Approaching Outside: Scaling Unsupervised 3D Object Detection from 2D Scene

Ruiyang Zhang[1], Hu Zhang[2], Hang Yu[3], and Zhedong Zheng[1]*

[1] FST and ICI, University of Macau, China
[2] CSIRO Data61, Australia
[3] Shanghai University, China
ruiyang.061x@gmail.com, Hu1.Zhang@csiro.au, yuhang@shu.edu.cn, zhedongzheng@um.edu.mo
https://github.com/Ruiyang-061X/LiSe

Abstract. The unsupervised 3D object detection is to accurately detect objects in unstructured environments with no explicit supervisory signals. This task, given sparse LiDAR point clouds, often results in compromised performance for detecting distant or small objects due to the inherent sparsity and limited spatial resolution. In this paper, we are among the early attempts to integrate LiDAR data with 2D images for unsupervised 3D detection and introduce a new method, dubbed LiDAR-2D Self-paced Learning (LiSe). We argue that RGB images serve as a valuable complement to LiDAR data, offering precise 2D localization cues, particularly when scarce LiDAR points are available for certain objects. Considering the unique characteristics of both modalities, our framework devises a self-paced learning pipeline that incorporates adaptive sampling and weak model aggregation strategies. The adaptive sampling strategy dynamically tunes the distribution of pseudo labels during training, countering the tendency of models to overfit easily detected samples, such as nearby and large-sized objects. By doing so, it ensures a balanced learning trajectory across varying object scales and distances. The weak model aggregation component consolidates the strengths of models trained under different pseudo label distributions, culminating in a robust and powerful final model. Experimental evaluations validate the efficacy of our proposed LiSe method, manifesting significant improvements of +7.1% AP_{BEV} and +3.4% AP_{3D} on nuScenes, and +8.3% AP_{BEV} and +7.4% AP_{3D} on Lyft compared to existing techniques.

Keywords: Unsupervised 3D Object Detection · 2D Scene Understanding · Self-paced Learning · Unsupervised Learning

1 Introduction

Unsupervised 3D object detection in the context of autonomous driving aims to discover potential 3D objects in an unsupervised manner [29, 49, 55]. The

* Corresponding author.

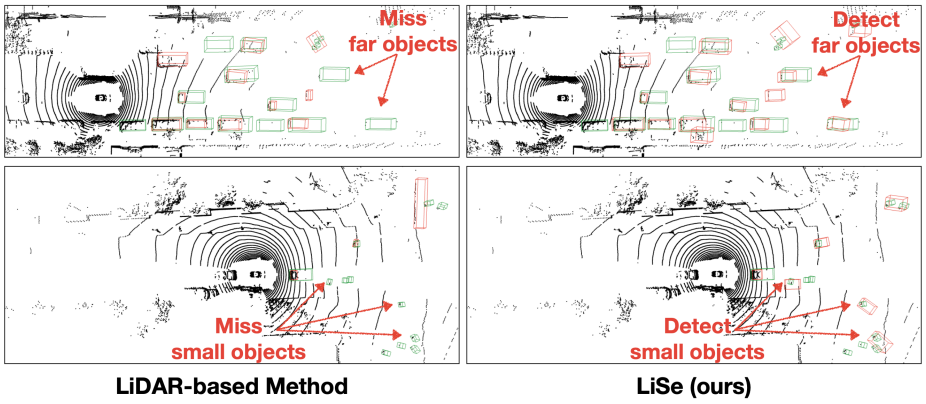

Fig. 1: We show typical limitations of LiDAR-based methods for unsupervised 3D object detection. Compared with the prevailing LiDAR-based method, *i.e.*, MODEST [49] generally misses objects in the distance and small objects (left), our proposed method LiSe successfully recalls such objects (right). Best viewed in color: the green boxes are ground truth labels and the red boxes are predictions.

key underpinning unsupervised 3D detection is to develop intelligent algorithms that can effectively reason about and adapt to the vast array of potential object classes and their various manifestations in real-world scenarios without explicit prior knowledge or labeled data. The process involves not only accurately estimating the 3D position of objects but also learning to identify previously unseen object types and handle unpredictable environmental conditions, which remains challenging to 2D-based methods [6, 14]. This technology is crucial for ensuring the safety and efficiency of autonomous vehicles as they navigate through complex and unpredictable road environments. The ability to accurately detect 3D objects allows these systems to anticipate potential hazards, make informed decisions, and react accordingly. It can be widely applied to real-world applications, including pedestrian protection [9,12], auto-driving assistance systems [15,17–19] and traffic management [27, 36]. The inherent challenge lies in designing models capable of extracting discriminative features from sparse and noisy sensor data, such as point clouds and images, while simultaneously overcoming issues related to class imbalance, scale variation, and partial occlusions.

Most existing works usually focus on mining the LiDAR data to discover unlabeled 3D objects [29, 49, 55]. For instance, some works [49, 55] utilize the rule-based generation of pseudo boxes followed by a self-training process, and another line of works [29] harness the motion cues provided by LiDAR scene flow to identify potential dynamic objects. While LiDAR data provides accurate depth information and a comprehensive perception of the surrounding environment, it is limited by the inherent sparsity and spatial resolution. This sparsity challenge becomes particularly pronounced in scenarios involving long-range or small-scale objects (see Figure 1), where the dearth of LiDAR points returned significantly compromises the discriminative power required to accurately segre-

gate foreground entities from their background context. Current methodologies often adopt an iterative training pipeline that commences with initial object proposals followed by pseudo-label refinement. However, overdependence on LiDAR alone can lead to a blind spot for detecting diminutive or distant objects, thus culminating in suboptimal overall detection capabilities.

In this paper, we thus propose a novel LiDAR-2D Self-paced Learning (**LiSe**) for unsupervised 3D detection, which integrates LiDAR data with 2D images. It aims to leverage the rich textual and RGB color information in 2D scenes to overcome the limitations of LiDAR in detecting distant and small objects. We adopt the off-the-self multi-traversal method for LiDAR-based 3D detection [49], while applying the 2D detection [28] and segmentation [20] with 3D lifting for image-based 3D detection. We observe that two modalities are complementary on the objects with different distances and resolution, and serve as good initialization seed. Then we apply the self-paced training strategy to propagate the object label and refine the box prediction. During training, we observe that the models tend to overfit to the common category, *e.g.*, cars, and gradually lose the ability to detect relatively rare objects, *e.g.*, bicycles. To alleviate the diminishing detection ability of such long-tail samples, we introduce an adaptive sampling strategy that dynamically adjusts the distribution of training data based on the feedback of the model. Therefore, we could obtain the snapshots trained with different data distributions during the training process, and thus the learned snapshots are inherently with complementary focuses. We further propose the weak model aggregation strategy to merge all snapshot weights along the self-paced learning process as the final model. We conduct extensive quantitative experiments and qualitative analyses to validate the effectiveness of our method. In conclusion, our contributions are summarized as follows:

- Considering the inherent sparsity of LiDAR data, we propose a new approach, called **LiSe**, jointly leveraging 2D images and 3D LiDAR to improve the pseudo label quality in all ranges. The rich texture in 2D images provides a straightforward discovery of small and distant objects.
- Considering the imbalanced object distribution in self-training, we propose an adaptive sampling strategy to explicitly emphasize the long-tailed objects, followed by the weak model aggregation, which iteratively fuses the strengths of different snapshots into a final stable model.
- Extensive experiments on nuScenes and Lyft verify the effectiveness of the proposed method, surpassing state-of-the-art by a clear margin on both AP_{3D} and AP_{BEV}. Especially, for long-range detection (50-80m), the AP_{BEV} metric even exceeds that of fully supervised model.

2 Related Work

Unsupervised 3D Object Detection. Unsupervised 3D object detection is attracting increasing interest within the research community [10, 43, 47, 54]. Cen *et al.* [5] utilize a fully-supervised detector to generate proposals for unknown classes. However, this approach struggles with overconfidence issues and

lacks the ability to generate proposals for semantically distinct classes. Studies like MODEST [49] and OYSTER [55] have explored a self-training pipeline to incrementally discover more objects from LiDAR data. Nevertheless, the inherent sparsity of LiDAR data hampers their ability to detect small objects. Najibi *et al.* [29] employ scene flow to identify objects in motion, training detection model with generated pseudo boxes. Najibi *et al.* [30] propose distilling VLM knowledge into 3D detection model, addressing detection task by initially segmenting based on text references, followed by 3D box fitting. Different from existing works, our work aims to leverage the information density in 2D scenes to enhance recognition of distant and small objects. We also develop adaptive sampling strategy to address data distribution imbalance problem during self-training process.

Open-vocabulary 2D Detection. Open-vocabulary 2D detection methods can be categorized into two groups: (1) Knowledge distillation methods [2, 8, 44] focus on transferring the extensive open-vocabulary knowledge from VLMs into closed-set 2D detectors. Therefore, the detection capability is limited by the scope of the teacher VLM. (2) Region-text pretraining methods [3, 23, 53] emphasize learning from region captions through pretraining at the region level, albeit with high computational costs due to the large scale of pretraining datasets. Similarly, our work leverages pretrained open-vocabulary detection models to recognize objects in 2D images. It helps us to instill 2D prior to the 3D detection. We also consider to eliminate the negative impact of the class imbalance and overfitting during the self-training process.

Image and LiDAR Fusion. In the realm of closed-set object detection, recent works have begun to study image and LiDAR data fusion. These works can be divided into two categories, *i.e.*, sequential fusion [33, 45] and parallel fusion [7, 31]. (1) Sequential based approaches [33,45] use 2D base models to initiate the 3D detection pipeline. The drawback here is that failures in 2D models accumulate in the 3D detection model, which inevitably degrades the overall performance. (2) Parallel fusion approaches integrate two modalities at different stages of the pipeline, including early fusion at the input stage [46], deep fusion at the feature stage [7, 13, 21, 26], and late fusion at the output stage [31, 58]. A primary challenge in parallel fusion is to ensure semantic alignment, due to substantial gap between characteristics of images and point clouds. Different from existing works, our method fuses two modalities by combining box seeds from both. We introduce self-paced learning to progressively update the box labels, offering robustness where the failure of 2D models does not halt the pipeline.

Self-paced Learning. Self-paced learning [22] describes a self-directed learning process where the distribution of training data is dynamically adjusted based on the model performance. This concept has seen broad application in computer vision field, including image classification [40, 48, 50], object detection and localization [37, 39, 51], scene segmentation [24, 34], and video processing [25, 52]. However, these strategies are predominantly focused on the image and video domains. Diverging from existing works, we have tailored a self-paced learning strategy specifically for the 3D unsupervised detection task, incorporating unique 3D attributes such as object distance and volume information.

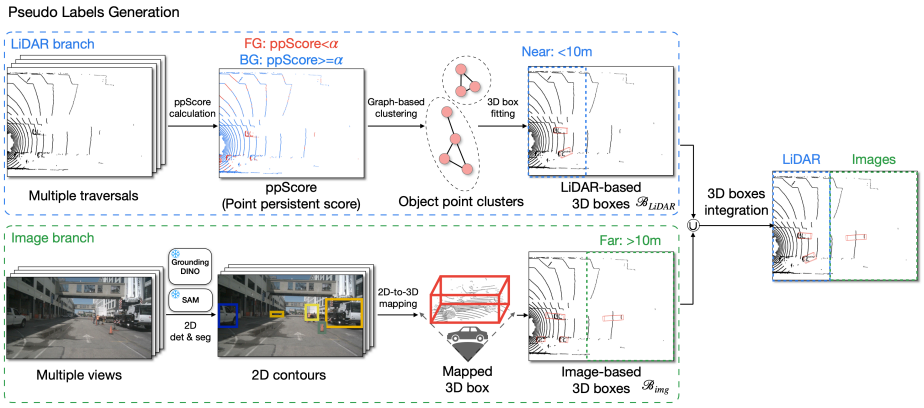

Fig. 2: Illustration of the pseudo label generation process in LiSe, which distinctively harnesses information density from 2D scenes to complement LiDAR data. Our approach involves a generation method tailored for each modality to obtain LiDAR-based and image-based 3D boxes. In the LiDAR branch, an off-the-shelf multi-traversal based method generates pseudo labels, primarily covering near-range objects. Concurrently, the image branch uses a pretrained open-vocabulary 2D detector and Segment-Anything-Model to generate 2D contours from images, which are then mapped into 3D space. Following this, a distance-aware 3D boxes integration process fuses boxes from both LiDAR and image modalities. Notably, image-based boxes at longer ranges (*e.g.*, $>$ 10m) are merged with LiDAR-based 3D boxes. This integration addresses the limitations of LiDAR-based method in detecting long-range and small objects. The resulting pseudo labels are proficient in originally challenging samples (*e.g.*, distant and small objects) for LiDAR-based methods, laying a solid foundation for enhancing detection model performance on these challenging cases.

3 Method

We present a detailed description of our method in this section and structure it into three parts: (1) integration of LiDAR data with 2D scene (see Figure 2), (2) adaptive sampling strategy (see Figure 3), and (3) weak model aggregation (see Figure 3).

3.1 Integration of LiDAR Data with 2D Scene

Pseudo-boxes from LiDAR. In our work, we apply a multi-traversal approach to extract significant objects from LiDAR data. Multi-traversal approach is based on the idea that when a vehicle traverses the same location multiple times, entities that remain unchanged in both position and state are likely static background elements, *e.g.* buildings. Conversely, items that shift in location are probable foreground objects, *e.g.* moving cars. Specifically, we first conduct the data processing for given LiDAR data. For locations visited more than once, the LiDAR scans collected at these locations are combined. We then use the GPS/INS data which provides accurate information on vehicle location and the

rotation matrix to calibrate the data so that different LiDAR scans are aligned into same coordinate system. After the alignment, we calculate the point persistency score (ppScore) [49] of each point $\tau(u)$ to quantify whether it belongs to unchanging or changing objects. The higher ppScore indicates a more static point and lower ppScore indicates a more dynamic one.

With the calculated ppScore, a clustering process that considers both the similarity of ppScore and the actual geometric distance between points is utilized to segment the entire point cloud into distinct clusters. A graph is constructed where points within radius threshold r_t of each other are connected by an edge, and the weight of this edge is calculated as the absolute difference of their ppScores $|\tau(u) - \tau(v)|$. Following graph construction, a variant of the DBSCAN [11] algorithm which is adapted for application on the graph is used on the constructed graph, resulting in numerous clusters of points with similar ppScores and proximate geometric distances. A filtering process that excludes clusters where the top K percent of points with ppScore above threshold α is applied, designating these as static clusters (*e.g.*, large building walls). The remaining clusters are treated as foreground objects. Finally, an off-the-shelf bounding box fitting algorithm [56] is applied to each cluster to create a 3D box.

Pseudo-boxes from Images. For 3D pseudo-boxes generation from images, we employ an off-the-shelf open-vocabulary 2D detector, *e.g.*, GroundingDINO [28] to first identify discriminative objects within the images. To realize this purpose, we construct the detection prompt by concatenating discriminative class names together, feeding it into GroundingDINO, and obtaining a collection of 2D boxes. Usually, the detected 2D boxes often contain substantial background areas, which do not accurately reflect the shapes of real-world objects. Direct employment of these 2D boxes for subsequent processing can result in imprecise 3D box estimations. Fortunately, the 2D boxes naturally serve as initial prompts for the Segment-Anything-Model (SAM) [20]. By inputting these 2D boxes into SAM as prompts, we instead obtain refined 2D masks. These masks reflect the actual contours of the targeted objects, significantly mitigating the drawbacks inherent in utilizing 2D boxes directly. To estimate 3D boxes from images, a projection process from 3D to 2D is then applied and can be formulated as:

$$\hat{u}_i = \mathbf{K} \cdot \mathbf{E} \cdot u_i, \ (i = 1, ..., m), \tag{1}$$

where u_i is one LiDAR point in the 3D space, $\mathbf{K}$ is the intrinsic matrix, $\mathbf{E}$ is the extrinsic matrix, and m is number of points in point cloud. For those projected 2D points lying in the masks, we reserve their corresponding 3D points. This process is equal to building a frustum extending from the ego center of the vehicle to 2D mask. This frustum is considered to contain the point cluster corresponding to the detected object. We then apply the region growth algorithm [1] to get the cluster with the most points. Subsequently, tight external 3D bounding box is estimated based on the cluster. All the generated 3D boxes are then consolidated to form the final pseudo labels for a single LiDAR point cloud. Ultimately, we achieve a comprehensive set of pseudo labels for the training dataset, entirely independent of any ground-truth 3D annotations.

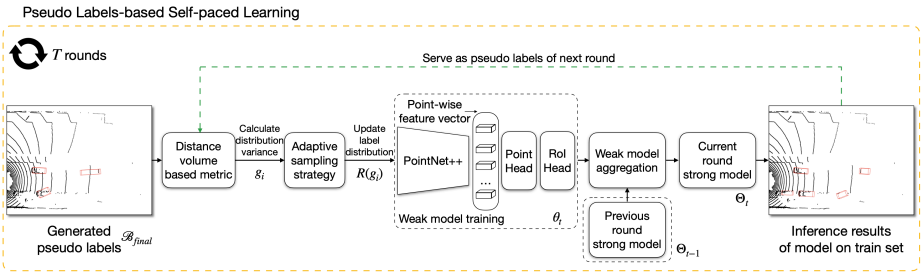

Fig. 3: Illustration of the self-paced learning process in LiSe. Initial distribution of objects and inference distribution after training are first calculated with the distance volume-based metric. Adaptive sampling strategy thus updates sampling rates for different objects based on changes in two distributions. We further consider weak model aggregation to combine newly trained model with previously aggregated model to obtain a stronger, more robust model for current round. Finally, we iteratively update distribution of pseudo labels and model weight for T rounds to obtain the final model.

It is worth noting that due to the rich texture information in 2D images and the strong detection ability of the employed open-vocabulary 2D detector, many distant and small objects, which are usually challenging to identify in LiDAR data, can be recognized. The pseudo labels derived from images can thus serve as a robust complement to those obtained from LiDAR, potentially enhancing the overall quality and coverage of the training data.

Integration between LiDAR and 2D scene. To enhance the integration of pseudo boxes from both LiDAR and images for training models, we employ a distance-aware strategy. This approach optimally leverages the complementary characteristics of two data sources. We begin by establishing a predefined range, and then we selectively include image-derived boxes that lie within this range, alongside LiDAR-generated boxes. Final bounding boxes can be derived from:

$$\mathcal{B}_{final} = \mathcal{B}_{LiDAR} \cup \{b_i \mid d(b_i) \geq d_{min}, b_i \in \mathcal{B}_{img}\}, \tag{2}$$

where $\mathcal{B}_{img}$ denotes all image-derived boxes and $\mathcal{B}_{LiDAR}$ denotes LiDAR-generated boxes. b_i is one image-derived 3D box, $d(b_i)$ is the distance between box b_i and the ego car. d_{min} is the determined range value. Considering that objects in close proximity typically exhibit a high density of LiDAR points, LiDAR data alone is often sufficient for precise estimations. Our distance-aware strategy allows for flexible exclusion of image-derived boxes in these near-range areas by adjusting range values. It helps to prevent possible conflicts with LiDAR-generated boxes.

3.2 Adaptive Sampling Strategy

Despite integrating 2D scenes into 3D pseudo-boxes is able to recall the missed distant and small objects, the model tends to be biased toward easier samples, *e.g.*, closer or larger objects in training. Such bias persists throughout all training

rounds and the reason behind is attributed to the imbalanced data distribution. We thus propose an adaptive sampling strategy, dynamically balancing different objects throughout the training phases (see Figure 3).

We first propose distance volume-based metric, which leverages general properties in 3D world, *i.e.*, distance and volume to categorize objects. For distance-based categorization, we adopt the criteria in MODEST [49], dividing objects into near- and far-range ones: objects within 0-30m are considered as near objects and those beyond 30m are categorized as far objects. For volume-based categorization, we consult GPT-4 for general information on the volume and size of common categories. We then classify objects with a volume smaller than $5m^3$ as small objects, and those larger than $5m^3$ as large objects. For example, common categories such as pedestrians, cyclists will be categorized as small objects, and typical cars or other vehicles will be attributed to large ones.

Based on the distance volume-based metric, we calculate the initial object distribution before training and inference distribution after training. We analyze the differences between two distributions: For object groups whose probability in inference distribution is significantly increased, we adaptively downsample these objects in the next round. Conversely, for object groups whose probabilities are decreased in the inference, we adaptively upsample these groups accordingly. Therefore, we introduce a sampling score, which can be formulated as:

$$R(g_i) = \begin{cases} 1 - (Q(g_i) - Q_{init}(g_i)) & \text{if } Q(g_i) > Q_{init}(g_i) \\ 1 + (Q_{init}(g_i) - Q(g_i)) & \text{if } Q(g_i) \leq Q_{init}(g_i) \end{cases}, \tag{3}$$

where g_i is one type of objects grouped by distance volume-based metric. $Q(g_i)$ is sampling probability in inference distribution and $Q_{init}(g_i)$ is sampling probability in initial distribution. $R(g_i)$ is new sampling score for objects in group g_i in next round. Adaptive resampling can calibrate model towards harder samples and away from easier ones, thus resulting in self-paced learning process.

3.3 Weak Model Aggregation

The models obtained in different rounds tend to be proficient in different object groups, with the adaptive sampling strategy assigning varied sampling ratios. For example, while a model trained in the t-th round excels at identifying large objects, the model in the $(t+1)$-th round takes more attention to detecting small objects with the increased sampling rate for small objects. The models obtained in different rounds have their unique bias and lack a comprehensive detection ability. Therefore, we refer to these models as "weak models", and introduce weak model aggregation, which combines these weak models to create a robust, stronger model (see Figure 3). We select a model as initialization starting from round T_s. Similar to weight-average approaches [41, 57], we average each weak model in subsequent rounds with the previous aggregated strong model, and the obtained model serves as the strong one for the current round. An aggregation coefficient λ is utilized to balance the influence of the previous strong model and

the current weak model. The calculation process can be formulated as:

$$\Theta_t = \begin{cases} \theta_t & \text{if } 1 \leq t < T_s \\ \lambda \cdot \Theta_{t-1} + (1-\lambda) \cdot \theta_t & \text{if } T_s \leq t \leq T \end{cases}, \tag{4}$$

where t is round number, T_s is start round to perform weak model aggregation, and T is total number of self-paced learning rounds. λ is aggregation coefficient. θ_t is current weak model in t-th round, Θ_{t-1} is aggregated strong model obtained in $(t-1)$-th round, and Θ_t is the aggregated strong model in t-th round.

3.4 Pseudo Labels-based Self-paced Learning

We unify integrated pseudo labels, adaptive sampling strategy, and weak model aggregation into a self-paced learning process (see Fig. 3). Specifically, it consists of two stages: seed training and self-training. In seed training, integrated pseudo labels $\mathcal{B}_{final}$ are used to train an initial detector Θ_0. Self-training is an iterative process repeated for T rounds. In t-th round, detector trained from previous round Θ_{t-1} first conducts inference on the training set to obtain pseudo training labels for the current round. The pseudo training labels are then redistributed with our proposed adaptive sampling strategy to counter the bias towards easier object groups, *e.g.*, near-range and large objects. Then the updated pseudo labels are harnessed to train a new detector θ_t. Weak model aggregation aggregates the weak model in current round θ_t and strong model from previous round Θ_{t-1} into a strong model for current round Θ_t. Different from vanilla self-training, in our process, distribution of pseudo training labels is adjusted based on model feedback, which results in a self-paced learning process.

4 Experiment

Dataset. We conduct unsupervised 3D detection experiments on nuScenes [4] and Lyft [16], two widely recognized benchmarks in autonomous driving. In our experiment, we follow the basic dataset configuration in MODEST [49]. Specifically, we only utilize LiDAR point clouds which are collected from locations with more than one sample in order to satisfy the multi-traversal requirement. For nuScenes, the final used data consists of 3,985 training keyframes and 2,412 testing ones. For Lyft, we use 11,873 training samples and 4,901 testing samples. We emphasize that the ground-truth 3D annotations are not used in our training and they are just involved in the testing to evaluate model performance.
Metric. Two different metrics AP_{BEV} and AP_{3D} are considered. AP_{BEV} focuses on accuracy from the Bird's Eye View (BEV), while AP_{3D} considers additional height information and evaluates detection results in 3D space, thus offering more comprehensive assessment. Furthermore, we consider objects *w.r.t.* the distance and report evaluation results for objects in the near range (0-30m), middle range (30-50m), far range (50-80m), and the full range (0-80m).
Implementation Details. In the training, we adopt PointRCNN [35] as the backbone. In each self-paced training round, we train the model for 80 epochs on

Table 1: Detection results on nuScenes. We report AP_{BEV} and AP_{3D} at $IoU = 0.25$ for objects across various distances. The results are shown in AP_{BEV} / AP_{3D} format. $T = 0$ is training from seed labels. $T = 2$ and $T = 10$ are the results for $2th$ and $10th$ round self-training, respectively. The supervised performance of model trained with ground-truth boxes is in the first row (Supervised). It is noticeable that the performance of LiSe significantly surpasses that of the state-of-the-art OYSTER [55] across all evaluated metrics. *: We present the results of our reimplementation, as official code for OYSTER is not available. Our reimplementation follows OYSTER settings, which conduct two rounds of self-training.

Method	**0-30m**	**30-50m**	**50-80m**	**0-80m**
Supervised	39.8 / 34.5	12.9 / 10.0	4.4 / 2.9	22.2 / 18.2
MODEST-PP ($T = 0$)	0.7 / 0.1	0.0 / 0.0	0.0 / 0.0	0.2 / 0.1
MODEST-PP ($T = 10$)	-	-	-	-
MODEST ($T = 0$)	16.5 / 12.5	1.3 / 0.8	0.3 / 0.1	7.0 / 5.0
MODEST ($T = 10$)	24.8 / 17.1	5.5 / 1.4	1.5 / 0.3	11.8 / 6.6
OYSTER ($T = 0$)	14.7 / 12.3	1.5 / 1.1	0.5 / 0.3	6.2 / 5.4
OYSTER ($T = 2$)*	26.6 / 19.3	4.4 / 1.8	1.7 / 0.4	12.7 / 8.0
LiSe ($T = 0$)	5.8 / 4.7	0.6 / 0.2	0.3 / 0.2	2.1 / 1.8
LiSe ($T = 10$)	**35.0** / **24.0**	**11.4** / **4.4**	**4.8** / **1.3**	**19.8** / **11.4**

nuScenes and 60 epochs on Lyft. We adopt AdamOneCycle [38] as the optimizer, with a default learning rate of 0.01, weight decay of 0.01, and momentum of 0.9. The learning rate is reduced at epochs 35 and 45 by a factor of 0.1, with a minimum learning rate clip of $1e^{-7}$. The random seed is set as 0. The total batch size is set at 8, uniformly distributed among 4 × A6000 (48G) GPUs. In pseudo label generation, we follow previous works [49] to set α as 0.7 and K as 20. Our code is based on OpenPCDet [42] and is implemented in PyTorch [32].

4.1 Main Results

We present nuScenes results in Table 1 and observe that LiSe significantly outperforms all existing methods. In particular, compared to the state-of-the-art OYSTER, LiSe achieves an improvement of +7.1% in AP_{BEV} and +3.4% in AP_{3D} within the 0-80m range. In other distances, such as 0-30m, 30-50m, and 50-80m, LiSe consistently surpasses OYSTER, demonstrating a universally enhanced detection capability. The improvement validates the effectiveness of our proposed integration with 2D scenes, adaptive sampling strategy, and weak model aggregation in enhancing overall detection ability of model. It is also noteworthy that AP_{BEV} of LiSe in the long range (50-80m) even exceeds that of fully supervised results. These results affirm that incorporating 2D scene understanding significantly augments the detection of distant and small objects.

We further conduct experiments on Lyft, using the same hyper-parameters on nuScenes (see Table 2). We observe that the proposed LiSe surpasses the

Table 2: Comparison on Lyft. We observe that the proposed LiSe significantly surpasses MODEST across all evaluated metrics, especially in long range (50-80m). T denotes self-training round.

Method	**0-30m**	**30-50m**	**50-80m**	**0-80m**
Supervised	82.8 / 82.6	70.8 / 70.3	50.2 / 49.6	69.5 / 69.1
MODEST-PP ($T = 0$)	46.4 / 45.4	16.5 / 10.8	0.9 / 0.4	21.8 / 18.0
MODEST-PP ($T = 10$)	49.9 / 49.3	32.3 / 27.0	3.5 / 1.4	30.9 / 27.3
MODEST ($T = 0$)	65.7 / 63.0	41.4 / 36.0	8.9 / 5.7	42.5 / 37.9
MODEST ($T = 10$)	73.8 / 71.3	62.8 / 60.3	27.0 / 24.8	57.3 / 55.1
LiSe ($T = 0$)	54.5 / 54.0	24.2 / 22.8	1.4 / 1.2	29.2 / 27.5
LiSe ($T = 10$)	**76.7** / **74.0**	**66.1** / **64.4**	**46.6** / **43.7**	**65.6** / **62.5**

Table 3: Ablation studies on the primary components of the proposed method, including integration with 2D scenes (3D and 2D), adaptive sampling strategy (ADS), and weak model aggregation (WMA).

3D	**2D**	**ADS**	**WMA**	**0-30m**	**30-50m**	**50-80m**	**0-80m**
✓				24.8 / 17.1	5.5 / 1.4	1.5 / 0.3	11.8 / 6.6
	✓			31.8 / 14.0	3.8 / 0.8	0.6 / 0.0	12.4 / 4.7
✓	✓			31.4 / 19.9	8.3 / 3.1	3.4 / 0.9	16.2 / 9.1
✓	✓	✓		32.8 / 22.3	11.1 / 3.8	3.9 / 0.9	18.4 / 10.2
✓	✓		✓	34.3 / 23.3	10.0 / 4.1	4.0 / 1.3	18.5 / 11.0
✓	✓	✓	✓	**35.0** / **24.0**	**11.4** / **4.4**	**4.8** / **1.3**	**19.8** / **11.4**

competitive MODEST across all evaluated metrics. More importantly, LiSe outperforms MODEST by +19.4% in AP_{BEV} and +18.9% in AP_{3D} in the long range (50-80m), which contributes most to overall improvement. These results validate the effectiveness and generalizability of our proposed method.

4.2 Ablation Studies and Analyses

Effect of Integration with 2D Scenes. Comparing the first three rows in Table 3, we observe that integrating 2D scenes into 3D-based pseudo boxes yields the best overall performance across both AP_{BEV} and AP_{3D}. The significant improvement over LiDAR-based methods highlights the unique advantages of 2D scenes in detecting distant and small objects. We further examine the way to integrate 2D scenes (see Table 4). Specifically, we only incorporate 3D boxes from images with distance over 5m, 10m, and 15m. Table 4 indicates by integrating image-based boxes with distance over 10m achieves best performance. Such results also suggest that image-based boxes and LiDAR-based boxes conflict with each other in range 0-10 meters. Consideration of object distance in integration avoids such conflicts, and make two modalities complementary.

Table 4: Ablation studies on integrating image-based 3D boxes according to the distance. We incorporate image-based 3D boxes with a distance greater than 5, 10, and 15 meters (>5m, >10m, >15m). The term "All" refers to the use of all image-based boxes. We find that >10m yields the optimal results. This threshold balancedly integrates the advantage of 2D scenes and avoids conflict with LiDAR-based 3D boxes in near range.

3D	2D	0-30m	30-50m	50-80m	0-80m
All	All	29.3 / 19.8	4.7 / 2.3	2.4 / 0.5	13.8 / 8.0
All	>5m	30.7 / 19.8	**8.6** / **3.3**	3.1 / 0.7	15.4 / 8.9
All	>10m	**31.4** / **19.9**	8.3 / 3.1	**3.4** / **0.9**	**16.2** / **9.1**
All	>15m	30.2 / 20.6	5.7 / 2.8	2.1 / 0.4	14.3 / 8.9

Table 5: Adaptive sampling strategy based on the volume of objects (Volume) or the distance of objects (Distance). We could observe that volume-based strategy facilitates the distant objects in 30-50m, while distance-focused sampling improves box detection, with remarkable improvement in median distance. After combining two factors into consideration, we arrive at a balanced strategy for all ranges.

Volume	Distance	0-30m	30-50m	50-80m	0-80m
		31.4 / 19.9	8.3 / 3.1	3.4 / 0.9	16.2 / 9.1
✓		31.5 / 20.1	10.1 / 3.0	3.4 / 0.6	17.1 / 8.6
	✓	**35.0** / 22.2	10.5 / **3.8**	3.2 / 0.7	18.2 / 9.6
✓	✓	32.8 / **22.3**	**11.1** / **3.8**	**3.9** / **0.9**	**18.4** / **10.2**

Effect of Adaptive Sampling Strategy. Comparing rows 3 and 4 in Table 3, it shows that adaptive sampling further enhances model performance, particularly in the 30-50m and 50-80m ranges, validating its effectiveness in enhancing long-range detection capabilities. We also conduct a comprehensive ablation study to evaluate the design of adaptive sampling strategy. According to Table 5, the inclusion of a single metric, such as volume or distance, significantly enhances model performance. Notably, when incorporating a distance-based metric, both AP_{BEV} and AP_{3D} in the 30-50m range exceed the performance achieved with a volume-based metric alone. The improvement underscores the effectiveness of distance-based adaptive sampling in improving long-range detection capabilities. The combination of both volume-based and distance-based metrics yields the best performance, demonstrating that combined metric can more comprehensively addresses overlooking on long-tailed object groups. These two metrics work in a complementary manner, further enhancing overall model efficacy.

Effect of Weak Model Aggregation. Comparing rows 3 and 5 in Table 3, we observe that weak model aggregation alone enhances model performance. This improvement demonstrates effectiveness of weak model aggregation in creating a more robust model. Furthermore, when comparing rows 4, 5, and 6 in Table 3, we find combination of adaptive sampling strategy and weak model aggregation yields the best performance. This validates crucial role of interaction between adaptive sampling strategy and weak model aggregation. We also examine im-

Table 6: Ablation study of starting round T_s and the aggregation coefficient λ selection during the weak model aggregation. (1) We fix λ and study T_s. We observe that initiating the aggregation process at a later round, when the performance of model is higher and fluctuating, yields better results. (2) If we fix T_s as 6, we can see the large λ with slow update speed has achieved the best results in all different ranges, which stabilizes the prediction result.

T_s	λ	0-30m	30-50m	50-80m	0-80m
3	0.999	31.7 / 21.5	8.7 / 4.2	2.7 / 0.7	16.3 / 10.2
6	0.999	35.4 / 21.3	10.0 / 4.0	3.2 / 0.6	18.5 / 10.1
8	0.999	**34.3** / **23.3**	**10.0** / **4.1**	**4.0** / **1.3**	**18.5** / **11.0**
6	0.999	**35.4** / **21.3**	**10.0** / **4.0**	**3.2** / **0.6**	**18.5** / **10.1**
6	0.99	31.9 / 20.7	8.8 / 3.1	2.9 / 0.5	16.5 / 9.0
6	0.9	31.4 / 21.0	7.4 / 3.3	2.3 / 0.5	16.0 / 9.5

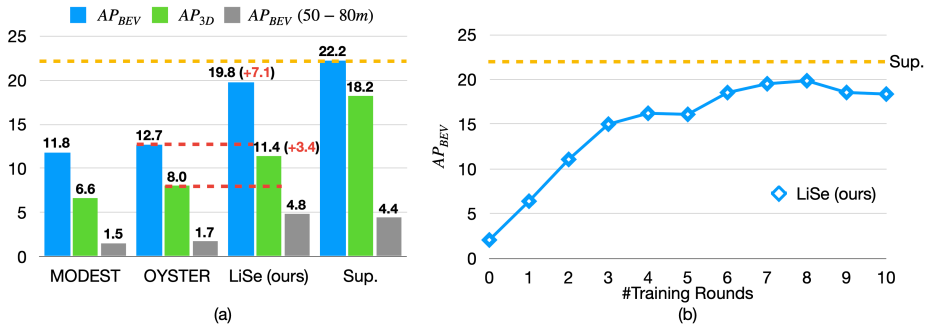

Fig. 4: Statistical analysis of performance of different models. (a) Visualization comparison of the performances of various methods. This comparison shows superiority of LiSe over purely LiDAR-based methods. (b) Visualization of performance changes throughout the training process. The trend shows combination of adaptive sampling strategy with weak model aggregation ensures a stable and effective training process.

pact of start round T_s and aggregation coefficient λ in Table 6. Initially, we fix λ at 0.999 and vary start round in 3, 6, and 8. The findings suggest that initiating aggregation process at a later round, when performance of model is higher and fluctuating, yields better results. During this period, models generally exhibit good performance with distinct strengths, making it an opportune time for weak model aggregation. Additionally, we fix start round and vary aggregation coefficient λ in 0.999, 0.99, and 0.9. Observations indicate larger coefficient, such as 0.999, leads to best performance. The ablation implies aggregation process benefits from being smoother and progressing in smaller steps.

Statistical analyses. We present statistical analyses of performance of different models in Figure 4. In Figure 4(a), LiSe significantly outperforms state-of-the-art model, OYSTER [55]. Notably, AP_{BEV} of LiSe at long ranges (50-80m) surpasses that of fully-supervised results. It verifies our integration with 2D scenes effec-

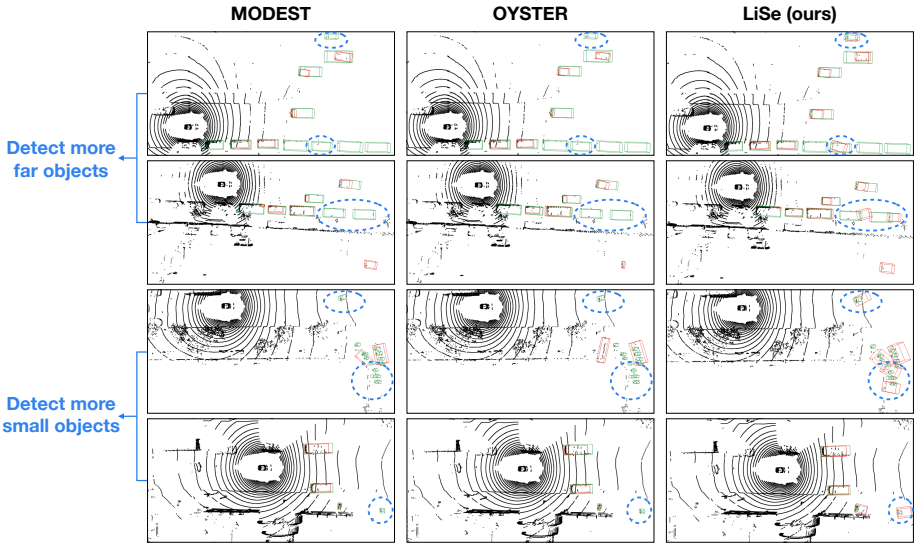

Fig. 5: Visualization comparison between MODEST [49], OYSTER [55], LiSe (ours), and ground truth boxes. The overall results indicate LiSe is superior in detecting distant and small objects. Green boxes represent ground truth labels, red boxes indicate predictions and blue circles highlight differences in predictions.

tively enhances capability of model to detect distant objects. In Figure 4(b), we can observe LiSe starts at a low point in initial training round, yet consistently achieves improved performance during self-paced training process.

Visualization. We present qualitative analysis in Figure 5. From rows 1 and 2, we observe LiSe excels at detecting distant objects, even when there are very limited points captured. Results in rows 3 and 4 indicate our LiSe framework performs significantly better than MODEST and OYSTER in detecting small objects. These results validate our obtained model is robust in detecting potentially existing objects, especially for challenging distant and small objects.

5 Conclusion

In this paper, we introduce a framework **LiSe** for unsupervised 3D detection. We propose integration with 2D scenes to improve detection ability in distant and small objects. In self-paced learning process, we further propose adaptive sampling strategy to continuously improve perception ability in challenging samples. Additionally, we introduce weak model aggregation, combining models trained under different distributions into a final, robust model. Extensive experiments affirm superior detection ability of our method. The comprehensive ablation studies and qualitative analyses also validate effectiveness of each proposed module. We hope our work will contribute to the fusion between 2D and 3D data for unsupervised 3D object detection and inspire future work in related fields.

Acknowledgement

The paper is supported by Start-up Research Grant at the University of Macau (SRG2024-00002-FST).

References

1. Adams, R., Bischof, L.: Seeded region growing. IEEE Transactions on pattern analysis and machine intelligence **16**(6), 641–647 (1994)
2. Bangalath, H., Maaz, M., Khattak, M.U., Khan, S.H., Shahbaz Khan, F.: Bridging the gap between object and image-level representations for open-vocabulary detection. Advances in Neural Information Processing Systems **35**, 33781–33794 (2022)
3. Buettner, K., Kovashka, A.: Enhancing the role of context in region-word alignment for object detection. arXiv:2303.10093 (2023)
4. 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)
5. Cen, J., Yun, P., Cai, J., Wang, M.Y., Liu, M.: Open-set 3d object detection. In: 2021 International Conference on 3D Vision (3DV). pp. 869–878. IEEE (2021)
6. Chen, M., Zheng, Z., Yang, Y., Chua, T.S.: Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1905–1914 (2023)
7. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017)
8. Cho, H.C., Jhoo, W.Y., Kang, W., Roh, B.: Open-vocabulary object detection using pseudo caption labels. arXiv:2303.13040 (2023)
9. Choi, S., Jang, J., Oh, C., Park, G.: Safety benefits of integrated pedestrian protection systems. International journal of automotive technology **17**, 473–482 (2016)
10. Dewan, A., Caselitz, T., Tipaldi, G.D., Burgard, W.: Motion-based detection and tracking in 3d lidar scans. In: 2016 IEEE international conference on robotics and automation (ICRA). pp. 4508–4513. IEEE (2016)
11. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD. vol. 96, pp. 226–231 (1996)
12. Gandhi, T., Trivedi, M.M.: Pedestrian protection systems: Issues, survey, and challenges. IEEE Transactions on intelligent Transportation systems **8**(3), 413–430 (2007)
13. Guo, Y., Yu, H., Ma, L., Luo, X., Xie, S.: Die-cdk: A discriminative information enhancement method with cross-modal domain knowledge for fine-grained ship detection. IEEE Transactions on Circuits and Systems for Video Technology (2024)
14. Guo, Y., Yu, H., Xie, S., Ma, L., Cao, X., Luo, X.: Dsca: A dual semantic correlation alignment method for domain adaptation object detection. Pattern Recognition **150**, 110329 (2024)
15. Horgan, J., Hughes, C., McDonald, J., Yogamani, S.: Vision-based driver assistance systems: Survey, taxonomy and advances. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. pp. 2032–2039. IEEE (2015)

16. Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V., Ondruska, P.: One thousand and one hours: Self-driving motion prediction dataset. In: Conference on Robot Learning. pp. 409–418. PMLR (2021)
17. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17853–17862 (2023)
18. Huang, Z., Liu, H., Lv, C.: Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. arXiv:2303.05760 (2023)
19. Jia, X., Gao, Y., Chen, L., Yan, J., Liu, P.L., Li, H.: Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7953–7963 (2023)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv:2304.02643 (2023)
21. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.L.: Joint 3d proposal generation and object detection from view aggregation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–8. IEEE (2018)
22. Kumar, M., Packer, B., Koller, D.: Self-paced learning for latent variable models. Advances in neural information processing systems **23** (2010)
23. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
24. Li, S., Zhu, X., Huang, Q., Xu, H., Kuo, C.C.J.: Multiple instance curriculum learning for weakly supervised object detection. arXiv:1711.09191 (2017)
25. Liang, J., Jiang, L., Meng, D., Hauptmann, A.G.: Learning to detect concepts from webly-labeled video data. In: IJCAI. vol. 1, pp. 3–1 (2016)
26. Liang, M., Yang, B., Wang, S., Urtasun, R.: Deep continuous fusion for multi-sensor 3d object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 641–656 (2018)
27. Liu, G., Shi, H., Kiani, A., Khreishah, A., Lee, J., Ansari, N., Liu, C., Yousef, M.M.: Smart traffic monitoring system using computer vision and edge computing. IEEE Transactions on Intelligent Transportation Systems **23**(8), 12027–12038 (2021)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv:2303.05499 (2023)
29. Najibi, M., Ji, J., Zhou, Y., Qi, C.R., Yan, X., Ettinger, S., Anguelov, D.: Motion inspired unsupervised perception and prediction in autonomous driving. In: European Conference on Computer Vision. pp. 424–443. Springer (2022)
30. Najibi, M., Ji, J., Zhou, Y., Qi, C.R., Yan, X., Ettinger, S., Anguelov, D.: Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8602–8612 (2023)
31. Pang, S., Morris, D., Radha, H.: Clocs: Camera-lidar object candidates fusion for 3d object detection. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10386–10393. IEEE (2020)

32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
33. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 918–927 (2018)
34. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7374–7383 (2019)
35. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 770–779 (2019)
36. Shi, Y., Lv, F., Wang, X., Xia, C., Li, S., Yang, S., Xi, T., Zhang, G.: Opentransmind: A new baseline and benchmark for 1st foundation model challenge of intelligent transportation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6327–6334 (2023)
37. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 761–769 (2016)
38. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV). pp. 464–472. IEEE (2017)
39. Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: Curriculum self-paced learning for cross-domain object detection. Computer Vision and Image Understanding **204**, 103166 (2021)
40. Tang, Y., Yang, Y.B., Gao, Y.: Self-paced dictionary learning for image classification. In: Proceedings of the 20th ACM international conference on Multimedia. pp. 833–836 (2012)
41. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
42. Team, O.D.: Openpcdet: An open-source toolbox for 3d object detection from point clouds. `https://github.com/open-mmlab/OpenPCDet` (2020)
43. Tian, H., Chen, Y., Dai, J., Zhang, Z., Zhu, X.: Unsupervised object detection with lidar clues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5962–5972 (2021)
44. Wang, L., Liu, Y., Du, P., Ding, Z., Liao, Y., Qi, Q., Chen, B., Liu, S.: Object-aware distillation pyramid for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11186–11196 (2023)
45. Wang, Z., Jia, K.: Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1742–1749. IEEE (2019)
46. Wei, Y., Su, S., Lu, J., Zhou, J.: Fgr: Frustum-aware geometric reasoning for weakly supervised 3d vehicle detection. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 4348–4354. IEEE (2021)
47. Wong, K., Wang, S., Ren, M., Liang, M., Urtasun, R.: Identifying unknown instances for autonomous driving. In: Conference on Robot Learning. pp. 384–393. PMLR (2020)

48. Yang, L., Balaji, Y., Lim, S.N., Shrivastava, A.: Curriculum manager for source selection in multi-source domain adaptation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 608–624. Springer (2020)
49. You, Y., Luo, K., Phoo, C.P., Chao, W.L., Sun, W., Hariharan, B., Campbell, M., Weinberger, K.Q.: Learning to detect mobile objects from lidar scans without labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1130–1140 (2022)
50. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems **34**, 18408–18419 (2021)
51. Zhang, D., Han, J., Zhao, L., Meng, D.: Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. International Journal of Computer Vision **127**, 363–380 (2019)
52. Zhang, D., Yang, L., Meng, D., Xu, D., Han, J.: Spftn: A self-paced fine-tuning network for segmenting objects in weakly labelled videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4429–4437 (2017)
53. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. Advances in Neural Information Processing Systems **35**, 36067–36080 (2022)
54. Zhang, H., Xu, J., Tang, T., Sun, H., Yu, X., Huang, Z., Yu, K.: Opensight: A simple open-vocabulary framework for lidar-based object detection. In: Proceedings of the European conference on computer vision (ECCV) (2024)
55. Zhang, L., Yang, A.J., Xiong, Y., Casas, S., Yang, B., Ren, M., Urtasun, R.: Towards unsupervised object detection from lidar point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9317–9328 (2023)
56. Zhang, X., Xu, W., Dong, C., Dolan, J.M.: Efficient l-shape fitting for vehicle detection using laser scanners. In: 2017 IEEE Intelligent Vehicles Symposium (IV). pp. 54–59. IEEE (2017)
57. Zheng, Z., Yang, Y.: Adaptive boosting for domain adaptation: Toward robust predictions in scene segmentation. IEEE Transactions on Image Processing **31**, 5371–5382 (2022)
58. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **29**(10), 3037–3045 (2018)