OWL: Unsupervised 3D Object Detection by Occupancy Guided Warm-up and Large Model Priors Reasoning

Xusheng Guo[1,2], Wanfa Zhang[1,2], Shijia Zhao[1,2], Qiming Xia[1,2], Xiaolong Xie[3], Mingming Wang[4], Hai Wu[5*], Chenglu Wen[1,2*]

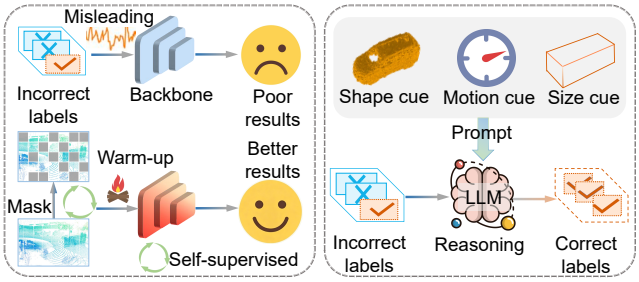
[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China
[3]Tongji University, China
[4]Guangzhou Automobile Group Co. R&D Center, China
[5]Pengcheng Laboratory, China

Abstract

Unsupervised 3D object detection leverages heuristic algorithms to discover potential objects, offering a promising route to reduce annotation costs in autonomous driving. Existing approaches mainly generate pseudo labels and refine them through self-training iterations. However, these pseudo-labels are often incorrect at the beginning of training, resulting in misleading the optimization process. Moreover, effectively filtering and refining them remains a critical challenge. In this paper, we propose **OWL** for unsupervised 3D object detection by occupancy guided warm-up and large-model priors reasoning. OWL first employs an Occupancy Guided Warm-up (OGW) strategy to initialize the backbone weight with spatial perception capabilities, mitigating the interference of incorrect pseudo-labels on network convergence. Furthermore, OWL introduces an Instance-Cued Reasoning (ICR) module that leverages the prior knowledge of large models to assess pseudo-label quality, enabling precise filtering and refinement. Finally, we design a Weight-adapted Self-training (WAS) strategy to dynamically re-weight pseudo-labels, improving the performance through self-training. Extensive experiments on Waymo Open Dataset (WOD) and KITTI demonstrate that OWL outperforms state-of-the-art unsupervised methods by over 15.0% mAP, revealing the effectiveness of our method.

Introduction

3D object detection aims to localize and classify objects within the environment, thereby providing essential contextual information for downstream planning and decision-making. This capability is crucial for ensuring safety in autonomous driving. With the advancement of deep learning, numerous learning-based 3D object detection approaches have emerged (Deng et al. 2021; Yin, Zhou, and Krahenbuhl 2021; Wu et al. 2023a; Wang et al. 2023; Deng et al. 2024; Huang et al. 2025). However, these methods heavily rely on high-quality manual labels, and as the scope of autonomous driving expands, the reliance on ever-larger datasets makes annotation prohibitively expensive and laborious.



(a) Different initialization strategies. (b) LLM reasoning for label filtering.

Figure 1: (a) Illustration of the effects of different backbone initialization strategies. (b) Illustration of how to refine pseudo-labels with reasoning based on instance cues.

To reduce annotation costs while maintaining detector performance, researchers have investigated paradigms that require less annotated data (Guo, Liu, and Song 2022; Zeng et al. 2024) or utilize weaker supervisory signals (Liu et al. 2022; Xia et al. 2023, 2024; Zheng et al. 2025). HSSDA (Guo, Liu, and Song 2022) annotates only a subset of scenes, while CoIn (Xia et al. 2023) and HINTED (Xia et al. 2024) further minimize full-scene annotation to a single object per scene. However, these approaches still depend on manually provided annotations as supervisory signals, and they exhibit significant performance degradation when the labeled data is either limited or of low quality.

Recently, some research has explored unsupervised methods that leverage commonsense knowledge priors to generate pseudo-labels instead of manual annotation (You et al. 2022; Zhang et al. 2023; Baur, Moosmann, and Geiger 2024; Wu et al. 2024). However, these unsupervised methods primarily suffer from two key challenges: (1) *Unstable network initialization*. During initialization, the network is highly susceptible to inaccurate pseudo-labels, which can divert it from the correct convergence trajectory, as illustrated in Figure 1(a). (2) *Pseudo-label filtering and refinement*. In self-training, errors in some pseudo-labels, such as misclassifications and false positives, can accumulate and amplify across

*Corresponding author.

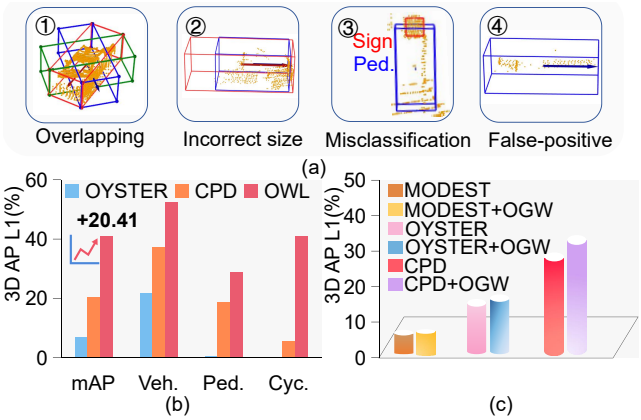


Figure 2: (a) Common errors frequently occur during the process of pseudo-label iteration. (b) The L1 3D AP results of OWL on the WOD test set. (c) The impact of OGW on the performance of unsupervised methods.

iterations (see Figure 2(a)).

In response to these issues, prior work has conducted relevant explorations. For the issue of initialization instability, some studies leverage self-supervised methods, such as masked encoder (Hess et al. 2023) pretraining and contrastive learning (Xia et al. 2023), to learn more robust features. However, these approaches are primarily designed for supervised settings and perform poorly on unsupervised tasks that lack ground-truth. For the second challenge, existing methods filter pseudo-labels with thresholding based on rule-specific or predictive scores (Wu et al. 2024; Xu and Waslander 2023), which may overlook potential objects and remain error-prone.

To address these challenges, this paper proposes an efficient framework, termed **OWL**, to improve the unsupervised 3D detection method with occupancy warm-up and large model priors reasoning. Specifically, OWL first employs an Occupancy-Guided Warm-up (**OGW**) strategy. It leverages an occupancy-prediction proxy task to capture the spatial-structural characteristics of the scene and warm up the network. This initialization strategy equips the network with a preliminary capacity for spatial feature extraction from the very beginning of training, thereby mitigating the adverse impact of erroneous pseudo-labels. Nevertheless, the strategy yields limited improvement when label errors are overly abundant (Figure 2(c)). Consequently, we introduce an Instance-Cued Reasoning (**ICR**) module that harnesses large model prior reasoning. By mining instance-level cues and exploiting the extensive prior knowledge and inferential prowess of large models, the module enables accurate and efficient pseudo-label selection and refinement. Finally, we design a Weight-adapted Self-training (**WAS**) to dynamically re-weight pseudo-labels, refining detection through iterative self-training.

Extensive experiments conducted on the KITTI and WOD datasets demonstrate that our approach surpasses both previous unsupervised and partially weakly supervised methods on the KITTI and WOD datasets, validating the efficiency of our method. The main contributions of this work are as follows:

- We introduce **OWL**, which leverages an occupancy proxy task to warm up the network and harnesses the knowledge and reasoning capabilities of large models to refine pseudo-labels, substantially enhancing the performance of unsupervised 3D object detection.
- We propose the OGW strategy for initializing the unsupervised network so that it possesses a certain perception ability at the beginning of training, mitigating the adverse influence of incorrect labels.
- We devise an ICR module that leverages the prior knowledge and reasoning ability of large models for pseudo-label selection and refinement, thereby markedly elevating the quality of pseudo-labels.

Related Work

Fully/weakly supervised 3D object detection. Fully supervised 3D object detection leverages all human-annotated labels as supervisory signals to regress the bounding boxes and categories of objects within point-cloud scenes. Point-based methods (Shi, Wang, and Li 2019; Qi et al. 2017) extract features by directly processing the raw point cloud, whereas voxel-based approaches (Deng et al. 2021; Yin, Zhou, and Krahenbuhl 2021; Wu et al. 2022) partition the spatial points into voxels for subsequent computation, yielding higher efficiency at the cost of some positional fidelity. Moreover, several studies (Wu et al. 2023b; Huang et al. 2025) enhance detection accuracy through multi-modal data fusion. Despite the strong performance of fully-supervised approaches, they rely on extensive annotations as supervisory signals. Since the annotation cost in full supervision is expensive, some weakly supervised works have attempted to utilize sparse annotations or semantic seed points to achieve the effect of label supervision (Xia et al. 2023; Zhao et al. 2025), reaching performance close to that of full supervision. Unlike the above, we focus more on obtaining a reliable detector without using any manual annotations.

Unsupervised 3D object detection. Unsupervised 3D detection, which does not employ real labels, faces the key challenge of generating high-quality pseudo-labels. Traditional methods that use ground point removal and point cloud clustering (Ester et al. 1996) can produce preliminary pseudo-labels. In recent years, methods such as MODEST (You et al. 2022) and OYSTER (Zhang et al. 2023) have improved pseudo-label quality by incorporating multi-pass knowledge and temporal consistency. LISO (Baur, Moosmann, and Geiger 2024) and CPD (Wu et al. 2024) generate pseudo-labels through motion tracking and common sense prototypes, and enhance the quality of pseudo-labels via iterative self-training. Union (Lentsch, Caesar, and Gavrila 2024) and Motal (Wu et al. 2025) leverage multi-modal images to obtain the correct categories of objects. However, the generated pseudo-labels in these methods often contain an amount of noise and incorrect labels, which severely affect detection performance. This issue is particularly evident in iterative self-training methods. Methods like CPD and LISO

iteratively filter pseudo-labels based on predicted threshold scores. However, the scores from unsupervised learning are inherently noisy, and such filtering and screening methods may miss potential objects and accumulate noise during the iterative process. Therefore, we aim to utilize more latent clues in the scene to reason the rationality of pseudo-labels and optimize the labels during the self-training process.

Self-supervised Learning. Self-supervised learning has emerged as a pivotal paradigm by dispensing with external labels and instead exploiting inherent data structure for representation acquisition. Early endeavors (Gidaris, Singh, and Komodakis 2018) learn discriminative representations via rotation prediction. Contrastive approaches such as CoIn (Xia et al. 2023) and PointContrast (Xie et al. 2020) distinguish positive and negative sample features. CPD (Wu et al. 2024) enhances detection capability by distilling high-quality commonsense prototypes, yet remains vulnerable to noisy pseudo-labels. Masked modeling methods including MAE (He et al. 2022), Voxel-MAE (Hess et al. 2023) and Occupancy-MAE (Min et al. 2023) refine perceptual representations through masked-token reconstruction. Prior unsupervised methods derive features exclusively from noisy pseudo-labels; to remedy this, we introduce the CRW strategy for initializing the network, mitigating the adverse influence of incorrect labels.

Methodology

Initial Label Generation

Motion Artifact Removal. MAR (Wu et al. 2024; Zheng et al. 2025) concatenates the continuous $2n+1$ LiDAR sweeps $\{f_{-n}, \ldots, f_n\}$ into the current frame f_0. Foreground and background points are separated by computing the Persistence Point Score (PP-Score) (You et al. 2022) between successive frames; only the points from the current sweep P_0 and the static points from the other sweeps are retained. After ground removal, this yields the dense point cloud P_0^*.

Dynamic Radius Clustering. DBSCAN (Ester et al. 1996) methods employed a fixed clustering radius, which can lead to inaccurate classification boundaries. Because distant point clouds are sparser and contain fewer instances, the clustering radius must be adjusted accordingly. We thus adopt a dynamic clustering radius:

$$r = \alpha \cdot \left(1 + \beta \cdot e^{-\rho}\right) \cdot r_0, \tag{1}$$

where α, β denote the scaling factor and density–sensitive factor, ρ denotes the density at the radius midpoint, and r_0 denotes the initial preset radius. We then fit bounding boxes using the dynamic clustering radius to train the initial detector. Finally, we feed the point cloud into the detector for inference and obtain the initial pseudo-labels $\mathcal{B}$ after NMS-Selection (Zheng et al. 2025).

Occupancy Guided Warm-Up

Previous unsupervised methods directly train on pseudo-labels, where the label noise degrades the network's feature learning and accumulates during self-training. To learn robust feature representations, we warm up the network weights by a self-supervised strategy.

Specifically, given a dense scene P_0^*, we first apply voxelization, a standard preprocessing step in 3D detection, to convert the scene into a voxel grid $V^{W\times H\times D}$ and thereby markedly enhance computational efficiency. Since the density of point clouds varies with distance, we adopt a strategy similar to Occupancy-MAE(Min et al. 2023), applying a random voxel mask from near to far for each voxel. For each voxel, we calculate the distance from the voxel center to the lidar origin by

$$(x_c, y_c, z_c) = \frac{1}{N}\sum_{i=1}^{N}(x_i, y_i, z_i), \tag{2}$$

$$d = \left\|(x_c, y_c, z_c)\right\|_2. \tag{3}$$

Considering the disparity in the number of foreground and background points and the variation in point cloud density with distance, we employ a dynamic mask strategy as:

$$p(d, w) = w \cdot (0.1 + 0.5 \cdot e^{-0.25\cdot\lfloor d/10\rfloor}), \tag{4}$$

where x, y, z denote the point cloud coordinates, d is the distance from the voxel center to the coordinate origin, N is the number of points in the cloud, and p is the sampling ratio. Specifically, when the region's point cloud is within the bounding box of the pseudo-label $\mathcal{B}$ (potential foreground), $w = w_{fr} = 1$; otherwise, $w = w_{bg} = 0.5$. Then, we perform Bernoulli sampling based on the ratio:

$$\mathcal{M} \sim \text{Bernoulli}\big(p(d)\big) = \{m_j\} \in \{0,1\}^{H\times W\times D}, \tag{5}$$

where $\mathcal{M}$ denotes the set of voxel masks, and each $m \in \mathcal{M}$ is a binary mask for an individual voxel.

For voxels that are not masked, they serve as the training data. The training supervision ground truth is the occupancy status of each voxel $T \in \{0,1\}^{N_f}$, and we conduct occupancy prediction training to predict the probability of each voxel being occupied as:

$$\hat{y} = f_\theta(\mathbf{V} \odot \mathcal{M}) \in \{0,1\}^{H\times W\times D}, \tag{6}$$

where N_f denotes the number of voxels not masked out, T represents the voxel occupancy status, f_θ stands for the backbone network, and $\hat{y}$ signifies the predicted occupancy probability.

Finally, we calculate the occupancy loss $\mathcal{L}_{occ}$. In this way, we essentially warm up the 3D backbone of the detector, enabling it to learn noise-free features and better detect incomplete and distant objects, thereby guiding the learning of noisy unsupervised labels more effectively.

$$\mathcal{L}_{occ} = -\frac{1}{|\Omega|}\sum_{(h,w,d)\in\Omega}\Big[y\log\hat{y} + (1-y)\log(1-\hat{y})\Big]. \tag{7}$$

Instance-Cued Reasoning for label refinement

Self-training is a crucial process for improving the accuracy of weakly supervised and unsupervised methods. However, we observe that performance plateaus after a certain number of self-training iterations. This is because errors in pseudo-labels accumulate over successive iterations, and relying

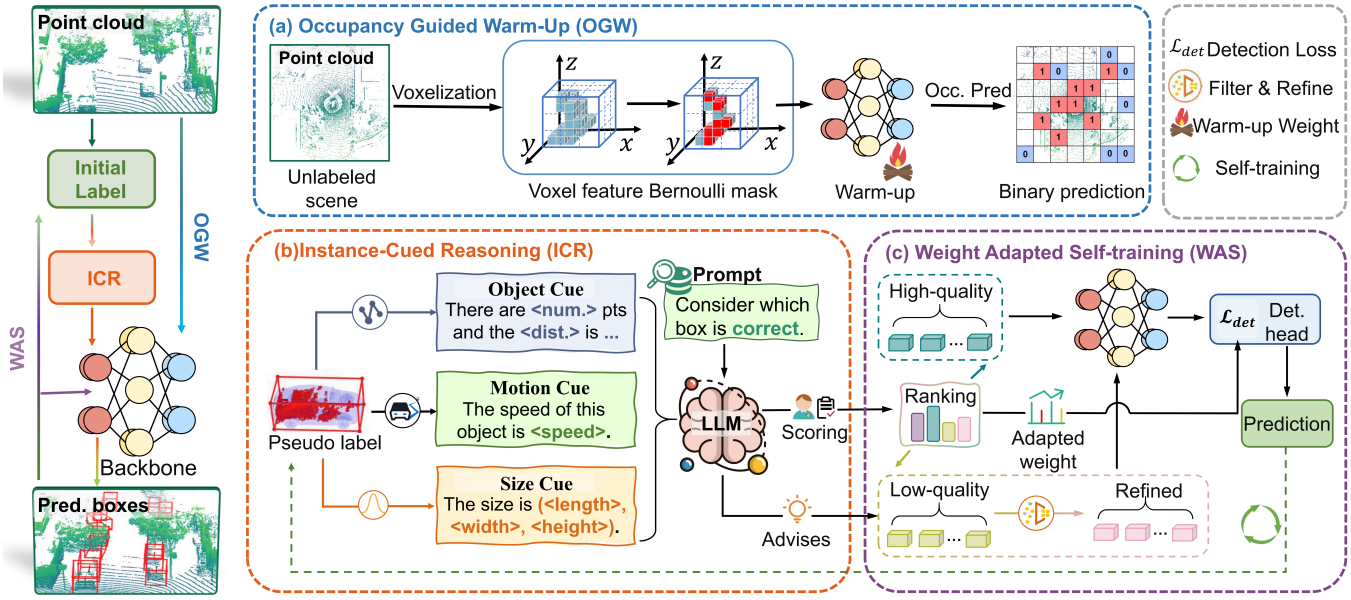


Figure 3: OWL framework. (a) Occupancy Guided Warm-Up (OGW) warms up the 3D backbone, guiding the network to learn spatial and semantic context features of the scene through the self-supervised Occupancy proxy task. (b) Instance-Cued Reasoning (ICR) performs reasoning and judgment on various instant cues based on the knowledge of LLMs, and then filters and refines the pseudo-labels. (c) Weight Adapted Self-training (WAS) adaptively reweights each pseudo-label's loss, letting the model down-weight low-confidence samples.

solely on the OGW module is insufficient to correct these labeling errors. As shown in Figure 2(a), the detector is indeed capable of discovering latent objects; however, the resulting pseudo-labels often suffer from the following issues: (1) Incorrect orientations or overlapping bounding boxes. (2) Inaccurate bounding boxes. (3) Misdetections—including wrong categories and false positives. Previous approaches typically filter pseudo-labels based solely on the model's predicted confidence, which risks discarding potential foreground objects. To address this limitation, we introduce the Instance-Cued Reasoning (ICR) module that leverages the prior knowledge and reasoning ability of large models for pseudo-label selection and refinement.

Instance Cue Mining. After label initialization, we obtain the initial pseudo-label $\mathcal{B} = \{b_j\}_j$, where $b_j = [x, y, z, l, w, h, \alpha, c]_j$ denotes position, length, width, height, orientation and class name respectively of the $j\,th$ box as the prior for the size cue. We apply category-agnostic tracking $\mathcal{T}$ to potential targets based on the detection boxes and adjust the inappropriate sizes of tracked objects and missed targets according to temporal consistency. We also estimate the motion of instances v_j based on this, and this motion cue helps to distinguish between foreground and background targets. Subsequently, we calculate a variety of instance attributes $\mathcal{O}$ for the targets, including the number of points in the point cloud and the average intensity information. Finally, we calculate the prior reference confidence by distribution score s_{dis} and consistency score s_{cons} for each label:

$$s_{dis} = [1 - \mathcal{N}(\|(x_j, y_j, z_j)\|)] + \frac{N_j}{r \times r}, \tag{8}$$

$$s_{cons} = min(0.05, \sum_k \mathcal{S}_k log(\frac{\mathcal{S}_k}{s_k}))/0.05, \tag{9}$$

where $\mathcal{N}(\|(x_j, y_j, z_j)\|)$ denotes the distance normalization of b_j, N_j is the number of grids occupied by points in the box, and r is the resolution. S_k is the size set for each category by LLMs, and $s_j = \{l_j, w_j, h_j\}$ denotes the length, width, and height of the detected bounding box, respectively. Unlike other methods, we use only this score as a cue for reference.

Large Model Priors Reasoning Refiner. After extracting a substantial amount of instance cues, we need to gauge the plausibility of each bounding box based on these cues. Large-scale reasoning models present a promising approach for this task. We feed the label information into a cue reasoner $\mathcal{G}$. For every bounding box $\mathcal{B} = \{b_j\}$, we prompt the discriminator to assess the validity of the pseudo-label:

$$m_k, s_{rea}, \Delta L, cls_{new} = \mathcal{G}(\mathcal{B}, \mathcal{P}, \mathcal{V}, \mathcal{O}), \tag{10}$$

where m_k, s_{rea}, and ΔL denote the box mask, the reasoning score, and the correction vector for width, height, and length, respectively. And $\mathcal{P}$, $\mathcal{V}$, $\mathcal{O}$ denotes the prompt, motion cue and instance cue.

After cue reasoning, we perform filtering and refinement of the pseudo-labels. Bounding boxes b_k that pass the filtering stage have their dimensions and class labels updated

Method	Reference	Vehicle 3D AP@0.7		Pedestrian 3D AP@0.5		Cyclist 3D AP@0.5		mAP	
		L1	L2	L1	L2	L1	L2	L1	L2
MODEST(You et al. 2022)	CVPR2022	6.46	5.80	0.17	0.10	1.14	1.10	2.59	2.33
OYSTER(Zhang et al. 2023)	CVPR2023	14.66	14.10	0.18	0.14	0.33	0.32	5.06	4.85
CPD(Wu et al. 2024)	CVPR2024	37.40	32.13	16.31	13.22	5.06	4.87	19.59	16.74
OWL (Ours)	-	**48.08**	**41.51**	**29.66**	**24.68**	**32.27**	**31.14**	**36.91**	**32.44**
Improvement		*+10.68*	*+9.38*	*+13.35*	*+11.46*	*+27.21*	*+26.27*	*+17.32*	*+15.70*

Table 1: Unsupervised 3D object detection results on WOD validation set. AP@0.7 means the average precision at $IoU_{0.7}$. The best performance is highlighted in bold.

Method	Vehicle AP@0.7				Pedestrian AP@0.5				Cyclist AP@0.5			
	L1		L2		L1		L2		L1		L2	
	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH
MODEST	7.58	7.22	6.57	6.26	0.02	0.01	0.02	0.01	0.00	0.00	0.00	0.00
OYSTER	21.66	20.81	18.79	18.05	0.64	0.32	0.57	0.28	0.01	0.00	0.01	0.00
CPD	37.26	35.00	32.49	30.51	18.65	8.34	16.57	7.40	5.71	3.69	5.55	3.55
OWL (Ours)	**52.53**	**48.22**	**45.97**	**42.19**	**28.69**	**12.54**	**25.53**	**11.13**	**41.52**	**23.42**	**39.92**	**22.52**
Improvement	*+15.27*	*+13.22*	*+13.48*	*+11.68*	*+10.04*	*+4.20*	*+8.96*	*+3.73*	*+35.81*	*+19.73*	*+34.37*	*+18.97*

Table 2: Unsupervised 3D object detection results on WOD test set.

according to the network predictions. Predictions that are filtered out are nevertheless retained only if their consistency score exceeds a threshold η, yet their confidence weights are down-scaled, thereby reducing their contribution during training. L_{com} denotes the average size of the corresponding category provided by the large model.

$$b'_k = \begin{cases} b_k + \Delta L, & m_k = 1, \\ L_{com}, & m_k = 0 \ \wedge \ s_{cons} > \eta, \\ \varnothing, & otherwise. \end{cases} \tag{11}$$

Weight-adapted Self-training

After obtaining the filtered pseudo-labels, we train the detection network F. Following the OGW module, we warm up the 3D backbone of the detector to guide the network toward preliminary learning of noise-free global features. Subsequently, following standard 3D detection practices, we use the pseudo-labels as supervisory signals to regress the object bounding boxes and predict their classes. The regression and classification tasks are optimized using the smooth absolute error loss and the focal loss(Lin et al. 2017), respectively:

$$\mathcal{L}_{\text{total}} = \frac{1}{Nf} \sum_i \omega_i [\alpha \mathcal{L}_{\text{reg}} + \beta \mathcal{L}_{\text{cls}}], \tag{12}$$

where ω_i denotes the weight assigned to each object's pseudo-label, reflecting its quality:

$$\omega_i = \lambda_1 s_{\text{cons}} + \lambda_2 s_{\text{rea}}, \tag{13}$$

N_f, $\mathcal{L}_{reg}$, and $\mathcal{L}_{cls}$ denote the number of samples, regression loss, and classification loss. $\lambda_1 = \lambda_2 = 1$, $\alpha = 2\beta = 1$.

After several training epochs, we perform test-time augmentation (TTA) inference on the training set using the current detector. The resulting predictions undergo a similar filtering and refinement process to generate the initial labels $\mathcal{B}_{\text{ST1}}$ for the next round.

Method	BEV AP		
	Vehicle@0.7	Pedestrian@0.5	Cyclist@0.5
MODEST	13.87	0.10	0.30
OYSTER	23.67	0.20	0.10
CPD	51.56	17.84	4.95
OWL (Ours)	**57.76**	**30.85**	**33.72**

Table 3: BEV results on WOD validation set.

Experiments

Datasets and Metrics

KITTI Dataset. The KITTI (Geiger, Lenz, and Urtasun 2012) dataset covers diverse scenarios such as urban, rural, and highway environments, and is one of the earliest and most widely used datasets in autonomous driving. It is divided into 3,712 frames for training and 3,769 for validation.

Waymo Open Dataset (WOD). The WOD (Sun et al. 2020) consists of 5 RGB cameras and 5 LiDAR sensors, capturing 1,150 scenes, each lasting 20 seconds, and containing over 200k frames of diverse and complex scene data. The WOD is divided into 798 training, 202 validation, and 150 testing sequences, respectively. No annotations were used during the training process.

Comparison with SoTA Methods

Method	Setting	Label rate	3D AP Veh.@0.7	3D AP Ped.@0.5	3D AP Cyc.@0.5	3D AP Veh.@0.4	3D AP Ped.@0.4	3D AP Cyc.@0.4
Centerpoint	Fully.	100%	67.25	68.18	67.40	89.07	80.50	76.49
CoIn	Weakly.	2%	45.04	26.02	62.85	-	-	-
SP3D			43.54	13.27	60.65	-	-	-
MODEST	Unsup.	0%	6.13	0.14	1.12	21.50	0.50	1.40
OYSTER			14.38	0.16	0.33	32.80	0.20	0.00
LISO			-	-	-	54.30	3.70	1.60
CPD			34.77	14.77	4.970	58.90	19.20	5.30
OWL (Ours)			**44.80**	**27.17**	**31.71**	**64.90**	**34.64**	**34.24**

Table 4: Comparison with weakly/fully supervised and unsupervised methods on WOD validation set. 'Fully.', 'Weakly.', and 'Unsup.' mean fully supervised, weakly supervised, and unsupervised settings, respectively.

Method	Modality	Setting	Label ratio	3D AP@IoU0.5 Car	Ped.	Cyc.	Car 3D AP@IoU0.7 Easy	Mod.	Hard
Centerpoint	L	Fully Sup.	100%	94.77	52.11	69.32	89.07	80.50	76.49
Voxel-RCNN	L			96.88	63.73	76.47	92.38	85.29	82.86
CoIn	L	Weakly Sup.	2%	83.89	31.69	49.10	89.17	75.32	62.98
SP3D	L+I			80.37	59.10	77.91	91.12	75.94	66.46
MODEST	L	Unsup.	0%	37.19	1.96	0.10	12.65	11.14	10.60
OYSTER	L			54.58	3.33	1.76	23.22	20.31	19.97
LISO	L			41.10	9.70	5.30	-	-	-
CPD	L			83.89	15.48	8.30	72.98	55.07	53.94
OWL (Ours)	L			**86.10**	**30.23**	**50.98**	**79.80**	**62.67**	**58.11**

Table 5: Comparison with weakly/fully supervised and unsupervised methods on KITTI validation set. 'L' and 'I' denote lidar and image modality.

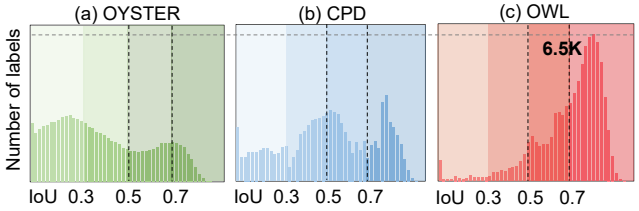


Figure 4: The IoU distribution between pseudo-labels and ground truth is presented.

Method	3D Recall $IoU_{0.5}$	3D Recall $IoU_{0.7}$	3D Precision $IoU_{0.5}$	3D Precision $IoU_{0.7}$
MODEST	12.04	4.89	22.81	10.05
OYSTER	21.01	11.12	21.09	9.45
CPD	39.33	20.54	28.22	14.74
OWL (Ours)	**56.77**	**33.19**	**37.35**	**20.16**

Table 6: Pseudo-label comparison results on WOD validation set.

Comparison with Unsupervised Methods. The comparison results between OWL and unsupervised methods on the WOD validation set and test set are presented in Table 1 and 2, respectively. The results demonstrate that our approach significantly outperforms prior unsupervised SoTA methods. For instance, on WOD validation set, OWL achieves relative improvements of 17.32% and 15.70% in L1 and L2 mAP over CPD. We also report the results of BEV AP on WOD in Table 3. This improvement derives from the design of our ICR, OGW, and WAS modules, which generate more accurate pseudo-labels and a more robust network initialization.

Comparison with Fully/weakly Methods. As KITTI lacks continuous-frame sequences, we follow CPD's protocol: we train on the WOD dataset and evaluate on the KITTI validation set, comparing with a variety of unsupervised and weakly-supervised methods. As shown in Table 4 and 5, OWL surpasses prior unsupervised methods on both KITTI and WOD over 15% mAP, outperforms several weakly-supervised and multimodal approaches, and approaches fully-supervised performance on KITTI, demonstrating the advancement of our method.

Pseudo-label Comparison. To evaluate the quality of pseudo-labels, we analyzed OWL's Recall and Precision performance on the WOD validation set. As shown in Table 6, OWL boosts Recall by 12.65% and Precision by 5.42%

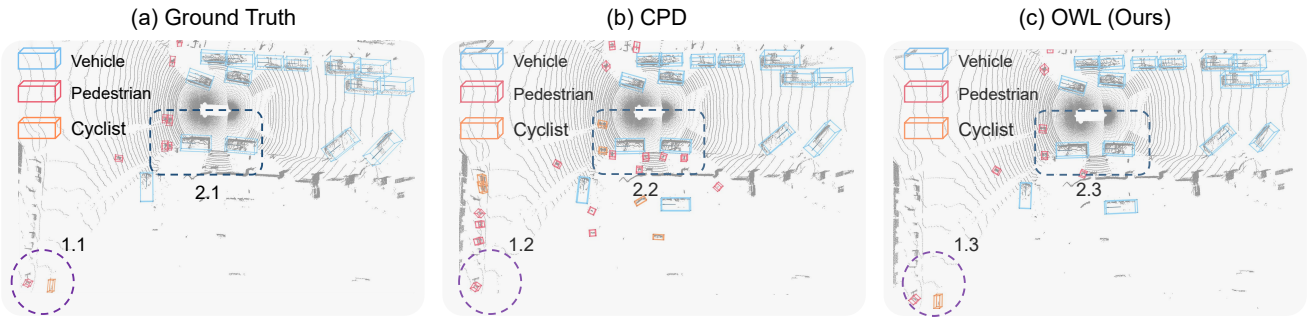


Figure 5: Visualization comparison of different methods.

at $IoU_{0.7}$. Figure 4 displays the IoU distribution between our pseudo-labels and the ground truth, demonstrating that our pseudo-labels outperform those of prior methods. This improvement stems from the effective screening and refinement performed by our ICR module.

Components				3D AP L1	3D AP L2
PLI	ICR	OGW	WAS		
✓				19.59	16.74
✓	✓			30.25	26.38
✓	✓	✓		34.16	30.81
✓	✓	✓	✓	**36.67**	**32.44**

Table 7: Ablation study of our method.

Warm-up methods	L1		L2	
	AP	APH	AP	APH
Train directly	30.25	24.01	26.38	21.16
Contrastive learning	29.10	23.98	25.54	20.97
Point completion	31.06	24.65	27.08	21.53
OGW	**34.16**	**26.63**	**30.81**	**23.40**

Table 8: Comparison with different warm-up methods on WOD validation set.

Ablation Study

Components analysis of OWL. We evaluated each component and assessed its impact using the WOD validation set, with the results shown in Table 7. Compared with previous pseudo-label initialization (PLI), our ICR module yields a 10.66% AP improvement, attributed to leveraging large models and instance-level cues to filter and refine a substantial number of erroneous pseudo-labels. OGW further improves detection performance by 3.91% by guiding stable initialization, mitigating the misleading impact of incorrect labels, and enhancing spatial awareness.

Influence of different warm-up methods. As shown in Table 8, we examine the impact of different warm-up strategies on detector performance. Compared to direct training, contrastive learning underperforms because it typically relies on labels to distinguish between positive and negative samples, whereas point completion yields only a limited gain. This experiment demonstrates the effectiveness of our OGW strategy.

Method	3D mAP		
	[0,30) m	[30,50) m	[50,inf)
OYSTER	15.30	3.26	0.28
CPD	31.49	17.84	5.14
OWL (Ours)	**57.74**	**37.16**	**13.56**

Table 9: Range detection results on WOD validation set.

Performance along different distances. Table 9 presents the detection performance of unsupervised methods across varying distances, showing that OWL achieves superior object detection at longer ranges.

Visualization Analysis

In Figure 5, OWL's improvements are visually demonstrated. Through label filtering and screening, OWL significantly reduces pseudo-label annotation errors, as shown by the marked decrease in False-Negative and False-Positive samples in Figure 5 (2.2, 2.3). Additionally, detection performance for under-represented categories, like Cyclist in Figure 5 (1.3), is significantly enhanced.

Conclusion

In this paper, we propose OWL, a new framework for enhancing unsupervised 3D object detection. We introduce the OGW module to guide network initialization, mitigating the impact of incorrect labels during early training stages. To address the prevalence of noisy pseudo-labels, we propose the ICR module to extract instance cues and leverage the knowledge and reasoning capabilities of large models to refine the pseudo-labels. Finally, we design the WAS strategy to learn the loss adaptively. These designs mutually reinforce pseudo-label quality to improve the detection performance.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.42571514).

References

Baur, S. A.; Moosmann, F.; and Geiger, A. 2024. Liso: Lidar-only self-supervised 3d object detection. In *European Conference on Computer Vision*, 253–270. Springer.

Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; and Li, H. 2021. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1201–1209.

Deng, J.; Ye, W.; Wu, H.; Huang, X.; Xia, Q.; Li, X.; Fang, J.; Li, W.; Wen, C.; and Wang, C. 2024. Cmd: A cross mechanism domain adaptation dataset for 3d object detection. In *European Conference on Computer Vision*, 219–236. Springer.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Guo, X.; Liu, R.; and Song, D. 2022. HSSDA: Hierarchical relation aided Semi-Supervised Domain Adaptation. *AI Open*, 3: 156–161.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hess, G.; Jaxing, J.; Svensson, E.; Hagerman, D.; Petersson, C.; and Svensson, L. 2023. Masked autoencoder for self-supervised pre-training on lidar point clouds. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 350–359.

Huang, X.; Xu, Z.; Wu, H.; Wang, J.; Xia, Q.; Xia, Y.; Li, J.; Gao, K.; Wen, C.; and Wang, C. 2025. L4dr: Lidar-4dradar fusion for weather-robust 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3806–3814.

Lentsch, T.; Caesar, H.; and Gavrila, D. 2024. Union: Unsupervised 3d object detection using object appearance-based pseudo-classes. *Advances in Neural Information Processing Systems*, 37: 22028–22046.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, C.; Gao, C.; Liu, F.; Liu, J.; Meng, D.; and Gao, X. 2022. Ss3d: Sparsely-supervised 3d object detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8428–8437.

Min, C.; Xiao, L.; Zhao, D.; Nie, Y.; and Dai, B. 2023. Occupancy-mae: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders. *IEEE Transactions on Intelligent Vehicles*, 9(7): 5150–5162.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–779.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Wang, H.; Shi, C.; Shi, S.; Lei, M.; Wang, S.; He, D.; Schiele, B.; and Wang, L. 2023. Dsvt: Dynamic sparse voxel transformer with rotated sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13520–13529.

Wu, H.; Deng, J.; Wen, C.; Li, X.; Wang, C.; and Li, J. 2022. CasA: A cascade attention network for 3-D object detection from LiDAR point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.

Wu, H.; Lin, H.; Guo, X.; Li, X.; Wang, M.; Wang, C.; and Wen, C. 2025. Motal: Unsupervised 3D Object Detection by Modality and Task-specific Knowledge Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6284–6293.

Wu, H.; Wen, C.; Li, W.; Li, X.; Yang, R.; and Wang, C. 2023a. Transformation-equivariant 3d object detection for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2795–2802.

Wu, H.; Wen, C.; Shi, S.; Li, X.; and Wang, C. 2023b. Virtual sparse convolution for multimodal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21653–21662.

Wu, H.; Zhao, S.; Huang, X.; Wen, C.; Li, X.; and Wang, C. 2024. Commonsense prototype for outdoor unsupervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14968–14977.

Xia, Q.; Deng, J.; Wen, C.; Wu, H.; Shi, S.; Li, X.; and Wang, C. 2023. Coin: Contrastive instance feature mining for outdoor 3d object detection with very limited annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6254–6263.

Xia, Q.; Ye, W.; Wu, H.; Zhao, S.; Xing, L.; Huang, X.; Deng, J.; Li, X.; Wen, C.; and Wang, C. 2024. Hinted: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15321–15330.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, 574–591. Springer.

Xu, J.; and Waslander, S. L. 2023. HyperMODEST: Self-Supervised 3D Object Detection with Confidence Score Filtering. In *2023 20th Conference on Robots and Vision (CRV)*, 217–224. IEEE.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11784–11793.

You, Y.; Luo, K.; Phoo, C. P.; Chao, W.-L.; Sun, W.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2022. Learning to detect mobile objects from lidar scans without labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1130–1140.

Zeng, S.; Zheng, W.; Lu, J.; and Yan, H. 2024. Hardness-aware scene synthesis for semi-supervised 3D object detection. *IEEE Transactions on Multimedia*, 26: 9644–9656.

Zhang, L.; Yang, A. J.; Xiong, Y.; Casas, S.; Yang, B.; Ren, M.; and Urtasun, R. 2023. Towards unsupervised object detection from lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9317–9328.

Zhao, S.; Xia, Q.; Guo, X.; Zou, P.; Zheng, M.; Wu, H.; Wen, C.; and Wang, C. 2025. SP3D: Boosting Sparsely-Supervised 3D Object Detection via Accurate Cross-Modal Semantic Prompts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29374–29384.

Zheng, M.; Xu, Z.; Xia, Q.; Wu, H.; Wen, C.; and Wang, C. 2025. Seg2Box: 3D Object Detection by Point-Wise Semantics Supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10591–10598.