\mathbf{P}_{2D,k} = \{b_{k,j}\}_{j=1}^{N_k}$, where $b_{k,j}$ specifies position and size. The complete set for a given frame is $\mathbf{P}_{2D} = \bigcup_{k=1}^K \mathbf{P}_{2D,k}$. SAM [17] then pairs instance masks $\mathbf{M} = \{m_{k,j}\}_{k=1,j=1}^{K,N_k}$ with each proposal, where $m_{k,j} \in \{0, 1\}^{H \times W}$ denotes a binary mask for each $b_{k,j}$, with $m_{k,j}(x, y) = 1$ indicating instance pixels.

Cross-Modal Association. During cross-modal association, novel object proposals are paired with object clusters to form cross-modal proposals. To generate object clusters, we first deploy Patchwork++ [19] to segment and remove ground points. HDBScan [3] then clusters the remaining points into object clusters. To associate 2D bounding boxes with these 3D clusters, each 2D proposal is back-projected into a corresponding 3D frustum. Using the center frustum ray, 2D proposals are paired with object clusters exhibiting strong spatial correspondence. The pairs form the set of cross-modal proposals. Additional details are available in

Sec. 6.2 of the supplementary.

Adaptive 3D Box Search. For each cross-modal proposal, an amodal 3D bounding box is generated to accurately capture the object’s position, dimensions, and orientation in 3D. To mitigate the impact of 3D annotation noise, it is crucial to explore the parameter space effectively and identify the optimal configuration for novel objects. Find n’ Propagate [10] employs an exhaustive greedy search over a frustum-defined area to estimate box parameters for novel object proposals. However, this approach inefficiently distributes a large number of search candidates to suboptimal configurations due to its expansive search space and limited precision. Our method addresses this inefficiency by reformulating the task as a continuous nonlinear optimization problem within a localized search area. Additionally, while their cost function considers only point density and multi-view alignment, our method extends it to incorporate surface alignment, further mitigating annotation errors.

Given a cross-modal proposal, comprising a 2D bounding box b_{img} and a set of object points \mathcal{P}_{obj} , the objective is to determine the 3D bounding box parameters $\theta = (x, y, z, l, w, h, r_y)$ that best represent the object. Following similar works [10, 14, 49], the search is guided by constraining the bounding box dimensions to a predefined set of box anchors $\mathbf{A} = \{(\mathbf{A}_{min}^i, \mathbf{A}_{max}^i)\}_{i=1}^C$, corresponding to each novel class. Our method relies on CLIP to classify the object into its corresponding novel class c . To reduce ambi-

guity in orientation, the yaw angle r_y is constrained to the range $[0, \pi]$. The continuous nonlinear optimization problem is then formulated in standard form:

$$\begin{aligned} \min_{\theta} \mathcal{J}(\theta, \mathcal{P}_{\text{obj}}, \mathbf{e}, b_{\text{img}}) &= \mathcal{J}_{3\text{D}}(\theta, \mathcal{P}_{\text{obj}}, \mathbf{e}) + \mathcal{J}_{2\text{D}}(\theta, b_{\text{img}}) \\ \text{subject to } \mathbf{A}_{\text{min}} &\leq (l, w, h) \leq \mathbf{A}_{\text{max}} \\ 0 &\leq r_y \leq \pi, \end{aligned} \quad (1)$$

where $\mathcal{J}_{3\text{D}}$ and $\mathcal{J}_{2\text{D}}$ are the cost functions detailed below, b_{img} is the 2D bounding box coordinates, \mathcal{P}_{obj} represents object points in 3D space, and \mathbf{e} is the ego's position in 3D space.

To evaluate the fit of the box with respect to the object points, $\mathcal{J}_{3\text{D}}$ is defined as:

$$\begin{aligned} \mathcal{J}_{3\text{D}}(\theta, \mathcal{P}_{\text{obj}}, \mathbf{e}) &= \lambda_1 \mathcal{J}_{\text{density}}(\theta, \mathcal{P}_{\text{obj}}) \\ &\quad + \lambda_2 \mathcal{J}_{\ell\text{-shape}}(\theta, \mathcal{P}_{\text{obj}}) \\ &\quad + \lambda_3 \mathcal{J}_{\text{surface}}(\theta, \mathbf{e}). \end{aligned} \quad (2)$$

The point density term $\mathcal{J}_{\text{density}}$ quantifies the proportion of the object's points that are enclosed within the 3D box. This encourages the parameters to adjust so that all points are captured by the box. $\mathcal{J}_{\text{density}}$ is defined as:

$$\mathcal{J}_{\text{density}}(\theta, \mathcal{P}_{\text{obj}}) = -\frac{\mathcal{G}_{\text{inside}}(\theta, \mathcal{P}_{\text{obj}})}{|\mathcal{P}_{\text{obj}}|}, \quad (3)$$

where

$$\mathcal{G}_{\text{inside}}(\theta, \mathcal{P}_{\text{obj}}) = \sum_{p_i \in \mathcal{P}_{\text{obj}}} \mathbb{1}(p_i \in B_{3\text{D}}(\theta)). \quad (4)$$

The ℓ -shape fitting term $\mathcal{J}_{\ell\text{-shape}}$ measures how well points within the 3D bounding box align with two anchoring 2D edges, E_1 and E_2 . Of the eight corners defining the box, the top four corners form edges, and the two closest to the ego reference point are selected. Minimizing the distance of object points p_i to the closest edge $E \in \{E_1, E_2\}$ ensures the bounding box aligns with the most relevant structural edges. In this case, $d(p_i, E)$ is the euclidean distance from p_i to E . $\mathcal{J}_{\ell\text{-shape}}$ is defined as:

$$\begin{aligned} \mathcal{J}_{\ell\text{-shape}}(\theta, \mathcal{P}_{\text{obj}}) &= \\ \sum_{p_i \in \mathcal{P}_{\text{obj}}} &\frac{\min_{E \in \{E_1, E_2\}} d(p_i, E) \cdot \mathbb{1}(p_i \in B_{3\text{D}}(\theta))}{\mathcal{G}_{\text{inside}}(\theta, \mathcal{P}_{\text{obj}})}. \end{aligned} \quad (5)$$

The surface fitting term $\mathcal{J}_{\text{surface}}$ assigns object points to the box faces nearest to the ego, aligning surfaces closer to the ego reference point \mathbf{e} . This is achieved by introducing a bias that pushes the bounding box away from the ego, clipped by a constant C_{surface} in the cost function. $\mathcal{J}_{\text{surface}}$ is defined as:

$$\mathcal{J}_{\text{surface}}(\theta, \mathbf{e}) = -\min(\|(\theta_x, \theta_y) - (\mathbf{e}_x, \mathbf{e}_y)\|_2, C_{\text{surface}}). \quad (6)$$

Finally, consistency within the image space is measured by applying calibration transformations T_{ext} and T_{int} . These transformations project the 3D bounding box θ into its 2D counterpart in the image space using $P_{3\text{D} \rightarrow 2\text{D}}$. The resulting projection is evaluated against the original 2D bounding box using the IoU metric:

$$\mathcal{J}_{2\text{D}}(\theta, b_{\text{img}}) = \gamma \cdot \text{IoU}_{2\text{D}}(P_{3\text{D} \rightarrow 2\text{D}}(\theta), b_{\text{img}}). \quad (7)$$

The optimization is governed by a cost function that balances multiple objectives. The objectives are weighted by control constants $(\lambda_1, \lambda_2, \lambda_3, \gamma)$.

Our method employs an evolutionary algorithm to iteratively refine the parameter space with increasing granularity. Box candidates are initialized within localized regions around object clusters to enhance exploration efficiency. Each scene frame is annotated, and results are compiled into a novel object bank. Duplicate 3D bounding boxes, arising from multi-view overlaps or multiple clusters for the same object, are resolved using non-maximum suppression (NMS). Additionally, ImmortalTracker [39] performs multi-object tracking to filter noisy annotations and estimate novel object velocities.

Selective Alignment. Our method motivates the idea of having only high-quality alignment pairs contribute to the alignment loss during training, while all 3D annotations remain included in the box-related losses. Prior works [4, 10] restrict alignment filtering to multi-view alignment methods, as formulated in Eq. (7). However, as illustrated in Fig. 3a, such filtering is insufficient in cases of partial occlusion or low resolution, where multi-view alignment appears nearly perfect. To address this, SC-NOD deploys additional filters to detect and exclude unreliable alignments, ensuring they do not contribute to the alignment loss during training.

To mitigate *occlusion-induced corruption*, the instance mask m of each novel object proposal is used to filter out pairs with inadequate object representation. Specifically, the proportion of pixels occupied by the object within the 2D crop is thresholded by τ_{occ} :

$$\frac{\sum_{x=1}^{W_{\text{crop}}} \sum_{y=1}^{H_{\text{crop}}} m(x, y)}{H_{\text{crop}} W_{\text{crop}}} \leq \tau_{\text{occ}}. \quad (8)$$

This process identifies and excludes occluded objects from cross-modal alignment. Different values of τ_{occ} are selected depending on the novel class, outlined in Sec. 6.4.

Next, SC-NOD filters out samples associated with image crops having an insufficient number of pixels, as these are prone to *resolution-induced* noise. Samples with a resolution below the threshold τ_{res} are marked as having insufficient semantic resolution for discriminative 2D features:

$$H_{\text{crop}} W_{\text{crop}} \leq \tau_{\text{res}}. \quad (9)$$

These strategies lead to more robust and accurate cross-modal alignment.

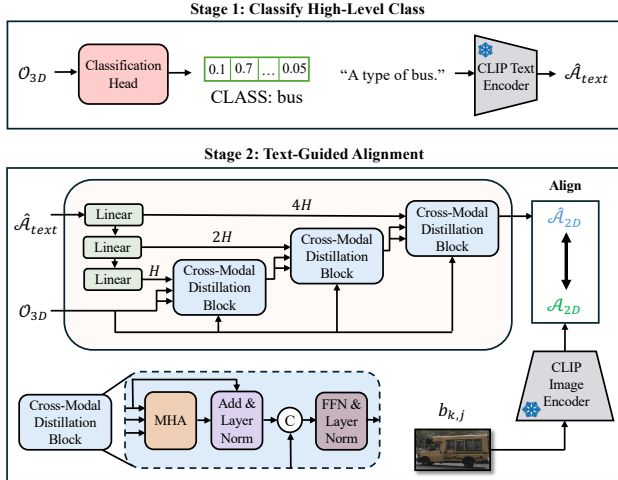


Figure 5. **Illustration of the Hierarchical Two-Stage Alignment (H2SA) Head.** H2SA first predicts the high-level novel classes, then derives class-based text prompts. H2SA then uses text prototypes to incrementally map 3D features to their 2D counterparts.

3.3. Model Architecture

Overview. The architecture of OV-SCAN comprises the off-the-shelf TransFusion-L [1] 3D detector and our Hierarchical Two-Stage Alignment (H2SA) head. TransFusion-L processes a LiDAR point cloud to generate 3D object embeddings for novel objects and decodes their bounding boxes via its regression head. Meanwhile, H2SA transforms the 3D object embeddings into predicted alignment features for prompt-based classification.

Hierarchical Two-Stage Alignment (H2SA) Head. To enable strong semantic alignment, H2SA employs a hierarchical two-stage alignment process. TransFusion-L first predicts the high-level novel class c as an auxiliary task. A high-level text prompt $\hat{\mathcal{A}}_{\text{text}} \in \mathbb{R}^D$ derived from the predicted c is then used to guide the alignment between 3D object embeddings $\mathcal{O}_{3D} \in \mathbb{R}^H$ and target 2D image embeddings $\mathcal{A}_{2D} \in \mathbb{R}^D$. The target embeddings are generated using the CLIP ViT-H/14 model, trained on DFN-5B [11], which has strong zero-shot performance on ImageNet [9] even at low resolutions. Since CLIP’s image and text encoders produce high-dimensional embeddings with $8H = D$, H2SA upscales the 3D object embeddings to match this dimension by employing three embedding scales $s \in \{H, 2H, 4H\}$, as illustrated in Fig. 5. Each scale transcribes the object’s semantics into a scale-specific text prototype, progressively incorporating details for enhanced fine-grained alignment. Lastly, we note that our lightweight alignment head seamlessly integrates with TransFusion-L without compromising its deployability or real-time inference performance.

Multi-Sensor Fusion. To show that OV-SCAN is also ef-

fective in the multi-sensor setting, we introduce OV-SCAN-Fusion, which integrates an image backbone following BEVFusion [23]. We adopt pretrained InternImage-T [40] to prevent distribution bias, since BEVFusion’s SWIN-T [22] backbone is pretrained on NuImages [2].

3.4. Training

To train OV-SCAN, our method employs the TransFusion-L loss [1]. We incorporate an additional cross-modal alignment loss term into the overall loss function, where S_c is cosine similarity.

$$\ell_{\text{align}} = \frac{\sum_{i=1}^{N_{\text{pos}}} \left(1 - S_c \left(\text{H2SA}(\mathcal{O}_{3D}^{(i)}, \hat{\mathcal{A}}_{\text{text}}^{(i)}), \mathcal{A}_{2D}^{(i)}\right)\right) \cdot \mathbb{1}_{\text{align}, i}}{\sum_{i=1}^{N_{\text{pos}}} \mathbb{1}_{\text{align}, i}} \quad (10)$$

All N_{pos} positively matched prediction-target pairs contribute to each loss term, except for the alignment loss. Only pairs with targets selected by the selective alignment filter, $\mathbb{1}_{\text{align}, i}$, contribute to the alignment loss. We adopt the same class-based data augmentation strategies as prior LiDAR-only works [51] by using novel class c . Specifically, we balance each training sample by keeping a database of object point clouds and ensuring class distribution balance through the targeted injection of objects during training.

3.5. Prompt-based Classification

To classify objects, OV-SCAN employs a prompt-based classification strategy by computing similarities between predicted alignment embeddings and text embeddings of fine-grained subclasses. Using this approach, high-level novel classes can be inferred by reverse mappings from fine-grained subclasses (e.g., a “minivan” is mapped to “car”). The fine-grained sub-class to high-novel class mappings are outlined in Sec. 6.6 from the supplementary material.

4. Experiments

4.1. Experimental Setup

Datasets. Our OV-3D object detection experiments are conducted on the nuScenes [2] and KITTI [12] datasets. We consider all classes as novel and utilize human-annotated labels solely for evaluation.

Novel Object Discovery Setup. In our experiments, novel classes (open-set classes) are based on the nuScenes and KITTI class labels. We use particle swarm optimization [16] for SC-NOD’s adaptive 3D box search, balancing the cost function with $(\lambda_1, \lambda_2, \lambda_3, \gamma) = (5.0, 1.0, 1.0, 3.0)$ and performing 150,000 iterations per cross-modal proposal. For more details on the implementation, please refer to Sec. 6.3 in the supplementary material.

Training Setup. OV-SCAN is trained on 8 NVIDIA V100 GPUs with a batch size of four for 20 epochs. OV-SCAN

Table 1. **Results on nuScenes.** We report the overall mAP, NDS, and individual class APs. All classes are novel (*i.e.*, no human annotations are used to train except in CoDA). CoDA uses “car” as a base class. Best in **bold** and “*” denotes omitted methods for fair comparison.

Method	mAP	NDS	Car	Ped.	Truck	Motorcyc.	Bicyc.	T. Cone	Bus	Barrier	Con.V.	Trailer
OV-3DET [25]	5.7	12.0	16.1	21.1	2.9	2.0	1.8	11.0	1.3	0.3	0.3	0.0
CoDA [4]	10.3	16.3	85.5	3.5	1.9	2.0	5.9	3.9	0.0	0.0	0.0	0.0
UP-VL [30]	11.3	17.2	19.4	30.8	7.9	16.8	14.1	15.6	3.1	0.5	4.3	0.4
Find n’ Prop. [10]	16.7	22.4	24.3	22.8	8.6	35.8	34.7	21.3	11.1	4.4	4.1	0.1
OpenSight [49]	22.9	23.3	25.3	52.5	11.6	30.4	26.1	25.6	5.1	42.2	8.7	0.8
OV-SCAN	31.1	32.8	61.6	60.1	30.3	39.8	31.0	39.6	22.0	18.8	6.8	0.6
OV-SCAN-Fusion*	33.8	34.4	62.0	57.6	34.2	44.7	40.2	44.6	24.4	18.9	10.8	0.8
Supervised w/ Boxes*	49.8	60.8	81.4	75.4	34.7	55.1	39.4	62.5	52.7	70.8	13.6	12.4

Table 2. **Results on KITTI.** We report the overall AP_{3D@50} for each class at medium difficulty. All classes are novel.

Method	Car	Ped.	Cyclist	mAP _{3D@50}
OV-3DET [25]	42.65	15.71	18.20	25.52
ImOV3D [42]	45.51	19.53	28.00	31.01
OV-SCAN	45.10	27.61	29.81	34.17

Table 3. **Ablations on Adaptive 3D Box Search.** We evaluate performance on nuScenes using different box search methods and varying numbers of search iterations per proposal.

Method	Iter.	mAP	mAP _n	mAP _m	mAP _f
Greedy	150,000	25.4	38.3	29.2	22.4
Adaptive	37,500	28.9	47.8	33.8	25.4
Adaptive	75,000	29.2	47.5	34.3	25.8
Adaptive	150,000	31.1	48.1	35.6	27.7

Table 4. **Ablation Studies.** Selected settings are marked in gray.

(a) 3D Box Search Cost Weights ($\lambda_1, \lambda_2, \lambda_3, \gamma$)			(b) Resolution Filter	
	mAP	Filter	mAP	
(5.0, 0.0, 0.0, 3.0)	26.1	w/o	30.4	
(1.0, 1.0, 1.0, 1.0)	27.4	$\tau_{res} = 1000$	30.8	
(5.0, 2.5, 2.5, 3.0)	27.8	$\tau_{res} = 4000$	31.1	
(5.0, 1.0, 1.0, 3.0)	31.1	$\tau_{res} = 10000$	30.6	
(c) Occlusion Filter			(d) H2SA Head	
Filter	mAP	Alignment Head	mAP	
w/o	29.4	One-Stage + FFN	25.2	
fixed ($\tau_{occ} = 0.35$)	30.7	Two-Stage + FFN	29.4	
class-based τ_{occ}	31.1	H2SA	31.1	

adopts the AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.01. The CLIP image and text encoders are kept frozen. For OV-SCAN-Fusion, we add an image backbone on top of pretrained OV-SCAN and fine-tune. OV-SCAN-Fusion is trained for five additional epochs

using a cosine annealing schedule initialized at a learning rate of 0.0001. This work is built on OpenPCDet [38], an open-source PyTorch-based 3D object detection toolbox.

Evaluation Metrics. For nuScenes, we evaluate performance using mAP and NDS, and 10 class APs. For KITTI, we compute APs for three classes using a stricter 3D IoU at a 0.5 matching threshold. Following OpenSight [49], our ablation studies also report mAP_{n/m/f} at three ranges: near (0-18m), mid (0-34m), and far (0-54m), respectively.

4.2. Main Results

NuScenes. The OV-3D object detection results for top-performing methods on the nuScenes dataset can be seen in Tab. 1. For a fair comparison to the other methods, which are all LiDAR-based, we only consider OV-SCAN and not OV-SCAN-Fusion. Our LiDAR-based method surpasses previous benchmarks by achieving the best mAP and NDS, thus surpassing OpenSight by 8.2 mAP and 9.5 NDS. Without being given 3D human-annotations, OV-SCAN achieves an AP score above 60 for both car and pedestrian categories. We also note that our method exceeds previous benchmarks by more than 5 AP for 6 out of the 10 classes. Furthermore, we show that simply adding camera as an additional input modality to OV-SCAN and then fine-tuning can improve the overall performance. Finally, as an upper bound, we report the performance of our method using ground truth boxes. CLIP is still used to derive high-level class labels and align. **KITTI.** In Tab. 2, we present our results on KITTI, demonstrating the applicability of our method across multiple datasets. OV-SCAN outperforms OV-3DET [25] and ImOV3D [42] in the overall metric, achieving comparable results to ImOV3D [42] in the car category while surpassing both in the other two classes.

Novel Object Discovery. SC-NOD generates 319,028 3D annotations for training, a fraction of the 797,179 available in the nuScenes dataset. While all annotations contribute to the box loss, only 171,532 (54%) of those generated are utilized for cross-modal alignment. The remainder of generated annotations are excluded as a result of filtering due to significant occlusion (39%) or insufficient resolution (7%).

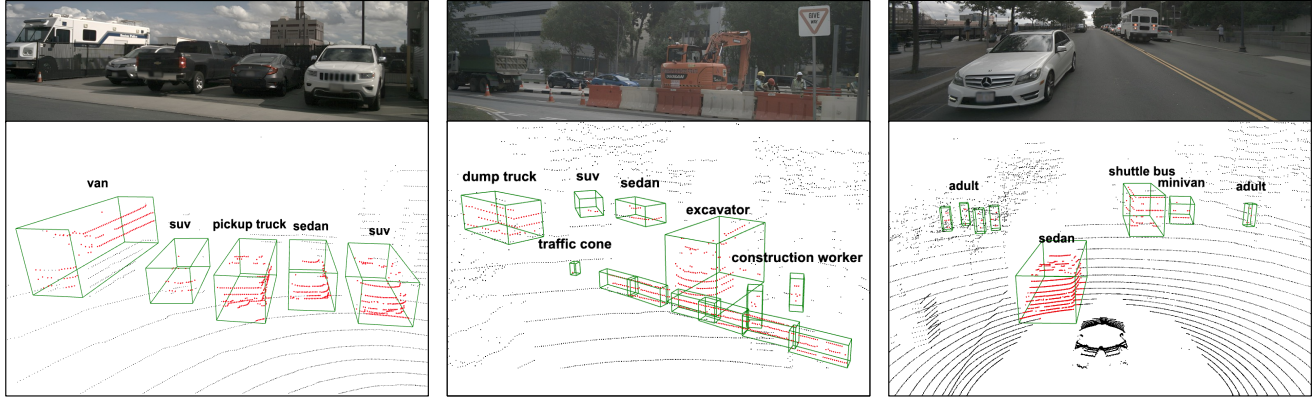


Figure 6. **Visualization results using Open-Vocabulary.** OV-SCAN performs inference on a set of urban scenes identifying a diverse set of objects. To demonstrate OV capabilities, objects are classified into more fine-grained classes as opposed to traditional closed-set classes.

As shown in Fig. 7b, the relative annotation count of SC-NOD labels exhibits a comparable distribution.

4.3. Ablation Studies

Adaptive 3D Box Search. To assess the effectiveness of our adaptive 3D box search in SC-NOD, we evaluate its performance on the nuScenes dataset, comparing it to the greedy search approach used in Find N’ Propagate [10]. Tab. 3 shows that our method consistently outperforms the greedy search even with fewer iterations per novel object. We further ablate various parameterizations of the cost function during the adaptive 3D box search, as seen in Tab. 4a, highlighting the importance of balancing the heuristic cost terms for effective 3D box search.

Selective Alignment. Tab. 4b and Tab. 4c demonstrates the effectiveness of filtering techniques used in selective cross-modal alignment. A simple occlusion filter with a fixed threshold τ_{occ} yields a notable performance gain, while class-based thresholds achieve the highest improvement (+1.7 mAP). For the resolution filter, optimizing the threshold to balance resolution quality and the number of filtered samples proves most effective (+0.7 mAP).

Alignment Head. To assess the impact of the H2SA head, we introduce a one-step baseline in Tab. 4d for comparison. This variant removes the classification loss term, merges TransFusion-L’s class heatmaps into a single class-agnostic heatmap, and replaces the text-guided alignment network with a simple feed-forward network. This simplification results in a performance drop (-5.9 mAP). Furthermore, incorporating class-based text prompts to guide cross-modal alignment further enhances performance (+1.7 mAP).

4.4. Limitations

The primary limitation of SC-NOD is its limited annotation recovery (Fig. 7a), due to reliance on 2D proposals. Objects clearly visible in the point cloud may remain un-

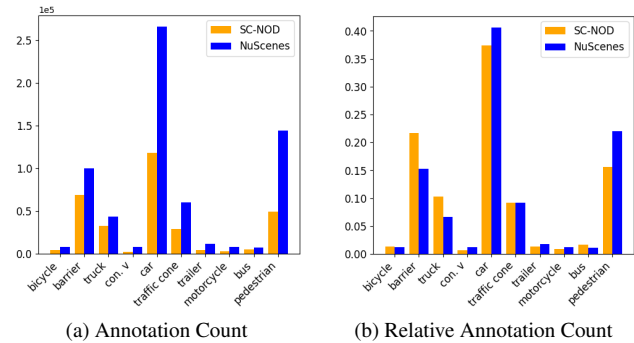


Figure 7. **Distribution of training labels.**

labeled if absent in multi-view images, resulting in unlabeled objects. Additionally, labeling sparse objects at long distances remains challenging, significantly reducing performance at greater ranges (Tab. 3). Furthermore, for 3D box search, SC-NOD adopts class-based anchors following prior works. While suitable for novel classes with low dimensional variance, this method struggles with highly shape-variable objects such as trailers, buses, and construction vehicles. These insights motivate future work exploring alternative methods less dependent on 2D proposals and anchor-free box-parameterization strategies.

5. Conclusion

We introduce OV-SCAN, an OV-3D detector that achieve detection through enhanced cross-modal alignment. Without human-provided 3D annotations, SC-NOD accurately generates 3D boxes and also carefully guides cross-modal alignment. By adapting H2SA, we further strengthen alignment and enable robust open-set classification. Experiments on nuScenes demonstrate state-of-the-art OV-3D detection.

Acknowledgment

This work was supported by the NSERC-CSE Research Community project entitled “An End-to-End Approach to Safe and Secure AI Systems.” Researchers funded through the NSERC-CSE Research Communities Grants do not represent the Communications Security Establishment Canada or the Government of Canada. Any research, opinions or positions they produce as part of this initiative do not represent the official views of the Government of Canada.

References

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1080–1089. IEEE, 2022. 3, 6, 4
- [2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multi-modal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6, 1
- [3] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II*, pages 160–172. Springer, 2013. 4
- [4] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. CoDA: Collaborative Novel Box Discovery and Cross-modal Alignment for Open-vocabulary 3D Object Detection. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 2, 3, 5, 7
- [5] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3d object detection, 2024. 1
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. 3
- [7] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. VoxNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21674–21683. IEEE, 2023. 3
- [8] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16901–16911. IEEE, 2024. 1, 3
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 6
- [10] Djamahl Etchegaray, Zi Huang, Tatsuya Harada, and Yadan Luo. Find n’ Propagate: Open-Vocabulary 3D Object Detection in Urban Environments. In *Computer Vision – ECCV 2024*, pages 133–151, Cham, 2025. Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science. 1, 2, 3, 4, 5, 7, 8
- [11] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T. Toshev, and Vaishaal Shankar. Data Filtering Networks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 6
- [12] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 6, 1
- [13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3
- [14] Rui Huang, Henry Zheng, Yan Wang, Zhuofan Xia, Marco Pavone, and Gao Huang. Training an open-vocabulary monocular 3d detection model without 3d data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 4
- [15] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. <https://github.com/ultralytics/ultralytics>, 2023. 3
- [16] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN’95 - International Conference on Neural Networks*, pages 1942–1948 vol.4, 1995. 6, 1
- [17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment Anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023. 4, 1
- [18] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection From Point Clouds. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12697–12705. Computer Vision Foundation / IEEE, 2019. 3
- [19] Seungjae Lee, Hyungtae Lim, and Hyun Myung. Patchwork++: Fast and Robust Ground Segmentation Solving Partial Under-Segmentation Using 3D Point Cloud. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022*, pages 13276–13283. IEEE, 2022. 4

- [20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded Language-Image Pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965. IEEE, 2022. 1, 3
- [21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv:2303.05499, 2024. 1, 3, 4
- [22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 2021. 6
- [23] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird’s-Eye View Representation. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 2774–2781. IEEE, 2023. 3, 6
- [24] Zhijian Liu, Xinyu Yang, Haotian Tang, Shang Yang, and Song Han. FlatFormer: Flattened Window Attention for Efficient Point Cloud Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 1200–1211. IEEE, 2023. 3
- [25] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-Vocabulary Point-Cloud Object Detection without 3D Annotation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 1190–1199. IEEE, 2023. 1, 2, 3, 7
- [26] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel Transformer for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3144–3153. IEEE, 2021. 3
- [27] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *European Conference on Computer Vision (ECCV)*, pages 728–755, Cham, 2022. Springer Nature Switzerland. 1, 3
- [28] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1
- [29] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2886–2897. IEEE, 2021. 3
- [30] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R. Qi, Xinchun Yan, Scott Ettinger, and Dragomir Anguelov. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 8568–8578. IEEE, 2023. 7, 1
- [31] Yaqian Ning, Jie Cao, Chun Bao, and Qun Hao. DVST: Deformable Voxel Set Transformer for 3D Object Detection from Point Clouds. *Remote. Sens.*, 15(23):5612, 2023. 3
- [32] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85. IEEE Computer Society, 2017. 3
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 1
- [34] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 770–779. Computer Vision Foundation / IEEE, 2019. 3
- [35] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [36] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN++: Point-Voxel Feature Set Abstraction With Local Vector Representation for 3D Object Detection. *Int. J. Comput. Vis.*, 131(2):531–551, 2023. 3
- [37] Pei Sun, Mingxing Tan, Weiyue Wang, Chenxi Liu, Fei Xia, Zhaoqi Leng, and Dragomir Anguelov. SWFormer: Sparse Window Transformer for 3D Object Detection in Point Clouds. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X*, pages 426–442. Springer, 2022. 3
- [38] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. <https://github.com/open-mmlab/OpenPCDet>, 2020. 7
- [39] Qitai Wang, Yuntao Chen, Ziqi Pang, Naiyan Wang, and Zhaoxiang Zhang. Immortal Tracker: Tracklet never dies. arXiv:2111.13672, 2021. 5
- [40] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022. 6

- [41] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 18(10):3337, 2018. 3
- [42] Timing Yang, Yuanliang Ju, and Li Yi. Imov3d: Learning open-vocabulary point clouds 3d object detection from only 2d images. *NeurIPS 2024*, 2024. 3, 7
- [43] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. DetCLIP: Dictionary-Enriched Visual-Concept Paralleled Pre-training for Open-world Detection. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 1
- [44] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. IS-Fusion: Instance-Scene Collaborative Fusion for Multimodal 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14905–14915. IEEE, 2024. 3
- [45] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-Based 3D Object Detection and Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11784–11793. Computer Vision Foundation / IEEE, 2021. 3
- [46] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-Vocabulary DETR with Conditional Matching. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, pages 106–122. Springer, 2022. 3
- [47] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-Vocabulary Object Detection Using Captions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14393–14402. Computer Vision Foundation / IEEE, 2021. 3
- [48] Dongmei Zhang, Chang Li, Renrui Zhang, Shenghao Xie, Wei Xue, Xiaodong Xie, and Shanghang Zhang. FM-OV3D: Foundation Model-Based Cross-Modal Knowledge Blending for Open-Vocabulary 3D Detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 16723–16731. AAAI Press, 2024. 2, 3, 4
- [49] Hu Zhang, Jianhua Xu, Tao Tang, Haiyang Sun, Xin Yu, Zi Huang, and Kaicheng Yu. OpenSight: A Simple Open-Vocabulary Framework for LiDAR-Based Object Detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIV*, pages 1–19. Springer, 2024. 1, 2, 3, 4, 7
- [50] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4490–4499. Computer Vision Foundation / IEEE Computer Society, 2018. 3
- [51] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced Grouping and Sampling for Point Cloud 3D Object Detection. *CoRR*, abs/1908.09492, 2019. arXiv: 1908.09492. 6