CSOT: Cross-Scan Object Transfer for Semi-Supervised LiDAR Object Detection

Jinglin Zhan[1], Tiejun Liu[1], Rengang Li[1], Zhaoxiang Zhang[2], and Yuntao Chen[3⋆]

[1] IEIT Systems
[2] Institute of Automation, CAS
[3] Centre for Artificial Intelligence and Robotics, HKISI, CAS
{zhanjinglin,liutj,lirg}@ieisystem.com,
zhaoxiang.zhang@ia.ac.cn, chenyuntao08@gmail.com

Abstract. Large-scale 3D bounding box annotation is crucial for LiDAR object detection but comes at a high cost. Semi-supervised object detection (SSOD) offers promising solutions to leverage unannotated data, but the predominant pseudo-labeling approach requires careful hyperparameter tuning for training on noisy teacher labels. In this work, we propose a Cross-Scan Object Transfer (CSOT) paradigm for LiDAR SSOD. Central to our approach is HotspotNet, a transformer-based network that predicts possible placement locations and the object-place fitness scores for inserting annotated objects into unlabeled scans in a semantic coherence manner. Based on HotspotNet, CSOT successfully enables object copy-paste in LiDAR SSOD for the first time. To train object detectors on partially annotated scans generated by CSOT, we adopt a spatial-aware classification loss throughout our partial supervision to handle false negative issues caused by treating all unlabeled objects as background. We conduct extensive experiments to verify the efficacy and generality of our method. Compared to other state-of-the-art label-efficient methods used in LiDAR detection, our approach requires the least amount of annotation while achieves the best detector. Using only 1% of the labeled data on the Waymo dataset, our semi-supervised detector achieves performance on par with the fully supervised baseline. Similarly, on the nuScenes dataset, our semi-supervised CenterPoint reaches 99% of the fully supervised model's detection performance in terms of NDS score, while using just 5% of the labeled data. Code is released at https://github.com/JinglinZhan/CSOT

Keywords: Semi-Supervised Learning · LiDAR Object Detection · Autonomous Driving

1 Introduction

Large-scale annotated datasets have catalyzed the success of LiDAR object detection, but at a significant cost in terms of money and time [25,33]. For instance,

⋆ Corresponding author.

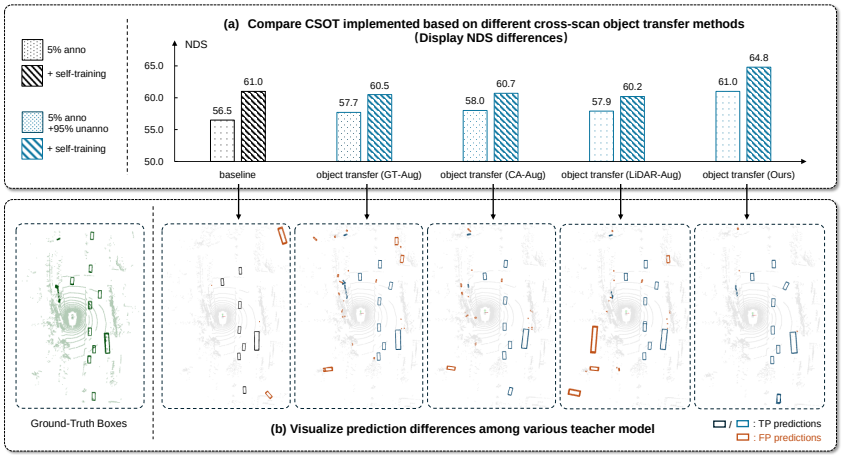

Fig. 1: Preliminary experiments of CSOT pipeline for LiDAR SSOD. (a) Compare CSOT implemented based on different cross-scan object transfer methods and display the NDS differences. (b) Visualize prediction differences among various teacher model that used to generate pseudo labels for self-training.

annotating the 12 million 3D boxes in the Waymo Open [38] dataset may have consumed millions of dollars and tens of thousands of human hours. In contrast, autonomous driving fleets capture thousands, and even millions, of hours of unannotated sensor data every day. The huge gap between the speed of manual labeling and that of data accumulation makes labeling a major slowdown. This calls for a more scalable approach to leverage the abundant data collected by autonomous fleets and continually advance object detection capabilities.

Semi-Supervised Learning (SSL) presents a promising way for reducing annotation demands by leveraging both limited labeled data and vast quantities of unlabeled data during training. Various SSL methods have been proposed, including pseudo-labeling (a.k.a. self-training) [23], co-training [2], consistency regularization [20] and generative modeling [19]. Among these SSL approaches, pseudo-labeling has gained significant traction, particularly in the field of semi-supervised object detection (SSOD) [4, 12, 31, 37, 39, 40, 42, 52, 62, 65]. However, generating reliable pseudo-labels without human intervention remains a significant challenge. Noisy pseudo-labels can negatively impact model performance, leading to suboptimal results. To address this, researchers have proposed various techniques, including sophisticated post-processing mechanisms [4,5,26,28,40,52, 60], iterative training procedures [39,45,49], and noise-tolerant detectors [12,65] and loss objectives [31, 32, 62, 64]. Nevertheless, these approaches often involve complex pipelines and require careful parameter tuning, making them inefficient and difficult to adapt to new datasets.

In contrast with noisy pseudo-labeling, we leverage a novel semi-supervised learning approach for 3D object detection that transfers annotated objects to generate noise-free training data. Specifically, we transfer point clouds and bound-

ing boxes of annotated objects to unannotated LiDAR scans to create partially labeled data. Object transfer via copy-paste is a commonly employed data augmentation [6,7,16,47,51] in LiDAR object detection. But naively applying object copy-paste [11, 15, 47] to semi-supervised LiDAR object detection yields sub-optimal results. As shown in Fig. 1, SSOD built upon vanilla object transfer does improve performance over the baseline, but combining it with self-training results in a negative net effect on detection performance compared to using self-training alone. The key insight is that object copy-paste without careful contextual consistency maintenance improves detector performance at the cost of generating a large number of false positives. Although the commonly used object detection metric mAP is not sensitive to false positives if the recall improves, the self-training pipeline accumulates these false positives over iterations.

In order to achieve coherent object copy-paste, we propose Hotspot Network, a learning-based approach to insert objects with strong contextual consistency, enabling object-transfer-based semi-supervised LiDAR object detection to achieve its full potential. HotspotNet takes unannotated LiDAR scans as inputs, along with point clouds and bounding boxes of annotated objects. It then predicts possible placement locations for these objects and assigns scores to each location, indicating the likelihood of the object fitting well within the context of the scanned environment. This allows us introducing noise-free supervision signal into unlabeled point clouds and use these newly created partially-labeled data to optimize detector performance. Based on HotspotNet, we design a novel Cross-Scan Object Transfer(CSOT) pipeline for LiDAR SSOD. CSOT enjoys precise partial annotations that maintain great contextual coherence with unannotated scans rather than noisy pseudo labels for training detectors on unlabeled data. It's important to note that training detectors on partially annotated point clouds, as opposed to fully labeled ones, can lead to a significant issue: unlabeled objects being miscategorized as background, resulting in a high number of false negatives. To mitigate this problem, we propose a spatial-aware classification loss and employ it throughout our partial supervision. Our CSOT paradigm features a streamlined training pipeline and requires minimal hyper-parameter tuning, enabling it to quickly generalize across various datasets and detectors. Moreover, this approach is orthogonal and complementary to commonly used self-training techniques. We conducted extensive experiments on the Waymo [38] and nuScenes [3] datasets to confirm the effectiveness of our methods.

Our contributions are summmarized below:

- We propose a Hotspot Network for instructing object placement in a semantic coherence manner, and employ an asymmetric supervision for training HotspotNet on both annotated and unannotated data.
- Based on HotspotNet, we design a novel Cross-Scan Object Transfer paradigm for LiDAR semi-supervised object detection. To the best of our knowledge, this is the first successful attempt of object copy-paste in LiDAR SSOD.
- We conduct extensive experiments on nuScenes and Waymo datasets to verify the efficacy and generality of our method. Our approach achieves competitive performance compared to fully-supervised baselines while using 20x

to 100x fewer annotations, significantly outperforming other state-of-the-art label-efficient approaches in LiDAR object detection.

2 Related Work

2.1 Semi-Supervised Object Detection

Unsupervised learning eliminates data annotation but significantly degrades model performance [48, 55, 61]. Semi-supervised methods maintain low annotation costs meanwhile greatly bridges the gap to fully-supervised approaches [43]. As a prior work, STAC [37] proposes a pseudo labeling SSOD paradigm and catalyzes extensive follow-on researches. It trains a teacher with all annotated data and generates pseudo labels for supervising the student on unlabeled data. Producing reliable pseudo labels is critical in this paradigm and numerous methods have been explored. 3DIoUMatch [40] utilizes the estimated 3D IoU to filter poorly localized predictions. LabelMatch [4] leverages adaptive label-distribution-aware confidence thresholds for generating unbiased pseudo labels. Proficient Teacher [52] employs clustering-based box voting for improving the precision of final predictions. HSSDA [29] designs a dynamic dual-threshold strategy to generates more reasonable supervision for training the student model. Besides various filter mechanism, many other approaches, such as iterative training procedures [39, 62], noise-tolerant detectors [12, 65] and a variety of loss objectives with consistency or reweighting terms [31,42,62], are also proposed for alleviating negative effects caused by noisy pseudo labels. Despite some progress has been made, unreliable pseudo-labels still remain an open challenge. The sophisticated pipelines with extensive manual threshold tuning also impede these approaches' quick adaptation to new datasets.

2.2 Copy-Paste For LiDAR Object Detection

Copy-paste method is developed by Jaderberg et.al. for text recognition [18], then extended by Dwibedi et.al. [9] and Yun et.al. [57] for 2D object detection. Yan et.al. [47] introduces copy-paste into LiDAR object detection and develops the well known ground-truth augmention approach (so called GT-Aug), which is extensively adopted in many later studies [1,6,16,17,21,51,54,66]. Since then, numerous approaches have been proposed to optimize GT-Aug. Part-Aware [8] and Shape-Aware [63] GT-Aug divide objects into partitions and apply different augmentation methods to each partition. Pattern-Aware [14] GT-Aug downsamples the points of objects to create a new one with farther distance. PointCutMix [22] replaces part of the sample with shape-preserved subsets from another one. Some researchers notice placement issues in GT-Aug and develop hand-craft valid maps to restrict the insertion of objects [11, 15]. Up to now, researches on copy-paste is primarily focused on exploring and developing its capabilities for augmenting data to support fully-supervised learning tasks. Ghiasi et.al. introduce copy-paste into SSL [13] for augmenting both pseudo labeled and supervised labeled images. The above trial still limited in the field of data augmentation and the copy-paste technique remains largely untapped.

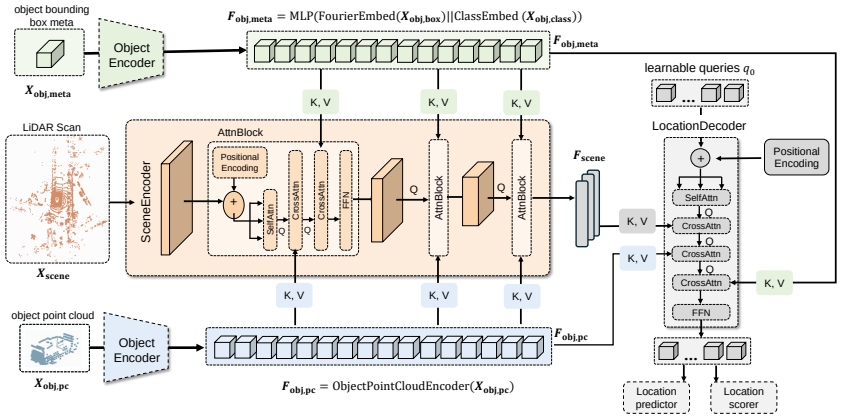

Fig. 2: The overall architecture of HotspotNet. The Object Encoders extract features from object point cloud $X_{obj,pc}$ and bounding box meta data $X_{obj,meta}$. The scene encoder featurize the scene point cloud and cross-attends to object features $\mathcal{F}_{\text{obj,pc}}$ and $\mathcal{F}_{\text{obj,meta}}$. The location decoder predicts the placeable locations and their scores.

3 Method

We provide details of Cross-Scan Object Transfer SSOD paradigm in this section. We present the overall architecture of HotspotNet in Sec. 3.1, and describe the Asymmetric Supervision for training HotspotNet in Sec. 3.2. The pipeline of our newly proposed SSOD paradigm, CSOT, is elaborated in Sec. 3.3.

3.1 Hotspot Network

The key to LiDAR copy-paste is identifying semantically coherent locations within background scans to paste objects. Placing objects at unreasonable locations, such as inside walls, floating in sky or with weird pose, will cause a large amount of false positives and thus deteriorate the detector performance.

We propose a learning-based approach for object placement that provides sufficient supervision signal and maintain a high signal-to-noise ratio. Specifically, we employ a Hotspot Network to predict the placeable location $\mathbf{b}$ and the object-place fitness scores o for inserting object in a semantic coherence manner.

We abbreviate Hotspot Network as HotspotNet and visualize the overall architecture in Fig. 2. The HotspotNet comprises object encoders for object point cloud $X_{\text{obj,pc}}$ and meta data $X_{\text{obj,meta}}$, respectively. The scene encoder featurize the scene point cloud and cross-attends to object features $\mathcal{F}_{\text{obj,pc}}$ and $\mathcal{F}_{\text{obj,meta}}$. A location decoder predicts the placeable location $\mathbf{b} = \{x, y, z, w, l, h, \theta, c\}$ and the location fitness score o.

Object Encoder We process the object point cloud $X_{\text{obj,pc}}$ with object point cloud encoders like PointNet [34] or PointNeXt [35] and obtain object features $\mathcal{F}_{\text{obj,pc}}$. We encode the bounding box meta $X_{\text{obj,meta}}$ with appropriate meta feature encoders like Fourier embedding for boxes, and CLIP or one-hot embedding

for object classes. We then transform the initial box embedding with a MLP to obtain bounding box meta feature $\mathcal{F}_{\text{obj,meta}}$.

$$\mathcal{F}_{\text{obj,pc}} = \texttt{ObjectPointCloudEncoder}(X_{\text{obj,pc}}) \tag{1}$$

$$\mathcal{F}_{\text{obj,meta}} = \texttt{MLP}(\texttt{FourierEmbed}(X_{\text{obj,box}}) \mid\mid \texttt{ClassEmbed}(X_{\text{obj,class}})) \tag{2}$$

Scene Encoder Given a scene LiDAR scan X_{scene}, we leverage off-the-shelf point cloud encoders like CenterPoint [54] or SST [10] for extracting features $\mathcal{F}_{\text{scene}}$. We use spatial cross-attention in the first convolution block of each stage of the scene point cloud encoder. This allows the scene encoder aggregating features from both the object point clouds and the box metadata for assessing the object fitness of each spatial location.

Location Decoder We randomly initialize a set of learnable queries $\mathbf{q}_0$. These queries are combine with position embeddings to obtain location queries $\mathbf{q}$. Our location decoder leverages both scene features and object features for predicting possible inserting locations and scores .

3.2 Training HotspotNet with Asymmetric Supervision

The HotspotNet consists of a location predictor and a location scorer that work jointly to suggest object placements. The location predictor outputs candidate object positions, while the scorer assigns fitness scores to evaluate how semantically consistent each candidate is. We employ a multi-task learning approach that features Asymmetric Supervision for optimizing different predictions. The proposed losses supervise the location predictor and scorer to accurately position objects within a new environment.

To train the location predictor, we utilize the annotated LiDAR scans. For each annotated scan, we randomly extract an object point cloud from the original point clouds along with its corresponding bounding box. The regressor is then supervised with an ℓ_1 loss between the predicted location $\mathbf{b}$ and the extracted bounding box $\mathbf{b}_{\text{cut-out}}$. We constrain object distance and perspective from the ego car in Eq. 3 to preserve perspective and point density. We employ dropout regularization on the bounding box meta data to prevent the location predictor from directly copying the meta data to the prediction output. It encourages the predictor to learn meaningful features from the point clouds itself, rather than relying solely on the provided meta data.

$$\mathcal{L}_{\text{loc}}(\mathbf{b}) = \|\mathbf{b} - \mathbf{b}_{\text{cut-out}}\|_1 \tag{3}$$

To train the location scorer, we utilize an asymmetric supervision approach for positive and negative samples. Positive samples are obtained by matching location queries with the extracted object point clouds. These positive samples encourage the scorer to assign high scores to queries that accurately position objects within the scan.

$$\mathcal{L}^{p}_{\text{score}}(o) = -\log o \tag{4}$$

To avoid treating all queries that do not match the extracted objects as negative samples, a careful selection process is employed. Instead of using all non-matching queries, negative samples are chosen based on queries that have low fitness scores, denoted by s. This approach ensures that the scorer learns to differentiate between suitable and unsuitable object placements, focusing on the most informative negative samples.

$$\mathcal{L}^n_{\text{score}}(o) = -(1 - \mathbb{1}[s_p \cdot s_o \cdot s_c > \tau]) \log(1 - o) \tag{5}$$

The fitness scores s_p, s_o, and s_c are used to measure the accuracy and coherence of an object's position, orientation, and category within the scene. The position score $s_p = \exp(-(\mathbf{b}_{xyz} \cdot \mathbf{n}_g + d_g)/\sigma_p^2)$[1] penalizes the predicted location $\mathbf{b}_{xyz}$ deviating from the local ground plane specified by $(\mathbf{n}_g, d_g)$. Local ground planes here could be obtained via non-learning ground detection methods like Patch-Works++ [24]. Similarly, the orientation score $s_o = \cos((\mathbf{b}_\theta - \hat{\mathbf{b}}_\theta)/2)$ encourages consistency between the heading of the new object location and the predominant heading $\hat{\mathbf{b}}_\theta$ of objects in the same category within the scene. Finally, the category compatible score $s_c = \mathbb{1}[\mathbf{b}_c \in \mathcal{C}]$ prevents pasting object of category that is not in the current scene(e.g. bicycle on highway). Notably, these scores can be computed not only between the extracted object and its original scan but also between the extracted object and other unannotated LiDAR scans, in which $\hat{\mathbf{b}}_\theta$ can be simplified as ego orientation and s_c can be ignored directly. This adaptability opens up the possibility of learning object placement scoring from a more extensive dataset through unsupervised learning, which assigns positive and negative samples solely based on the fitness scores.

Combining these asymmetric losses and carefully selecting positive and negative samples, our HotspotNet learns to accurately position objects within a new scan while maintaining semantic coherence.

3.3 Cross-Scan Object Transfer for LiDAR-based SSOD

Under the guidance of HotspotNet (described in Sec. 3.1), we inject precise supervision signals into unannotated point clouds and train detectors on these partially-labeled data. Fig. 3 illustrates the differences between our Cross-Scan Object Transfer SSOD paradigm and the commonly used self-training technique. Our method bypasses the iterative self-training procedures, cumbersome pseudo-label filtering mechanisms, and complex loss objective designs.

Partially Supervised Object Detection Our approach for introducing supervision into unlabeled LiDAR scans is straightforward. We denote all objects in labeled point clouds as $\{\mathbf{O}^L\}$, and all unlabeled scans as $\{\mathbf{S}^U\}$. For an unlabeled LiDAR scan $\mathbf{S}^U$, we randomly sample a subset of $\{\mathbf{O}^L\}$ and follow the instruction of HotspotNet to inject $\{\mathbf{O}^L\}$ into $\mathbf{S}^U$ in a semantically coherent manner. Since inserted objects come with ground-truth classes, poses and dimensions,

[1] $\mathbf{b}_\square$ denotes the $\square$ component for location prediction $\mathbf{b}$.

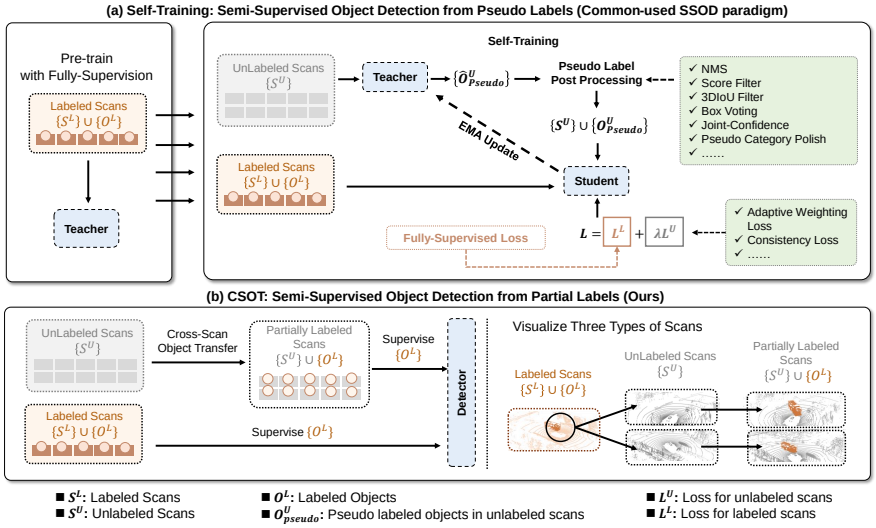

Fig. 3: (a) Common-used pseudo-labeling SSOD paradigm. (b) Cross-Scan Object Transfer SSOD paradigm, which bypasses the iterative self-training procedures, cumbersome pseudo-labels filter mechanisms and complex loss objectives designs.

these newly created scans allow for supervision. However, it is important to note that the above synthesized point clouds are only partially annotated, as we lack information about the original unlabeled objects in these unannotated scans. Learning a detector from partially labeled data is known as Sparsely Annotated Object Detection(SAOD) [30, 41]. The differences of training a detector from fully and partially annotated data will be discussed in the next part.

Spatial-Aware Classification Loss An object detector generally includes a classifier to differentiate foreground objects from background clutter, and a localizer to predict the accurate location, dimension and orientation of an object. The localizer is typically supervised by regressing the offset from current voxel location to actual bounding boxes of foreground objects, so it can be naturally supervised by new point clouds generated via cross-scan object transfer. However, the classifier requires both positive and negative samples for learning a decision boundary. Once again, the inserted objects could provide positive samples for the classifier but we can not treat all other voxels as background in SAOD just like vanilla object detection. This is due to the existence of the so-called hard negatives as reported in other SAOD works [41].

In our approach, hard negatives refer to the original unlabeled objects intrinsically present in the unannotated LiDAR scans. Training the classifier in our semi-supervised detector with these hard negatives will result in an unacceptable number of false negatives. Instead of resorting to sophisticated re-weighting/recalibration strategies [44, 59] or teacher-student siamese networks [41], we simply

choose negative samples in a spatial-aware manner to avoid sampling unlabeled objects as negatives. We propose a spatial-aware classification loss for our partial supervision as Eq.6:

$$L_{\text{cls}}^{\text{n}}(\mathbf{x}) = \mathbb{1}[(\min_i \|\mathbf{x} - \mathbf{c}_i\|_2) \leq \sigma] \cdot L_{\text{cls}}^{\text{n}}(\mathbf{x}) \tag{6}$$

where $L_{\text{cls}}^{\text{n}}(\mathbf{x})$ denotes focal loss for negative samples at the heatmap location $\mathbf{x}$. The indicator function $\mathbb{1}(\cdot)$ equals 1 when the $\ell 2$ distance between a location $\mathbf{x}$ and the center location $\mathbf{c}$ of any foreground object is less than a threshold σ, and 0 otherwise. We set σ as twice of the Gaussian radius used for positive samples assignment in CenterPoint [54] for unlabeled scans, and infinity for fully annotated data. An immediate drawback of our spatial-aware classification loss is that there are too few negative samples for the classifier. To alleviate this issue, we randomly sample 1% samples in the heatmap predictions as negative samples, and set the corresponding value of $\mathbb{1}(\cdot)$ as 1 to make the classification boundary between positive and negative samples more clear.

Synergy with Self-Training Techniques A deeper analysis of the proposed CSOT SSOD paradigm reveals that the primary mechanism driving improved detection performance is the extrapolation of labeled object detection to novel contextual environments. This stands in contrast to self-training methods, which learn from pseudo-labeled objects in unannotated scans.

As a result, the proposed CSOT pipeline for semi-supervised LiDAR object detection is orthogonal to, yet complementary with, existing self-training techniques [37], allowing for additional performance gains. Deliberately designing a pseudo-labeling SSOD framework is beyond the scope of this paper. To verify the complementary effect of our approach when combined with self-training techniques, we implement a basic teacher-student paradigm. We employ a single universal confidence score threshold τ_{pseudo}, which is invariant across categories, to filter low-confidence predictions. We then utilize a fully-supervised loss, as in CenterPoint [54], for optimizing the student detector.

4 Experiment

The proposed methods are verified in this section. We show implementation details in Sec. 4.1 and provide all experimental results in Sec. 4.2 and Sec. 4.3. More ablations and discussions are shown in our supplementary materials.

4.1 Experimental Setup

Datasets We validate our method on two widely used autonomous driving datasets: nuScenes [3] and Waymo [38]. The nuScenes dataset has 700 training, 150 validation, and 150 test sequences. Official nuScenes detection metrics are NuScenes Detection Score (NDS) and mean Average Precision (mAP). The Waymo Open Dataset consists of 798 training and 202 validation sequences. Official Waymo detection metrics are mean Average Precision (mAP) and mean Average Precision weighted by Heading (mAPH).

Training Details We use one-stage CenterPoint [54] as our baseline detector to validate the effectiveness of our methods. Voxel sizes are set as [0.075m, 0.075m, 0.2m] for nuScenes and [0.1m, 0.1m, 0.15m] for Waymo. The training epochs for one cycle of learning procedure, including fully-supervised learning, partially-supervised learning and self-training, are set as 20 for nuScenes and 7 for Waymo. We follow many previous work [6, 51, 54, 66] to adopt Adam optimizer with one-cycle learning rate policy. Labeled data are uniformly sampled from the entire dataset. Then objects and their annotations are extracted to build the ground-truth database for injecting precise supervision signals into unlabeled scans. We finetune partially-supervised detectors on labeled scans with several epochs. We set a score threshold of 0.3 for nuScenes and 0.6 for Waymo to filter low-quality pseudo labels, and verify the orthogonality of our methods with self-training.

4.2 Benchmarks

Benchmarks on nuScenes dataset We compare our methods against other label-efficient LiDAR detection approaches on the nuScenes dataset. As shown in Tab. 1, our methods obtain the significant performance improvements over previous works. For the sake of fairness, we choose the fully supervised detector trained by Yin et.al. [54] (the designer of CenterPoint backbone) as the baseline detector, and list the results achieved from other label-efficient LiDAR object detection approaches for comparison [27, 43, 56, 58]. Without leveraging models pre-trained on external datasets and using only 5% of the annotations, we achieve competitive detector which reaches 97.2% of the fully-supervised performance on mAP and 98.9% on NDS. We significantly shrink the mAP and NDS gaps between the baseline detector and the one trained with 20x less annotations from around 10∼20 to merely 1.6 and 0.7 respectively. Our semi-supervised detector, trained with only 5% manual annotations, obtains over 90% detection performance on 8 of the 10 categories defined in nuScenes dataset, and over 95% detection performance on 6 categories.

Benchmarks on Waymo dataset We extend our methods on Waymo dataset and achieve similar outstanding results. we choose the fully supervised detector trained by Shi et.al. [36], Yan et.al. [47], Fan et.al. [10] and Yin et.al. [54] as the baseline detector, and list the results achieved from other label-efficient LiDAR object detection approaches for comparison [29, 46, 50, 52, 53, 56]. As shown in Tab. 2, our methods are valid on both anchor-based (SECOND) and anchor-free (CenterPoint) detectors. Taking CenterPoint as an example, with merely 1% annotations, we achieve a competitive detector which matches fully supervised performance on vehicle and pedestrian, and exceeds it on cyclist. The 100x reduction in annotations leads to only a minor 1.8 APH decrease for L2 Vehicle, a negligible 0.4 APH drop for L2 Pedestrian, and boosts the APH of L2 Cyclist by 4.1. The abnormal enhanced detection performance on Cyclist should be attributed to our learning-based approach for cross-scan object transfer. Instructed by the proposed HotspotNet, we paste objects in a semantic coherent

Table 1: Compare with other labeled efficient LiDAR object detection approaches on nuScenes Dataset validation split. Performances of the baseline detector reported by its original designer [54] are marked with gray backgrounds.

Method	Ref	Annos	External	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Mot.	Byc.	Ped.	T.C.
Baseline	CVPR 2021 [54]	100%		58.0	65.5	84.6	51.0	17.5	60.2	53.2	70.9	53.7	28.7	83.4	76.7
GCC3D	ICCV 2021 [27]	5%		41.1	46.8	—	—	—	—	—	—	—	—	—	—
SceneFlow	ECCV 2022 [58]	5%		36.0	48.3	—	—	—	—	—	—	—	—	—	—
SSDA3D	AAAI 2023 [43]	5%	✓	—	—	76.2	—	—	—	—	—	—	—	—	—
AD-PT	NeurIPS 2023 [56]	5%	✓	45.0	53.0	78.9	43.8	11.1	55.2	21.2	55.1	39.0	17.8	72.3	55.4
Ours	—	**5%**		**56.4**	**64.8**	**81.7**	**53.5**	**16.5**	**58.6**	**32.8**	**66.0**	**60.0**	**47.8**	**82.0**	**65.5**

Table 2: Compare with other labeled efficient LiDAR object detection approaches on Waymo Dataset validation split. Performances of the baseline detector reported by its original designers [10, 36, 47, 54] are mark with gray backgrounds.

Detector	Method	Ref	Annos	Ext	Vehicle				Pedestrian				Cyclist				All
					L1		L2		L1		L2		L1		L2		L2
					AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	APH
PV-RCNN	Baseline	CVPR 2020 [36]	100%		77.5	76.9	69.0	68.4	75.0	65.7	66.0	57.6	67.8	66.4	65.4	64.0	63.3
PV-RCNN	HSSDA	CVPR 2023 [29]	1%		56.4	53.8	49.7	47.3	40.1	20.9	33.5	17.5	29.1	20.9	27.9	20.0	28.3
PV-RCNN	MIXSUP	ICLR 2024 [50]	1%		—	—	—	55.5	—	—	—	52.0	—	—	—	62.3	56.6
SECOND	Baseline	Sensors 2018 [47]	100%		72.3	71.7	63.9	63.3	68.7	58.2	60.7	51.3	60.6	59.3	58.3	57.1	57.2
SECOND	MIXSUP	ICLR 2024 [50]	10%		—	—	—	55.0	—	—	—	49.6	—	—	—	58.1	54.2
SECOND	ProficientTeacher	ECCV 2022 [52]	5%		—	—	53.0	52.5	—	—	50.3	38.7	—	—	49.9	46.0	45.7
SECOND	AD-PT	NeurIPS 2023 [56]	3%	✓	—	—	60.5	59.9	—	—	54.9	45.8	—	—	50.8	49.7	51.8
SECOND	ProposalContrast	ECCV 2022 [53]	1%		—	—	37.6	36.9	—	—	39.7	31.7	—	—	37.7	35.7	34.8
SECOND	**Ours**	—	**1%**		**69.4**	**68.9**	**61.2**	**60.8**	**64.0**	**60.7**	**55.1**	**52.3**	**61.1**	**60.4**	**58.9**	**58.2**	**57.1**
SST	Baseline	CVPR 2022 [10]	100%		74.2	73.8	65.5	65.1	78.7	69.6	70.0	61.7	70.7	69.6	68.0	66.9	64.6
SST	MIXSUP	ICLR 2024 [50]	10%		—	—	—	59.1	—	—	—	60.0	—	—	—	63.1	60.7
SST	MV-JAR	CVPR 2023 [46]	5%		—	—	56.5	56.0	—	—	57.7	47.7	—	—	37.4	36.3	46.7
CenterPoint	Baseline	CVPR 2021 [54]	100%		76.6	76.0	68.9	68.4	79.0	73.4	71.0	65.8	72.1	71.0	69.5	68.5	67.6
CenterPoint	MIXSUP	ICLR 2024 [50]	10%		—	—	—	61.8	—	—	—	57.7	—	—	—	67.5	62.3
CenterPoint	AD-PT	NeurIPS 2023 [56]	3%	✓	—	—	60.4	59.8	—	—	60.6	54.0	—	—	62.7	61.6	58.5
CenterPoint	**Ours**	—	**1%**		**75.1**	**74.7**	**67.0**	**66.6**	**75.9**	**72.9**	**68.2**	**65.4**	**75.9**	**75.1**	**73.4**	**72.6**	**68.2**

manner, which enables our CSOT SSOD to reach its full potential and meanwhile can be used as a powerful data augmentation for fully-supervised detector. Related ablation results are discussed in Sec. 4.3. Notably, under the most stringent constraints with minimal labeled point clouds and without leveraging any external pre-training, our approach stands alone as the only label-efficient technique which is able to achieve equivalence detection performance to the baseline detector on the average L2 APH of all categories defined in Waymo dataset.

4.3 Ablations

Ablate training stage on nuScenes dataset. Our semi-supervised detectors trained with 20x fewer nuScenes annotations are compared to both fully supervised and pseudo-labeling semi-supervised models. Results in Tab. 3 indicate that our newly proposed CSOT is a valid label-efficient approach, and it almost completely maintains the performance gains of self-training when used together. The fully supervised detector trained with 5% labeled scans shows substantial performance degradation compared to that trained on the entire labeled dataset, with a mAP decline from 59.5 to 43.8 and a NDS decrease from 66.0 to 56.5. As shown in Lines 3-5 of Tab. 3, self-training technique is able to boost the mAP from 43.8 to 51.0 and NDS from 56.5 to 61.0. Our CSOT, which allows

Table 3: Compare with fully supervised and pseudo-labeling semi-supervised detectors. All results are achieved based on our implementation code of CenterPoint and evaluated on nuScenes dataset validation split.

	Annos	Scans	CSOT	Self-Training	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Mot.	Byc.	Ped.	T.C.
1	100%	100%			59.5	66.0	83.7	56.6	18.7	66.9	34.8	67.1	63.7	50.4	84.5	68.7
2	5%	5%			43.8	56.5	75.8	39.7	12.3	44.1	20.5	56.9	36.3	18.4	76.7	56.9
3	5%	100%		✓	51.0	61.0	80.1	49.1	14.6	55.6	26.7	62.3	45.4	30.8	81.9	63.1
4	5%	100%	✓		50.4	61.0	79.6	46.4	11.7	50.1	25.6	59.1	52.1	37.2	80.6	62.0
5	5%	100%	✓	✓	**56.4**	**64.8**	81.7	53.5	16.5	58.6	32.8	66.0	60.0	47.8	82.0	65.5

Table 4: Compare with fully supervised and pseudo-labeling semi-supervised detectors. All results are achieved based on our implementation code of CenterPoint and evaluated on Waymo dataset validation split (mean L1/L2 AP/APH over all categories defined in Waymo dataset).

	Annos	Scans	CSOT	Self-Training	Object Transfer	L1 AP/APH	L2 AP/APH	mAP/mAPH
1	100%	100%			—	74.4 / 72.7	68.6 / 67.0	71.5 / 69.9
1a	100%	100%			GT-Aug	74.7 / 73.1	68.8 / 67.3	71.8 / 70.2
1b	100%	100%			Ours	77.5 / 76.0	71.6 / 70.2	74.6 / 73.1
2	1%	1%			—	68.3 / 66.6	62.3 / 60.8	65.3 / 63.7
3	1%	100%		✓	—	71.9 / 70.4	65.8 / 64.5	68.9 / 67.5
4	1%	100%	✓		Ours	71.4 / 69.8	65.4 / 63.9	68.4 / 66.9
5	1%	100%	✓	✓	Ours	75.7 / 74.2	69.5 / 68.2	72.6 / 71.2

supervision on unannotated point clouds by converting unlabeled scans into partially labeled scans via learning-based copy-paste, achieves a similar detection performance (with a mAP of 50.4 and a NDS of 61.0). It is noteworthy that iterative employment of self-training techniques does not lead to sustained gains in detection performance (see supplementary material), while incorporating just one cycle of self-training on our partially supervised detector is able to further optimize the mAP from 50.4 to 56.4 and NDS from 61.0 to 64.8.

Ablate training stage on Waymo dataset We extend the above ablations on Waymo dataset and summarize all results in Tab. 4, which further validate the effectiveness of CSOT, as well as its orthogonality with the common used pseudo-labeling techniques. 100x fewer labeled point clouds lead to a degradation of detection performance, reducing the mean AP and APH from 71.5, 69.9 to 65.3, 63.7. Our Cross-Scan Object Transfer paradigm brings a boost of around 3 APH for both LEVEL_1 and LEVEL_2 objects. CSOT is proved to be complimentary to self-training, and the performance gains can be completely maintained. As shown in the Line 5 of Tab. 4, the performance of detector trained with 1% annotations can be finally optimized to match even surpass the fully-supervised baseline model. The abnormal enhanced detection performance with 100x reduced labeled data should be attributed to the learning-based copy-paste operation employed in this work. As shown in Lines 1a-1b of Tab. 4, replacing the commonly used copy-paste augmentation (GT-Aug) with our learning based approaches boosts the fully-supervised detector with an increase of 2.8 mAP. Notably, even regarding the fully-supervised detector augmented via our object

Table 5: Ablate our learning-based cross-scan object transfer which is instructed by HotspotNet. All results are evaluated on nuScenes dataset validation split.

(a) Compare our learning-based cross-scan object transfer (instructed by HotspotNet) with other similar methods.

	Annos	CSOT	Obj Transfer	Self	Obj Appearance r	Obj Appearance α	Local Context (x, y, z)	Global Context r_z	Global Context cls	rule /learned	mAP	NDS
1	5%				—	—	—	—	—	—	43.8	56.5
2	5%			✓	—	—	—	—	—	—	51.0	61.0
3	5%	✓	GT-Aug [47]		✓	✓					46.0	57.7
4	5%	✓	CA-Aug [15]		✓	✓	✓			rule	46.0	58.0
5	5%	✓	LiDAR-Aug [11]		✓	✓	✓			rule	46.2	57.9
6	5%	✓	Ours		✓	✓	✓	✓	✓	learned	**50.4**	**61.0**
6a	5%	✓	Ours		✓	✓	✓	✓	✓	rule	49.7	60.6
7	5%	✓	GT-Aug [47]		✓	✓					49.7	60.5
8	5%	✓	CA-Aug [15]		✓	✓	✓			rule	49.7	60.7
9	5%	✓	LiDAR-Aug [11]		✓	✓	✓			rule	49.7	60.2
10	5%	✓	Ours		✓	✓	✓	✓	✓	learned	**56.4**	**64.8**
10a	5%	✓	Ours		✓	✓	✓	✓	✓	rule	53.6	63.3

(b) Ablate contribution of each fitness score defined in $\mathcal{L}^n_{\text{score}}(o)$ (Eq. 5).

	Annos	CSOT	Self	position s_p	orientation s_o	category s_c	mAP	NDS
11	5%	✓	✓				49.7	60.5
12	5%	✓	✓	✓			52.1	62.6
13	5%	✓	✓		✓		49.4	60.8
13a	5%	✓	✓		✓	✓	52.2	62.3
13b	5%	✓	✓	✓		✓	55.1	64.3
14	5%	✓	✓			✓	51.6	62.0
15	5%	✓	✓	✓	✓	✓	**56.4**	**64.8**

transfer implementation as the baseline, reducing annotations by 100x causes only a minimal degradation in performance (remains 97.3%/97.4% mAP/mAPH of our fully-supervised detector).

Ablate our learning-based object transfer. We compare our learning-based object transfer which instructed by the proposed HotspotNet with other similar techniques in Tab. 5(a), and ablate negative sample strategy employed for calculating $\mathcal{L}^n_{\text{score}}(o)$ (Eq. 5) in Tab. 5(b). As shown in Lines 3-10 of Tab. 5, the proposed learning-based object transfer is able to maximize the benefits of CSOT. It also stands alone in its ability to seamlessly integrate with self-training technique without compromising their performance gains. CSOT built upon our learning-based object transfer enhances the detection performance with a considerable 6.6 mAP and 4.5 NDS. The improvement surpasses the one built upon previous similar approaches (with a margin of about 4 mAP and 3 NDS). The poor partially-supervised detector achieved from other object transfer approaches produce enormous false positive predictions (visualized in Fig. 1) and greatly influence the downstream self-training. Replacing our learning-based object transfer with previous methods reduces the mAP gains of self-training from 6.0 to around 3.5, and degrades the final mAP from 56.4 to 49.7. The necessity of learning-based object transfer are verified in Tab. 5, Lines 6-6a,10-10a. Learn-based approach enables better generalization and can help reduce noise, which introduces an increased final mAP of 2.8. We further ablate negative sample strategy employed for calculating $\mathcal{L}^n_{\text{score}}(o)$ and list results in Lines 11-15 of Tab. 5. Pocation and category provide the most important contextual informa-

Table 6: Ablate robustness of our methods across various network architectures. All results are evaluated on nuScenes dataset validation split.

	Annos	Detector	Encoder	Anchor-base	Anchor-free	Voxel Size	CSOT	Self	mAP	NDS
1	100%	CenterPoint	VFE		✓	(0.075,0.075,0.2)			59.5	66.0
2	5%	CenterPoint	VFE		✓	(0.075,0.075,0.2)			43.8	56.5
3	5%	CenterPoint	VFE		✓	(0.075,0.075,0.2)		✓	51.0	61.0
4	5%	CenterPoint	VFE		✓	(0.075,0.075,0.2)	✓		50.4	61.0
5	5%	CenterPoint	VFE		✓	(0.075,0.075,0.2)	✓	✓	**56.4**	**64.8**
6	100%	CenterPoint	PFE		✓	(0.1,0.1,8.0)			49.2	60.1
7	5%	CenterPoint	PFE		✓	(0.1,0.1,8.0)			30.2	48.3
8	5%	CenterPoint	PFE		✓	(0.1,0.1,8.0)		✓	37.8	53.1
9	5%	CenterPoint	PFE		✓	(0.1,0.1,8.0)	✓		40.0	54.9
10	5%	CenterPoint	PFE		✓	(0.1,0.1,8.0)	✓	✓	**45.0**	**57.9**
11	100%	SECOND	VFE	✓		(0.1,0.1,0.2)			55.4	64.7
12	5%	SECOND	VFE	✓		(0.1,0.1,0.2)			38.4	53.2
13	5%	SECOND	VFE	✓		(0.1,0.1,0.2)		✓	45.2	58.1
14	5%	SECOND	VFE	✓		(0.1,0.1,0.2)	✓		45.4	59.1
15	5%	SECOND	VFE	✓		(0.1,0.1,0.2)	✓	✓	**48.6**	**61.3**

tion, orientation provides additional cues to narrow down the possible identity. Specifically, combining orientation constraint with category consistency improve mAP by 0.6 (Tab. 5, Lines 13a and 14). Removing orientation constraint from all reduces mAP by 1.3 (Tab. 5, Lines 13b and 15).

Extend our method to different detectors We confirm the generality of our methods on various detectors, which utilizes diverse point cloud feature encoder or target assigner strategy. All results are listed in Tab. 6. For an anchor-free detector CenterPoint with a pillar-based feature encoder, the proposed CSOT bring a boost of 9.8 mAP and 6.6 NDS, which outperforms the self-training technique by 2.2 mAP and 1.8 NDS. Using our partially-supervised detector as the teacher model to generate pseudo-labels for unlabeled point clouds, and then conduct one-cycle of self-training. The performance of semi-supervised detector is ultimately reaches 45.0 mAP and 57.9 NDS. The mAP gap introduced by 20x fewer annotations is narrowed from 19.0 to 4.2. Experiments on anchor-based detector (SECOND) also shows similar results. The effectiveness of CSOT and the additive effect of combining it with self-training techniques are clearly demonstrated in Lines 11-15 of Tab. 6.

5 Conclusion

In this work, we propose a novel Cross-Scan Object Transfer paradigm for LiDAR semi-supervised object detection. It features a HotspotNet for transferring objects semantic coherence manner, which reduces false positive predictions and enables the proposed CSOT to reach its full potential. To the best of our knowledge, this work is the first successful attempt of object transfer in LiDAR SSOD. The proposed CSOT paradigm is also orthogonal to the self-training techniques. We achieve competitive detectors on par with fully-supervised baselines while using 20x and 100x reduced annotations on nuScenes and Waymo dataset.

Acknowledgements

This work was supported by the InnoHK funding.

References

1. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: CVPR. pp. 1090–1099 (2022)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT. pp. 92–100 (1998)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11621–11631 (2020)
4. Chen, B., Chen, W., Yang, S., Xuan, Y., Song, J., Xie, D., Pu, S., Song, M., Zhuang, Y.: Label matching semi-supervised object detection. In: CVPR. pp. 14381–14390 (2022)
5. Chen, B., Li, P., Chen, X., Wang, B., Zhang, L., Hua, X.S.: Dense learning based semi-supervised object detection. In: CVPR. pp. 4815–4824 (2022)
6. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Largekernel3d: Scaling up kernels in 3d sparse cnns. In: CVPR. pp. 13488–13498 (2023)
7. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: CVPR. pp. 21674–21683 (2023)
8. Choi, J., Song, Y., Kwak, N.: Part-aware data augmentation for 3d object detection in point cloud. In: IROS. pp. 3391–3397 (2021)
9. Dwibedi, D., Misra, I., Hebert, M.: Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: ICCV. pp. 1301–1310 (2017)
10. Fan, L., Pang, Z., Zhang, T., Wang, Y.X., Zhao, H., Wang, F., Wang, N., Zhang, Z.: Embracing single stride 3d object detector with sparse transformer. In: CVPR. pp. 8458–8468 (2022)
11. Fang, J., Zuo, X., Zhou, D., Jin, S., Wang, S., Zhang, L.: Lidar-aug: A general rendering-based augmentation framework for 3d object detection. In: CVPR. pp. 4710–4720 (2021)
12. Gao, J., Wang, J., Dai, S., Li, L.J., Nevatia, R.: Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In: ICCV. pp. 9508–9517 (2019)
13. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR. pp. 2918–2928 (2021)
14. Hu, J.S.K., Waslander, S.L.: Pattern-aware data augmentation for lidar 3d object detection. In: IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 2703–2710 (2021)
15. Hu, X., Duan, Z., Ma, J.: Context-aware data augmentation for lidar 3d object detection. In: ICIP. pp. 11–15 (2023)
16. Hu, Y., Ding, Z., Ge, R., Shao, W., Huang, L., Li, K., Liu, Q.: Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In: AAAI. pp. 969–979 (2022)
17. Huang, D., Chen, Y., Ding, Y., Liao, J., Liu, J., Wu, K., Nie, Q., Liu, Y., Wang, C., Li, Z.: Rethinking dimensionality reduction in grid-based 3d object detection. arXiv preprint arXiv:2209.09464 (2022)

18. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014)
19. Kingma, D.P., Rezende, D.J., Mohamed, S., Welling, M.: Semi-supervised learning with deep generative models. In: Adv. Neural Inform. Process. Syst. (2014)
20. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)
21. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J.: Pointpillars: Fast encoders for object detection from point clouds. In: CVPR. pp. 12697–12705 (2019)
22. Lee, D., Lee, J., Lee, J., Lee, H., Lee, M., Woo, S., Lee, S.: Regularization strategy for point cloud via rigidly mixed sample. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15900–15909 (2021)
23. Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. In: ICML. p. 896 (2013)
24. Lee, S., Lim, H., Myung, H.: Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3d point cloud. In: IROS. pp. 13276–13283 (2022)
25. Li, E., Wang, S., Li, C., Li, D., Wu, X., Hao, Q.: Sustech points: A portable 3d point cloud interactive annotation platform system. IEEE Intelligent Vehicles Symposium (IV) pp. 1108–1115 (2020)
26. Li, H., Wu, Z., Shrivastava, A., Davis, L.S.: Rethinking pseudo labels for semi-supervised object detection. In: AAAI. pp. 1314–1322 (2022)
27. Liang, H., Jiang, C., Feng, D., Chen, X., Xu, H., Liang, X., Zhang, W., Li, Z., Gool, L.V.: Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In: ICCV. pp. 3293–3302 (2021)
28. Liu, C., Zhang, W., Lin, X., Zhang, W., Tan, X., Han, J., Li, X., Ding, E., Wang, J.: Ambiguity-resistant semi-supervised learning for dense object detection. In: CVPR. pp. 15579–15588 (2023)
29. Liu, C., Gao, C., Liu, F., Li, P., Meng, D., Gao, X.: Hierarchical supervision and shuffle data augmentation for 3d semi-supervised object detection. In: CVPR. pp. 23819–23828 (2023)
30. Liu, C., Gao, C., Liu, F., Liu, J., Meng, D., Gao, X.: Ss3d: Sparsely-supervised 3d object detection from point cloud. In: CVPR. pp. 8428–8437 (2022)
31. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: ICLR (2021)
32. Liu, Y.C., Ma, C.Y., Kira, Z.: Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In: CVPR. pp. 9819–9828 (2022)
33. Meng, Q., Wang, W., Zhou, T., Shen, J., Jia, Y., Gool, L.V.: Towards a weakly supervised framework for 3d point cloud object detection and annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **44**(8), 4454–4468 (2021)
34. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 652–660 (2017)
35. Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H.A.A.K., Elhoseiny, M., Ghanem, B.: Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In: Adv. Neural Inform. Process. Syst. pp. 23192–23204 (2022)
36. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: CVPR. pp. 10529–10538 (2020)

37. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
38. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR. pp. 2446–2454 (2020)
39. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Adv. Neural Inform. Process. Syst. (2017)
40. Wang, H., Cong, Y., Litany, O., Gao, Y., Guibas, L.J.: 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In: CVPR. pp. 14615–14624 (2021)
41. Wang, T., Yang, T., Cao, J., Zhang, X.: Co-mining: Self-supervised learning for sparsely annotated object detection. In: AAAI. pp. 2800–2808 (2021)
42. Wang, X., Yang, X., Zhang, S., Li, Y., Feng, L., Fang, S., Lyu, C., Chen, K., Zhang, W.: Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In: CVPR. pp. 3240–3249 (2023)
43. Wang, Y., Yin, J., Li, W., Frossard, P., Yang, R., Shen, J.: Ssda3d: Semi-supervised domain adaptation for 3d object detection from point cloud. In: AAAI. pp. 2707–2715 (2023)
44. Wu, Z., Bodla, N., Singh, B., Najibi, M., Chellappa, R., Davis, L.S.: Soft sampling for robust object detection. arXiv preprint arXiv:1806.06986 (2018)
45. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: ICCV. pp. 3060–3069 (2021)
46. Xu, R., Wang, T., Zhang, W., Chen, R., Cao, J., Pang, J., Lin, D.: Mv-jar: Masked voxel jigsaw and reconstruction for lidar-based self-supervised pre-training. In: CVPR. pp. 13445–13454 (2023)
47. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
48. Yang, J., Shi, S., Wang, Z., Li, H., Qi, X.: St3d: Self-training for unsupervised domain adaptation on 3d object detection. In: CVPR. pp. 10368–10378 (2021)
49. Yang, Q., Wei, X., Wang, B., Hua, X.S., Zhang, L.: Interactive self-training with mean teachers for semi-supervised object detection. In: CVPR. pp. 5941–5950 (2021)
50. Yang, Y., Fan, L., Zhang, Z.: Mixsup: Mixed-grained supervision for labelefficient lidar-based 3d object detection. In: ICLR (2024)
51. Ye, D., Zhou, Z., Chen, W., Xie, Y., Wang, Y., Wang, P., Foroosh, H.: Lidarmultinet: Towards a unified multi-task network for lidar perception. In: AAAI. pp. 3231–3240 (2023)
52. Yin, J., Fang, J., Zhou, D., Zhang, L., Xu, C.Z., Shen, J., Wang, W.: Semi-supervised 3d object detection with proficient teachers. In: ECCV. pp. 727–743 (2022)
53. Yin, J., Zhou, D., Zhang, L., Fang, J., Xu, C.Z., Shen, J., Wang, W.: Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In: ECCV. pp. 17–33 (2022)
54. Yin, T., Zhou, X., Krähenbühl, P.: Center-based 3d object detection and tracking. In: CVPR. pp. 11784–11793 (2021)

55. 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: CVPR. pp. 1130–1140 (2022)
56. Yuan, J., Zhang, B., Yan, X., Chen, T., Shi, B., Li, Y., Qiao, Y.: Ad-pt: Autonomous driving pre-training with large-scale point cloud dataset. In: Adv. Neural Inform. Process. Syst. p. 36 (2023)
57. Yun, S., Han, D., Oh, S.J., Chun, S., Cho, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: CVPR. pp. 6023–6032 (2019)
58. Yurtsever, E., Erçelik, E., Liu, M., Yang, Z., Zhang, H., Topçam, P., Listl, M., Çaylı, Y.K., Knoll, A.: 3d object detection with a self-supervised lidar scene flow backbone. In: ECCV. pp. 247–265 (2022)
59. Zhang, H., Chen, F., Shen, Z., Hao, Q., Zhu, C., Savvides, M.: Solving missing-annotation object detection with background recalibration loss. In: ICASSP. pp. 1888–1892 (2020)
60. Zhang, L., Sun, Y., Wei, W.: Mind the gap: Polishing pseudo labels for accurate semi-supervised object detection. In: AAAI. pp. 3463–3471 (2023)
61. Zhang, L., Yang, A.J., Xiong, Y., Casas, S., Yang, B., Ren, M., Urtasun, R.: Towards unsupervised object detection from lidar point clouds. In: CVPR. pp. 9317–9328 (2023)
62. Zhao, N., Chua, T.S., Lee, G.H.: Sess: Self-ensembling semi-supervised 3d object detection. In: CVPR. pp. 11079–11087 (2020)
63. Zheng, W., Tang, W., Jiang, L., Fu, C.W.: Se-ssd: Self-ensembling single-stage object detector from point cloud. In: CVPR. pp. 14494–14503 (2021)
64. Zhou, H., Ge, Z., Liu, S., Mao, W., Li, Z., Yu, H., Sun, J.: Dense teacher: Dense pseudo-labels for semi-supervised object detection. In: ECCV. pp. 35–50 (2022)
65. Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-teaching: An end-to-end semi-supervised object detection framework. In: CVPR. pp. 4081–4090 (2021)
66. Zhu, B., Jiang, Z., Zhou, X., Li, Z., Yu, G.: Class-balanced grouping and sampling for point cloud 3d object detection. arXiv preprint arXiv:1908.09492 (2019)