# HINTED: Hard Instance Enhanced Detector with Mixed-Density Feature Fusion for Sparsely-Supervised 3D Object Detection

Qiming Xia[1] Wei Ye[1] Hai Wu[1] Shijia Zhao[1] Leyuan Xing[1]
Xun Huang[1] Jinhao Deng[1] Xin Li[2] Chenglu Wen[1]* Cheng Wang[1]
[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China
[2]Section of Visual Computing and Interactive Media, Texas A&M University, Texas, USA

## Abstract

*Current sparsely-supervised object detection methods largely depend on high threshold settings to derive high-quality pseudo labels from detector predictions. However, hard instances within point clouds frequently display incomplete structures, causing decreased confidence scores in their assigned pseudo-labels. Previous methods inevitably result in inadequate positive supervision for these instances. To address this problem, we propose a novel Hard INsTance Enhanced Detector (**HINTED**), for sparsely-supervised 3D object detection. Firstly, we design a self-boosting teacher (**SBT**) model to generate more potential pseudo-labels, enhancing the effectiveness of information transfer. Then, we introduce a mixed-density student (**MDS**) model to concentrate on hard instances during the training phase, thereby improving detection accuracy. Our extensive experiments on the KITTI dataset validate our method's superior performance. Compared with leading sparsely-supervised methods, HINTED significantly improves the detection performance on hard instances, notably outperforming fully-supervised methods in detecting challenging categories like cyclists. HINTED also significantly outperforms the state-of-the-art semi-supervised method on challenging categories. The code is available at* https://github.com/xmuqimingxia/HINTED.

## 1. Introduction

In autonomous driving scenarios, the significant success in 3D object detection tasks relies heavily on accurate annotation information [1, 6, 10, 15, 28, 29]. However, acquiring high-quality annotation data leads to a substantial cost, especially when dealing with large-scale outdoor scenarios. As an effective strategy for reducing annotation costs, sparsely-supervised learning has gained widespread attention. The sparsely-supervision setting involves selecting only a subset from the entire training dataset for partial annotation, where "partial annotation" means labeling only one instance per frame.

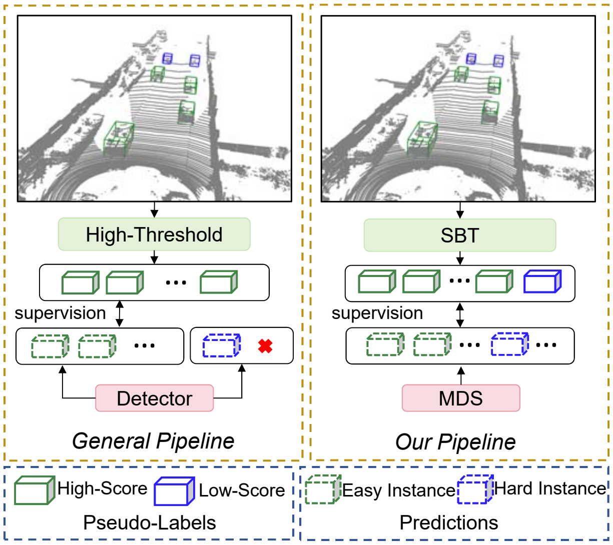

Figure 1. Comparison of sparsely-supervised 3D object detection methods. The current sparsely-supervised 3D object detection methods [12, 30] heavily rely on high-quality pseudo-labels generated by high thresholds for self-training, which inevitably results in a lack of positive supervision signals for hard instances. We propose HINTED, which introduces a teacher-student framework for sparse supervision settings. More specifically, the self-boosting teacher (SBT) generates more potential pseudo-labels, and the mixed-density student (MDS) incorporates different feature representations to perceive predictions of simulated hard instances. HINTED encourages consistent detection performance across different density representations within the same scene.

As shown in Figure 1, the current mainstream sparsely-supervised 3D object detection methods [12, 30] primarily adopt self-training strategy, using pre-training detector mining unlabeled instance. Especially, SS3D [12] generates high-quality pseudo-labels from predictions with high

*Corresponding author

score thresholds. This operation can make the later self-training more reliable. CoIn++ [30] employs the Test-Time Augmentation (TTA) strategy to get more accurate pseudo-labels. These methods achieve good performance for easy objects in sparsely-supervised settings. However, hard objects often have lower scores, leading to the missing of many positive predictions.

It is worth noting that the detection of hard objects has also gained attention in recent fully supervised 3D object detection methods. FocalFormer3D [3] continually masks out easy samples during the network training process, increasing the training focus on hard samples to enhance its ability to detect hard instances. SST [5] and FSD [6] maintain the original resolution to ensure that information about hard objects is not lost. These fully supervised methods enable the network to focus on training hard objects based on the ground-truth locations. However, sparse supervision lacks full annotation; these methods cannot be directly applied to sparse supervision strategies.

In the situation where many scenes are unlabeled, an intuitive solution is the teacher-student network commonly used in semi-supervised methods. DetMatch [19] produces more accurate pseudo-labels by aligning 2D and 3D detections from two modalities. HSS3D [13] uses hierarchical supervision to retain some of the low-score promising pseudo-labels in the teacher model's output. These teacher-student frameworks have succeeded in efficiently mining pseudo-labels from unlabeled scenes [14, 23, 24]. However, the abundance of unlabeled instances within sparsely labeled scenes, as indicated by [12], hampers efficient information transfer, impeding the direct application of the teacher-student framework in sparsely-supervised 3D object detection. Despite this, teacher-student networks still provide a new perspective for solving the problem of sparsely labeled 3D object detection.

Based on the above methods, we propose to enhance the information transfer efficiency of teacher-student networks for sparsely labeled scenes and pay more attention to hard instances during the training process. Specifically, our approach mainly comprises two components: (1) a novel self-boosting teacher (SBT); and (2) a mixed-density student (MDS). Unlike traditional teacher-student networks, our proposed SBT network can continuously update pseudo-labels for sparsely labeled frames, thereby avoiding the provision of incorrect supervision from sparsely labeled frames. Furthermore, we observed that hard instances are primarily distributed in distant, sparse point cloud scenes. To increase the attention to hard instances during training, MDS encourages attention to the consistent prediction of mixed-density features to enhance the detection performance of hard instances.

In summary, our contributions are three-fold:

- We design a novel self-boosting teacher (SBT), which addresses the issue of traditional teacher networks struggling to transfer information from sparsely labeled scenes effectively.
- We introduced a mixed-density student, which encourages consistent prediction of mixed-density features, thereby increasing the performance of hard instances.
- Our method has achieved a state-of-the-art performance under a sparsely-supervision setting, particularly in significantly improving detection results for hard instances, with a 32% improvement for pedestrians and a 25% improvement for cyclists.

## 2. Related Work

### 2.1. Sparsely-supervised 3D object detection

Recently, sparsely-supervised 3D object detection has attracted increasing attention due to the low cost of bounding box annotations. Unlike fully-supervised 3D object detection, sparsely-supervised 3D object detection only requires a sparse number of precise annotations. Therefore, sparse supervision requires the generation of pseudo labels to assist the detector in further training. For example, SS3D [12] utilized missing-annotated instance mining and background mining to obtain pseudo-labels. Similarly, CoIn [30] conducts contrastive instance feature mining to generate feature-level pseudo-labels. Furthermore, CoIn++ [30], which is obtained by combining CoIn with a self-training framework, achieves performance comparable to fully-supervised methods.

However, exciting methods are based on high-threshold filtering algorithms, leading to the lack of positive supervision information, which is not conducive to detecting hard instances. In this work, we aim to improve the performance of the detector on hard objects, making the sparsely-supervised 3D object detection algorithm more mature.

### 2.2. Weakly/semi-supervised 3D object detection

In addition, the weakly/semi-supervised setting also explores the research of object detection algorithms under low annotation cost. Unlike sparse supervision, the weakly-supervised methods [17, 26, 35] use weak annotation instead of 3D box annotation to reduce annotation costs. However, this strategy still requires a large number of instance-level annotations. In the semi-supervised setting, all instances in the labeled frames are annotated, and training often adopts student-teacher networks to mine pseudo-labels from unlabeled frames. Pioneering works like SESS [36] and 3DIoUMatch [24] first introduced semi-supervised learning to 3D object detection. SESS proposes a triple consistency regularization approach to refine 3D proposals, while 3DIoUMatch designs a filtering strategy for high-quality pseudo-labels and eliminates duplicates via an IoU estimation branch. Subsequent efforts aim to further improve pseudo-label quality without simply thresholding. [34] directly optimizes the pseudo-label fidelity over iterations of re-training. DetMatch [19] enforces multi-modal consistency across views to correct pseudo-labels. Most recently, HSSDA [13] takes a hierarchical supervision approach to mine missing objects and introduces a shuffle data augmentation. However, due to the different experimental settings, the weakly labeled frames in sparse supervision limit the effectiveness of information transmission between the teacher-student network. Therefore, it is challenging to directly introduce the teacher-student network to the sparsely-supervised 3D object detection.

### 2.3. Hard instance probing in 3D object detection

With the in-depth research of 3D object detection methods [2, 22, 25, 27, 31, 33, 37], more work is focusing on detecting hard instances. One direct approach is to adjust the network architecture to pay more attention to hard objects. SST [5] adopts sparse region attention to avoid down-sampling small objects, while FSD [6] further identifies sparse instances for long-range detection. Another line of work uses multi-stage networks to progressively refine bounding boxes and optimize hard instance mining. [3] proposes a module specialized for hard instance probing on top of a multi-stage framework. The methods above are all based on fully supervised strategies. With supervision from ground truth annotation information, algorithms can easily locate the positions of hard instances during training. Nevertheless, these strategies are not friendly to sparsely-supervised 3D object detection.

Overall, the study of hard instances is an active area in 3D object detection, and there has been significant progress in fully supervised 3D object detection. It is also important to explore the detection of hard instances in a sparsely-supervised setting.

## 3. Method

### 3.1. Preliminary

**Problem definition.** We start by introducing the definition of sparsely-supervised 3D object detection. Specially, the detector is trained with a sparsely labeled scene set $\mathcal{D}^l = \left\{(P_i^s, Y_i^s)\mid_{i=1}^{N^l}\right\}$ and an unlabeled scene set $\mathcal{D}^u = \left\{P_i^u \mid_{i=1}^{N^u}\right\}$, where $N^l$ and $N^u$ are the numbers of sparsely labeled scenes and unlabeled scenes. For each sparsely labeled scene $P_i^s$, the annotation $Y_i^s$ is just the annotation of a random instance in the scene, which includes eight-dimensional information, namely the three-dimensional spatial position, three-dimensional size, orientation, and category of the object.

**Teacher-student framework.** Inspired by mainstream semi-supervised methods [11, 13, 24], we also adopt a teacher-student framework for conducting sparsely-supervised 3D object detection. This framework involves two detectors with identical configurations, and we follow previous work by selecting PV-RCNN [20] and VoxelR-CNN [4]. The information transfer between the teacher network and the student network is accomplished through exponential moving average (EMA).

### 3.2. Overview

The pipeline of our HINTED framework is depicted in Figure 2, derived from the fundamental teacher-student mutual learning framework. During the pre-training stage, we employ the training strategy of CoIn [30] to obtain the initial weights for both the teacher and student models.

In the pseudo-label generation stage, to address the issue of sparse labels interfering with the training of the student network, we introduce a self-boosted teacher network. Following [13], we employ a dual-threshold strategy to retain more promising pseudo-labels. During the training stage of the student network, we propose a mixed-density student network to focus on feature learning for hard instances.

### 3.3. Self-boosting Teacher

Recently, the teacher-student framework has demonstrated significant potential in exploring unannotated scenes. However, unlike strongly annotated scenes where all instances are labeled, each sparsely annotated scene has only one instance labeled. The unlabeled instances within annotated scenes can lead to indistinguishable features, hindering effective information transfer [30]. To solve this problem, we develop a self-boosting teacher network (SBT). In contrasted to traditional teacher-student framework, SBT performs pseudo-label updates for labeled scenes, fully exploiting the instance information to eliminate indistinguishable features.

As shown in Figure 2, the proposed SBT takes sparsely labeled point cloud $P_i^s$ and unlabeled point cloud $P_i^u$ as inputs. As same as traditional 3D object detection pipeline [21, 32], SBT first extracts 3D features from $P_i^s$ and $P_i^u$ by backbone network, and then maps 3D features to 2D BEV features$\{B_i^s\}$ and $\{B_i^u\}$. After BEV features pass through the detector, the detector generates pseudo-labels $\bar{Y}_i^s$ and $\bar{Y}_i{}^u$, respectively. For $\bar{Y}_i{}^u$, it is directly designated as the supervision signal for the subsequent iteration of training in the student network. For $\bar{Y}_i{}^s$, we utilize sparse annotation $Y_i^s$ for filtering, retaining ground truth annotations while discarding inaccurate redundant pseudo-labels. Therefore, the pseudo-labels output by the teacher network are given by

$$\bar{Y} = \left\{\bar{Y}_i^u \cup \varphi(\bar{Y}_i^s, Y_i^s \mid \tau)\right\} \tag{1}$$

where $\varphi(\cdot \mid \tau)$ is inspired by the NMS algorithm [18], we calculate the IOU between the pseudo-label set $\bar{Y}_i^s$ and the

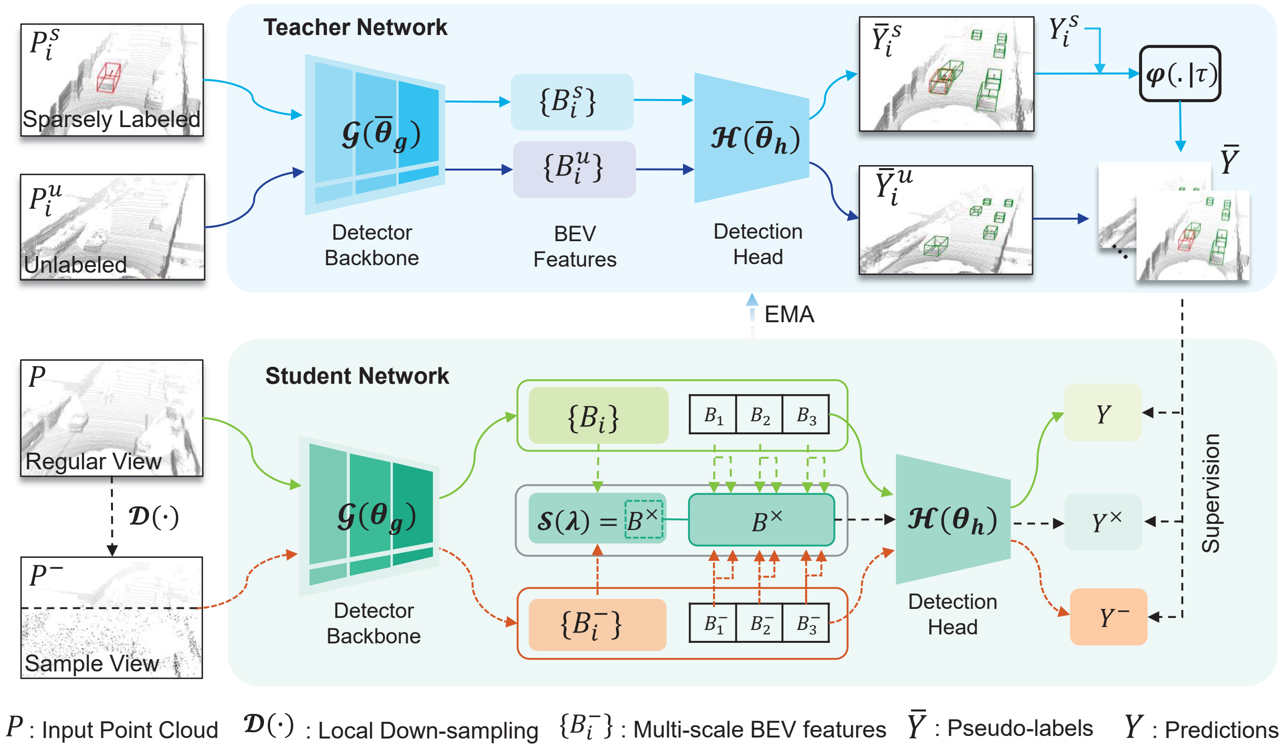

Figure 2. The overview of proposed HINTED, which consists of a self-boosting teacher (SBT) and a mixed-density student (MDS). In SBT, the teacher network provides the student network with rich pseudo-labels containing valuable information. In MDS, the model first extracts two multi-scale BEV features $\{B_i\}$ and $\{B_i^-\}$ for a regular view and local down-sample view with a detector backbone module $\mathcal{G}(\theta_g)$, respectively. And then, a mixed-density feature $B^\times$ is generated by a feature fusion module $\mathcal{S}(\lambda)$. Finally, the detection head module $\mathcal{H}(\theta_h)$ encourages consistent predictions of different density presentations. The weights $\bar{\theta}$ in the teacher are updated by EMA of the weights $\theta$ in the student. During testing, the model with the original architecture and regular input view is utilized.

sparse ground truth $Y_i^s$, and filter out pseudo-labels with an IOU greater than $\tau = 0.01$. Through this straightforward operation, we can efficiently handle sparsely labeled point clouds.

### 3.4. Mixed-Density Student

With the assistance of the SBT network, we obtain pseudo-labels for all scenes, which can support the supervised training of the student network. Additionally, enhancing the focus on hard instances during training has been proven beneficial for the detection of hard instances [3]. However, the pseudo-labels $\bar{Y}$ inevitably lack positive supervision for some hard instances. Therefore, resorting to fully-supervised methods by directly utilizing ground truth to enhance focus on hard instances can't yield optimal results. We observe that hard instances are primarily distributed in point cloud spaces farther away from the LiDAR acquisition vehicle. The main difference in point cloud spaces, in terms of distance, is the substantial variance in point cloud density. Building upon this observation, we down-sample the nearby point clouds to generate more point cloud spaces containing hard instances. Subsequently, we leverage the mixed-density feature to enhance the detection head's capability in handling the features of hard instances.

Given a point cloud scene, most student networks typically extract multi-scale BEV features $\{B_i \,|_{i=1,2,3}\} = \{B_1, B_2, B_3\}$ with backbone network. In $\{B_i\}$, the spatial sizes of the features decrease layer by layer. Then, the predictions are generated through the detection head from the last layer BEV feature $B_3$.

In our work, we first extract two multi-scale BEV features $\{B_i \,|_{i=1,2,3}\} = \{B_1, B_2, B_3\}$ and $\{B_i^- \,|_{i=1,2,3}\} = \{B_1^-, B_2^-, B_3^-\}$ from regular view point cloud $P$ and the nearby down-sampled view input point cloud $P^-$ , respectively. Then, we build a mixed-density feature $B^\times$ by adaptive fusing two multi-scale features.

The feature in $B_i$ shares the same spatial size as $B_i^-$, and they are also aligned in feature space. Inspired by [14], for feature-aligned feature maps, adaptive linear weighting is a simple yet effective strategy for multi-scale feature fusion. Based on this, we generate the mixed-density feature $B^\times$:

$$B^\times = \mathcal{S}(\{B_i\}_i, \{B_i^-\}_i, \lambda) \tag{2}$$

where $\mathcal{S}$ is the fusion function for multi-scale features, and $\lambda$ is the adaptive weight. Specifically:

$$B^{\times} = \sum_{i=1}^{3} \left[ \lambda_i Avg\left(B_i\right) + \bar{\lambda}_i Avg\left(B_i^{-}\right) \right] \tag{3}$$

where $Avg(\cdot)$ denotes the average pooling operation. The purpose of this operation is to ensure that the sizes of all feature maps are consistent with the bottom-level feature map, e.g. $B_3$, facilitating subsequent linear weighted summation. Moreover, inspired by SE block [9], we obtain suitable adaptive weights $\lambda_i$ and $\bar{\lambda}_i$. They are adaptive weights corresponding to the feature layers, and calculated by:

$$\lambda = \delta\left(f\left(Avg\left(B\right)\right)\right) \tag{4}$$

where $\delta(\cdot)$ denotes the sigmoid activation function, and $f(\cdot)$ is a fully connected layer. More detailed fusion specifics are provided in the supplementary materials.

Note that the spatial sizes of $B_3$, $B_3^{-}$, and $B^{\times}$ are same, representing feature expressions of different densities. To encourage consistent prediction results across features of different densities, we input the three types of features separately into the detection head module, generating three corresponding sets of predictions. Consequently, the training objective for MDS network:

$$\mathcal{L}_{MDS} = \gamma_1 \mathcal{L}_{det}\left(\mathcal{H}\left(B^{\times}\right), \bar{Y}\right) + \gamma_2 \mathcal{L}_{det}\left(\mathcal{H}\left(B_3^{-}\right), \bar{Y}\right) + \gamma_3 \mathcal{L}_{det}\left(\mathcal{H}\left(B_3\right), \bar{Y}\right) \tag{5}$$

where $\mathcal{L}_{det}$ follows the calculation method of the baseline approach, maintaining consistency with it. And, $\mathcal{H}(\cdot)$ denotes the detection head module. $\gamma_1$, $\gamma_2$, and $\gamma_3$ are hyper-parameters, we study them in Table 7 of ablation study section. Unlike traditional student networks, we adopt a unified approach to manage both unlabeled scenes and sparsely labeled scenes. Consequently, the loss is no longer computed separately for sparsely labeled and unlabeled scenes; instead, it's uniformly calculated, removing the necessity to adjust weights individually.

Although the introduction of multi-level mixed-density features does incur some additional computational overhead. It's worth noting that, built upon the original network structure, the memory consumption and added parameters can be deemed negligible. Specifically, we only employ simple operations such as global average pooling, linear layers, sigmoid activation functions to implement the feature fusion module.

Once the model has undergone sufficient training, only the modules belonging to the original detector can be retained. Specifically, the steps related to mixed-density feature extraction and fusion in the student network can be discarded. This approach ensures fairness in comparisons by adhering to the original detector architecture and input density during inference. Moreover, the final detection model has the same number of parameters as the original detector, and there is no increase in inference latency.

# 4. Experiments

## 4.1. Dataset and Evaluation Metrics

The KITTI dataset [7] is the most widely used dataset for sparsely-supervised 3D object detection. We adhere to the approach presented in recent studies [30], which partition the KITTI training dataset (comprising 7481 scenes) into a *train* split with 3712 scenes and a *val* split with 3769 scenes. Subsequently, we randomly choose 10% of scenes from the *train* split and keep only one object annotation within each selected scene. This procedure results in the creation of the *limited* split. In contrast to the original *train* split, the *limited* split necessitates only 2% cost of full annotations [30]. To ensure a fair comparison, we follow the principal official evaluation metric, which involves calculating the 3D Average Precision (AP) across 40 recall thresholds (R40).

## 4.2. Implementation Details

Our HINTED uses a pre-trained model from CoIn [30]. We train the entire network with a batch size of 8 and a learning rate of 0.003 for 80 epochs on 4 RTX 3090 GPUs. For local sampling, we split the scene into two parts along the X-axis, $[0 < x < 30]$ and $[30 < x < 70]$, defining $[0 < x < 30]$ as the nearby dense point cloud. We use random sampling for the nearby dense point cloud with a sampling rate of 20%. Following previous 3D object detection methods [13, 16], we also apply data augmentation during training. Specifically, for weak augmentation in the teacher network, we randomly flip along the X and Y-axes with a 50% probability, uniformly sample scale factors for scaling within the range $[0.91, 1.12]$, and rotate the entire scene along the Z-axis with random angles selected from $[-\pi/4, \pi/4]$. For strong augmentation in the student network, we perform shuffle augmentation on the point cloud. The hyper-parameters and strategy choices designed during the implementation process will be further discussed in the ablation study.

## 4.3. Comparison with State-of-the-art Methods

**Comparison with spasely-supervised methods.** We compare the proposed HINTED with state-of-the-art sparsely-supervised methods. For a fair comparison, all detectors adopted the VoxelRCNN [4] as the base architecture. Table 1 illustrates a performance comparison of different methods. Following the common evaluation metric in 3D object detection [16], the IOU thresholds of three categories are assessed at 0.7 (car), 0.5 (pedestrian), and 0.5 (cyclist), respectively.

| Setting | Cost | Method | Car | | | Ped | | | Cyc | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Easy | Mod. | Hard | Easy | Mod. | Hard | Easy | Mod. | Hard |
| *Fully-supervised* | *100%* | *VoxelRCNN[4]* | *92.3* | *84.9* | *82.6* | *69.6* | *63.0* | *58.6* | *88.7* | *72.5* | *68.2* |
| Sparsely-supervised | 3% | SS3D*[12] | 88.8 | 78.5 | 76.9 | - | - | - | - | - | - |
| | 2% | VoxelRCNN[4] | 70.5 | 54.9 | 44.8 | 42.6 | 38.5 | 32.1 | 73.3 | 47.8 | 43.2 |
| | | CoIn[30] | 89.1 | 70.2 | 55.6 | 50.8 | 45.2 | 39.6 | 80.2 | 52.3 | 48.6 |
| | | CoIn++[30] | 92.0 | 79.5 | 71.5 | 46.7 | 36.1 | 31.2 | 82.0 | 58.4 | 54.6 |
| | | Our HINTED | **94.3** | **82.5** | **78.7** | **66.5** | **59.9** | **53.7** | **94.6** | **76.3** | **73.0** |

Table 1. Comparsion with state-of-the-art sparsely-supervised methods on KITTI *val* split. All methods are based on VoxelRCNN, and we report the 3D AP results of full cost (100%) and limited cost (3%, 2%). The best sparsely-supervised methods are highlighted in **bold** and * indicates result with R11.

| Setting | Method | Modality | 1% | | | | 2% | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Car | Ped. | Cyc. | Avg. | Car | Ped. | Cyc. | Avg. |
| Semi-supervised | PV-RCNN [20] | LIDAR | 73.5 | 28.7 | 28.4 | 43.5 | 76.6 | 40.8 | 45.5 | 54.3 |
| | 3DIoUMatch [24] | LIDAR | 76.0 | 31.7 | 36.4 | 48.0 | 78.7 | 48.2 | 56.2 | 61.0 |
| | DetMatch [19] | LIDAR+RGB | 77.5 | **57.3** | 42.3 | 59.0 | 78.2 | 54.1 | 64.7 | 65.6 |
| | HSSDA [13] | LIDAR | **80.9** | 51.9 | 45.7 | 59.5 | **81.9** | 58.2 | 65.8 | 68.6 |
| | Our HINTED | LIDAR | 79.9 | 52.1 | **50.5** | **60.5** | 80.4 | **58.9** | **73.2** | **70.8** |

Table 2. Comparison with state-of-the-art semi-supervised methods on KITTI *val* split. All methods are built upon PV-RCNN and trained on randomly selected 2% or 1% fully annotated frames from *train* split.

| Data | Method | Car-3D Detection | | |
|---|---|---|---|---|
| | | Easy | Mod. | Hard |
| weakly* + 534 precisely# | WS3D [17] | 84.0 | 75.1 | 73.2 |
| weakly* + 2%Fully | WSS3D [35] | 84.5 | 75.8 | 71.1 |
| 2%Fully | Our HINTED | **90.6** | **80.4** | **77.6** |

Table 3. Comparison with state-of-the-art weakly-supervised methods on KITTI *val* split. *: point annotations. #: high-quality annotated instance.

As shown in Table 1, the proposed HINTED outperforms all other sparsely-supervised methods. For the more challenging categories, pedestrians and cyclists, our method shows particularly significant improvements. When compared to CoIn, which previously achieved the best performance on pedestrians, we achieve an average 32% improvement in detection performance for the pedestrian category. In comparison to CoIn++, which previously achieved the best performance on cyclists, we achieve an average 25% improvement in detection performance.

Furthermore, we observe that compared to the fully supervised VoxelRCNN [4], our proposed HINTED not only achieves comparable performance in the car and pedestrian categories but even surpasses it in the cyclist category. This indicates that our method not only generates high-quality pseudo-labels but also successfully reinforces the learning of hard instances during the network's training process.

**Comparison with semi-supervised methods.** Our proposed HINTED also adopts a teacher-student network similar to semi-supervised methods, so we further explore the performance of our method in semi-supervised setting. Similar to previous semi-supervised approaches [13, 24], we randomly select 1% and 2% of fully annotated frames from the *train* split for training and verifying on the *val* split. All the results are obtained within the framework of PV-RCNN [20].

Table 2 presents a comparison of our results with different semi-supervised methods. Compared to existing methods, our method shows significant improvement in the challenging cyclist category. However, for the car category, our method does not demonstrate an advantage. This may be because focusing on learning from difficult instances loses some relatively simple structural features. Overall, our method achieves the best Avg. performance among all methods, with moderate gains in average precision under both 1% and 2% labeling rates. This validates that our strategy also helps detect hard instances under semi-supervised settings.

**Comparison with weakly-supervised methods.** We also compare our method with state-of-the-art weakly-supervised methods. In WS3D[17] and WSS3D[35], they not only utilize a small number of annotated bounding boxes as supervision signals but also incorporate a large amount of center-click annotations. Given that our HINTED method doesn't incorporate a specifically designed center-click annotation, it directly utilizes outcomes derived from semi-supervision (with only 2% full supervision). As shown in Table 3, our method still demonstrates

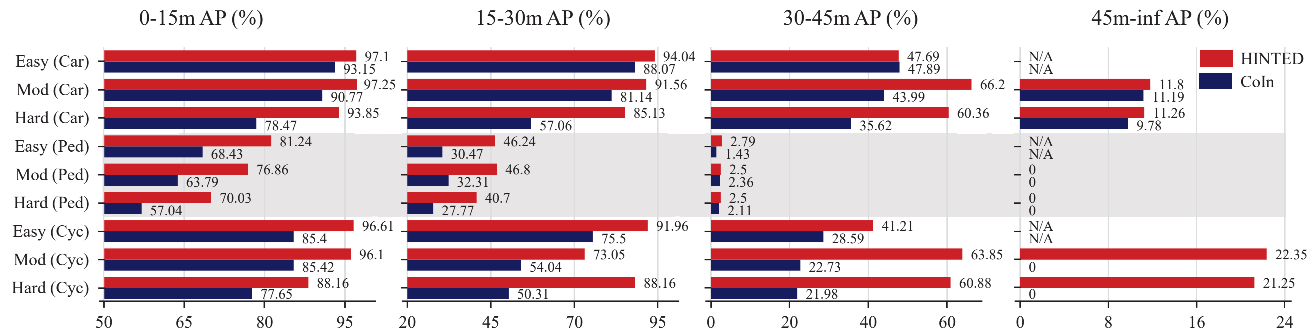

Figure 3. Comparison of three categories at different thresholds. The red bars represent the performance of our HINTED, while the blue bars represent the performance of CoIn [30]. N/A indicates that there are no samples at this difficulty level.

significant performance advantages without the help of additional point annotations.

### 4.4. Ablation Study

In this section, we conduct ablation experiments to validate the effectiveness of each module in HINTED. All the experiments in this section are based on VoxelRCNN [4].

| Density | | | SBT | MDS | Car-3D Detection | | |
|---|---|---|---|---|---|---|---|
| $B$ | $B^-$ | $B^\times$ | | | Easy | Mod. | Hard |
| ✓ | | | | | 89.5 | 79.2 | 72.3 |
| ✓ | | | ✓ | | 93.8 | 81.0 | 74.6 |
| ✓ | ✓ | | ✓ | | 93.7 | 81.8 | 75.8 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **94.3** | **82.5** | **78.7** |

Table 4. Effect of each component designed in HINTED on the KITTI *val* split.

**Effect of each component.** We systematically verify the efficacy of each component, and the results are illustrated in Table 4. In the $1^{st}$ row of Table 4, the initial network aims at training directly through a teacher-student framework. However, as only one instance is labeled in the labeled scenes, the presence of the remaining unlabeled instances interferes with the learning process of the student network, ultimately leading to detection performance weaker than that achieved by self-training (CoIn++ in Table 1).

In the $2^{nd}$ row, by replacing the traditional teacher-student network with our proposed SBT, labeled frames transfer information more effectively, achieving a superior baseline. The validation of the subsequent two modules is carried out directly on the basis of SBT. As shown in $3^{rd}$ and $4^{th}$ rows, MDS both can further improve the performance. The last row of Table 4 is a combination of SBT, MDS achieving optimal performance. This demonstrates that our proposed HINTED can generate high-quality pseudo-labels and effectively transfer information within the teacher-student network.

**Comparison at different distance thresholds.** Similar to our HINTED, CoIn also leverages unannotated scene information during training. HINTED, as a new teacher-student network, further enhances the performance of CoIn [30]. Fig. 3 specifically illustrates the comparison of the two methods across all categories at various difficulty levels. It can be seen from the figure that our method improves the performance across all categories. This demonstrates that our teacher network can provide accurate pseudo-labels, enhancing the detection capability. In long-range and hard object detection, our HINTED also demonstrates significant advantages, especially for the cyclist category, where we can produce correct predictions even beyond 45 meters.

| Sampling method | ratio | Car-3D Detection | | |
|---|---|---|---|---|
| | | Easy | Mod. | Hard |
| FPS | 40% | 93.89 | 81.67 | 76.21 |
| | 20% | 94.09 | 81.71 | 76.48 |
| | 10% | 93.97 | 81.59 | 76.23 |
| RS | 40% | 94.26 | 81.66 | 76.19 |
| | 20% | **94.33** | **82.56** | **78.75** |
| | 10% | 94.25 | 82.34 | 78.51 |

Table 5. Comparison of different sample strategies on KITTI *val* split.

**Comparison with other point cloud sample strategies.** We compare the impact of different sampling strategies on performance. We chose the most commonly used farthest point sampling (FPS) and random sampling (RS) methods, each with three different sampling rates: 10%, 20%, and 40%. The experimental results are presented in Table 5. Compared to FPS, RS shows overall better performance. This demonstrates that FPS, being a relatively stable sampling method, is not conducive to emulating distant point clouds from nearby ones, and it doesn't provide the optimal learning capability for hard instances. In the experiments involving RS of nearby point clouds, it is observed that the detector's performance reached its peak when the sampling rate was set to 20%.

| Fusion Strategy | Car-3D Detectin | | |
|---|---|---|---|
| | Easy | Mod | Hard |
| Baseline | 92.0 | 79.7 | 74.2 |
| ROI-based Fusion | 87.9 | 77.4 | 72.7 |
| Average-based Fusion | 91.2 | 80.3 | 73.5 |
| Attension-basd Fusion | **94.2** | **82.3** | **78.5** |

Table 6. The influence of different feature fusion approaches.

**Comparison of feature fusion approaches.** In this section, we investigate the impact of different feature fusion strategies on detection performance. The baseline in Table 6 represents the detection results obtained without using mixed features. Building upon this, we design three different feature fusion methods. In the $2^{nd}$ row of the table, "ROI-based Fusion" denotes generating regions of interest (ROI) based on the network's predictions and blending the dense-sparse features within those ROI. However, the misalignment between the two types of feature regions of interest can result in this fusion module being unhelpful in generating high-quality pseudo-labels and may even hinder the network's learning process. The $3^{rd}$ row reports the result of "Average-based Fusion", which directly combines the features of the two types of point clouds using a mean operation. This rough operation doesn't yield significant benefits. Compared to the other three methods, the attention-based fusion method adopted in our method achieves the best results, demonstrating that the adaptive weight-based feature fusion can enhance the detector's performance.

| | | | | | | |
|---|---|---|---|---|---|---|
| $\gamma_1$ ($\gamma_2$=$\gamma_3$=1) | 0 | 0.1 | 0.2 | 0.5 | 0.7 | 1.0 |
| Avg. | 63.1 | **72.9** | 70.9 | 68.4 | 68.3 | 67.5 |
| $\gamma_2$($\gamma_1$=0.1, $\gamma_3$=1) | 0 | 0.1 | 0.2 | 0.5 | 0.7 | 1.0 |
| Avg. | 62.9 | 66.7 | 66.5 | 68.9 | 69.3 | **72.9** |
| $\gamma_3$($\gamma_1$=0.1, $\gamma_2$=1) | 0 | 0.1 | 0.2 | 0.5 | 0.7 | 1.0 |
| Avg. 62.5 | 64.3 | 67.6 | 67.9 | 68.7 | 71.1 | **72.9** |

Table 7. The influence of $\gamma_1$, $\gamma_2$ and $\gamma_3$.

**Choice of hyper-parameters.** In this section, we refer to [8] and employ the method of controlling variables to obtain the optimal hyper-parameters. We initialize the weights $\gamma_2$ and $\gamma_3$ to 1, gradually changing the $\gamma_1$. Following this approach, we identify the optimal $\gamma_2$ and $\gamma_3$. And the experimental results are shown in the Table 7. The results in the table are the mean values of the three categories in terms of 'mod' difficulty. The best performance is achieved when the $\gamma_1$, $\gamma_2$ and $\gamma_3$ are set to $0.1, 1, 1$, respectively.

### 4.5. Quality Analysis

In this section, we compare the predictions of the pre-trained model CoIn with the predictions of our final student model generated by HINTED. As shown in Figure 4 (a) and (b), it is noticeable that although the pre-trained model initially couldn't provide positive supervision signals for distant hard instances (region of the yellow dashed line), our MDT approach successfully enhances the detection capabilities for these hard instances, resulting in more accurate detection outcomes.

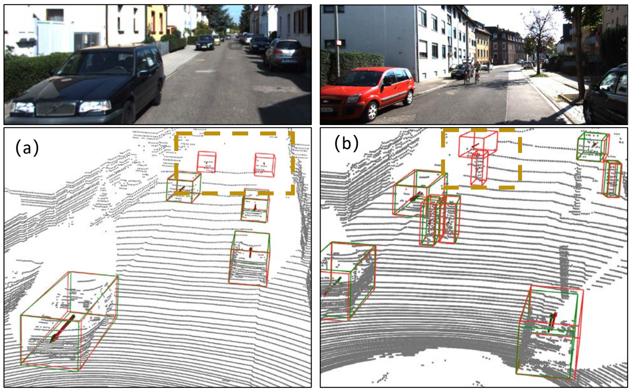

Figure 4. Qualitative analysis comparison. The red bounding boxes represent the results obtained from our HINTED, while the green bounding boxes are from the pre-trained CoIn model. The point cloud inside the yellow dashed line represents the hard instances.

## 5. Conclusion

This paper proposed a novel hard instance mining method, HINTED, for sparsely-supervised outdoor 3D object detection. In addition, benefiting from the self-boosting teacher network and mixed-density student, the effectiveness of the 3D detector has been significantly improved. Sufficient experiments on the KITTI dataset have demonstrated the effectiveness of our method under sparsely-supervised settings, and we have also demonstrated that under semi-supervised settings, our method can also help improve the detection performance for hard instances. In future work, we will attempt to continue verifying the effectiveness of our method on larger datasets.

**Limitation:** Due to limited memory, we cannot parallelize the generation of feature maps in the student network using shared weights, leading to longer training times. Nevertheless, during the inference process, the local down-sampled view and mixture density view will be discarded, thus not increasing the model parameters or reducing the inference speed.

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

## References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[2] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21674–21683, 2023. 3

[3] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: Focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8394–8405, 2023. 2, 3, 4

[4] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wen gang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 3, 5, 6, 7

[5] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8458–8468, 2022. 2, 3

[6] Lue Fan, Feng Wang, Naiyan Wang, and ZHAO-XIANG ZHANG. Fully sparse 3d object detection. *Advances in Neural Information Processing Systems*, 35:351–363, 2022. 1, 2, 3

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 5

[8] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11873–11882, 2020. 8

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[10] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21643–21652, 2023. 1

[11] Zhaoqi Leng, Shuyang Cheng, Benjamin Caine, Weiyue Wang, Xiao Zhang, Jonathon Shlens, Mingxing Tan, and Dragomir Anguelov. Pseudoaugment: Learning to use unlabeled data for data augmentation in point clouds. In *European conference on computer vision (ECCV)*, pages 555–572. Springer, 2022. 3

[12] Chuandong Liu, Chenqiang Gao, Fangcen Liu, Jiang Liu, Deyu Meng, and Xinbo Gao. Ss3d: Sparsely-supervised 3d object detection from point cloud. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8428–8437, 2022. 1, 2, 6

[13] Chuandong Liu, Chenqiang Gao, Fangcen Liu, Pengcheng Li, Deyu Meng, and Xinbo Gao. Hierarchical supervision and shuffle data augmentation for 3d semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23819–23828, 2023. 2, 3, 5, 6

[14] Liang Liu, Boshen Zhang, Jiangning Zhang, Wuhao Zhang, Zhenye Gan, Guanzhong Tian, Wenbing Zhu, Yabiao Wang, and Chengjie Wang. Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7370–7379, 2023. 2, 4

[15] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1

[16] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, pages 1–55, 2023. 5

[17] Qinghao Meng, Wenguan Wang, Tianfei Zhou, Jianbing Shen, Luc Van Gool, and Dengxin Dai. Weakly supervised 3d object detection from lidar point cloud. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 515–531, 2020. 2, 6

[18] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, pages 850–855. IEEE, 2006. 3

[19] Jinhyung Park, Chenfeng Xu, Yiyang Zhou, Masayoshi Tomizuka, and Wei Zhan. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. In *European Conference on Computer Vision*, pages 370–389. Springer, 2022. 2, 3, 6

[20] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10526 – 10535, 2020. 3, 6

[21] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43:2647–2664, 2021. 3

[22] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131(2):531–551, 2023. 3

[23] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017. 2

[24] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J. Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14610–14619, 2021. 2, 3, 6

[25] Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. DSVT: Dynamic Sparse Voxel Transformer With Rotated Sets. In *CVPR*, 2023. 3

[26] Yi Wei, Shang Su, Jiwen Lu, and Jie Zhou. Fgr: Frustum-aware geometric reasoning for weakly supervised 3d vehicle detection. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4348–4354. IEEE, 2021. 2

[27] Hai Wu, Chenglu Wen, Wei Li, Ruigang Yang, and Cheng Wang. Learning transformation-equivariant features for 3d object detection. 2022. 3

[28] Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, and Cheng Wang. Virtual sparse convolution for multimodal 3d object detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21653–21662, 2023. 1

[29] Qiming Xia, Yidong Chen, Guorong Cai, Guikun Chen, Daoshun Xie, Jinhe Su, and Zongyue Wang. 3-d hanet: A flexible 3-d heatmap auxiliary network for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13, 2023. 1

[30] Qiming Xia, Jinhao Deng, Chenglu Wen, Hai Wu, Shaoshuai Shi, Xin Li, and Cheng Wang. Coin: Contrastive instance feature mining for outdoor 3d object detection with very limited annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6254–6263, 2023. 1, 2, 3, 5, 6, 7

[31] Qiangeng Xu, Yiqi Zhong, and Ulrich Neumann. Behind the curtain: Learning occluded shapes for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 3

[32] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18, 2018. 3

[33] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7652–7660, 2018. 3

[34] Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3d object detection with proficient teachers. In *European Conference on Computer Vision*, pages 727–743. Springer, 2022. 3

[35] Dingyuan Zhang, Dingkang Liang, Zhikang Zou, Jingyu Li, Xiaoqing Ye, Zhe Liu, Xiao Tan, and Xiang Bai. A simple vision transformer for weakly semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8373–8383, 2023. 2, 6

[36] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[37] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *European Conference on Computer Vision*, pages 496–513. Springer, 2022. 3