ORACLE-3D: Open-world Region-aligned Cross-modal Learning for Label-efficient 3D Scene Understanding

Yuru Wang[1*], Pei Liu[2,3*], Songtao Wang[1], Zehan Zhang[1†], Xinyan Lu[1], Changwei Cai[1], Hao Li[1], Haipeng Liu[1‡], Qingtian Ning[1], and Jun Ma[2,3†]

[1] Li Auto Inc., Beijing, China
[2] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[3] The Hong Kong University of Science and Technology, Hong Kong SAR, China
jun.ma@ust.hk

Abstract. The recent surge in vision-language models (VLMs) has revolutionized open-world perception, yet transferring this success to 3D scene understanding remains impeded by the scarcity of large-scale point-text pairs and the computational overhead of existing frozen-VLM approaches. To bridge this gap, we present ORACLE-3D (**O**pen-world **R**egion-**A**ligned **C**ross-modal **LE**arning for Label-efficient **3D** Scene Understanding), a unified framework that effectively co-embeds point clouds, images, and text into a shared latent space without relying on labor-intensive 3D-text pair construction. ORACLE-3D leverages the image modality as a semantic bridge, employing a novel cross-modal distillation strategy that comprises logit distillation and feature distillation. These components transfer the rich, open-vocabulary capabilities of pre-aligned 2D VLMs directly to the 3D domain. Furthermore, to mitigate the semantic noise arising from inevitable calibration errors and motion artifacts during 3D-2D projection, we propose a vision-point matching module that dynamically rectifies point-pixel misalignments. To ensure robust convergence amidst the gradient bias inherent in heterogeneous multi-modal learning, we introduce a two-stage optimization strategy coupled with four task-specific weighted losses. Extensive evaluations on the nuScenes, Waymo, SemanticKITTI, and ScanNet datasets demonstrate the superiority of our paradigm. Specifically, ORACLE-3D outperforms state-of-the-art methods by an average of 11.6 and 14.8 in mIoU on base-annotated and annotation-free tasks, respectively. Project Page: https://ocean-luna.github.io/ORACLE.github.io/.

Keywords: Scene understanding · Multi-modal information fusion · Vision-language models

[*] These authors contributed equally to this work.
[†] Corresponding authors.
[‡] Project Leader.

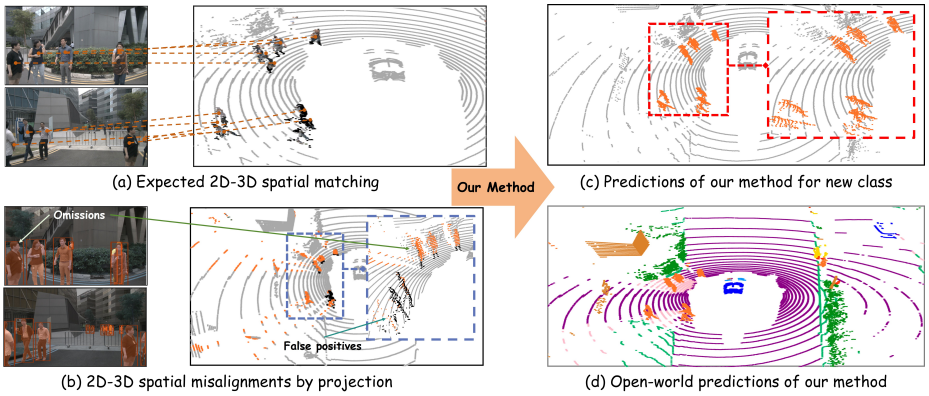

Fig. 1: The matching example of 2D images and 3D point clouds. (a) Expected matching between images and point clouds. (b) Image semantic labels are assigned to the corresponding point clouds by projection, which has many false-positive mismatched points. (c) The semantic predictions of new classes without any manual annotations by our method. (d) Semantic prediction of our method for all points of an input point cloud.

1 Introduction

Open-world 3D scene understanding is pivotal for enabling autonomous systems to perceive and interact with complex, unstructured environments. Unlike traditional closed-set approaches [9, 29] that rely on fixed vocabularies and extensive manual annotations, open-world models aim to recognize and distinguish arbitrary objects, including those unseen during training, without explicit supervision [12, 28]. This capability is imperative for safety-critical applications such as autonomous driving, mobile robotics, and virtual reality, where the diversity of real-world objects far exceeds the capacity of any annotated dataset.

Recent advances in Vision-Language Models (VLMs) have catalyzed progress in this domain, primarily through two paradigms. Frozen-VLM methods [28, 30, 47, 52] attempt to lift the open-vocabulary capabilities of 2D models directly into 3D. However, these approaches often neglect region-level feature distillation and necessitate resource-intensive processes for feature extraction and fusion [20, 37, 49], limiting their scalability to advanced 3D architectures. Alternatively, pair-construction methods [6, 12, 13, 42] seek to align point clouds with text by generating massive 3D-text pairs for contrastive training. While recent works like [50] achieve state-of-the-art performance, constructing such datasets is inherently tedious and expensive. Unlike the abundance of image-text pairs on the internet, high-quality 3D-text pairs are scarce [11, 38], creating a significant bottleneck for scalability.

To mitigate the scarcity of 3D-text data, we observe that point clouds are naturally acquired alongside calibrated images. This presents an opportunity to utilize images as a semantic bridge. A straightforward strategy involves applying

2D open-world models [7, 32] to images and projecting the resulting labels into 3D. However, as illustrated in Fig. 1(b), direct projection is often inaccurate due to sensor calibration errors, temporal misalignment, and parallax [17, 18], leading to noisy 3D semantics. Consequently, establishing a unified framework that leverages images to bridge the gap between text and point clouds requires addressing three critical challenges: (i) Co-embedding point cloud features with pre-aligned image-text representations despite limited data [15,24]; (ii) Correcting geometric misalignments inherent in 2D-to-3D projection [15]; and (iii) Stabilizing the multi-modal training process to prevent gradient dominance by any single modality.

In this work, we propose ORACLE-3D, a unified multi-modal learning framework designed to surmount these challenges. Instead of relying on expensive 3D-text pairs, we leverage 2D foundation models to generate regional semantic supervision. Our key insight is to treat the image branch as a reliable semantic anchor. We introduce Logit Distillation (LD) and Feature Distillation (FD) mechanisms to align point cloud features with the shared image-text embedding space. To tackle projection noise, we devise a Vision-Point Matching (VPM) module that explicitly refines the correspondence between pixels and points. Furthermore, to address the optimization difficulties arising from long-tail novel classes and multi-modal gradient imbalance, we implement a two-stage training strategy coupled with four task-specific weighted losses. This ensures that the model first establishes robust 2D semantic priors before jointly optimizing for 3D understanding. During inference, ORACLE-3D enables open-vocabulary 3D recognition using only point clouds and text prompts. Our main contributions are summarized as follows:

- We present a unified multi-modal contrastive learning framework for open-world 3D scene understanding that eliminates the need for constructing expensive point cloud-text pairs.
- We devise a novel VPM module alongside dual-level distillation (logit and feature) to effectively co-embed point, image, and text modalities into a shared semantic space.
- We introduce a robust two-stage optimization strategy and a set of task-specific weighted losses, ensuring stable convergence and balanced multi-modal learning.
- Extensive experiments demonstrate that ORACLE-3D achieves state-of-the-art performance on the nuScenes benchmark and exhibits strong generalization across diverse outdoor (Waymo, SemanticKITTI) and indoor (ScanNet) environments.

2 Related Works

2.1 Open-vocabulary 2D Scene Understanding

Recent progress in large vision-language pretraining has significantly advanced open-vocabulary recognition, primarily through two paradigms: CLIP-based alignment and text-image grounding. CLIP-based methods [27, 39, 44–46, 51] leverage

contrastive encoders to replace task-specific classifiers with text embeddings, enabling zero-shot transfer while maintaining compatibility with standard detection pipelines. Conversely, grounding-based approaches like GLIP [48] and Grounding DINO [25, 34] directly optimize region-text correspondence, offering fine-grained localization for free-form queries. Leveraging these complementary strengths, we adopt GLEE [39] and Grounding DINO as our 2D label generators. They produce high-quality, category-conditioned bounding boxes and masks, which are subsequently projected to 3D to supervise our framework, balancing robust open-vocabulary semantics with precise localization.

2.2 Open-vocabulary 3D Scene Understanding

Open-vocabulary 3D scene understanding aims to recognize objects beyond predefined categories. Early methods [3, 8, 26, 41] focused on learning discriminative features for unseen classes but were limited by conventional representations. The emergence of VLMs like CLIP [31] shifted the focus to cross-modal knowledge transfer. Projection-based approaches [4, 5, 16, 19] lift 2D semantic predictions or features onto point clouds to guide segmentation. Others, such as OpenMask3D [36] and OpenScene [28], align region-level features with text queries via cross-modal distillation. However, these methods often suffer from resource-intensive feature extraction or rely on frozen backbones that hinder scalability. Alternatively, pair-construction methods like RegionPLC [43] and PLA [12] train on large-scale point-text datasets, which are costly to curate. In contrast, our unified framework achieves scalable open-vocabulary understanding without generating explicit point-text pairs, offering a lightweight and architecture-agnostic solution that seamlessly integrates recent advances in VLMs.

3 Methodology

In this section, we present ORACLE-3D, a unified framework designed to effectively transfer the rich semantic knowledge of 2D VLMs to 3D scene perception. We first introduce our automated data generation pipeline. Subsequently, we elaborate on the network architecture, highlighting our multi-modal alignment mechanisms and a joint optimization strategy tailored for robust learning from noisy pseudo-labels.

3.1 Automated Label Generation Pipeline

To mitigate the scarcity of dense 3D-text annotations, we leverage the open-vocabulary generalization capabilities of state-of-the-art 2D foundation models. Formally, given a sequence of synchronized images $\mathcal{I} = \{I_n\}_{n=1}^{N}$ and a set of candidate text prompts $\mathcal{T} = \{t_c\}_{c=1}^{C}$ (e.g., *"car", "cyclist"*), our pipeline automatically generates instance-level supervision triplets $\mathcal{S} = \langle \hat{\mathbf{b}}, \hat{\mathbf{m}}, \hat{\mathbf{t}} \rangle$, consisting of bounding boxes, segmentation masks, and semantic tags, respectively.

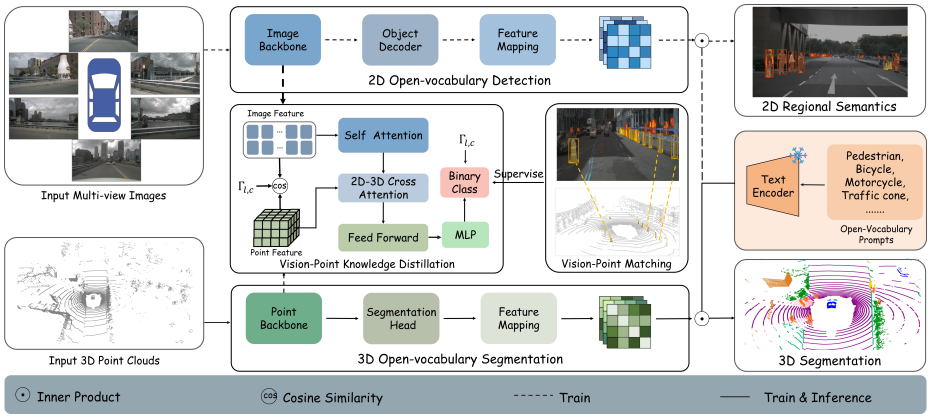

Fig. 2: Overview of our proposed multi-modal unified learning framework, ORACLE-3D. We leverage 2D foundation models to generate regional pseudo-labels on images, serving as supervision for novel categories. The framework processes spatially-temporally aligned point clouds and multi-view images through dual branches, aligning extracted features with text embeddings derived from category prompts. To bridge the modality gap, we distill image logits and features into the point cloud branch and employ a VPM module to learn explicit cross-modal correspondences. During inference, ORACLE-3D achieves open-world 3D understanding using only point clouds and text descriptions.

We employ GLEE [39] as our primary labeler, owing to its robust zero-shot performance across diverse object categories. For comparative analysis, we also explore a cascaded pipeline combining Grounding DINO [25] for open-set detection with SAM2 [33] for instance segmentation.

The generated 2D triplets are subsequently lifted into 3D space via the LiDAR-to-image projection matrix Γ_{lc}, associating the point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ with semantic attributes. We explicitly acknowledge that this 2D-to-3D projection inevitably introduces noise and geometric misalignment. Consequently, our training paradigm is specifically designed to be resilient to these imperfections, refining the alignment through multi-modal consistency constraints.

3.2 ORACLE-3D Architecture

As illustrated in Fig. 2, the ORACLE-3D framework integrates three parallel streams: a frozen text encoder, a learnable image encoder-decoder, and a point cloud segmentation network. The text encoder processes category prompts to produce embeddings emb_t, utilizing global average pooling to aggregate sequence information, consistent with the GLEE protocol.

To facilitate open-set recognition, we replace standard fixed-dimension classifiers with contrastive similarity modules in both the 2D and 3D branches. These modules align perception features with the text embedding space. The classification logits for image pixels (L_I) and 3D points (L_P) are formulated as:

$$L_I = \phi(F_I(I) \cdot W_{i2t}, emb_t), \tag{1}$$
$$L_P = \phi(F_p(P) \cdot W_{p2t}, emb_t), \tag{2}$$

where ϕ represents the dot product similarity, and W_{i2t} and W_{p2t} are learnable projection matrices that map the output of the image extractor F_I and point extractor F_p to the text feature dimension.

During each training iteration e, the model ingests a batch of multi-view images $\{I_e^k\}_{k=1}^K$, the synchronized point cloud P_e, and the calibration matrix Γ_{lc}. The 2D branch is initialized with pre-trained weights from GLEE to preserve its rich vision-language priors. We fine-tune the image branch using the generated pseudo-labels $\langle \hat{\mathbf{b}}, \hat{\mathbf{m}}, \hat{\mathbf{t}} \rangle$ while simultaneously optimizing the 3D branch. This joint training strategy serves a dual purpose: it aligns 3D features with the text-image space and acts as a semantic clustering mechanism. By learning consistent feature representations for specific categories, the model effectively filters out transient noise and false positives inherent in the pseudo-labels, resulting in robust open-vocabulary 3D segmentation.

3.3 Cross-modal Distillation and Matching

To operationalize the image-as-bridge strategy, we align 3D point cloud representations with the pre-established joint embedding space of 2D images and text. We achieve this through two complementary mechanisms: Vision-point Knowledge Distillation, which transfers semantic priors from 2D to 3D, and a VPM module, which learns robust dense correspondences between the modalities.

Vision-point Knowledge Distillation We design a dual-level distillation strategy to transfer the rich open-vocabulary knowledge encapsulated in the image-text domain to the point cloud network. This transfer is implemented through two complementary mechanisms: LD at the output level and FD at the latent level.

We first employ LD to leverage the semantic predictions from the image branch to supervise the point cloud branch, specifically targeting novel categories. Let L_I and L_P denote the classification logits for image pixels and 3D points, respectively, computed via dot-product similarity with text embeddings (as detailed in Eq. (1) and Eq. (2)). We generate pseudo-labels for the 3D points by projecting the image-based predictions into 3D space. The projected semantic map for novel classes, S_{C_n}, is derived as:

$$S_{C_n} = \text{Proj}(\sigma(L_I), \Gamma_{lc}), \tag{3}$$

where σ represents the sigmoid activation, and $\text{Proj}(\cdot, \Gamma_{lc})$ denotes the projection operation from the image plane to the point cloud using the calibration matrix

Γ_{lc}. The logit distillation objective, $\mathcal{L}_{distill_L}$, enforces consistency between the point cloud logits and these projected 2D predictions:

$$\begin{aligned}\mathcal{L}_{distill_L} = \mathcal{L}_{CE}(L_P, S_{C_n} \cup I_{g_{C_b}}) \\ + \mathcal{L}_{Dice}(L_P, S_{C_n} \cup I_{g_{C_b}}).\end{aligned} \tag{4}$$

Here, $\mathcal{L}_{CE}$ and $\mathcal{L}_{Dice}$ refer to the Cross-Entropy and Dice losses, respectively. The term $I_{g_{C_b}}$ represents the mask for base category points, ensuring that the distillation process focuses on novel categories while ignoring points already covered by ground-truth base annotations.

To further enhance representation learning, we incorporate FD to constrain the latent feature space, ensuring the 3D encoder learns representations compatible with text embeddings. We directly align the latent features of point clouds with their corresponding image features. This alignment is enforced strictly on 2D-3D pairs that demonstrate both spatial projection consistency and semantic agreement. The feature distillation loss, $\mathcal{L}_{distill_F}$, minimizes the cosine distance between the projected features:

$$\mathcal{L}_{distill_F} = 1 - \text{CoDist}(F_I(I_m) \cdot W_{i2t}, F_p(P_m) \cdot W_{p2t}), \tag{5}$$

where $\text{CoDist}(\cdot)$ computes the cosine similarity. I_m and P_m denote the subsets of image and point features, respectively, that satisfy the spatial and semantic filtering criteria.

Vision-point Matching Module Geometric projection alone can be noisy due to calibration errors or occlusion. To mitigate this, we introduce the VPM module, which learns to verify fine-grained correspondences between image pixels and 3D points. We formulate this as a binary classification task to distinguish between matched (positive) and unmatched (negative) pairs.

The VPM module employs an attention-based architecture. We designate the image features as the query Q_I, and the point cloud features as both the key K_P and value V_P. First, the image features undergo self-attention to encode contextual information, resulting in refined queries Q_{atten_I}. These are then integrated with the point cloud features via a cross-attention mechanism:

$$\xi_i = \text{softmax}\left(\frac{Q_{atten_I} K_P^\top}{\sqrt{d^{\xi_i}}}\right) V_P, \tag{6}$$

where ξ_i represents the output of the i-th attention head, and d^{ξ_i} is the scaling factor based on the feature dimension. The outputs from multiple heads are concatenated and fused through a Feed-Forward Network (FFN) to produce the final encoding:

$$E_{vpm} = \mathcal{FFN}(\text{Concat}(\xi_1, \ldots, \xi_h)). \tag{7}$$

This encoding E_{vpm} serves as a joint representation map $G_{attenIP}$ for the projected pairs. Finally, a Multi-Layer Perceptron (MLP) acts as a binary classifier to predict the matching probability $\hat{f}_{map}$:

$$\hat{f}_{map} = \mathcal{MLP}(G_{attenIP}), \quad \hat{f}_{map} \in \mathbb{R}^{r \times 2}, \tag{8}$$

Algorithm 1 Unified Multi-modal Training Procedure

Input: Multi-view Images $\mathbf{I}$, Point Cloud P, Calibration Γ_{lc}, Base Labels Y_b, Text Prompts T.
Output: Trained Network Parameters $\theta_{img}, \theta_{pts}, \theta_{vpm}$.
Initialize: Image Branch θ_{img} (from GLEE), Point Branch θ_{pts}, VPM Module θ_{vpm}.
Hyperparameters: Max iterations N_{max}, Warm-up ratio $r = 0.5$.
for $iter = 1$ to N_{max} **do**
  Data Generation: Obtain pseudo-labels $\langle \hat{\mathbf{b}}, \hat{\mathbf{m}}, \hat{\mathbf{t}} \rangle$ from $\mathbf{I}$ via 2D Foundation Model.
  Text Encoding: $emb_t \leftarrow \text{TextEncoder}(T)$.
  if $iter < r \cdot N_{max}$ **then**
    {*Phase I: 2D Semantic Warm-up*}
    Forward Image Branch: $L_I, F_I \leftarrow \text{Net}_{img}(\mathbf{I}, emb_t)$.
    Compute Loss: $\mathcal{L} = \beta \mathcal{L}_{image}$.
    Update: θ_{img} via AdamW (with gradient clipping).
  else
    {*Phase II: Joint Multi-Modal Optimization*}
    Forward Image Branch: $L_I, F_I \leftarrow \text{Net}_{img}(\mathbf{I}, emb_t)$.
    Forward Point Branch: $L_P, F_P \leftarrow \text{Net}_{pts}(P, emb_t)$.
    Projection: Map 2D predictions and features to 3D space using Γ_{lc}.
    Loss Computation:
      $\mathcal{L}_{image}, \mathcal{L}_{point} \leftarrow$ Supervision from Y_b and pseudo-labels.
      $\mathcal{L}_{distill} \leftarrow$ Logit + Feature distillation.
      $\mathcal{L}_{VPM} \leftarrow$ Binary matching verification.
      Total Loss: $\mathcal{L} = \beta \mathcal{L}_{image} + \delta \mathcal{L}_{point} + \gamma \mathcal{L}_{distill} + \eta \mathcal{L}_{VPM}$.
    Update: $\theta_{img}, \theta_{pts}, \theta_{vpm}$ via AdamW (decoupled learning rates).
  end if
end for

where r indicates the number of candidate pixel-point pairs. The VPM module is trained using ground-truth correspondences from both base and novel categories to ensure robust generalization across the entire open-vocabulary domain.

3.4 Joint Optimization Objective

The training of ORACLE-3D is orchestrated through a composite loss function that simultaneously optimizes intra-modal segmentation accuracy and inter-modal semantic alignment. The total objective, $\mathcal{L}$, is a weighted summation of four distinct components:

$$\mathcal{L} = \beta \mathcal{L}_{\text{image}} + \delta \mathcal{L}_{\text{point}} + \gamma \mathcal{L}_{\text{distillation}} + \eta \mathcal{L}_{\text{VPM}}, \tag{9}$$

where the hyperparameters $\beta, \delta, \gamma, \eta$ balance the contribution of the 2D supervision, 3D supervision, knowledge transfer, and geometric verification, respectively.

3.5 Training Protocol and Inference

To ensure stable optimization and effective knowledge transfer, we implement a two-stage curriculum learning strategy, as summarized in Alg. 1.

Algorithm 2 Image-free Open-vocabulary Inference

Input: Input Point Cloud $P \in \mathbb{R}^{N\times 3}$, Arbitrary Category List T_{open}.
Output: Per-point Semantic Predictions $Y_{pred} \in \mathbb{R}^{N}$.
{*Note: Inference relies solely on point cloud data.*}
Step 1: Text Embedding
$emb_t \leftarrow \text{TextEncoder}(T_{open})$. {Encode arbitrary queries}
Step 2: 3D Feature Extraction
$F_P \leftarrow \text{Net}_{pts}(P)$. {Extract point features}
Step 3: Semantic Similarity Calculation
Project point features: $F'_P = F_P \cdot W_{p2t}$.
Compute Similarity Matrix: $S = \phi(F'_P, emb_t)$. {Dot product}
Step 4: Category Assignment
for each point $i \in \{1, \ldots, N\}$ **do**
    $Y^{(i)}_{pred} = \text{argmax}_{c\in T_{open}} S^{(i,c)}$.
end for
return Y_{pred}

Phase I: 2D Semantic Warm-up. The initial phase, occupying the first 50% of training iterations, focuses exclusively on pre-training the image branch. This establishes a robust 2D feature extractor capable of generating reliable pseudo-labels. To prevent instability during initialization, we apply gradient clipping specifically to the image branch parameters.

Phase II: Joint Multi-modal Optimization. Subsequently, we activate the point cloud branch for full joint training using the composite loss function defined in Eq. 9. We employ the AdamW optimizer with decoupled hyperparameters, assigning distinct learning rates and weight decay settings to the image and point cloud networks. This tailored configuration accounts for the differing convergence rates of visual and geometric architectures, ensuring optimal learning dynamics. Empirical results demonstrate that this staged approach significantly enhances performance, particularly for novel categories.

Inference Mechanism. The inference procedure, formalized in Alg. 2, operates as a streamlined, point-only framework (Fig. 2). The reliance on image data is entirely removed, as visual-semantic knowledge is distilled into the 3D network during training. Specifically, arbitrary category names are encoded into text embeddings via the frozen text encoder, and feature representations are extracted for each point in the input point cloud. Semantic predictions are then generated by computing the cosine similarity between point features and text embeddings, assigning the category with the maximal score. This design ensures efficient 3D scene understanding without the computational overhead of processing concurrent image streams during deployment.

4 Experiment

4.1 Dataset

We evaluate our framework across four standard benchmarks, encompassing large-scale outdoor autonomous driving scenarios (nuScenes, SemanticKITTI, Waymo) and complex indoor environments (ScanNet).

nuScenes [2]: A comprehensive multi-modal dataset designed for urban autonomous driving. It captures diverse weather conditions and complex traffic scenarios using a full sensor suite (LiDAR, cameras, and radar). The dataset provides extensive 3D annotations, serving as a rigorous benchmark for robust perception systems.

SemanticKITTI [1]: Built upon the KITTI Vision Odometry Benchmark, this dataset offers dense, point-wise semantic annotations for sequential LiDAR scans. Its fine-grained labeling is essential for evaluating detailed segmentation capabilities in dynamic road environments.

Waymo [35]: Distinguished by its massive scale and high-fidelity sensor data, Waymo contains high-resolution LiDAR and camera inputs from a vast array of driving conditions. It serves as a critical benchmark for assessing the scalability and generalization of perception models.

ScanNet [10]: Representing the indoor domain, ScanNet consists of reconstructed RGB-D scans with dense semantic annotations for meshes and point clouds. It covers a diverse set of architectural layouts and common household objects, facilitating the evaluation of indoor 3D scene understanding.

4.2 Implementation Details and Metrics

Our proposed ORACLE-3D integrates a multi-modal architecture tailored for open-vocabulary 3D understanding. For the 2D branch, we adopt components from GLEE [39], utilizing the CLIP [31] text encoder for robust language processing and MaskDINO [23] as the image encoder-decoder. To demonstrate the versatility of our framework, the 3D backbone is instantiated with three representative architectures: MinkUNet [9], SparseUNet32 [14], and Point Transformer V3 (PTv3) [40].

All models are trained using the Adam optimizer on a cluster of 16 NVIDIA A800 GPUs. The training process is structured into two stages: an initial alignment phase for the image-text branches, followed by joint training of the full unified framework. We evaluate performance using two standard metrics: mean Intersection-over-Union (mIoU) and harmonic mean IoU (hIoU). While mIoU measures average performance across all classes, hIoU is critical for open-vocabulary settings as it balances performance between base and novel categories, offering a more robust indicator of generalization capability. Comprehensive details are provided in the Supplementary Material.

Table 1: Open-world 3D semantic segmentation comparison on the nuScenes dataset. "Bx/Ny" indicates a data split comprising x base categories and y novel categories (e.g., B12/N3 for 12 base and 3 novel classes). Full Supervision denotes the upper-bound performance where models are trained using ground-truth annotations for both base and novel categories. In the open-world setting, the **best** results are highlighted in bold. Higher values indicate better performance.

	Method	B12/N3			B10/N5		
		hIoU	mIoU_B	mIoU_N	hIoU	mIoU_B	mIoU_N
Full supervision	SparseUnet32 [14]	75.2	76.6	73.8	75.7	76.6	74.8
	MinkUNet [9]	70.9	74.6	67.7	72.3	74.8	70.0
	PTv3 [40]	77.2	79.9	74.8	77.6	78.20	77.1
Open-world	3DGenZ [26]	1.6	53.3	0.8	1.9	44.6	1
	3DTZSL [8]	1.2	21.0	0.6	6.4	17.1	3.9
	Ovseg-3D [22]	0.6	74.4	0.3	0	71.5	0
	PLA [12]	47.7	73.4	35.4	24.3	73.1	14.5
	RegionPLC [43] (SparseUnet32)	64.4	75.8	56.0	49.0	75.8	36.3
	ORACLE-3D (SparseUnet32)	68.3	75.9	62.2	65.3	75.9	57.3
	ORACLE-3D (MinkUNet)	64.1	74.3	56.3	64.5	74.2	57.0
	ORACLE-3D (PTv3)	**71.3**	**76.9**	**66.5**	**69.6**	**76.3**	**64.0**

4.3 Base-annotated Open-world Results

Performance on nuScenes Table 1 details the comparative results on nuScenes. When utilizing the same SparseUNet32 backbone as RegionPLC, ORACLE-3D significantly boosts the mIoU for novel categories by 6.2 to 21.0 across various partitions. This gain is achieved without explicit point-text pair construction, highlighting our framework's efficiency. Furthermore, integrating stronger backbones such as PTv3 yields even more pronounced improvements, surpassing RegionPLC by up to 27.7 on novel classes. Crucially, ORACLE-3D maintains competitive performance on base classes, exhibiting minimal degradation within 3 compared to fully supervised baselines, while outperforming language agnostic methods like 3DGenZ [26] by over 60.

Indoor Scene Understanding on ScanNet Extending to the indoor domain, Table 2 illustrates that ORACLE-3D establishes a new state-of-the-art on ScanNet. It outperforms RegionPLC by 3.2 to 11.6 in novel category mIoU, with a remarkably narrow gap ranging from 1.8 to 4.7 relative to fully supervised methods. These results collectively confirm ORACLE-3D's strong generalization capabilities across both outdoor autonomous driving scenarios and complex indoor environments.

Generalization to Waymo and SemanticKITTI As shown in Table 3, our method demonstrates robust scalability on large-scale outdoor datasets. On the Waymo B15/N2 partition, ORACLE-3D achieves exceptional performance,

Table 2: Open-world 3D semantic segmentation comparison on the ScanNet dataset.

	Method	B15/N4		B12/N7		B10/N9	
		$mIoU_B$	$mIoU_N$	$mIoU_B$	$mIoU_N$	$mIoU_B$	$mIoU_N$
Full supervision	SparseUnet32 [14]	68.4	79.6	71.2	71.6	76.5	71.9
Open-world	3DGenZ [26]	56.0	12.6	35.5	13.3	63.6	6.6
	3DTZSL [8]	36.7	6.1	36.6	2.0	55.5	4.2
	Ovseg-3D [22]	64.4	0.0	55.7	0.1	68.4	0.9
	PLA [12]	68.3	62.4	69.5	45.9	76.2	40.8
	RegionPLC [43] (SparseUnet32)	68.2	70.7	69.9	66.6	76.3	55.6
	ORACLE-3D (SparseUnet32)	**68.4**	**75.7**	**69.9**	**69.8**	**76.6**	**67.2**

Table 3: Open-world evaluation on the SemanticKITTI and Waymo datasets. We compare our method against the fully supervised upper bound.

Dataset	Method	$mIoU_B\uparrow$	$mIoU_N\uparrow$
Waymo [35]	Full Supervision (PTv3)	74.3	81.8
	ORACLE-3D (PTv3)	74.0	75.4
SemanticKITTI [1]	Full Supervision (PTv3)	57.7	77.4
	ORACLE-3D (PTv3)	57.4	62.3

Table 4: Comparison for annotation-free 3D semantic segmentation on the nuScenes dataset.

Method	mIoU↑
CLIP2Scene [5]	20.8
OpenScenes [28]	42.1
MSeg Voting [21]	31.0
2D-3D Projection	28.6
ORACLE-3D	**56.9**

trailing the fully supervised upper bound by only 6.4 in novel class mIoU. On SemanticKITTI, despite the inherent limitation of restricted forward-facing 120° sensor coverage compared to full 360° supervision, our method remains competitive, validating its applicability across diverse data acquisition modalities.

4.4 Annotation-free Open-world Results

We evaluate the annotation-free capabilities of our framework on the nuScenes validation set comprising 16 categories. By freezing the image branch initialized with GLEE weights, we transfer 2D open-world knowledge to the 3D domain via distillation and matching, entirely without relying on 2D or 3D annotations. As shown in Table 4, we benchmark against leading zero-shot methods including CLIP2Scene, OpenScene, and MSeg Voting. ORACLE-3D demonstrates superior performance, surpassing CLIP2Scene by 36.1, OpenScene by 14.8, and MSeg Voting by 25.9, thereby validating the effectiveness of our cross-modal transfer strategy in the absence of supervision.

4.5 Qualitative Analysis

Open-world Detection Capabilities As illustrated in Fig. 3, our method effectively localizes novel categories while preserving high recognition accuracy for base classes. By synergizing robust 2D priors with adaptive 3D filtering, the

framework suppresses projection-induced artifacts and refines object boundaries. Notably, the model demonstrates resilience in detecting diminutive or distant targets, such as traffic cones and pedestrians, despite the challenge of limited point density.

Segmentation Performance Fig. 4 further validates the segmentation quality in an annotation-free context, exhibiting strong concordance with ground truth. Our method accurately identifies a diverse spectrum of categories, ranging from ubiquitous background elements like vegetation and sidewalks to long-tail instances such as construction vehicles. We observe minor inter-class confusion between semantically similar categories like terrain and sidewalk, which we attribute to the inherent ambiguity in the broad vocabulary of the pre-trained text encoder. Additional qualitative results and comprehensive visualizations are provided in the Supplementary Material.

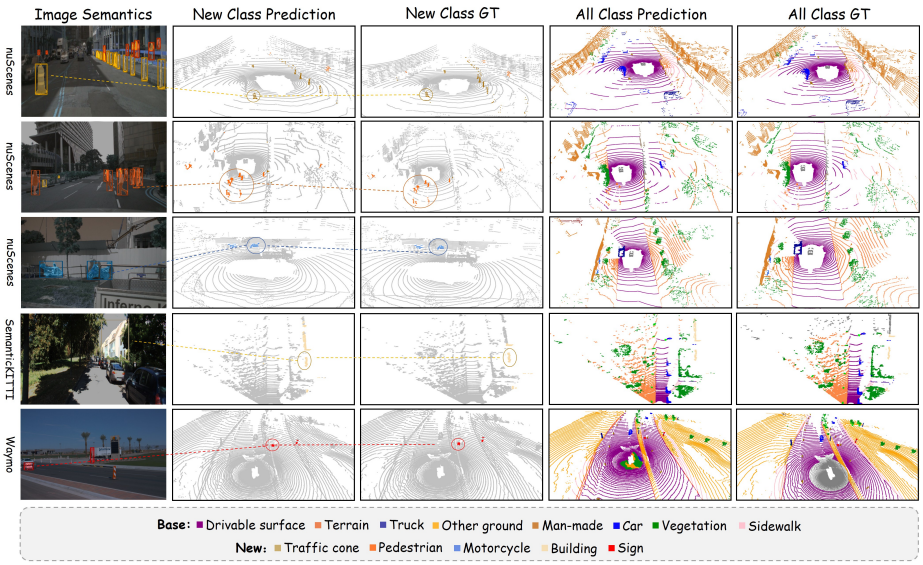

Fig. 3: Qualitative results of the base-annotated task

4.6 Ablation Studies

To thoroughly investigate the contributions of each proposed component and optimization strategy to our framework's overall performance, a series of comprehensive ablation studies was systematically conducted. All experiments were performed on the nuScenes dataset, utilizing the B12/N3 category partition.

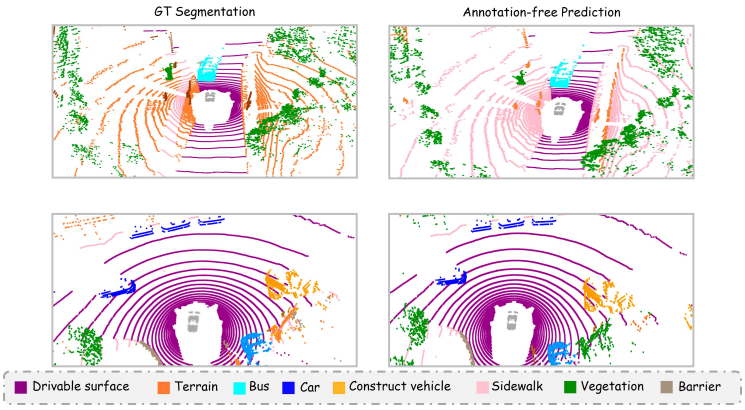

Fig. 4: Qualitative results of the annotation-free task.

Efficacy of the Multi-modal Learning Framework To validate the effectiveness of our multi-modal integration, we benchmark against a naive baseline that directly projects 2D pixel semantics from a pre-trained foundation model onto the 3D point cloud without further refinement. As shown in Table 4, our complete framework significantly outperforms this projection-only approach, achieving a substantial 28.3 mIoU improvement on novel classes. This performance differential confirms the critical role of our integrated design in fusing image, text, and point cloud modalities. It particularly highlights the necessity of a dedicated 3D network to intelligently refine projected semantics and handle the inherent sparsity of point clouds for novel categories.

Table 5: Evaluation of the two-stage optimization strategy and the efficacy of specific loss functions.

One-stage	Two-stage	$\mathcal{L}_{logits}$	$\mathcal{L}_{distill_N}$	$\mathcal{L}_{distill_F}$	$\mathcal{L}_{VPM}$	$\text{mIoU}_B\uparrow$	$\text{mIoU}_N\uparrow$
✓		✓				73.2	57.3
✓		✓	✓	✓	✓	74.1	58.2
	✓	✓				76.6	60.1
	✓	✓	✓			76.4	62.6
	✓	✓	✓	✓		76.8	65.2
	✓	✓	✓	✓	✓	76.9	66.5

Analysis of Alignment and Optimization Strategies We systematically analyze the impact of individual modules and training strategies in Table 5. First, the impact of the training paradigm: Comparing the one-stage and two-stage baselines reveals that the two-stage strategy alone yields a significant

improvement of 3.4 and 2.8 in mIoU on base and novel classes, respectively. When integrating all loss components, the two-stage approach demonstrates even greater efficacy, boosting mIoU on the novel class by a substantial 8.3. This confirms that a progressive training regimen is crucial for stabilizing multi-modal alignment. Second, the contribution of specific losses: Building upon the two-stage baseline, the introduction of the novel class distillation loss increases mIoU on the novel class by 2.5, effectively guiding the model to recognize unseen categories. Subsequently, adding feature distillation contributes a further 2.6 gain in mIoU on the novel class by transferring fine-grained visual semantics. Finally, the VPM loss provides an incremental 1.3 improvement, culminating in our best performance. Collectively, these results affirm the synergistic contributions of each component in our framework. Additional ablation studies are provided in the Supplementary Material.

5 Conclusion and Future Work

In this paper, we presented ORACLE-3D, a unified framework for open-world 3D scene understanding that effectively bridges the modality gap by leveraging images as a semantic intermediary. By circumventing the reliance on expensive 3D-text pairs, our approach enables robust open-vocabulary transfer through multi-level distillation and vision-point matching. Extensive evaluations across nuScenes, SemanticKITTI, Waymo, and ScanNet demonstrate that ORACLE-3D achieves state-of-the-art performance and exhibits strong generalization across diverse backbones and environments.

Looking ahead, we identify two promising directions for future research. First, we aim to enhance the image branch to achieve synergistic 2D and 3D open-world understanding within a single, unified framework for holistic scene comprehension. Second, we plan to extend our architecture by replacing the point cloud branch with an occupancy prediction network, thereby enabling richer, volumetric 3D representations for complex environments.

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9297–9307 (2019)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
3. Cen, J., Yun, P., Zhang, S., Cai, J., Luan, D., Tang, M., Liu, M., Yu Wang, M.: Open-world semantic segmentation for lidar point clouds. In: European Conference on Computer Vision. pp. 318–334. Springer (2022)
4. Chen, R., Liu, Y., Kong, L., Chen, N., Zhu, X., Ma, Y., Liu, T., Wang, W.: Towards label-free scene understanding by vision foundation models. Advances in Neural Information Processing Systems **36**, 75896–75910 (2023)

5. Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., Wang, W.: Clip2scene: Towards label-efficient 3d scene understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7020–7030 (2023)
6. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26428–26438 (2024)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
8. Cheraghian, A., Rahman, S., Campbell, D., Petersson, L.: Transductive zero-shot learning for 3d point cloud classification. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 923–933 (2020)
9. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
11. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36**, 35799–35813 (2023)
12. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: Language-driven open-vocabulary 3d scene understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7010–7019 (2023)
13. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Lowis3d: Language-driven open-world instance-level 3d scene understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
14. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9224–9232 (2018)
15. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
16. Guo, J., Ma, X., Fan, Y., Liu, H., Li, Q.: Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. arXiv preprint arXiv:2403.15624 (2024)
17. He, J., Deng, X., Gui, J., Zhang, T., He, X.: Mdnet: Multimodal cooperative perception via spatial alignment of modal decision-making. IEEE Internet of Things Journal **12**(11), 16142–16154 (2025)
18. Hegde, S., Gangisetty, S.: Pig-net: Inception based deep learning architecture for 3d point cloud segmentation. Computers & Graphics **95**, 13–22 (2021)
19. Jiang, L., Shi, S., Schiele, B.: Open-vocabulary 3d semantic segmentation with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21284–21294 (2024)

20. Keetha, N., Karhade, J., Jatavallabhula, K.M., Yang, G., Scherer, S., Ramanan, D., Luiten, J.: Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21357–21366 (2024)
21. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: Mseg: A composite dataset for multi-domain semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2879–2888 (2020)
22. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)
23. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3041–3050 (2023)
24. Liu, D., Huang, X., Hou, Y., Wang, Z., Yin, Z., Gong, Y., Gao, P., Ouyang, W.: Uni3d-llm: Unifying point cloud perception, generation and editing with large language models. arXiv preprint arXiv:2402.03327 (2024)
25. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
26. Michele, B., Boulch, A., Puy, G., Bucher, M., Marlet, R.: Generative zero-shot learning for semantic segmentation of 3d point clouds. In: 2021 International Conference on 3D Vision (3DV). pp. 992–1002. IEEE (2021)
27. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36** (2024)
28. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 815–824 (2023)
29. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
30. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
32. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024)
33. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
34. Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., Huang, H., Ma, Z., Jiang, X., Chen, Y., et al.: Grounding dino 1.5: Advance the" edge" of open-set object detection. arXiv preprint arXiv:2405.10300 (2024)
35. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous

driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
36. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631 (2023)
37. Tang, G., Agarwal, A., Han, W., Darrell, T., Bai, Y.: Vector quantized feature fields for fast 3d semantic lifting. arXiv preprint arXiv:2503.06469 (2025)
38. Thengane, V., Zhu, X., Bouzerdoum, S., Phung, S.L., Li, Y.: Foundational models for 3d point clouds: A survey and outlook. arXiv preprint arXiv:2501.18594 (2025)
39. Wu, J., Jiang, Y., Liu, Q., Yuan, Z., Bai, X., Bai, S.: General object foundation model for images and videos at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3783–3795 (2024)
40. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler faster stronger. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4840–4851 (2024)
41. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8256–8265 (2019)
42. Yang, J., Ding, R., Deng, W., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding pp. 19823–19832 (2024)
43. Yang, J., Ding, R., Deng, W., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19823–19832 (2024)
44. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23497–23506 (2023)
45. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. Advances in Neural Information Processing Systems **35**, 9125–9138 (2022)
46. Yao, L., Pi, R., Han, J., Liang, X., Xu, H., Zhang, W., Li, Z., Xu, D.: Detclipv3: Towards versatile generative open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27391–27401 (2024)
47. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15244–15253 (2023)
48. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L.H., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: unifying localization and vl understanding. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 36067–36080 (2022)
49. Zhang, J., Dong, R., Ma, K.: Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2048–2059 (2023)
50. Zhang, Y., Jian, Y., Fan, H., Yang, Y., Zimmermann, R.: Uni3d-moe: Scalable multimodal 3d scene understanding via mixture of experts. arXiv preprint arXiv:2505.21079 (2025)

51. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16793–16803 (2022)
52. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21676–21685 (2024)