OV-Uni3DETR: Towards Unified Open-Vocabulary 3D Object Detection via Cycle-Modality Propagation

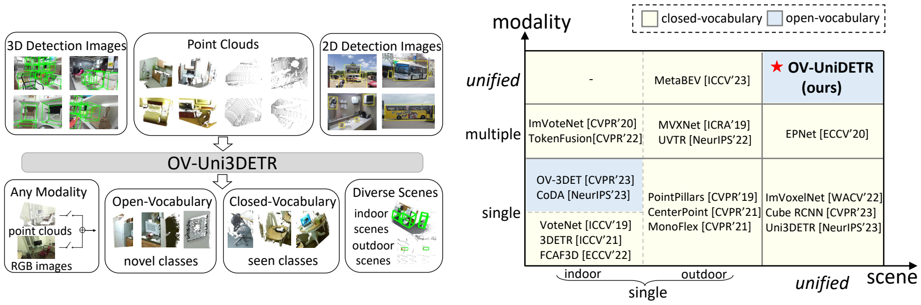
Zhenyu Wang[1], Yali Li[1], Taichi Liu[2], Hengshuang Zhao[3], and Shengjin Wang[1]

[1] Department of Electronic Engineering, Tsinghua University
[2] Rutgers University
[3] The University of Hong Kong
wangzy20@mails.tsinghua.edu.cn {liyali13, wgsgj}@tsinghua.edu.cn
tl821@rutgers.edu hszhao@cs.hku.hk

Abstract. In the current state of 3D object detection research, the severe scarcity of annotated 3D data, substantial disparities across different data modalities, and the absence of a unified architecture, have impeded the progress towards the goal of universality. In this paper, we propose **OV-Uni3DETR**, a unified open-vocabulary 3D detector via cycle-modality propagation. Compared with existing 3D detectors, OV-Uni3DETR offers distinct advantages: 1) Open-vocabulary 3D detection: During training, it leverages various accessible data, especially extensive 2D detection images, to boost training diversity. During inference, it can detect both seen and unseen classes. 2) Modality unifying: It seamlessly accommodates input data from any given modality, effectively addressing scenarios involving disparate modalities or missing sensor information, thereby supporting test-time modality switching. 3) Scene unifying: It provides a unified multi-modal model architecture for diverse scenes collected by distinct sensors. Specifically, we propose the cycle-modality propagation, aimed at propagating knowledge bridging 2D and 3D modalities, to support the aforementioned functionalities. 2D semantic knowledge from large-vocabulary learning guides novel class discovery in the 3D domain, and 3D geometric knowledge provides localization supervision for 2D detection images. OV-Uni3DETR achieves the state-of-the-art performance on various scenarios, surpassing existing methods by more than 6% on average. Its performance using only RGB images is on par with or even surpasses that of previous point cloud based methods. Code and pre-trained models will be released later.

1 Introduction

3D object detection aims to predict the oriented 3D bounding boxes and the semantic category tags for the real scenes given point clouds or RGB images. Recently, 2D object detection for universality has developed rapidly [22,28,59,74]. These methods leverage various forms of available data for training, enhancing the detector's universality, and enabling it to detect any category in any scene.



(a) What can OV-Uni3DETR do? **(b)** How does OV-Uni3DETR distinguish?

Fig. 1: Illustration for OV-Uni3DETR. (a): It utilizes various available data for training, including 3D point clouds, 3D detection images (with 3D box annotated and aligning with point clouds) and 2D detection images (only 2D box annotated). For inference, it can predict 3D boxes using any modality data, for both open-vocabulary and closed-vocabulary, both indoor and outdoor 3D detection. (b): Compared with existing 3D detectors, OV-Uni3DETR achieves modality unifying (modality-switchable during inference), scene unifying, and open-vocabulary learning simultaneously.

However, related research on 3D universal detection remains significantly lagging behind that of 2D. Currently, most of 3D object detection methods still rely on fully-supervised learning, restricted by data fully annotated for specific input modalities, and can only recognize categories that appear during training for either indoor or outdoor scenes.

The problem of 3D universal object detection is challenging because of the following reasons. First, existing 3D detectors work in a closed-vocabulary manner, thus can only detect seen classes. *Open-vocabulary 3D object detection* is urgently demanded to recognize and localize object instances with novel categories that are not acquired during training. However, the generalization ability in localizing novel objects is restricted by existing 3D detection datasets [9,13,53], where both size and categories are limited compared to those 2D ones [15,20,27,47]. Additionally, the lack of pre-trained image-text models [18,40,68] in the 3D domain further exacerbates the challenges associated with open-vocabulary 3D detection. Second, a *unified architecture for multi-modal 3D detection* is absent. Existing 3D detectors are predominantly designed for specific input modalities (either point clouds, RGB images, or both) and scenes (either indoor or outdoor ones). For open-vocabulary 3D detection, especially under the constraints posed by limited data, the lack of a unified multi-modal structure prevents the effective utilization of data from various modalities and sources. Consequently, the detector cannot generalize to novel objects effectively. Besides, without a unified architecture, 3D detectors struggle to adapt to different input modalities or handle situations where sensor modalities are missing. A unified solution to address the open-vocabulary challenge in a multi-modal context, therefore, is a crucial step for current research in 3D detection.

We propose **OV-Uni3DETR**, a unified multi-modal 3D detector for open-vocabulary 3D object detection, as is in Fig. 1a. During training, it is endowed with the capability to leverage multi-modal and multi-source data, including point clouds, 3D detection images with precise 3D box annotations and aligning with point clouds, and 2D detection images with only 2D box annotations. A critical enhancement is the integration of 2D detection images, which is particularly advantageous for open-vocabulary 3D detection due to the significantly greater number of annotated classes. With these multi-modal data, we further adopt a switched-modality training scheme. Benefiting from the above multi-modal learning manner, OV-Uni3DETR accommodates data from any modality during inference, thus achieving the function of test-time modality switching. It excels in detecting both base classes and novel classes. The unified structure further equips OV-Uni3DETR with the ability to detect in both indoor and outdoor scenes. As can be seen in Fig. 1b, OV-Uni3DETR achieves scene and modality unifying and possesses the open-vocabulary capability, thus greatly advancing the universality of 3D detectors across categories, scenes, and modalities.

With multi-modal learning, two challenging problems should be further addressed. The first is about *how to generalize the detector to novel classes*, and the second is about *how to learn from the extensive 2D detection images without 3D box annotations*. We propose the approach of cycle-modality propagation - knowledge propagation between 2D and 3D modalities to address the two challenges. For 2D to 3D propagation, we extract 2D bounding boxes with a 2D open-vocabulary detector, and project them into the point cloud space to approximate 3D boxes. In this way, the abundant semantic knowledge from the 2D detector can be propagated to the 3D domain to assist novel box discovery. For 3D to 2D, we leverage the geometric knowledge from a class-agnostic 3D detector to localize objects within 2D detection images, and assign category tags by Hungarian matching. Such geometric knowledge can compensate for the absence of 3D supervision information in 2D detection images.

Our main contributions can be summarized as follows:

- We propose OV-Uni3DETR, a unified open-vocabulary 3D detector with multi-modal learning. It excels in detecting objects of any class across various modalities and diverse scenes, thus greatly advancing existing research towards the goal of universal 3D object detection.
- We present a unified multi-modal architecture for both indoor and outdoor scenes. By eliminating modality inconsistencies and switched-modality training, it becomes test-time modality-switchable: utilizing any modality data.
- We propose the concept of a knowledge propagation cycle between 2D and 3D modalities. By leveraging 2D large-vocabulary semantic knowledge and precise 3D geometric knowledge, the training diversity can be guaranteed.

Extensive experiments demonstrate the strong ability of OV-Uni3DETR. It achieves state-of-the-art performance on various tasks of 3D detection. In the open-vocabulary setting, it surpasses previous methods by more than 7% on SUN RGB-D [53] and 8% on ScanNet [9]. For traditional close-vocabulary experiments, OV-Uni3DETR is also more than 3% higher than previous methods. With universality in modality, scene and category, we believe OV-Uni3DETR can become a significant step towards the future of 3D foundation models.

2 Related Work

3D Object Detection aims to predict category tags and oriented 3D bounding boxes for the scene. Some works take point cloud data as input. Because of the significant distinction in point cloud data, these models are usually separated into indoor [29,39,44,55,72] and outdoor [48–50,64,65] scenarios. A unified structure [60] for both indoor and outdoor point clouds is presented recently. RGB-based 3D detectors [2, 24, 30, 45, 57, 71] utilize only RGB images for 3D bounding box prediction. Constrained by the limited spatial information in RGB images, the performance of these detectors significantly lags behind their counterparts using point clouds. In comparison, multi-modal 3D detectors [12, 25, 38, 52, 58] utilize both point cloud data and RGB images. With multi-modal data, these detectors make better performance. However, these models can only work in the closed-vocabulary setting, thus are restricted from the limited scale of 3D detection data. A unified structure for all modalities and scenes is also absent currently.

Open-Vocabulary Object Detection aims to recognize and localize novel classes that are not annotated in datasets. Benefiting from large-scale image-text pre-training models [18, 40, 68], 2D open-vocabulary detection researches have been forwarded significantly. [11, 14, 35, 36, 59, 66] adopt such pre-trained parameters and detect novel classes with the help of text features. [1, 74] involve image-level supervision to help expand the vocabulary of the detectors. [22, 28, 69] unify object detection and visual grounding for pre-training and adopt text queries for detection. These methods focus on 2D detection. In the 3D field, restricted by the limited scare of 3D data, there are still no pre-trained point-text models, which makes 3D open-vocabulary detection a challenging problem.

Open-Vocabulary 3D Object Detection targets at novel class recognition and localization for 3D bounding boxes. Some works [10, 63, 67, 70] borrow ideas from 2D image-text pre-training and utilize text embeddings for 3D novel classes. These works are mainly about classification and semantic segmentation, and cannot be adopted in 3D detection. Recently, [6, 33] have forwarded open-set 3D detection, [34] conducts open-vocabulary 3D detection with the help of a pre-trained 2D detector, and [4] recognizes novel classes by novel object discovery and cross-modal alignment. However, these models only utilize point cloud data for inference, and can only work in indoor scenes. In comparison, we design a unified open-vocabulary detector for multi-modal data and different scenes.

3 OV-Uni3DETR

We present the overview of OV-Uni3DETR in Fig. 2. It takes both point clouds and RGB images during training. Once trained, it can be switchable and use either of them for inference. We propagate semantic knowledge from 2D to 3D for novel classes to fulfill open-vocabulary 3D detection learning. Extra 2D detection images are further introduced to promote novel class recognition, with a class-agnostic 3D detector propagating geometric knowledge from 3D to 2D.

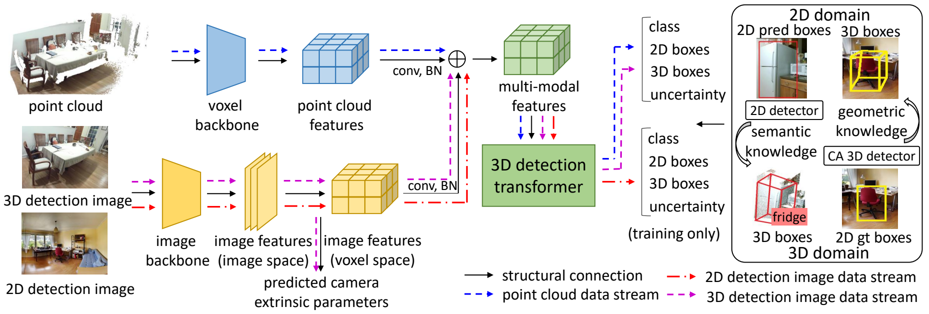


Fig. 2: Overview of OV-Uni3DETR. We extract features for point clouds and images. After converted into the same voxel space, they are added for the multi-modal features. The 3D detection transformer is finally utilized for class and box prediction. We perform semantic knowledge propagation from 2D to 3D for novel class discovery. To use 2D detection images, we predict the camera extrinsic parameters and propagate geometric knowledge from 3D to 2D through a class-agnostic (CA) 3D detector.

3.1 Multi-Modal Learning

Our multi-modal architecture takes point cloud X_P and images X_I for training, and can handle situations when missing sensor modalities exists for inference, *i.e.*, test-time modality-switchable. We first extract 3D point cloud features $F_P \in \mathbb{R}^{C\times X\times Y\times Z}$ with the voxel-based backbone, and 2D image features $F_I \in \mathbb{R}^{C\times H\times W}$ with the image backbone, where (H, W) denotes the size of extracted image features. F_I is then projected into the 3D voxel space for the image features in the voxel space $F'_I \in \mathbb{R}^{C\times X\times Y\times Z}$ through the camera parameters. Specifically, denote the camera intrinsic matrix as K and the extrinsic matrix as R_t, then the corresponding positions in the 2D image can be obtained by projecting 3D positions in the voxel space through KR_t.

Compared with 2D images, 3D point cloud data usually consist of more spatially rich information, which is crucial for 3D detection. As a result, the trained multi-modal detector is easy to strongly rely on point cloud features for detection, while simply treating image features as auxiliary information. Under this manner, if point cloud data are absent at the inference time, the performance will deteriorate seriously. To avoid this problem, we first utilize a 3D convolution layer with batch normalization. The multi-modal features are thus obtained by: $F_M = \mathrm{BN}(\mathrm{conv}(F_P)) + \mathrm{BN}(\mathrm{conv}(F'_I))$. With the 3D convolution and BN layers, different modality features are regularized. This prevents feature discrepancy and the suppression of image features. We then adopt the switched-modality training scheme to further avoid image features collapsing. Specifically, the 3D detection transformer randomly receives features from aforementioned modalities - F_M, F_P, F'_I, with pre-determined probabilities, which makes it possible for the model to detect with only single-modal data. By random switching, the model accepts image-only features during training. This also prevents the image features from being ignored. We directly use the 3D transformer in [60] for detection.

The multi-modal architecture finally predicts the category tags, the 4-dim 2D boxes and 7-dim 3D boxes for both 2D and 3D object detection. The L1 loss and decoupled IoU loss [60] are utilized for 3D box regression, and the L1 loss and GIoU loss [42] are utilized for 2D box regression. In the open-vocabulary setting, novel class samples exist, which increases the difficulty of training samples. We therefore introduce an uncertainty prediction μ following [2,32], and utilize it to weight the L1 regression loss. The loss for object detection learning is as:

$$L = L_{cls} + \sqrt{2} \cdot \exp(-\mu) \cdot (L_1^{3D} + L_1^{2D}) + L_{IoU}^{3D} + L_{IoU}^{2D} + \mu \tag{1}$$

For some 3D scenes, there may exist multi-view images, rather than the single monocular one. We extract image features for each of them and project to the voxel space using their own projection matrices. The multiple image features in the voxel space are summed for the multi-modal features.

3.2 Knowledge Propagation: 2D → 3D

We perform open-vocabulary 3D detection based on multi-modal learning introduced before. The core issue of open-vocabulary learning is recognizing novel classes that are not human-annotated during training. Due to the difficulty of acquiring point cloud data, the pre-trained vision-language models [18, 40, 68] do not exist in the point cloud field. The modality difference between point cloud data and RGB images restricts the performance of these models in 3D detection. We propose to leverage semantic knowledge from a pre-trained 2D open-vocabulary detector, and generate the corresponding 3D bounding boxes for novel classes. The generated 3D boxes will supplement 3D ground-truth labels with limited classes available at the training time.

Specifically, we first generate 2D bounding boxes or instance masks with a 2D open-vocabulary detector. Considering that the available data and annotations are much more abundant in the 2D field, these generated 2D boxes can achieve higher localization accuracy and cover a significantly wider range of categories. We then project these 2D boxes to the 3D space through $(KR_t)^{-1}$ to obtain the corresponding 3D boxes. The specific operation is that we project 3D points into the 2D space using KR_t, finding points within the 2D boxes, then clustering these points inside the 2D boxes for eliminating outliers to obtain the corresponding 3D boxes. Benefiting from the pre-trained 2D detector, novel objects that are not annotated can be available in the generated 3D box set. In this way, the rich semantic knowledge deriving from large-vocabulary 2D detection learning can be propagated from the 2D domain to the generated 3D boxes, thus greatly boosting 3D open-vocabulary detection. For multi-view images, 3D boxes are generated separately and ensembled together for the final usage.

During inference, when both point clouds and images are available, we can extract 3D boxes in the similar way. These generated 3D boxes can also be viewed as a form of 3D open-vocabulary detection results. We add these 3D boxes to the predictions of the multi-modal 3D transformer to supplement the potentially missing objects, and filter the overlapped bounding boxes through

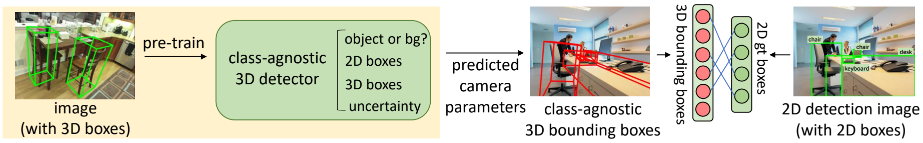


image
(with 3D boxes)
pre-train
class-agnostic 3D detector
object or bg?
2D boxes
3D boxes
uncertainty
predicted camera parameters
class-agnostic 3D bounding boxes
3D bounding boxes
2D gt boxes
chair
chair
desk
keyboard
2D detection image (with 2D boxes)

Fig. 3: Illustration of knowledge propagation from 3D to 2D. 3D detection images first train a class-agnostic 3D detector, which is then used to generate class-agnostic 3D bounding boxes with the predicted camera parameters for the 2D detection images. Hungarian matching is finally conducted between 2D boxes and 3D ones for the class-specific 3D bounding boxes.

3D NMS. The confidence scores assigned by the pre-trained 2D detector are systematically divided by a predetermined constant, and then reinterpreted as the confidence scores of the corresponding 3D boxes.

3.3 Knowledge Propagation: 3D → 2D

Compared with 3D object detection data with 3D bounding boxes annotated, 2D detection images with just 2D bounding boxes annotated are much more abundant, with significantly more scenes, objects and categories included. Motivated by this, we introduce these 2D detection images for training to boost open-vocabulary 3D detection, especially in novel class recognition. However, 3D bounding boxes are not annotated for these 2D detection images. Learning totally without 3D box annotations will have a negligible impact on 3D box prediction training, making 2D detection images hard to be fully utilized. Besides, the camera parameters are absent, making it infeasible to transform image features to the voxel space. We propose to propagate geometric knowledge from 3D to 2D through a class-agnostic 3D detector. The 3D bounding boxes and camera parameters will be predicted for these 2D detection images.

Camera parameter prediction. The camera parameters mainly include the camera intrinsic matrix K and the extrinsic matrix R_t. The intrinsic matrix K is mainly about the the camera's focal lengths (f_x, f_y) and the principal point (p_x, p_y). Denote the image resolution as (h, w), we simply treat $f_x = f_y = h$, $p_x = \frac{1}{2}w$, $p_y = \frac{1}{2}h$. The intrinsic matrix is thus estimated as: $K = \begin{bmatrix} h & 0 & \frac{1}{2}w \\ 0 & h & \frac{1}{2}h \\ 0 & 0 & 1 \end{bmatrix}$.

The camera extrinsic matrix $R_t \in \mathbb{R}^{3\times4}$ denotes the transformation from the 3D world coordinates to 3D camera coordinates. It can be written in $R_t = [R|T]$, where $R \in \mathbb{R}^{3\times3}$ is the rotation matrix and $T \in \mathbb{R}^{3\times1}$ is the position of the origin of the world coordinate system. The rotation matrix R is about rotating the angle θ around the axis $u = (u_x, u_y, u_x)$, and T is specifically expressed in $T = [t_x, t_y, t_z]^T$. In total, there are 7 parameters about the extrinsic matrix R_t to be estimated. We add a 8-dim branch on the image features in the image space to predict these parameters, specifically, $[\sin\theta, \cos\theta, u_x, u_y, u_x, t_x, t_y, t_z]$. We use the L1 loss for regression to learn these parameters. This camera parameter prediction module is pre-trained on 3D detection images, then utilized to predict missing camera parameters for 2D detection images.

Generating 3D bounding boxes. For 2D detection images, 2D annotated bounding boxes $\{c_i, bb_i^{2D}\}_{i=1}^{M}$ are available for object instances, where c_i and bb_i^{2D} are the category tag and 2D bounding box of the i-th object, M is the number of 2D boxes. We aim to generate 3D bounding boxes for these objects.

We first pre-train a class-agnostic 3D detector on 3D detection images with only seen class objects. Here we discard the point cloud branch for the class-agnostic 3D detector. The semantic category tags are directly removed for class-agnostic learning. Then, this pre-trained detector conducts inference on 2D detection images to obtain the predicted class-agnostic boxes $\{\hat{bb}_i^{2D}, \hat{bb}_i^{3D}\}_{i=1}^{N}$, where N is the box number, using the predicted camera parameters from the pre-trained camera parameter prediction module. Since class-agnostic classification has the generalization ability for novel classes [14,19,46,59], the predicted class-agnostic 3D boxes can include novel class objects, even only base classes participating in training. With the class-agnostic 3D detector, geometric knowledge from the 3D domain can be propagated to 2D detection images, and provide 3D localization information for them. As a result, the issue of the lack of 3D supervision information can be addressed.

We then need to assign category tags to the class-agnostic boxes. We formulate this problem as finding a bipartite matching between the ground-truth set and the extracted 3D box set. We propose that the optimal bipartite matching should minimize the overlaps between the ground-truth 2D boxes and extracted class-agnostic 3D boxes. We calculate the IoU between bb_i^{2D} and predicted 2D class-agnostic boxes $\hat{bb}_i^{2D}$, and the bipartite matching problem is written as:

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i}^{N} \mathrm{IoU}(bb_i^{2D}, \hat{bb}_{\sigma(i)}^{2D}) \tag{2}$$

We utilize the Hungarian algorithm to solve it [5]. After bipartite matching, the category tags can be assigned to the 3D boxes to obtain the class-specific 3D bounding boxes $\{c_i, bb_i^{2D}, \hat{bb}_{\sigma(i)}^{3D}\}_{i=1}^{M}$. They are treated as the ground-truth labels of these 2D detection images for training. With the generated class-specific 3D boxes, these 2D detection images can be used in the same way as 3D detection images. However, there exists noise in 3D boxes $\hat{bb}_i^{3D}$, which will hurt the 3D box prediction performance. Therefore, we design a dual-branch structure for the final output layers of the OV-Uni3DETR. It specifically refers to two distinct parameter sets of output layers with the same structure and loss. 3D detection images and point cloud data pass into one branch for classification and regression, and 2D detection images use another one. In this way, the noisy 3D boxes will not interfere with the accurate ones. At the inference time, the branch for 2D detection images will be discarded for accurate inference.

4 Experiments

We conduct extensive experiments under various conditions for open-vocabulary and traditional closed-vocabulary settings to demonstrate the strong detection ability of our OV-Uni3DETR in this section.

Datasets. For indoor 3D detection, we adopt two challenging datasets, SUN RGB-D [53] and ScanNet V2 [9]. We mainly follow the setting of [4] for open-vocabulary experiments. SUN RGB-D contains 5,285 training and 5,050 validation scenes, with 46 classes in total and oriented 3D bounding boxes annotated. Each scene consists of one single-view RGB image. The categories with the top 10 most training samples are selected as base (seen) categories, while the rest 36 are novel classes. ScanNet V2 contains 1,201 reconstructed training scans and 312 validation scans, with 200 object categories for axis-aligned bounding boxes in total [43]. We adopt the same setting as [4] for training and inference, where point clouds are generated from the single-view depth images. The top 10 classes are used for base classes, and the other 50 ones are novel classes. We also conduct experiments on the ScanNet setting from [34], which includes 20 classes as novel ones, no seen classes, for further comparison. For 2D detection images, we randomly select 6,000 images from the Objects365 [47] dataset. We mainly use the mean average precision (mAP) under IoU thresholds of 0.25.

For outdoor 3D detection, we mainly conduct experiments on the KITTI [13] and nuScenes [3] dataset. For KITTI, we split its official training set into 3,712 training samples and 3,769 validation samples for training and evaluation. Each scene contains one monocular image. We treat the car and cyclist classes as base classes, and the pedestrian category for the novel one. For nuScenes, we train on the 28,130 frames of samples from the training set and evaluate on the 6,010 validation samples. Each scene contains 6 images with different directions here. The classes of "car, trailer, construction vehicle, motorcycle, bicycle" are treated as seen and the rest five are unseen ones.

Implementation details. We implement OV-Uni3DETR mainly with mmdetection3D [8], and train it with the AdamW [31] optimizer. The point cloud branch is the same as [60], and we use ResNet50 [16] and FPN [26] for the image feature extractor without specially mentioned. For main experiments we generate 3D bounding boxes of the 365 classes of the Objects365 dataset for training. We use a pre-trained Detic [74] for the 2D open-vocabulary detector. All generated labels are filtered by the 0.4 confidence threshold.

4.1 Open-Vocabulary 3D Object Detection

Indoor 3D open-vocabulary detection. We evaluate OV-Uni3DETR on the indoor SUN RGB-D dataset for the 46 class setting, and list the AP_{25} metric in Tab. 1. OV-Uni3DETR obtains 9.66% AP_{25} for the 36 novel classes with point clouds only during inference, which surpasses CoDA [4] by 2.95% with the same used data. This demonstrates that our method recognizes novel classes well with the knowledge propagation cycle. Meanwhile, AP_{base} is even 9.57% higher, which

Table 1: The performance of OV-Uni3DETR on the SUN RGB-D and ScanNet dataset for open-vocabulary 3D object detection. P denotes the point cloud inputs and I denotes the image inputs. The experimental setting is totally the same as CoDA, and the utilized data are downloaded from CoDA officially released code.

Method	Inputs	SUN RGB-D			ScanNet		
		AP_{novel}	AP_{base}	AP_{all}	AP_{novel}	AP_{base}	AP_{all}
Det-PointCLIP [70]	P	0.09	5.04	1.17	0.13	2.38	0.50
Det-PointCLIPv2 [76]	P	0.12	4.82	1.14	0.13	1.75	0.40
Det-CLIP^2 [67]	P	0.88	22.74	5.63	0.14	1.76	0.40
3D-CLIP [40]	P+I	3.61	30.56	9.47	3.74	14.14	5.47
CoDA [4]	P	6.71	38.72	13.66	6.54	21.57	9.04
OV-Uni3DETR (ours)	P	9.66	48.29	18.06	12.09	30.47	15.15
	I	5.41	29.51	10.65	8.10	20.87	10.23
	P+I	**12.96**	**49.25**	**20.85**	**15.21**	**31.86**	**17.99**

Table 2: Comparison with methods in the same setting of [34] on ScanNet. Mean represents the average value of all 20 categories. OV-Uni3DETR is evaluated with the same setting, with only point clouds utilized for both training and inference.

Methods	**Mean**	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink
OV-3DET [34]	18.02	57.29	42.26	27.06	31.50	8.21	14.17	2.98	5.56	23.00	31.60
CoDA [4]	19.32	68.09	44.04	28.72	44.57	3.41	20.23	5.32	0.03	27.95	45.26
OV-Uni3DETR (ours)	**25.33**	86.05	50.49	28.11	31.51	18.22	24.03	6.58	12.17	29.62	54.63
Methods		bathtub	refrigerator	desk	nightstand	counter	door	curtain	box	lamp	bag
OV-3DET		56.28	10.99	19.72	0.77	0.31	9.59	10.53	3.78	2.11	2.71
CoDA		50.51	6.55	12.42	15.15	0.68	7.95	0.01	2.94	0.51	2.02
OV-Uni3DETR (ours)		63.73	14.41	30.47	2.94	1.00	1.02	19.90	12.70	5.58	13.46

should be credited to the multi-modal architecture. When only RGB images are available for inference, AP_{novel} is 5.41%, which is comparable to the performance of CoDA using point clouds. This demonstrates that the RGB image features do not collapse during training. As a result, RGB-only inference achieves an equally excellent 3D detection result, which well illustrates that OV-Uni3DETR can take any modality data and can be switchable for different modalities during inference. When both point clouds and RGB images are available, the detector obtains 12.96% AP_{novel}, 6.25% higher than CoDA. This demonstrates that our **cycle-modality propagation effectively facilitates the integration of multi-modal knowledge and the comprehensive utilization of information from different modalities**, thus assisting in achieving superior performance for open-vocabulary 3D detection.

We then evaluate OV-Uni3DETR on ScanNet and list the AP_{25} metric in Tab. 1. We utilize the same setting as CoDA, where one single-view image corresponds to one point cloud scene. Meanwhile, since many images share high similarity in content in this setting, with only slight difference in perspective, we do not use extra 2D detection images here. We observe that the superiority of our method becomes more remarkable on the ScanNet dataset. We achieve 5.55% higher AP_{novel} than CoDA using only point cloud data, and the base category AP_{25} is even 8.9% higher. It is noteworthy that the *image-only AP_{25} is even higher than that achieved by CoDA with point clouds*. OV-Uni3DETR obtains 8.10% AP_{novel} and 10.23% AP_{all} using RGB images only, which sur-

Table 3: The performance of OV-Uni3DETR on the KITTI and nuScenes dataset for open-vocabulary 3D object detection. For KITTI, the car and cyclist classes are seen during training while the pedestrian class is novel. We report AP_{25} with 11 recall positions on the moderate difficulty. For nuScenes, "Car, trailer, construction vehicle, motorcycle, bicycle" are seen and the rest five are unseen ones.

Method	Inputs	KITTI			nuScenes					
		$\mathrm{AP}_{Ped.}$	AP_{Car}	$\mathrm{AP}_{Cyc.}$	AP_{novel}	AP_{base}	AP_{all}	NDS_{novel}	NDS_{base}	NDS_{all}
Det-PointCLIP [70]	P	0.32	3.67	1.32	0.59	2.13	1.36	1.92	5.86	3.89
Det-PointCLIPv2 [76]	P	0.32	3.58	1.22	0.61	2.05	1.33	1.97	5.74	3.86
3D-CLIP [40]	P+I	1.28	42.28	21.99	2.74	12.60	7.67	8.98	23.81	16.39
OV-Uni3DETR (ours)	P	19.57	92.44	56.67	15.48	61.28	38.39	15.61	44.71	30.16
	I	9.98	75.14	18.44	12.54	55.30	33.93	14.67	39.43	27.05
	P+I	**23.04**	**92.55**	**58.21**	**18.96**	**63.34**	**41.15**	**17.05**	**46.69**	**31.87**

passes CoDA by 1.56% and 1.19% separately using point clouds. This strongly demonstrates that our **multi-modal architecture and associated switch-modality training effectively prevent the collapse of single-modal information, thus achieving modality unifying**. Meanwhile, the integrated utilization of different modalities is further boosted. For multi-modal inference, OV-Uni3DETR achieves 15.21% AP_{novel}, 8.67% higher than previous methods. The superiority of OV-Uni3DETR can thus be further demonstrated.

Additionally, we train and evaluate OV-Uni3DETR in the same setting as OV-3DET [34] on the ScanNet dataset, where 20 categories are evaluated. All categories are not annotated with 3D bounding boxes (*i.e.*, there are no seen classes and all classes are novel ones). For a fair comparison, we train and evaluate our model with point clouds only, and without RGB images participating in. The AP_{25} metric is listed in Tab. 2. We obtain the 25.33% AP_{25}, which surpasses OV-3DET by 7.31% and CoDA by 6.01%. This further demonstrates the effectiveness of our OV-Uni3DETR and its ability in different settings, even if no seen classes are available. For 16 classes out of the total 20 ones, we achieve the best 3D detection performance among the three methods. This validates the superiority of our method over different categories.

Outdoor 3D open-vocabulary detection. We then evaluate on the outdoor KITTI dataset. For the simplicity of training, we do not utilize the ground-truth sampling augmentation [64] here. We report the AP_{25} metric with 11 recall positions on the moderate difficulty objects from the validation set, and list the results in Tab. 3. To the best of our knowledge, we are the first to conduct open-vocabulary experiments on the outdoor 3D detection datasets. Outdoor point clouds are usually collected by the LiDAR sensor, where background points dominate the scene, and foreground objects are small and sparse, with significantly less points. The gap between outdoor LiDAR points and 2D images becomes larger, making detecting novel classes quite challenging in the outdoor scenes. Therefore, directly implementing CLIP-based methods to outdoor point clouds results in the limited performance. In this situation, OV-Uni3DETR still achieves 19.57% AP_{25} of the pedestrian class (the novel class) for 3D box prediction, which surpasses 3D-CLIP by 18.29%. For multi-modal results, AP_{novel}

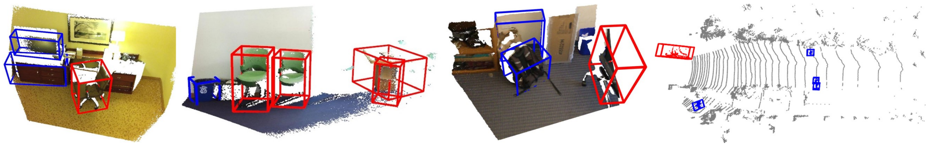


Fig. 4: Visualization of OV-Uni3DETR for open-vocabulary 3D detection on the SUN RGB-D (the first and the second), ScanNet (the third) and KITTI (the fourth) dataset. The red boxes are base classes and blue boxes are novel classes.

Table 4: The 2D detection performance of OV-Uni3DETR on SUN RGB-D compared to 2D detectors with image-text pre-training.

Method	Inputs	AP^{2D}_{novel}	AP^{2D}_{base}	AP^{2D}_{all}
CLIP [18]	I	0.52	11.78	2.97
RegionCLIP [73]	I	2.24	18.35	5.74
OV-Uni3DETR (ours)	P	0.40	3.05	0.98
	I	5.21	19.25	8.26
	P+I	**6.52**	**19.68**	**9.38**

Table 5: The performance of OV-Uni3DETR on SUN RGB-D for closed-vocabulary monocular 3D detection. The AP^{3D} metric is adopted from OMNI3D [2].

Method	Backbone	Trained on	AP_{25}	AP_{50}	AP^{3D}
ImVoxelNet [45]	ResNet34	SUN RGB-D	34.1	12.8	30.6
Cube RCNN [2]		SUN RGB-D	-	-	34.7
Cube RCNN		$OMNI3D_{IN}$	-	-	35.4
OV-Uni3DETR (ours)		SUN RGB-D	**41.6**	**15.4**	**37.7**
ImVoxelNet	ResNet50	SUN RGB-D	40.9	13.5	36.3
OV-Uni3DETR (ours)		SUN RGB-D	**44.6**	**16.1**	**39.7**

is further improved to 23.04%. This validates the ability of our OV-Uni3DETR to detect in both indoor and outdoor scenes, thus achieving scene unifying.

For nuScenes, we report mAP and nuScenes detection score (NDS), with the match thresholds of 15 meters in Tab. 3. The CBGS [75] strategy is adopted for training. For such a setting, where more categories appear in the outdoor scenes, OV-Uni3DETR achieves the performance that is equally good - 18.96% AP_{novel} and 63.34% AP_{base}. It achieves the 17.05% NDS_{novel}, which demonstrates that the detector can also predict object attributes like velocity well for novel classes, which is critical for the model applications in outdoor scenes.

Visualization. We provide visualized results in Fig. 4, where novel class objects are in blue boxes. As can be seen, OV-Uni3DETR well recognizes and localizes novel classes in indoor and outdoor scenes. This further validates its ability.

2D open-vocabulary detection. Besides the 3D boxes, OV-Uni3DETR predicts the 4-dim 2D boxes. Here we compare its 2D performance with the Faster RCNN [41] detector trained on the same SUN RGB-D dataset. To recognize novel classes, we initialize Faster RCNN with CLIP [18] and RegionCLIP [73] parameters for comparison. With only RGB images as input, we achieve the 5.21% AP_{novel} for the 2D boxes, which is 2.97% higher than RegionCLIP. Although point cloud data struggle to produce detection results that are comparable to that from RGB only, partially because of the lack of 2D information, they can contribute to further improvement for multi-modal inference. We achieve the 6.52% AP_{novel} for the multi-modal performance, 4.28% higher than RegionCLIP. This demonstrates that OV-Uni3DETR can also predict 2D boxes well.

Table 6: The performance of OV-Uni3DETR on KITTI test for closed-vocabulary monocular 3D object detection. *: AP^{3D} on the moderate car is the most important metric. §: the method uses more images (OMNI3D) for training.

Method	AP^{3D}			AP^{BEV}		
	Easy	Mod.*	Hard	Easy	Mod.	Hard
SMOKE [30]	14.03	9.76	7.84	20.83	14.49	12.75
PGD [56]	19.05	11.76	9.39	26.89	16.51	13.49
MonoRCNN [51]	18.36	12.65	10.03	25.48	18.11	14.10
MonoFlex [71]	19.94	13.89	12.07	28.23	19.75	16.89
GUPNet [32]	20.11	14.20	11.77	-	-	-
Cube RCNN § [2]	23.59	15.01	12.56	31.70	21.20	18.43
OV-Uni3DETR (ours)	20.81	**15.95**	**13.95**	29.80	**22.74**	**20.20**

Table 7: Effect of cycle-modality propagation on the SUN RGB-D dataset. 2D→3D and 3D→2D knowledge propagation are studied when RGB images exist.

Inputs	2D→3D	3D→2D	AP$_{novel}$	AP$_{base}$	AP$_{all}$
I			N/A	28.12	N/A
	✓		3.86	28.17	9.14
		✓	2.02	29.24	7.94
	✓	✓	**5.41**	**29.51**	**10.65**
P+I			N/A	48.45	N/A
	✓		11.21	48.42	19.30
		✓	3.72	49.23	13.61
	✓	✓	**12.96**	**49.25**	**20.85**

4.2 Closed-Vocabulary 3D Object Detection

In this subsection, we evaluate in the traditional closed-vocabulary 3D detection setting. As the point cloud branch is adopted from Uni3DETR, here we evaluate with only RGB images as input, and compare with existing RGB-only methods.

Indoor 3D closed-vocabulary detection. We first evaluate on SUN RGB-D for the 10 category setting from VoteNet [39], and list 3D AP in Tab. 5. With the ResNet34 backbone, OV-Uni3DETR obtains 37.7% AP^{3D}, which surpasses ImVoxelNet by 7.1% and Cube RCNN by 3%. Besides, it is 2.3% higher than Cube RCNN trained on OMNI3D$_{\text{IN}}$, with significantly more RGB images participating in training. With the ResNet50 backbone, the AP^{3D} is further improved to 39.7%, 3.4% higher than ImVoxelNet. This illustrates that OV-Uni3DETR can also achieve an excellent performance for the close-vocabulary setting.

Outdoor 3D closed-vocabulary detection. We then evaluate on KITTI and compare it with previous monocular 3D detection methods on the test set. We list the AP$_{70}$ metric with 40 recall positions in Tab. 6. We obtain the 15.95% AP for the moderate difficulty, which surpasses MonoFlex by 2.06% and GUPNet by 1.75%. Compared with Cube RCNN, which is trained on the OMNI3D benchmark with more images, the 3D detection AP is 0.94% higher. The ability in the outdoor closed-vocabulary setting can thus be validated.

4.3 Ablation Study

Finally, we conduct ablation studies in this subsection. We mainly analyze the effect of the cycle-modality propagation and multi-modal learning.

Cycle-modality propagation. Tab. 7 analyzes the effect of the cycle-modality propagation in open-vocabulary learning. We conduct such a study when RGB images exist, since the 3D→2D propagation requires images as input. Without the cycle-modality propagation, the detector does not have the open-vocabulary ability, thus cannot detect novel class objects. The *2D→3D propagation helps leverage semantic knowledge in the 2D domain, thus helps expand the 3D vocabulary size.* The detector can thus perform open-vocabulary detection. The 3D→2D

Table 8: Effect of the details in multi-modal learning on SUN RGB-D. "Switched", "3D-conv", and "dualb" are short for the switched-modality training strategy, 3D convolution with BN for feature regularization, and the dual-branch structure. These designs reduce the strong dependency on the point cloud modality. The image-only AP can thus be on par with the single-modality baseline.

	switched	3D-conv	dualb	AP_{novel}			AP_{base}			AP_{all}		
				P	I	P+I	P	I	P+I	P	I	P+I
single-modality				9.64	5.35	-	47.83	30.15	-	17.94	10.74	-
multi-modality		✓	✓	6.36	1.12	12.04	43.27	10.74	47.89	14.38	3.21	19.81
	✓		✓	9.57	3.37	10.92	47.86	25.82	49.01	15.54	8.25	19.20
	✓	✓		6.34	4.41	9.65	41.30	26.73	42.84	13.94	9.26	16.86
	✓	✓	✓	**9.66**	**5.41**	**12.96**	**48.29**	**29.51**	**49.25**	**18.06**	**10.65**	**20.85**

propagation further helps improve the performance. When only images are available, AP_{novel} is enhanced by 1.55%, and AP_{all} is increased by 1.51%. Meanwhile, the multi-modal AP_{novel} is improved by 1.75%. The *3D→2D propagation introduces geometric knowledge into 2D detection images, where more annotated categories and diversified contained scenes boost open-vocabulary learning.* Through our proposed cycle-modality propagation, the 2D semantic and 3D geometric knowledge can be leveraged, which contributes to the comprehensive utilization of knowledge from different modalities. Therefore, OV-Uni3DETR can perform open-vocabulary 3D detection with modality unifying.

Multi-modal learning. Tab. 8 analyzes the effect of our designed multi-modal learning manner. When either point clouds or RGB images are used for inference, the AP_{25} metrics are basically equal to or surpass those from single-modality training. Without switched-modality training, the detection performance deteriorates severely for single-modality inference. Especially when only images are available, switched-modality training contributes to the AP_{novel} improvement from 1.12% to 5.41%. By random switching, the model can accept image-only features during training, thus preventing image features from being ignored and reducing the strong dependency on point clouds. Equally, 3D convolution with BN helps improve the image-only AP_{novel} from 3.37% to 5.41%. The reason is that 3D convolution with batch normalization regularizes different modality features, reducing the modality gap between the point cloud features and the image features. This prevents feature discrepancy and the suppression of image features compared to more information-abundant point cloud features. Furthermore, the dual-branch structure helps distinguish information flows of different modalities, thus avoiding the effects of generated noisy 3D boxes. It thus helps improve the performance with any modality inputs: point cloud only AP_{novel} from 6.34% to 9.66%, and image-only AP_{novel} from 4.41% to 5.41%. The capacity of our multi-modal learning for modality unifying can be demonstrated.

5 Conclusion

In this paper, we mainly propose OV-Uni3DETR, a unified multi-modal open-vocabulary 3D detector. With the help of multi-modal learning and the cycle-modality knowledge propagation, our OV-Uni3DETR recognizes and localizes novel classes well, achieving modality unifying and scene unifying. Experiments demonstrate its strong ability in both open-vocabulary and close-vocabulary settings, both indoor and outdoor scenes, and with any modality data inputs. Addressing unified open-vocabulary 3D detection in the multi-modal context, we believe our research will stimulate following research along the promising but challenging universal 3D computer vision direction.

A Overview of 3D Object Detection Research State

Table 9: The overview of existing 3D object detectors on their capability of scene unifying, modality unifying and category unifying. Scene unifying mainly includes detecting in indoor and outdoor scenes. For modality unifying, the main aspects include utilizing both point clouds and images for training ("P-train", "I-train"), and being modality-switchable during inference ("switch"). For category unifying, the main process is about open-vocabulary learning ("open").

Methods			Scene		Modality			Category	
			indoor	outdoor	P-train	I-train	switch	closed	open
indoor	VoteNet [39]	ICCV'19	✓	✗	✓	✗	✗	✓	✗
	3DETR [37]	ICCV'21	✓	✗	✓	✗	✗	✓	✗
	FCAF3D [44]	ECCV'22	✓	✗	✓	✗	✗	✓	✗
	NeRF-Det [62]	ICCV'23	✓	✗	✗	✓	✗	✓	✗
outdoor	PointPillars [21]	CVPR'19	✗	✓	✓	✗	✗	✓	✗
	CenterPoint [65]	CVPR'21	✗	✓	✓	✗	✗	✓	✗
	MonoFlex [71]	CVPR'21	✗	✓	✗	✓	✗	✓	✗
	BEVFormer [24]	ECCV'22	✗	✓	✗	✓	✗	✓	✗
	VoxelNeXt [7]	CVPR'23	✗	✓	✓	✗	✗	✓	✗
multi-modal	MVXNet [52]	ICRA'19	✗	✓	✓	✓	✗	✓	✗
	ImVoteNet [38]	CVPR'20	✓	✗	✓	✓	✗	✓	✗
	TokenFusion [58]	CVPR'22	✓	✗	✓	✓	✗	✓	✗
	UVTR [23]	NeurIPS'22	✗	✓	✓	✓	✗	✓	✗
	VirConvNet [61]	CVPR'23	✗	✓	✓	✓	✗	✓	✗
scene-unified	EPNet [17]	ECCV'20	✓	✓	✓	✓	✗	✓	✗
	ImVoxelNet [45]	WACV'22	✓	✓	✗	✓	✗	✓	✗
	Cude RCNN [2]	CVPR'23	✓	✓	✗	✓	✗	✓	✗
	Uni3DETR [60]	NeurIPS'23	✓	✓	✓	✗	✗	✓	✗
modality-unified	MetaBEV [12]	ICCV'23	✗	✓	✓	✓	✓	✓	✗
open-vocabulary	OV-3DET [34]	CVPR'23	✓	✗	✓	✗	✗	✓	✓
	CoDA [4]	NeurIPS'23	✓	✗	✓	✗	✗	✓	✓
⋆ OV-Uni3DETR (ours)			✓	✓	✓	✓	✓	✓	✓

We list the overview of existing 3D object detectors about their unifying capability in Tab. 9. Related research on 3D universal detection still remains lagging behind that of 2D. A universal 3D detector, we believe, should at least have the ability to utilize any modality data to detect any category in any scene. As such, it should achieve unifying across three levels: scene, modality, and category. Early research primarily focuses on the specific domains. Some recent works achieve unifying on one aspect, but fail to address all three ones. In comparison, OV-Uni3DETR performs open-vocabulary 3D detection with modality unifying and scene unifying, and well achieves the targets of unifying along the modality, scene, category levels. Therefore, it greatly advances existing research towards the goal of universal 3D object detection. We believe OV-Uni3DETR can become a significant step towards the future of 3D foundation models.

B More Implemental Details

SUN RGB-D [53]. For training OV-Uni3DETR on the SUN RGB-D dataset, we adopt the initial learning rate of 2e-5 and the batch size of 32 for 90 epochs, and the learning rate is decayed by 10x on the 70th and 80th epoch. For point clouds, we filter the input point clouds in the range [-3.2m, 3.2m] for the x axis, [-0.2m, 6.2m] for the y axis and [-2m, 0.56m] for the z axis. We randomly flip the point clouds along the x axis and randomly sample 20,000 points for data augmentation. Common global translation, rotation and scaling strategies are also adopted here. For RGB images, we only adopt pixel-level data augmentation strategies, including brightness, contrast, saturation, and channel swapping. No spatial-level augmentations are utilized.

ScanNet [9]. For the ScanNet dataset, we adopt the totally same setting as [4] for training and inference, where point clouds are generated from the single-view depth images. We first normalize the point clouds to the center of the view, then adopt the range of [-5.12m, 5.12m] for the x and y axis and [-1.28m, 1.28m] for the z axis. We train OV-Uni3DETR with the initial learning rate of 1e-5 and the batch size of 24 for 40 epochs, and the learning rate is decayed by 10x on the 32nd and 38th epoch. As we can see, the training epochs we require is significantly less than the previous method CoDA [4].

Besides the experimental setting in our main paper, we also utilize a multi-view setting of the ScanNet dataset in this supplementary material. Specifically, each scene consists of plenty of multi-view images, and the point clouds are reconstructed from the multiple multi-view images for a panoramic, full-angle large scene. We also choose the top 10 classes for base classes, and the other 50 ones for novel classes. For point clouds, we adopt the range of [-6.4m, 6.4m] for the x and y axis and [-0.1m, 2.46m] for the z axis after global alignment. The input point clouds are randomly flipped along both the x and y axis. For RGB images, we randomly select 8 multi-view images for training. For inference, 24 input views are randomly selected for the multi-modal inference and 40 input views are utilized for RGB-only inference. Other hyper-parameters and operations are the same as the SUN RGB-D dataset.

KITTI [13]. For the KITTI dataset, we do not use ground-truth sampling and the object-level noise strategy augmentations during training. Instead, we randomly flip the point clouds along the x axis and randomly sample 18,000 points for data augmentation. For RGB images, besides the pixel-level data augmentation strategies, we also randomly flip the input images along the x axis.

nuScenes [3]. Compared to KITTI, the nuScenes dataset covers a larger range, with 360 degrees around the LiDAR instead of only the front view. Each scene consists of 6 multi-view images. We train OV-Uni3DETR for 20 epochs with the cyclic schedule. The CBGS [75] strategy is adopted for training point clouds.

C More Method Details

Algorithm 1 The overall procedure for training OV-Uni3DETR.

```
Input:
 1. 3D data ($X_P$, $X_I$): point clouds and corresponding 3D detection images, inherently
    associated with camera parameters $K$, $R_t$ and 3D annotations $\{c_i, bb_i^{3D}, bb_i^{2D}\}$.
 2. 2D data $X_I^{2D}$: 2D detection images inherently associated with 2D labels $\{c_i, bb_i^{2D}\}$,
    pre-trained 2D detector $\Phi^{2D}$.
Training:
  2D → 3D knowledge propagation:
    1. Apply $\Phi^{2D}$ on $X_I$: $\{\hat{c}_i, \hat{bb}_i^{2D}\} = \Phi^{2D}(X_I)$.
    2. Project 2D boxes to 3D for 3D boxes: $\hat{bb}_i^{3D} = \hat{bb}_i^{2D} \circ (KR_t)^{-1}$
  3D → 2D knowledge propagation:
    1. Pre-train a class-agnostic 3D detector $\Phi_{CA}^{3D}$ with the module for camera extrinsic
       parameter prediction using 3D detection images $X_I$, $R_t$, $\{bb_i^{3D}\}$.
    2. Apply $\Phi_{CA}^{3D}$ on $X_I^{2D}$: $\hat{R}_t, \{\hat{bb}_i^{2D}, \hat{bb}_i^{3D}\} = \Phi_{CA}^{3D}(X_I^{2D})$.
    3. Set camera intrinsic parameters $K$ according to image sizes for $X_I^{2D}$.
    4. Obtain class-specific 3D boxes $\{\hat{c}_i, bb_i^{2D}, \hat{bb}_i^{3D}\}$ for $X_I^{2D}$ by Hungarian matching.
  Network forwarding:
    1. Extract point cloud features $F_P$, image features $F_I$, and project image features
       to the voxel space through $KR_t$ for $F_I'$.
    2. Obtain multi-modal features: $F_M = \mathrm{BN}(\mathrm{conv}(F_P)) + \mathrm{BN}(\mathrm{conv}(F_I'))$.
    3. Randomly select $F_P$, $F_I'$, $F_M$ for category, 3D box and 2D box prediction.
    4. Supervise with $\{c_i, bb_i^{3D}, bb_i^{2D}\} + \{\hat{c}_i, \hat{bb}_i^{3D}, \hat{bb}_i^{2D}\}$ for training.
```

We summarize our method in Algorithm 1. Specifically, OV-Uni3DETR utilizes various sources of data, including point clouds, 3D detection images and 2D detection images for training. 3D detection images align with point clouds and are 3D box annotated, while 2D detection images are not aligned and only 2D box annotated.

For training OV-Uni3DETR, we introduce the concept of cycle-modality propagation. Specifically, for 2D → 3D knowledge propagation, we leverage a pre-trained 2D detector on 3D detection images, and project the generated 2D boxes to the 3D space to propagate semantic knowledge into the 3D domain. For 3D → 2D knowledge propagation, we first pre-train a class-agnostic 3D detector with a camera extrinsic parameter prediction branch, and apply it on 2D detection images to propagate geometric knowledge into the 2D domain.

With knowledge propagation, the input data are forwarded into the network for training, and the generated boxes from cycle-modality propagation are integrated into the ground-truth to supervise the training. The switch-modality training strategy is utilized for modality unifying.

Table 10: The performance of OV-Uni3DETR on the ScanNet multi-view setting for open-vocabulary 3D object detection. P denotes the point cloud inputs and I denotes image inputs.

Method	Inputs	AP_{novel}	AP_{base}	AP_{all}
Det-PointCLIP [70]	P	0.07	1.05	0.23
Det-PointCLIPv2 [76]	P	0.06	1.01	0.22
3D-CLIP [40]	P+I	2.52	11.21	3.97
OV-Uni3DETR (ours)	P	10.59	44.39	16.22
	I	7.70	28.39	11.15
	P+I	**13.72**	**48.05**	**19.44**

Table 11: The performance of OV-Uni3DETR on the ScanNet dataset for closed-vocabulary 3D object detection. Only multi-view RGB images are used for training and inference.

Method	# Training	# Inference	AP_{25}
ImVoxelNet [45]	20 views	20 views	44.1
		50 views	48.1
NeRF-Det [62]	20 views	20 views	44.9
		50 views	48.5
	50 views	50 views	51.3
		100 views	52.3
ImGeoNet [54]	50 views	50 views	54.8
OV-Uni3DETR (ours)	20 views	20 views	**52.3**
		50 views	**55.1**

D More Quantitative Results

D.1 ScanNet Multi-View Setting

3D open-vocabulary detection. We evaluate OV-Uni3DETR on the ScanNet dataset, multi-view setting, and list the AP_{25} metric in Tab. 10. We randomly select 8 multi-view images from a single scene for training. For inference, 24 input views are randomly selected for the multi-modal inference and 40 input views are utilized for RGB-only inference. We observe that the superiority of our method equally becomes remarkable on the ScanNet dataset. We achieve 10.59% AP_{novel} using only point cloud data, and 13.72% AP_{novel} using multi-modal data. The listed AP here is a little lower than that in our main paper for the ScanNet single-view setting, because the ScanNet scene here is reconstructed from multi-view images, thus covers a larger range and contains more objects. The performance of our OV-Uni3DETR in such the setting further demonstrates its ability for utilizing multi-view images.

3D closed-vocabulary detection. We then evaluate OV-Uni3DETR on the ScanNet dataset for the 18 category closed-vocabulary setting. With 20-view images participating in training and inference, our OV-Uni3DETR obtains the 52.3% AP_{25}, which surpasses NeRF-Det by 7.4% under the same conditions. When the images for inference increase to 50 views, our detector further achieves the 55.1% AP_{25}, 6.6% higher than NeRF-Det. The 3D detection performance is even higher than ImGeoNet, which uses more images (50-view) for training. This demonstrates that for the multi-view close-vocabulary indoor detection setting, OV-Uni3DETR can also obtain the state-of-the-art performance.

D.2 Ablation Study on Training Data

Here we further conduct the ablation study on the 2D detection images we used. We use 2D detection images from the COCO [27], ScanNet [9], Objects365 [47], OpenImages [20], and LVIS [15] dataset, together with the corresponding category vocabularies. According to our main paper, 2D detection images benefit

Table 12: Ablation study on the SUN RGB-D dataset about the 2D detection images utilized. The format of table entries is: dataset name - training vocabulary number. Only RGB images are used for inference.

2D detection images	AP_{novel}	AP_{base}	AP_{all}
coco-80c	3.29	32.19	9.57
scannet-100c	5.34	33.36	11.43
scannet-200c	5.65	28.53	10.62
objects365-365c	5.41	29.51	10.65
openimages-500c	5.79	32.61	11.62
lvis-1230c	5.89	31.03	11.35

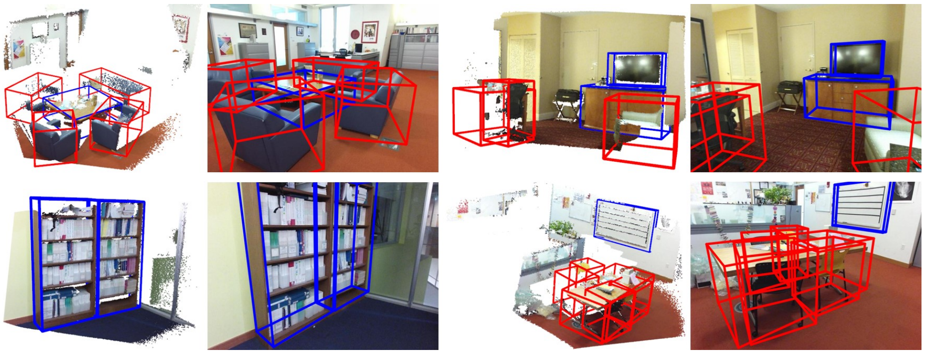


Fig. 5: Visualization of OV-Uni3DETR for open-vocabulary 3D detection on SUN RGB-D. The red boxes are base classes and blue boxes are novel classes.

RGB-only inference most significantly, so we just conduct experiments with only RGB images here. The 3D detection AP is listed in Tab. 12.

As we can see, when adopting 2D detection images from the COCO dataset, the performance of 3D open-vocabulary detection is limited, only 3.29% AP_{novel}. The reason is that the vocabulary size of the COCO dataset is still relatively small, only 80 categories. Many of them are outdoor classes, not overlapped with classes in the indoor SUN RGB-D dataset. When we adopt the ScanNet dataset, since it covers the similar indoor scenes, it benefits 3D open-vocabulary detection on the SUN RGB-D dataset more, contributing to the 5.65% AP_{novel} ultimately. When the vocabulary size of 2D detection images continues to increase, like the OpenImages dataset with 500 classes or the LVIS dataset with 1,230 classes, 3D open-vocabulary detection can be further boosted - 5.89% AP_{novel} after leveraging images from LVIS. This demonstrates that since 2D detection images contain more images and categories, more abundant information can be utilized after involving them in training. As a result, the performance of 3D open-vocabulary detection, especially in novel class recognition, can be boosted significantly.

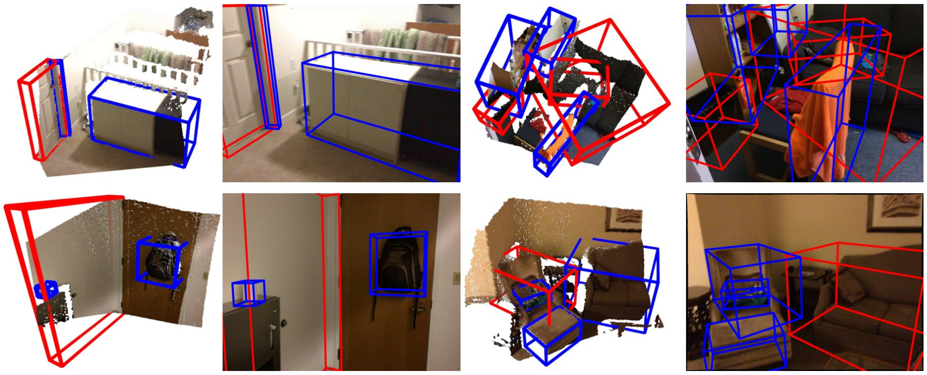


Fig. 6: Visualization of OV-Uni3DETR for open-vocabulary 3D detection on the **ScanNet single-view setting**. The point cloud scenes are reconstructed from one single-view depth image. This setting is the same as that from CoDA [4]. The red boxes are base classes and blue boxes are novel classes.

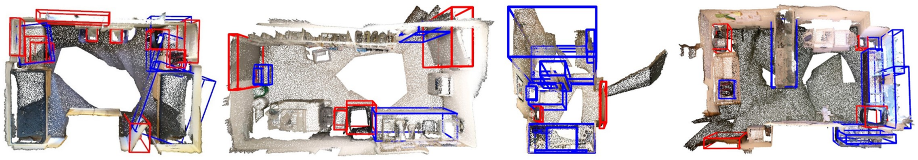


Fig. 7: Visualization of OV-Uni3DETR for open-vocabulary 3D detection on the **ScanNet multi-view setting**. Point cloud scenes are reconstructed from plenty of multi-view images. The red boxes are base classes and blue boxes are novel classes.

E Visualized Results

We provide more visualized results on the SUN RGB-D, the ScanNet and the KITTI dataset. OV-Uni3DETR detects novel classes equally well. On the SUN RGB-D dataset (Fig. 5), it recognizes and localizes novel classes like the coffee table in the first example, the tv and cabinet in the second example, the bookshelf and white board in the rest two examples. Besides, it detects seen classes like table, chair successfully. For both point clouds and RGB images, 3D bounding boxes can all be predicted well. On the ScanNet dataset single-view setting (Fig. 6), our OV-Uni3DETR also detects novel classes like backpack, tissue paper, ottoman well. Meanwhile, for ScanNet multi-view setting (Fig. 7), where the 3D scene is larger, contained objects are more thus 3D detection is more challenging, our method also detects both base and novel classes in the scene well. On the KITTI dataset (Fig. 8), it also detects the novel class pedestrian well. The visualization results further demonstrate the strong ability of OV-Uni3DETR.

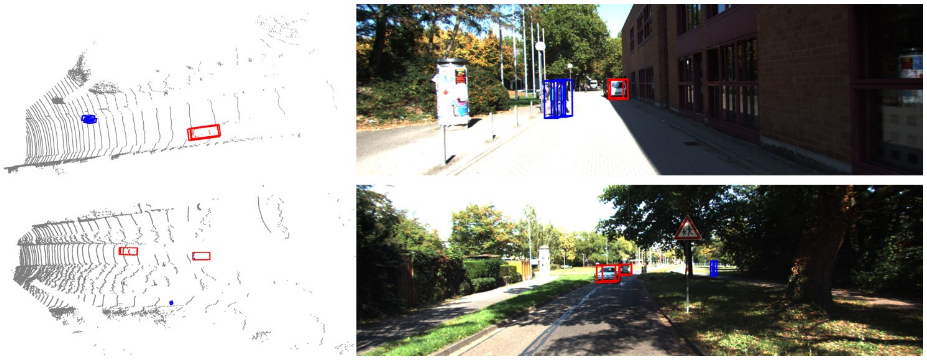


Fig. 8: Visualization of OV-Uni3DETR for open-vocabulary 3D detection on the KITTI dataset. The red boxes are base classes and blue boxes are novel classes.

References

1. Bangalath, H., Maaz, M., Khattak, M.U., Khan, S.H., Shahbaz Khan, F.: Bridging the gap between object and image-level representations for open-vocabulary detection. In: NeurIPS (2022) 4
2. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: A large benchmark and model for 3d object detection in the wild. In: CVPR (2023) 4, 6, 12, 13, 15
3. 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: CVPR (2020) 9, 16
4. Cao, Y., Zeng, Y., Xu, H., Xu, D.: Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. In: NeurIPS (2023) 4, 9, 10, 15, 16, 20
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 8
6. Cen, J., Yun, P., Cai, J., Wang, M.Y., Liu, M.: Open-set 3d object detection. In: 3DV (2021) 4
7. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: CVPR (2023) 15
8. Contributors, M.: Mmdetection3d: Openmmlab next-generation platform for general 3d object detection (2020) 9
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017) 2, 3, 9, 16, 18
10. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: Language-driven open-vocabulary 3d scene understanding. In: CVPR (2023) 4
11. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: CVPR (2022) 4
12. Ge, C., Chen, J., Xie, E., Wang, Z., Hong, L., Lu, H., Li, Z., Luo, P.: Metabev: Solving sensor failures for 3d detection and map segmentation. In: ICCV (2023) 4, 15

13. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR (2013) 2, 9, 16
14. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. ICLR (2022) 4, 8
15. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR (2019) 2, 18
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 9
17. Huang, T., Liu, Z., Chen, X., Bai, X.: Epnet: Enhancing point features with image semantics for 3d object detection. In: ECCV (2020) 15
18. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 2, 4, 6, 12
19. Kim, D., Lin, T.Y., Angelova, A., Kweon, I.S., Kuo, W.: Learning open-world object proposals without learning to classify. RAL (2022) 8
20. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. IJCV (2020) 2, 18
21. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: CVPR (2019) 15
22. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: CVPR (2022) 1, 4
23. Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., Jia, J.: Unifying voxel-based representation with transformer for 3d object detection. NeurIIPS (2022) 15
24. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022) 4, 15
25. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. In: NeurIPS (2022) 4
26. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017) 9
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 2, 18
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 1, 4
29. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: ICCV (2021) 4
30. Liu, Z., Wu, Z., Tóth, R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. In: CVPRW (2020) 4, 13
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. ICLR (2019) 9
32. Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., Yan, J., Ouyang, W.: Geometry uncertainty projection network for monocular 3d object detection. In: ICCV (2021) 6, 13
33. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-set 3d detection via image-level class and debiased cross-modal contrastive learning (2022) 4

34. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: CVPR (2023) 4, 9, 10, 11, 15
35. Ma, Z., Luo, G., Gao, J., Li, L., Chen, Y., Wang, S., Zhang, C., Hu, W.: Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In: CVPR (2022) 4
36. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection with vision transformers. In: ECCV (2022) 4
37. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: ICCV (2021) 15
38. Qi, C.R., Chen, X., Litany, O., Guibas, L.J.: Imvotenet: Boosting 3d object detection in point clouds with image votes. In: CVPR (2020) 4, 15
39. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: ICCV (2019) 4, 13, 15
40. 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: ICML (2021) 2, 4, 6, 10, 11, 18
41. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015) 12
42. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR (2019) 6
43. Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: ECCV (2022) 9
44. Rukhovich, D., Vorontsova, A., Konushin, A.: Fcaf3d: fully convolutional anchor-free 3d object detection. In: ECCV (2022) 4, 15
45. Rukhovich, D., Vorontsova, A., Konushin, A.: Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In: WACV (2022) 4, 12, 15, 18
46. Saito, K., Hu, P., Darrell, T., Saenko, K.: Learning to detect every thing in an open world. In: ECCV (2022) 8
47. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV (2019) 2, 9, 18
48. Sheng, H., Cai, S., Liu, Y., Deng, B., Huang, J., Hua, X.S., Zhao, M.J.: Improving 3d object detection with channel-wise transformer. In: ICCV (2021) 4
49. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: CVPR (2020) 4
50. Shi, S., Wang, Z., Shi, J., Wang, X., Li, H.: From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. TPAMI (2020) 4
51. Shi, X., Ye, Q., Chen, X., Chen, C., Chen, Z., Kim, T.K.: Geometry-based distance decomposition for monocular 3d object detection. In: ICCV (2021) 13
52. Sindagi, V.A., Zhou, Y., Tuzel, O.: Mvx-net: Multimodal voxelnet for 3d object detection. In: ICRA (2019) 4, 15
53. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR (2015) 2, 3, 9, 16
54. Tu, T., Chuang, S.P., Liu, Y.L., Sun, C., Zhang, K., Roy, D., Kuo, C.H., Sun, M.: Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In: ICCV (2023) 18
55. Wang, H., Dong, S., Shi, S., Li, A., Li, J., Li, Z., Wang, L., et al.: Cagroup3d: Class-aware grouping for 3d object detection on point clouds. NeurIPS (2022) 4

56. Wang, T., Xinge, Z., Pang, J., Lin, D.: Probabilistic and geometric depth: Detecting objects in perspective. In: CoRL (2022) 13
57. Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. In: ICCVW (2021) 4
58. Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., Wang, Y.: Multimodal token fusion for vision transformers. In: CVPR (2022) 4, 15
59. Wang, Z., Li, Y., Chen, X., Lim, S.N., Torralba, A., Zhao, H., Wang, S.: Detecting everything in the open world: Towards universal object detection. In: CVPR (2023) 1, 4, 8
60. Wang, Z., Li, Y., Chen, X., Zhao, H., Wang, S.: Uni3detr: Unified 3d detection transformer. In: NeurIPS (2023) 4, 5, 6, 9, 15
61. Wu, H., Wen, C., Shi, S., Li, X., Wang, C.: Virtual sparse convolution for multimodal 3d object detection. In: CVPR (2023) 15
62. Xu, C., Wu, B., Hou, J., Tsai, S., Li, R., Wang, J., Zhan, W., He, Z., Vajda, P., Keutzer, K., et al.: Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In: ICCV (2023) 15, 18
63. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: CVPR (2023) 4
64. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors (2018) 4, 11
65. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: CVPR (2021) 4, 15
66. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR (2021) 4
67. Zeng, Y., Jiang, C., Mao, J., Han, J., Ye, C., Huang, Q., Yeung, D.Y., Yang, Z., Liang, X., Xu, H.: Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In: CVPR (2023) 4, 10
68. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: CVPR (2022) 2, 4, 6
69. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. In: NeurIPS (2022) 4
70. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: CVPR (2022) 4, 10, 11, 18
71. Zhang, Y., Lu, J., Zhou, J.: Objects are different: Flexible monocular 3d object detection. In: CVPR (2021) 4, 13, 15
72. Zhang, Z., Sun, B., Yang, H., Huang, Q.: H3dnet: 3d object detection using hybrid geometric primitives. In: ECCV (2020) 4
73. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: CVPR (2022) 12
74. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022) 1, 4, 9
75. Zhu, B., Jiang, Z., Zhou, X., Li, Z., Yu, G.: Class-balanced grouping and sampling for point cloud 3d object detection. arXiv:1908.09492 (2019) 12, 17
76. Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P.: Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In: CVPR (2023) 10, 11, 18