Seg2Box: 3D Object Detection by Point-Wise Semantics Supervision

Maoji Zheng[1,2], Ziyu Xu[1,2], Qiming Xia[1,2], Hai Wu[1,2], Chenglu Wen[1,2]*, Cheng Wang[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
{zhengmaoji, xuziyu, xiaqiming, wuhai}@stu.xmu.edu.cn, {clwen, cwang}@xmu.edu.cn

Abstract

LiDAR-based 3D object detection and semantic segmentation are critical tasks in 3D scene understanding. Traditional detection and segmentation methods supervise their models through bounding box labels and semantic mask labels. However, these two independent labels inherently contain significant redundancy. This paper aims to eliminate the redundancy by supervising 3D object detection using only semantic labels. However, the challenge arises due to the incomplete geometry structure and boundary ambiguity of point-cloud instances, leading to inaccurate pseudo-labels and poor detection results. To address these challenges, we propose a novel method, named *Seg2Box*. We first introduce a Multi-Frame Multi-Scale Clustering ***(MFMS-C)*** module, which leverages the spatio-temporal consistency of point clouds to generate accurate box-level pseudo-labels. Additionally, the Semantic-Guiding Iterative-Mining Self-Training ***(SGIM-ST)*** module is proposed to enhance the performance by progressively refining the pseudo-labels and mining the instances without generating pseudo-labels. Experiments on the Waymo Open Dataset and nuScenes Dataset show that our method significantly outperforms other competitive methods by 23.7% and 10.3% in mAP, respectively. The results demonstrate the great label-efficient potential and advancement of our method.

Introduction

LiDAR-based 3D object detection and 3D semantic segmentation are widely applied in fields of autonomous driving (Zhu et al. 2024; Mao et al. 2023), robotics (Wisth, Camurri, and Fallon 2022; Beuchert, Camurri, and Fallon 2023) and smart cities (Wang et al. 2020, 2019). Traditional 3D object detection and semantic segmentation frameworks rely on their unique labels to supervise learning processes. However, the annotations for semantic masks and bounding boxes inherently contain considerable redundancy, as they both implicitly convey the semantic content and geometric structure of instances.

An intuitive solution to eliminate this redundancy is to use only bounding box labels and assign the semantic labels automatically to supervise the semantic segmentation task. Several studies have investigated the technique of semantics assignment. For example, Box2Mask (Chibane et al. 2022) assigns semantics to each foreground point through instance clustering and box voting. Box2seg (Kulharia et al. 2020) uses attention-based self-training and point class activation mapping to avoid using semantic labels. However, background information is missing in the bounding box annotation. In addition, bounding boxes mainly focus on a coarse level of instance layout. Using only bounding boxes to assign semantic labels inevitably introduces supervision noise, significantly decreasing the segmentation performance.

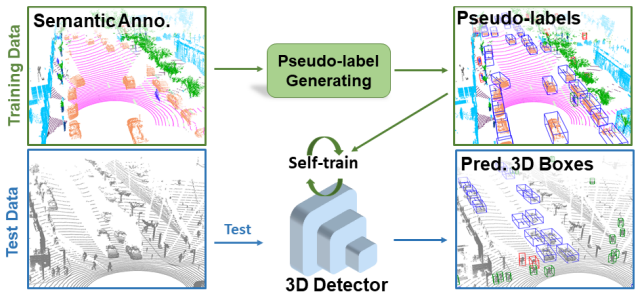

Figure 1: Our method uses only semantic annotation to train 3D object detection. It contains Pseudo-label Generation stage and Self-train Loop-improvement stage.

Compared to the coarse bounding box, the semantic label is more precise and detailed. Therefore, using only semantic labels possibly attains accurate detection and segmentation results and decreases label redundancy remarkably. However, the objects in 3D scenes are primarily sparse and self-occluded. Nowadays, how to recover the bounding boxes from the semantic labels to train the 3D detectors still remains a great challenge.

This paper mainly investigates two challenges of semantic label-supervised 3D object detection. (1) The incomplete geometry structure resulting from sparse point clouds (Fig. 2 - ①) and occlusion (Fig. 2 - ②) leads to pseudo-labels with erroneous size and position. (2) The boundary ambiguity arising from the truncated objects (Fig. 2 - ③) and the adjacent objects (Fig. 2 - ④) results in pseudo-labels with false boundaries. Consequently, the detector trained with the low-quality pseudo-labels can not achieve desirable performance.

To address these challenges, we develop a two-stage 3D

*Corresponding author.

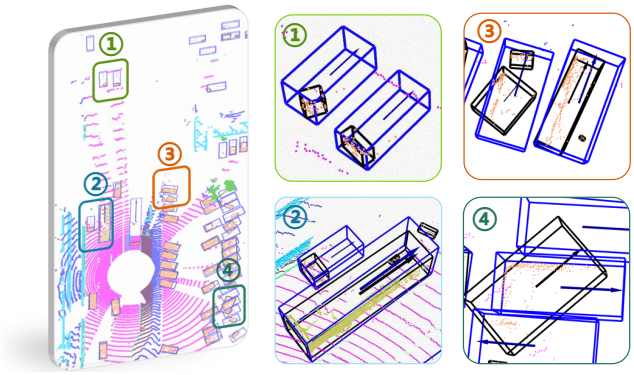

Figure 2: The challenges of generating box-level pseudo-labels from semantic labels. Blue boxes are the ground truth. And, black boxes are pseudo-labels generated by direct DBSCAN clustering. Points with different colors indicate different categories of objects. ①, ②: Incomplete objects due to sparse point cloud and occlusion. ③: Clustering one instance into multiple due to the truncated object. ④: Clustering multiple instances into one due to adjacent objects.

object detection method, named **Seg2Box**. In the first stage, we introduce a novel Multi-Frame Multi-Scale Clustering **(MFMS-C)** module to generate accurate pseudo-labels. MFMS-C adopts a density-based spatial clustering design and then fits bounding boxes to clustering results (Zhang et al. 2023; Wu et al. 2024). The key idea is to select the best one from a set of pseudo-label candidates. Specifically, MFMS-C first generates the pseudo-label candidates by clustering with different radii. MFMS-C then selects the best pseudo-label depending on a newly designed Meta Shape and Fitting Score (MSF-Score), which measures the pseudo-label quality from the aspect of completeness, distribution of points, and shape. In addition, the static instances in consecutive frames tend to be more complete, allowing the creation of accurate pseudo-labels with precise position and size. In the second stage, we develop a Semantic-Guiding Iterative-Mining Self-Training **(SGIM-ST)** module to enhance the detection performance. Specifically, We iteratively mine the miss annotated instances and refine the pseudo-labels using spatio-temporal consistency.

We validated our method on widely used Waymo Open Dataset (WOD) (Sun et al. 2020) and nuScenes Dataset (Caesar et al. 2020). Our method significantly outperforms baseline methods, achieving a 23.7% improvement in mAP L2 on WOD. Remarkably, compared to fully supervised methods by bounding boxes, our approach reached close to 95% accuracy in vehicle AP L2 with an IoU threshold of 0.5 while using only semantic labels as supervision.

The main contributions of this work include: (1) We proposed the first method for 3D object detection that relies solely on semantic label supervision. This innovation eliminates the redundancy between bounding box labels and semantic labels, offering a viable approach to reducing human labeling costs. (2) We proposed **MFMS-C** module for pseudo-label generation which significantly improves the accuracy of pseudo-labels. (3) We proposed the **SGIM-ST** module, which significantly enhances the detection performance by iteratively mining instances without pseudo-labels and refining pseudo-labels.

Related Work

3D Object Detection from Point Cloud. In 3D object detection, fully supervised methods (Yin, Zhou, and Krahenbuhl 2021; Li et al. 2023; Wu et al. 2022; Xia et al. 2023a) have been extensively researched and display outstanding performance. However, labor-intensive and time-consuming manual annotation limits their wide application. Weakly supervised methods aim to reduce the annotation burden by annotating only selected frames or instances. They identify unlabeled instances through teacher-student frameworks (Wang et al. 2021; Han et al. 2024) or feature-level instance mining (Xia et al. 2023b, 2024). Recently, unsupervised methods have been explored for the 3D object detection (Zhang et al. 2023; Wu et al. 2024). Despite these weakly supervised and unsupervised strategies vastly reducing the annotation cost, they are hard to obtain optimal performance. The primary obstacle is the insufficiency of information. Our method introduces cross-task supervision to reduce the annotation cost while ensuring sufficient supervised information.

Cross-Task Supervision. Cross-task supervision is a strategy that leverages shared knowledge across different tasks to train or enhance specific models. In 2D vision, Boxsup (Dai, He, and Sun 2015) employs box annotations to supervise the training of semantic segmentation models, effectively bridging the gap between 2D detection and segmentation. In 3D vision, Box2Mask (Chibane et al. 2022) uses bounding box voting and instance clustering to assign semantics to points inside boxes for semantic segmentation. However, it neglects background instances not labeled by the detection task. To address this issue, Box2Seg (Kulharia et al. 2020) introduces additional subcloud-level tags to estimate background pseudo-labels but get poor performance. Moreover, bounding boxes mainly focus on a coarse level of instance layout. Using only bounding boxes to assign semantic labels inevitably introduces supervision noise. Therefore, our method leverages more precise and detailed semantic labels to supervise 3D object detection. It not only attains accurate detection results but also decreases the label redundancy remarkably.

Pseudo-Label Generation for 3D Object Detection. Pseudo-label generation for 3D object detection is to estimate latent boxes for unlabeled data. Recent label-efficient methods (Xia et al. 2023b; Zhang et al. 2024; Liu et al. 2022) generate pseudo-labels by a pre-trained model to supplement the supervision data. WS3D (Qin, Wang, and Lu 2020) proposes an unsupervised 3D object proposal module to select high-confident boxes from 3D anchors. However, fixed-size anchors limit their effect and usage scenarios. Nowadays, unsupervised methods introduce the point-distribution-based strategy to generate pseudo labels. Modest (You et al. 2022) distinguishes dynamic instances from

multi-trip LIDAR sequences and then estimates the pseudo-labels by DBSCAN (Ester et al. 1996) clustering. Prototype and unsupervised tracking have been introduced by oyster (Zhang et al. 2023) and cpd (Wu et al. 2024) to refine the pseudo-labels.

Method

This paper introduces Seg2Box, a novel method for 3D object detection that uses only semantic labels as supervision. As illustrated in Fig.3, Seg2Box consists of two key stages: (1) Pseudo-labels generation by the MFMS-C module and (2) loop improvement by the SGIM-ST module.

Multi-Frame Multi-Scale Clustering for Pseudo-Label Generation

Traditional methods (Zhang et al. 2023; Ester et al. 1996) estimate box-level pseudo-labels from points by single radius-based DBSCAN (Zhang et al. 2023; Ester et al. 1996). However, these approaches fail to recover the labels of adjacent and truncated objects. The key reason is that the single radius-based clustering ignores the diversity of object density and location. In general, a small clustering radius is required for adjacent objects, while a large radius is needed for truncated or sparse objects. How to accurately estimate the bounding boxes for all objects is still an unsolved challenge. We observe that clustering semantic points with multiple radii to construct a set of candidate boxes, and then selecting the best one by Non-Maximum Suppression(NMS), can significantly avoid false boundaries. Additionally, stationary objects in consecutive frames with multi-angle scans typically have more complete structures (Wu et al. 2024). Those observations motivate us to design MFMS-C module to generate accurate pseudo-labels. Specifically, MFMS-C consists of Motion Artifact Removal, Multi-scale Clustering and NMS-Selection.

Motion Artifact Removal. Directly register continuous $2n+1$ frames $\{f_{-n}, ..., f_0, ..., f_n\}$ (i.e., past n, future n frames, and current frame) to build dense point cloud f_0^*, resulting in artifacts from moving objects, which will affect the accuracy of pseudo-labels. To issue this problem, we first divide the visible area of f_0^* into grids G from the BEV perspective. Then, for each grid $G_{i,j}$, we count its maximal time $G_{i,j}^{\ t}$ continuously occupied by foreground points. If $G_{i,j}^{\ t}$ is less than a certain threshold ε (related to the number of concatenated frames), $G_{i,j}$ is determined to be a moving area A_{mov}, otherwise, to be a static area A_{sta}. We then keep all the foreground points in current frame f_0 and remove the points in moving areas A_{mov} of other frames $f_{-n}, ..., f_{-1}, f_1, ..., f_n$ to found dense point cloud f_0^*.

Multi-Scale Clustering (MSC). We follow the idea of density-based spatial clustering DBSCAN (Ester et al. 1996), bounding box fitting (Zhang et al. 2017) on f_0^* to generate a set of bounding boxes $\bar{b}$. However, traditional methods use a single radius for clustering, which causes mistaken instance division, resulting in false boundaries. To address this issue, we design our MSC module. For detail, we sample each radius r from candidate radii $CAND_r$ to cluster f_0^* and gain the initial proposals $\hat{b}_i$. After that, we concentrate all $\hat{b}_i$ to build bounding box candidates $\hat{\mathcal{B}}$ as the input of the next NMS-Selection module.

Non-Maximum Suppression Selection (NMS-Selection). In the previous steps, we generate a lot of box candidates $\hat{\mathcal{B}}$ by MSC. However, each instance should only remain the best candidate at last. In object detection, NMS is widely used to eliminate redundant prediction boxes (Neubeck and Van Gool 2006). It suppresses the low-quality boxes through confident scores outputs by the network. Ground truth is needed to train the network, which is not available for our work. In contrast, we introduce a newly designed Meta Shape and Fitting Score (MSF-Score) to evaluate the quality of pseudo-labels from the aspect of object completeness, distribution of points, and shape. Intuitively, good scoring should keep consistent with IoU scoring. As shown in Fig. 4 (d), with increasing MSF-Score, the pseudo-label has a larger IoU with ground truth. MSF-Score consists of three sub-scores: Occupancy Score, Alignment Score, and Meta Shape Score.

Occupancy (OCC) Score. Assuming the pseudo-label from a complete object is likely to be more accurate. As shown in Fig. 4 - (a), we first divide the proposal box into grids with resolution $r \times r$ in BEV perspective (Wu et al. 2024). The occupancy score indicates the proportion of grids occupied by foreground points. $S_o(b)$ is calculated as follow:

$$S_o(b) = \frac{O}{r \times r}, \tag{1}$$

where O is the number of grids occupied by foreground points, and r is the resolution.

Alignment (ALG) Score. Due to the nature of LiDAR scanning, points tend to be concentrated on the surface of instances. Therefore, most of the points of the high-quality box should be close to the boundary (Fig. 4 - (b)). We assume a box b with angle α. To evaluate b, we first look for the area with the highest point density in BEV perspective and fit a linear line for the points P in this area, the angle of the line is θ. With an accurate pseudo-label, θ should align with α when points are concentrated along the length of the box, and it should be vertical to α when points are concentrated along the width of the box. Motivate by this observation, we calculate the Alignment Score $S_a(b)$ as follow:

$$S_a(b) = \begin{cases} 1 - sin(|\alpha - \theta|), if|\alpha - \theta| < \frac{\pi}{2} \\ 1 - sin(|\alpha + \frac{\pi}{2} - \theta|), otherwise \end{cases} \tag{2}$$

Meta Shape (MS) Score. In general, instances of the same category tend to have similar size ratios and fall within a certain size range (Fig. 4 - (c)). For each class C, we first construct its meta-shape $\mathcal{B} = \{l, w, h\}$, where l, w and h are the length, width and height, respectively. For these coarse-grained statistics, we can directly use category size information online (Luo et al. 2024). Shape Score $S_{ms}(b)$ is calculated as follow:

$$S_{ms}(b) = \begin{cases} 0 & , b \le 0.5 * \mathcal{B} \text{ or } b \ge 2 * \mathcal{B} \\ \min\left(0.05, \sum_k \mathcal{B}_k \log\left(\frac{\mathcal{B}_k}{b_k}\right)\right)/0.05, & else \end{cases} \tag{3}$$

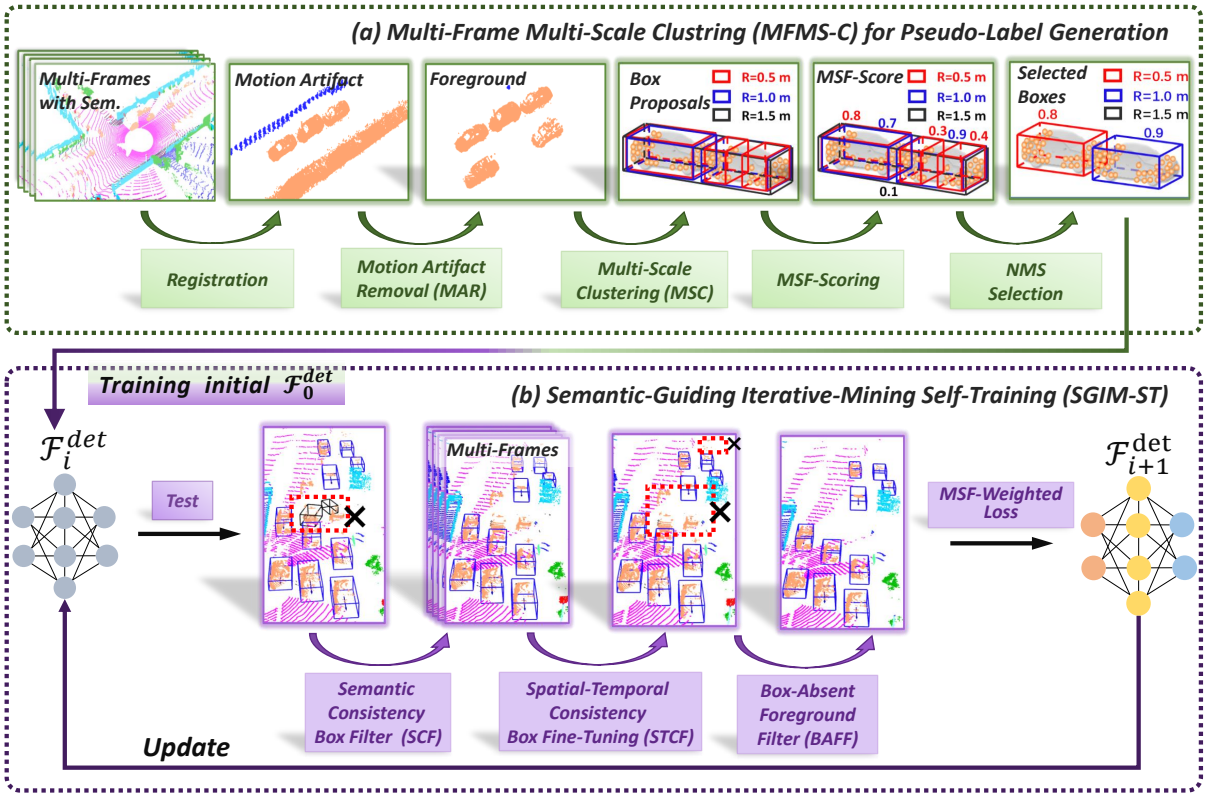

Figure 3: Illustration of Seg2Box framework. (a) MFMS-C generates box-level pseudo-labels from semantic points to train the initial detector $\mathcal{F}_0^{det}$. To address the challenges of pseudo-label generation due to incomplete geometry structure and boundary ambiguity, MFMS-C first generates numerous box proposals in consecutive frames using MSC. After that, NMS Selection remains the high-quality proposals depending on MSF-Scoring which measures the quality of pseudo-labels. (b) SGIM-ST enhances detection performance by iteratively mining the miss annotated instances and refining the pseudo-labels through SCF, STCF, BAF, and MSF-Weighted Loss.

it indicates the ratio difference between $\mathcal{B}$ and b (Wu et al. 2024), and b is the proposal box size $\{\hat{l}, \hat{w}, \hat{h}\}$.

Finally, We obtain the complete MFS-Score by linearly combining these three sub-scores:

$$MSF(b) = \begin{cases} \lambda_1 S_o(b) + \lambda_2 S_a(b) + \lambda_3 S_{ms}(b), \\ \lambda_1 + \lambda_2 + \lambda_3 = 1, \end{cases} \quad (4)$$

where $\lambda_i, i = \{1, 2, 3\}$ control the weight of each sub-score.

For each b in box candidates $\hat{\mathcal{B}}$, we compute its MSF-Score $S_{\hat{\mathcal{B}}}^{MSF}$. Combined with NMS, we suppress the low-quality pseudo-labels and keep the best candidate $\hat{\mathcal{B}}_{final}$ with the highest MSF-Score to train the initial detection model.

Semantic-Guiding Iterative-Mining Self-Training

Self-training enhances the performance of the model by employing the output from the previous iteration as the input of the next iteration until the model converges. Due to its ability to improve detection accuracy, self-training has been introduced to 3D object detection, especially for weakly supervised methods (Liu et al. 2022; Luo et al. 2024; Zhang et al. 2023). However, there are still some issues that need to be considered in weakly supervised 3D detection self-training. (1) The instances without generating pseudo-label would be regarded as the background during training (Meng et al. 2021). (2) False classification (False positive) of similar objects is difficult to filter out by the traditional score-threshold method. (3) The detection results of instances of sparse scanning are often unreliable. (4) Despite refinement, some pseudo-labels are still inaccurate. These issues may mislead the model training process and accumulate with the increase of training iterations, which causes poor detection performance. To address these issues, we introduce our SGIM-ST module. SGIM-ST consists of four designs, which are described in detail below.

Box-Absent Foreground Filter (BAF). Foreground points of objects without pseudo-labels will be falsely regarded as background, which will mislead the model training. Since our supervision signal is semantic labels, foreground points and background points are already distinguishable. For each foreground point p, if p is not inside any of the pseudo-labels, we filter it out from the point cloud. As self-training iterations progress, more and more precise pseudo-labels are mined, the filtered foreground points

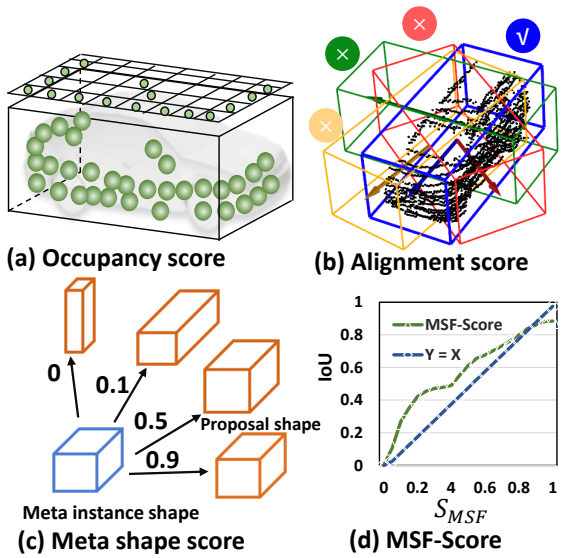

Figure 4: Meta Shape and Fitting Score (MSF-Score).

gradually decrease, and the scene becomes more and more complete.

Semantic Consistency Box Filter (SCF). False classification (False positive) of similar objects is a prevalent issue in 3D object detection. Traditional score-based filtering method often struggles to filter these errors. More critically, these errors can accumulate during iterative training, leading to reduced detection accuracy. For each predicted box b^*, we first extract the points P_{in} inside b^*. We then filter out b if the predicted category of b is inconsistent with the semantics of P_{in} or if multiple types of foreground semantic labels are found in P_{in}.

Spatial-Temporal Consistency Box Fine-Tuning (STCF). The quality of the predicted box decreases as the distance to the ego-car increases (Mao et al. 2023). And, the imprecise pseudo-labels may mislead the training process. Our key idea is to broadcast the near-range high-quality predictions of static objects to other frames since stationary objects should be consistent between frames. We have already segmented the visual area into the static area A_{sta} and the moving area A_{mov} in the first stage. Then, the predicted boxes $\mathcal{B}$ in A_{sta} are transformed into the global coordinate. After that, we use the NMS-Selection proposed in MSMF-C to score $\mathcal{B}$ and remain the best proposal $\mathcal{B}_{best}$ with the highest score. After that, we broadcast $\mathcal{B}_{best}$ back to a single frame by determining whether any foreground points in the single frame are inside $\mathcal{B}_{best}$. In this way, we refine the pseudo-labels for static objects even with spare points in some frames. We give up refining the moving objects since the poor tracking results.

MSF-Weighted Detection Loss (MSFLoss). To suppress the false supervision of label noise caused by inaccurate pseudo-labels, we assign different training loss weights to instances according to pseudo-label quality. Formally, we calculate ω_i based on the MSF-Score s_i^{msf} of pseudo-label:

$$\omega_i = \begin{cases} 0, & s_i^{msf} \le \theta_L \\ \frac{s_i^{msf} - \theta_L}{\theta_H - \theta_L}, & \theta_L < s_i^{msf} < \theta_H \\ 1, & s_i^{msf} \ge \theta_H \end{cases} \tag{5}$$

where θ_L and θ_H are high-quality and low-quality thresholds respectively. The final MSF-weighed detection loss is calculated as follows:

$$\mathcal{L}_{MSF} = \frac{1}{N} \sum_{i=1}^{N} \omega_i (\mathcal{L}_i^{hm} + \mathcal{L}_i^{reg}), \tag{6}$$

where N is the number of proposals, $\mathcal{L}_i^{hm}$, $\mathcal{L}_i^{reg}$ are heapmap loss and regression loss (Yin, Zhou, and Krahenbuhl 2021) between pseudo-labels and predictions.

Experiments

Datasets and Metricses

Waymo Open Dataset (WOD). For bounding box annotation, WOD (Sun et al. 2020) contains a total of ∼158k LiDAR frames for training and ∼40k LiDAR frames for validation. For semantic annotation, WOD annotates one frame every 7 frames on the top LiDAR scan, resulting in ∼23k frames for training. And, all our experiments are conducted on the frames with semantic annotation and following the official evaluation metrics.

NuScenes Dataset. nuScenes (Caesar et al. 2020) is a more challenging dataset since sparser scans and more categories. It contains 1,000 sequences, with 700, 150, and 150 for training, validation, and testing, respectively. nuScenes provides new metrics for 3D detection called NDS which comprehensively calculates mAP, Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). However, we ignore the last two submetrics, as they can't be obtained from semantic labels.

Implementation Details

In the pseudo-label generation stage, we used grid size $r = 7$ for Eq.1 to calculate the Occupancy-Score of the pseudo-label. We used $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$ in Eq.4 for the weights of MSF-Score. We used $\theta_L = 0.4$ and $\theta_H = 0.8$ in Eq.5 to calculate the loss weight of each pseudo-label. In the SGIM-ST stage, We used CenterPoint (Yin, Zhou, and Krahenbuhl 2021) as our based architecture and adopted the implementation of publicly available code from OpenPCDet (Team 2020) for all experiments. We trained both Waymo Open Dataset and nuScenes Dataset for 30 epochs and selected the best validation accuracy epoch as a result. All those experiments were trained on 2 NVIDIA GeForce RTX 3090 GPUs with the ADAM optimizer. For both datasets, we used the same detection ranges as fully supervised methods (Yin, Zhou, and Krahenbuhl 2021).

Baseline. Seg2Box supervises 3D detectors using only semantic labels, so no previously published baselines exist. We create baselines by pseudo-label generation methods

Method	St	Vehicle 3D AP				Pedestrian 3D AP				Cyclist 3D AP			
		L1		L2		L1		L2		L1		L2	
		$IoU_{0.5}$	$IoU_{0.7}$	$IoU_{0.5}$	$IoU_{0.7}$	$IoU_{0.25}$	$IoU_{0.5}$	$IoU_{0.25}$	$IoU_{0.5}$	$IoU_{0.25}$	$IoU_{0.5}$	$IoU_{0.25}$	$IoU_{0.5}$
CenterPoint (Boxes Sup)	×	93.75	72.49	88.13	64.45	89.24	74.28	83.61	66.34	76.67	71.20	75.20	68.58
DBSCAN (Ester et al. 1996) + Sem	×	40.71	9.98	35.60	8.52	67.66	31.20	58.67	26.15	42.42	31.95	40.91	30.77
DBSCAN + Sem + St	✓	43.07	13.09	37.45	11.65	63.99	33.47	56.76	28.16	56.49	32.86	54.47	31.67
OYSTER (Zhang et al. 2023) + Sem	✓	45.23	20.36	39.20	17.46	70.45	19.46	61.34	16.40	58.64	37.12	52.43	35.75
Seg2Box + Init	×	85.01	56.05	76.46	48.71	81.50	**51.47**	73.21	**43.93**	60.68	44.08	58.49	42.45
Seg2Box	✓	**90.56**	**62.28**	**83.73**	**54.71**	**82.35**	50.34	**74.31**	43.02	**66.56**	**46.66**	**64.27**	**44.96**

Table 1: 3D object detection results on WOD validation set. St means Self-training. Init means Init-training.

Method	St	mAP(↑)	Error(↓)		
			ATE	ASE	AOE
CenterPoint (Boxes Sup)	×	54.6	0.31	0.26	0.41
DBSCAN + Sem	×	28.9	0.42	0.51	1.41
DBSCAN + Sem + St	✓	36.0	0.44	0.50	1.51
OYSTER + Sem	✓	33.9	0.46	0.40	1.44
Seg2Box + Init	×	44.2	0.37	0.33	1.68
Seg2Box	✓	46.3	0.37	0.32	1.54

Table 2: 3D object detection results on nuScenes Dataset validation set. St means Self-training. Init means Init-training.

with semantic labels, and we named the modified method as Method-Name + Sem. We first found the baseline DBSCAN + Sem (Ester et al. 1996), since Seg2box follows the idea of density-based spatial clustering to obtain initial pseudo-labels (Zhang et al. 2023; You et al. 2022). DBSCAN + Sem + St adds two rounds of self-training. We also assigned semantics to the state-of-the-art unsupervised method OYSTER (Zhang et al. 2023), indicated as OYSTER + Sem. For a fair comparison, we used the MFF and SCF modules proposed in SGIM-ST for all training, which are intuitive but effective tricks with semantic labels.

Main Results

Results on WOD. The results on the WOD validation set are presented in TABLE 1. The DBSCAN + Sem performs poorly due to the inaccuracy of initial pseudo-labels generated in a single frame. With self-training, DBSCAN + Sem + St outperforms the init-training significantly but is still unsatisfactory due to the poor init-training performance. OYSTER + Sem attempts to refine the pseudo-labels by tracking and refining boxes in long trajectories. It still remains a poor performance since the poor tracking performance. Our Seg2Box + Init outperforms the best baseline method by 21.8% in L2 mAP although without self-training. These advancements come from our MFMS-C designs, which generate more accurate pseudo-labels. With SGIM-ST, Seg2Box improves init-training by 7.3% in 3D AP with an IoU threshold of 0.5 in Vehicle which reaches 95% of the performance of the fully supervised method. With more strict IoU threshold 0.7, Seg2Box can still reach 85% performance of fully supervised method.

Method	3D Recall			3D Precision		
	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$
DBSCAN	45.07	24.98	10.59	41.94	23.24	9.86
OYSTER	48.66	28.30	13.76	47.87	27.84	13.53
Ours	74.92	57.49	27.46	75.61	58.02	27.71

Table 3: Pseudo-label Comparison Results on WOD.

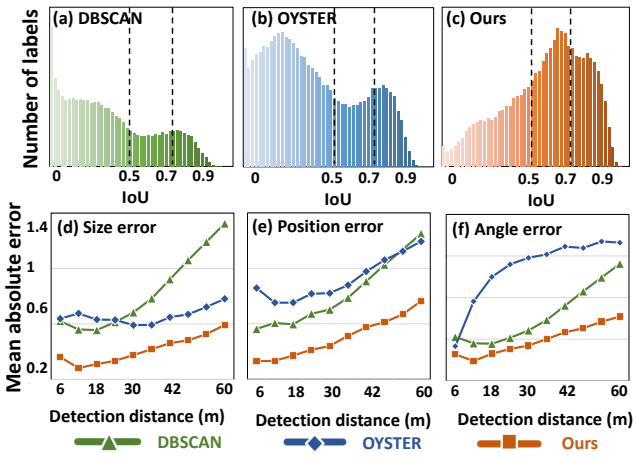

Figure 5: (a-c): The IoU distribution between pseudo-labels and ground truth. (d-f): The mean absolute errors (MAEs) for size, position, and angle of pseudo-labels.

Results on nuScenes Dataset. The results on the nuScenes validation set are presented in TABLE 2. Our method demonstrates a 10.3% improvement in mAP compared to the best baseline detector. In addition, Seg2Box also reduces errors in ATE and ASE. This improvement is attributed to Seg2Box's capacity to generate accurate initial pseudo-labels and its ability to mine and refine pseudo-labels throughout the self-training process continuously.

Pseudo-Label Comparison

To verify the quality of our pseudo-labels, we calculated the 3D recall and precision on WOD. As shown in TABLE 3, our method outperforms OYSTER (Zhang et al. 2023) with improvements of 13.7% and 14.18% in Recall and Precision, respectively, even under a strict 0.7 IoU threshold. To under-

Seg2Box Components				3D AP L1	3D AP L2
SFC	MFC	MSC	SGIM-ST		
✓				20.36	17.46
	✓			51.33	44.41
	✓	✓		56.05	48.71
	✓	✓	✓	62.28	54.71

Table 4: Seg2Box component analysis on WOD val. set.

SGIM-ST Components				3D AP L1	3D AP L2
BAF	MSFLoss	SCF	STCF		
				50.37	43.66
✓				56.05	48.70
✓	✓			57.08	49.71
✓	✓	✓		59.51	51.87
✓	✓	✓	✓	62.28	54.71

Table 5: SGIM-ST component analysis on WOD val. set.

stand the source of our improvements, we evaluated the IoU between the pseudo-labels and the ground truth. The IoU distributions are shown in Fig. 5 (a-c). Our method's IoU distribution is more concentrated near 1 with a larger number of high-quality pseudo-labels with IoU scores exceeding 0.7. Additionally, as shown in Fig. 5 (d-f), we also evaluated the size error, position error, and angle error of the pseudo-labels. Our method maintains more minor errors even as the detection distance increases. These results confirm that our MFMS-C module significantly reduces label errors and generates higher-quality pseudo-labels.

Ablation Study

Components Analysis of Seg2Box. To evaluate the individual contributions of Seg2Box, we incrementally added each component and assembled their impact on AP using the WOD validation set in Vehicle. The results are shown in TABLE 4. The comparison of the first row and the second row shows that Multi-Frame Clustering (MFC) significantly surpasses Singel-Frame Clustering (SFC) by 26.95% in AP L2. It indicates that more complete objects in consecutive frames are crucial to generating high-quality pseudo-labels. The third row shows that MSC further enhances performance by 4.3% in AP L2 since it avoids mistaken instance division to get precise boundaries for pseudo-labels. The last row indicates that SGIM-ST contributes a 6% increase in AP L2, demonstrating the ability of SGIM-ST to refine pseudo-labels and mine the instances without generating pseudo-labels.

Component Analysis of SGIM-ST. To prove the effect of of SGIM-ST, we incrementally added each component and evaluated their impact on AP using the WOD validation set in Vehicle. The results are shown on TABLE 5. The first row presents the result of init-training. The comparison of the first row and the second row shows that our BAF module significantly improves the performance by 5.04% in AP. It indicates that foreground points without generating pseudo-labels will seriously mislead the training of the model. The third row shows that assembled with MSFLoss, our method further enhances performance by 1.01% in AP, since it helps inhibit the influence of inaccurate pseudo-labels. With the help of semantic labels, SCF contributes a 2.16% improvement in AP by filtering the false classification (False positive) of predictions (the fourth row). The last row indicates that STCF contributes a 2.84% increase in AP, attributed to the refined pseudo-labels of static objects in long-distance.

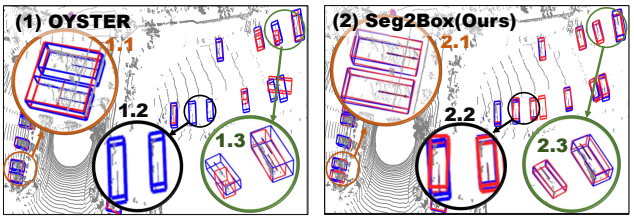

Figure 6: Visualization comparison of different detection results on WOD validation set. Blue boxes are the ground truth, and red boxes are the detection results.

Visualization Comparison of Detection Results

We visually compare the detection results of different methods in Fig. 6. OYSTER fails to detect distance, sparse instances (Fig. 6 (1.2)) or detects them with inaccurate sizes and positions (Fig. 6 (1.3)). Additionally, OYSTER also fail to detect the adjacent instances because of the false boundaries of initial pseudo-labels. In contrast, Seg2Box not only identifies these objects but also precisely predicts their sizes and positions. (Fig. 6 (2)).

Conclusion

This work presents Seg2Box, a novel framework for 3D object detection supervised by semantic labels. It indicates that cross-task supervision between 3D object detection and semantic segmentation is a feasible way to reduce annotation redundancy. First, the MFMS-C module estimates high-quantity pseudo-labels with correct instance divisions, positions and sizes. Furthermore, the SGIM-ST module, is a novel self-training framework designed to refine the pseudo-labels and mine the instances without pseudo-labels iteratively, thereby enhancing detection performance. Experimental results on the Waymo Open Dataset and the nuScenes Dataset demonstrated that Seg2Box outperforms other competitive methods by a large margin. Future work will explore multi-task learning of 3D object detection and 3D semantic segmentation, but using annotation from one task.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62171393), and the Fundamental Research Funds for the Central Universities (No.20720220064).

References

Beuchert, J.; Camurri, M.; and Fallon, M. 2023. Factor graph fusion of raw GNSS sensing with IMU and lidar for precise robot localization without a base station. In *ICRA*, 8415–8421. IEEE.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *CVPR*.

Chibane, J.; Engelmann, F.; Anh Tran, T.; and Pons-Moll, G. 2022. Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In *ECCV*, 681–699. Springer.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *ICCV*, 1635–1643.

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.

Han, Y.; Zhao, N.; Chen, W.; Ma, K. T.; and Zhang, H. 2024. Dual-Perspective Knowledge Enrichment for Semi-supervised 3D Object Detection. *AAAI*, 38(3): 2049–2057.

Kulharia, V.; Chandra, S.; Agrawal, A.; Torr, P.; and Tyagi, A. 2020. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *ECCV*, 290–308. Springer.

Li, Y.-Y.; Fan, L.; Liu, Y.; Huang, Z.; Chen, Y.; Wang, N.; Zhang, Z.; and Tan, T. 2023. Fully Sparse Fusion for 3D Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP.

Liu, C.; Gao, C.; Liu, F.; Liu, J.; Meng, D.; and Gao, X. 2022. Ss3d: Sparsely-supervised 3d object detection from point cloud. In *CVPR*, 8428–8437.

Luo, K.; Liu, Z.; Chen, X.; You, Y.; Benaim, S.; Phoo, C. P.; Campbell, M.; Sun, W.; Hariharan, B.; and Weinberger, K. Q. 2024. Reward Finetuning for Faster and More Accurate Unsupervised Object Discovery. *NeurIPS*, 36.

Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D object detection for autonomous driving: A comprehensive survey. *IJCV*, 131(8): 1909–1963.

Meng, Q.; Wang, W.; Zhou, T.; Shen, J.; Jia, Y.; and Van Gool, L. 2021. Towards a weakly supervised framework for 3d point cloud object detection and annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8): 4454–4468.

Neubeck, A.; and Van Gool, L. 2006. Efficient non-maximum suppression. In *ICPR'06*, volume 3, 850–855. IEEE.

Qin, Z.; Wang, J.; and Lu, Y. 2020. Weakly supervised 3d object detection from point clouds. In *ACM MM*, 4144–4152.

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 *CVPR*, 2446–2454.

Team, O. D. 2020. OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds. https://github.com/open-mmlab/OpenPCDet.

Wang, H.; Cong, Y.; Litany, O.; Gao, Y.; and Guibas, L. J. 2021. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *CVPR*, 14615–14624.

Wang, L.; Fan, X.; Chen, J.; Cheng, J.; Tan, J.; and Ma, X. 2020. 3D object detection based on sparse convolution neural network and feature fusion for autonomous driving in smart cities. *Sustainable Cities and Society*, 54: 102002.

Wang, Y.; Chen, Q.; Zhu, Q.; Liu, L.; Li, C.; and Zheng, D. 2019. A Survey of Mobile Laser Scanning Applications and Key Techniques over Urban Areas. *Remote. Sens.*, 11: 1540.

Wisth, D.; Camurri, M.; and Fallon, M. 2022. VILENS: Visual, inertial, lidar, and leg odometry for all-terrain legged robots. *IEEE Transactions on Robotics*, 39(1): 309–326.

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 Trans. Geosci. Remote Sens.*, 60: 1–11.

Wu, H.; Zhao, S.; Huang, X.; Wen, C.; Li, X.; and Wang, C. 2024. Commonsense Prototype for Outdoor Unsupervised 3D Object Detection. *arXiv preprint arXiv:2404.16493*.

Xia, Q.; Chen, Y.; Cai, G.; Chen, G.; Xie, D.; Su, J.; and Wang, Z. 2023a. 3-D HANet: A Flexible 3-D Heatmap Auxiliary Network for Object Detection. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–13.

Xia, Q.; Deng, J.; Wen, C.; Wu, H.; Shi, S.; Li, X.; and Wang, C. 2023b. CoIn: Contrastive Instance Feature Mining for Outdoor 3D Object Detection with Very Limited Annotations. In *ICCV*, 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 *CVPR*, 15321–15330.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *CVPR*, 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 *CVPR*, 1130–1140.

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 *CVPR*, 9317–9328.

Zhang, X.; Xu, W.; Dong, C.; and Dolan, J. M. 2017. Efficient L-shape fitting for vehicle detection using laser scanners. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, 54–59. IEEE.

Zhang, Z.; Chen, M.; Xiao, S.; Peng, L.; Li, H.; Lin, B.; Li, P.; Wang, W.; Wu, B.; and Cai, D. 2024. Pseudo Label Refinery for Unsupervised Domain Adaptation on Cross-dataset 3D Object Detection. In *CVPR*, 15291–15300.

Zhu, Y.; Hui, L.; Shen, Y.; and Xie, J. 2024. SPGroup3D: Superpoint Grouping Network for Indoor 3D Object Detection. In *AAAI*, volume 38, 7811–7819.