2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Learning Adaptive Pseudo-Label Selection for Semi-Supervised 3D Object Detection

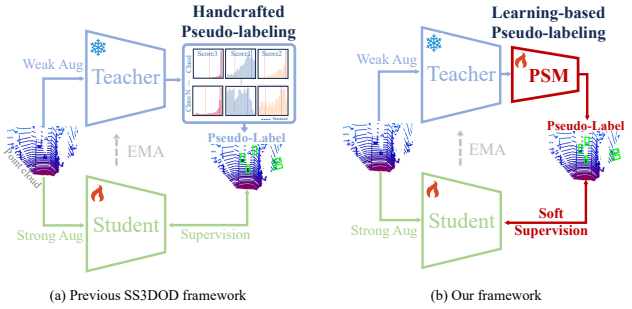
Taehun Kong[1,2] and Tae-Kyun Kim[1]

***Abstract*— Semi-supervised 3D object detection (SS3DOD) aims to reduce costly 3D annotations utilizing unlabeled data. Recent studies adopt pseudo-label-based teacher-student frameworks and demonstrate impressive performance. The main challenge of these frameworks is in selecting high-quality pseudo-labels from the teacher's predictions. Most previous methods, however, select pseudo-labels by comparing confidence scores over thresholds manually set. The latest works tackle the challenge either by dynamic thresholding or refining the quality of pseudo-labels. Such methods still overlook contextual information e.g., object distances, classes, and learning states, and inadequately assess the pseudo-label quality using partial information available from the networks. In this work, we propose a novel SS3DOD framework featuring a learnable pseudo-labeling module designed to automatically and adaptively select high-quality pseudo-labels. Our approach introduces two networks at the teacher output level. These networks reliably assess the quality of pseudo-labels by the score fusion and determine context-adaptive thresholds, which are supervised by the alignment of pseudo-labels over GT bounding boxes. Additionally, we introduce a soft supervision strategy that can learn robustly under pseudo-label noise. This helps the student network prioritize cleaner labels over noisy ones in semi-supervised learning. Extensive experiments on the KITTI and Waymo datasets demonstrate the effectiveness of our method. The proposed method selects high-precision pseudo-labels while maintaining a wider coverage of contexts and a higher recall rate, significantly improving relevant SS3DOD methods.**

I. INTRODUCTION

3D object detection in LiDAR point clouds is critical for scene understanding in autonomous driving, robotics, and AR/VR. However, existing methods require extensive 3D annotations, which are costly due to the need for accurate bounding boxes and cross-modal verification. As a result, labeled data is limited compared to the large amount of unlabeled data. Semi-supervised learning (SSL) addresses this by leveraging unlabeled data to improve performance.

Pseudo-label-based frameworks have become the standard for Semi-supervised 3D object detection (SS3DOD), effectively leveraging unlabeled data and achieving strong performance gains. In these methods, pseudo-label selection is critical and typically relies on thresholding scores (e.g., classification confidence) from the teacher network. Most works adopt fixed thresholds [1]–[6], while some, like ATF-3D [7] and HSSDA [8], explore adaptive thresholding strategies. However, recent pseudo-label selection methods are suboptimal for the following two reasons. First, predicting pseudo-label quality is challenging in SS3DOD. Detectors produce multiple scores, each showing different correlation with ground-truth quality (see Figure 4). This makes it difficult to set unified thresholds and assess reliability. Consequently, prior methods used only partial cues rather than leveraging all available indicators (see Figure 2d). Second, the optimal threshold depends on instance context, such as class, distance, and learning state. As shown in Figure 2, the score distributions vary, making fixed thresholds suboptimal. Effective thresholding should adapt to context and be updated throughout training to reflect dynamic learning states. Finding an optimal threshold that accounts for such contexts is complex, approaches such as [1], [2], [7], [8] only partially consider these contextual factors (see Figure 2d)

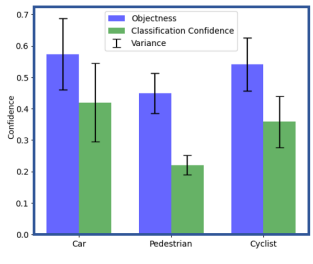


(a) Previous SS3DOD framework (b) Our framework

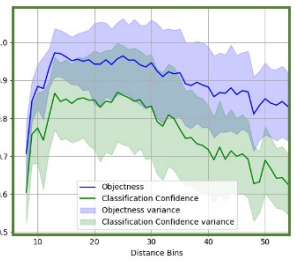
Fig. 1. Comparison of the proposed and previous pseudo-labeling methods in semi-supervised learning. (a) shows the handcrafted threshold-based approach. (b) illustrates our framework with a learnable Pseudo-label Selection Module and Soft Supervision for robust training.

To address the aforementioned limitations of prior work, we propose a novel learning-based pseudo-label selection method, named Pseudo-label Selection Module (PSM). The PSM leverages limited Ground Truth (GT) information to assess the quality of pseudo-labels and to determine a context-appropriate threshold. The PSM consists of the Pseudo-Label Quality Estimator (PQE) and Context-aware Threshold Estimator (CTE). The PQE encodes the teacher's various output scores to a single fusion score indicating reliable pseudo-label quality, while the CTE encodes the context to generate adaptive threshold values. During SSL, PSM is trained to dynamically select pseudo-labels considering the context, achieving a wide coverage of pseudo-labels while maintaining a high quality. Additionally, we introduce the Soft Supervision strategy to train robustly against pseudo-label noise. Our method combines the soft GT sampling augmentation and loss re-weighting to counteract pseudo-label noise given the coverage of labels. To the best of our knowledge, this is the first method to model pseudo-labeling using a neural network. Our contributions are summarized as follows:

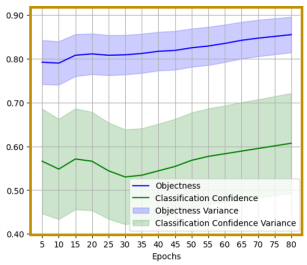
[1]School of Computing, KAIST, [2]AI Lab, LG Electronics



(a) Score dist. by classes



(b) Score dist. by distance



(c) Score dist. by learning states

	Thrshold	Context			Pseudo-label metric			
		Cls.	Dist.	State	Cls. Score	Obj. score	IoU consis.	Aux. score
3DIoUMatch	Manual	✓			✓	✓		✓
DDS3D				✓		✓		
DetMatch					✓		✓	
Reliable Student		✓				✓		
PTPM					✓			
A-Teacher		✓			✓			
ATF-3D	Auto	✓	✓		✓			
HSSDA		✓		✓	✓	✓	✓	
OURS		✓	✓	✓	✓	✓	✓	✓

(d) Comparison of pseudo-labeling methods

Fig. 2. (a), (b), and (c) show that classification confidence and objectness have different distributions depending on the context. (b) and (c) illustrate the distributions specifically for foreground objects. (d) compares previous pseudo-labeling methods in three aspects: the approach for determining score thresholds, the contexts considered, and the metrics used for evaluating pseudo-label quality. Auxiliary scores (Aux. score) refer to additional IoU predictions or objectness from different views.

- We introduce a novel learning-based pseudo-label selection method, Pseudo-label Selection Module (PSM), which better predicts the pseudo-label quality and considers the contexts for pseudo-label selection.
- We propose a noise-robust supervision strategy that prevents the student from being biased to pseudo-label noise.
- Extensive experiments on the KITTI and Waymo datasets show that the proposed framework significantly improves performance. Notably, in the limited labeled data scenario of KITTI 1%, we achieved around 20 mAP absolute improvement over the labeled-only 3D baseline.

II. RELATED WORK

A. 3D Object Detection

3D object detection involves predicting oriented 3D bounding boxes and types of objects from either monocular RGB images [9]–[12] or LiDAR point clouds. 3D object detection from LiDAR point clouds is generally categorized into point-based [13]–[15] and grid-based [16]–[19] based on data representation. Point-based detectors directly process raw point clouds to extract spatial features. For example, PointRCNN [13] uses a PointNet++ [20]-based backbone to extract point-level features and generates proposals in a two-stage process. Grid-based detectors convert point clouds into grids (voxels or pillars) and use CNNs. VoxelNet [21] applies PointNet and 3D CNNs, SECOND [16] uses sparse 3D convolutions for efficiency, and PointPillars [18], [19] encodes pillars with simplified PointNet. Point-voxel-based detectors combine both approaches. Methods like PV-RCNN [22], [23] integrate voxel and point operations for proposal generation or refinement. We conducted experiments using Voxel-RCNN [17] (grid-based) and PV-RCNN [22] (point-voxel-based).

B. Semi-Supervised Learning (SSL)

Semi-supervised learning is broadly divided into consistency regularization [24]–[26] and pseudo-labeling [27]–[29]. Consistency regularization enforces stable model outputs under different augmentations. Mean Teacher [25] uses a teacher updated by the student's EMA, applying a consistency loss between their outputs. Pseudo-labeling generates labels for unlabeled data to provide supervision. FixMatch [30] selects pseudo-labels above a fixed confidence threshold, while FlexMatch [31] adapts thresholds per class based on learning progress. While SemiReward [32] introduced two additional networks to measure the pseudo-label reliability via adversarial learning, it still resorts to predefined thresholds for pseudo-label selection. By contrast, our method tackles both reliable pseudo-label quality prediction and automatic pseudo-label selection.

C. Semi-Supervised 3D Object Detection

SSL is an active topic in 3D object detection. SESS [33] applies consistency regularization with asymmetric augmentations, while 3DIoUMatch [1] selects pseudo-labels using classification confidence, objectness, and IoU. The strong performance of 3DIoUMatch popularized pseudo-label-based teacher-student frameworks in SS3DOD [1]–[8], [34], [35]. Proficient Teacher [34] clusters bounding boxes from spatially and temporally augmented views for pseudo-labels. DDS3D [2] improves recall with dense pseudo-labels, decreasing thresholds during training. ATF-3D [7] searches thresholds by distance and class using fixed positive-negative ratios. DetMatch [3] uses a 2D-3D consistency cost with manual thresholding. Reliable Student [5] employs class-aware target assignment and loss softening with manual thresholds. A-Teacher [4] refines pseudo-labels using adjacent frames, and PTPM [6] divides scenes into patches to improve teacher performance. These methods, however, still rely on handcrafted pseudo-label selection. Recently, CSOT [36] proposed a specialized model that boosts performance by synthesizing scenes through copy-pasting labeled objects onto unlabeled ones. Note this technique is orthogonal to pseudo-labeling approaches. HSSDA [8] is the state-of-the-art pseudo-labeling method that clusters three different scores of teacher predictions exceeding an IoU threshold with labels, generating two thresholds per score for hierarchical supervision.

This study aims to improve pseudo-label selection by learning thresholding in SS3DOD, using HSSDA as the baseline.

III. METHODOLOGY

A. Teacher-student SSL Pipeline

Semi-supervised 3D object detection involves training on a limited labeled dataset D^l and an abundant unlabeled dataset D^u. The input point cloud consists of n points, and each point is characterized by 3D coordinates and additional information (e.g., color, intensity, and timestamps). The ground-truth annotation specifies objects in the labeled dataset using 7-dimensional parameters for 3D bounding boxes and a 1-dimensional object category.

While the 3D detector is trained with D^l in a supervised manner, the training process extends to semi-supervised learning (SSL) to incorporate D^u. Mainstream SSL frameworks for 3D object detection involve four stages: (1) **Burn-in Stage**: Train the 3D detector on D^l to initialize both the teacher and student models. (2) **Pseudo-labeling Stage**: Generate pseudo-labels by filtering the teacher's candidates on unlabeled data with weak augmentation α. (3) **Semi-supervision Stage**: Compute the supervised and unsupervised losses from the student's predictions on data with strong augmentation $\mathcal{A}$. (4) **Teacher Update Stage**: Update the teacher model using the EMA of the student model,

$$\theta_t = \rho \cdot \theta_t + (1 - \rho) \cdot \theta_s \tag{1}$$

θ_t represents the teacher parameters, which are updated based on the student parameters θ_s using ρ, the EMA momentum factor.

B. Method Overview

As illustrated in Figure 3, we introduce the Pseudo-Label Selection Module (PSM) within the teacher-student framework. The PSM reliably evaluates pseudo-label quality from various teacher outputs using the Pseudo-label Quality Estimator (PQE) and determines the threshold based on context-dependent score variations through the Context-aware Threshold Estimator (CTE). In the burn-in stage, the PSM is pre-trained using outputs from the teacher pipeline, which generates predictions for both original and weakly augmented scenes. From the teacher's outputs, we obtain instance-level predictions: objectness score s^{obj} and class distribution p^{cls} for the original scene, objectness score $\tilde{s}^{obj}$ for the weakly augmented scene, and predicted bounding boxes b and $\tilde{b}$ for original and weakly augmented scenes respectively. During the semi-supervision stage, the PSM is updated using the labeled data D^l to track the changes in the teacher's state and perform adaptive pseudo-labeling.

Additionally, we introduce a supervision strategy called Soft Supervision that is robust to pseudo-label noise. This prevents bias to pseudo-label noise by re-weighting the loss with a joint confidence score. We generalize the hierarchical supervision [8] that exploits dual-thresholds to a single threshold, and design Soft GT Sampling augmentation by modifying the GT sampling augmentation [16].

C. Pseudo-label Selection Module (PSM)

The proposed method aims to balance the quality and coverage of pseudo-labels using context-dependent multiple scores. If the ground truth (GT) labels are available, selecting pseudo-labels based on the Intersection over Union (IoU) with the GT bounding boxes is the most intuitive approach. GT-IoU provides a context-invariant measure of pseudo-label quality by indicating how close the pseudo-labels are to the actual ones. The PSM learns to predict or approximate GT-IoU-based pseudo-label selection using a labeled dataset D^l by two key components: the Pseudo-Label Quality Estimator (PQE) and the Context-Aware Threshold Estimator (CTE).

Pseudo-label Quality Estimator (PQE). Thresholding each score individually as in previous works [1], [8] often misses high-quality pseudo-labels and reduces the diversity of labels. Instead of using individual scores, aggregating them into a single score accounts for the importance and combination of different score values. Filtering based on this fusion score helps increase the pseudo-label coverage while preserving their qualities.

PQE takes as input the feature vector $x_i^s = [s_i^{obj}, \tilde{s}_i^{obj}, p_i^{cls}, v_i]$, which consists of four components: the objectness score s_i^{obj}, the auxiliary objectness score $\tilde{s}_i^{obj}$, the classification probability p_i^{cls}, and the IoU consistency $v_i = IoU(b_i, \tilde{b}_i)$ for the i-th pseudo-label candidate. PQE, $\mathcal{Q}$, encodes this score feature to $\mathcal{Q}(x_i^s) \in [0, 1]$, predicting the true quality of pseudo-labels, which is measured by GT-IoU i.e. $IoU(b_i, b_i^{GT})$, where b_i is the predicted pseudo-box and b_i^{GT} is the corresponding ground truth box.

The input data is passed onto a MLP module, yielding the predicted pseudo-label quality via a sigmoid function. The PQE is trained to minimize the mean squared error (MSE) loss between the GT-IoU and the predicted pseudo-label quality $\mathcal{Q}(x_i^s)$ over the pseudo-label candidates generated from the teacher before Non-Maximum Suppression (NMS). The training objective for PQE is as follows:

$$\mathcal{L}_{PQE} = \frac{1}{N_l} \sum_i ||\mathcal{Q}(x_i^s) - IoU(b_i, b_i^{GT})||_2^2 \tag{2}$$

Where N_l is the number of pseudo-label candidates.

Inspired by the late candidate fusion networks in 3D Object Detection [37], [38], PQE is designed to combine various scores of the teacher network and geometric associations at the output level. By aggregating diverse information, PQE provides a more reliable measure of pseudo-label quality. Figure 4 shows that the PQE score exhibits a higher positive correlation with GT-IoU than other scores. By contrast, the classification confidence often undervalues high-quality pseudo-labels, increasing the risk of losing valuable samples. The reliability of the PQE score reduces the loss of valued samples during filtering and allows for wider coverage while maintaining quality.

Context-Aware Threshold Estimator (CTE). While PQE provides a measure of pseudo-label quality, setting an appropriate threshold value for pseudo-label selection remains crucial. Since score distributions are context-dependent, the predicted pseudo-label quality $\mathcal{Q}(x_i^s)$ is also context-dependent. We consider the object class c_i and distance d_i as the context that influences the threshold, given the

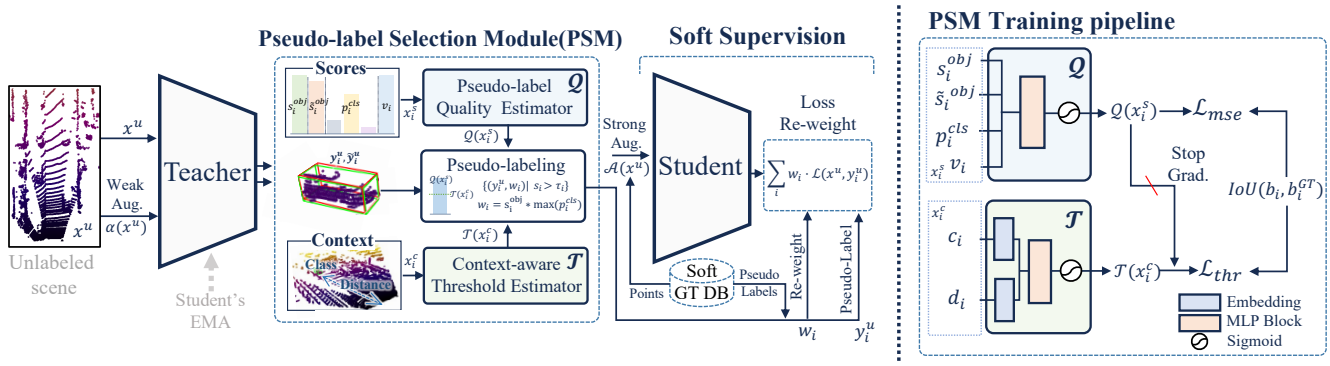


Fig. 3. Overview of the proposed framework, with two main components: the Pseudo-label Selection Module (PSM), which selects pseudo-labels using the detector's outputs and contexts, and Soft Supervision, which enhances robustness to pseudo-label noise. The PSM includes two neural networks, $\mathcal{Q}$ and $\mathcal{T}$, that predict pseudo-label quality and context-aware thresholds.

teacher's current learning state θ_t. The goal of Context-Aware Threshold Estimator (CTE) is to learn a context-aware threshold determination function $\mathcal{T}(c_i, d_i \mid \theta_t)$ that mimics GT-IoU-based thresholding:

$$\mathcal{Q}(x_i^s) > \mathcal{T}(c_i, d_i|\theta_t) \quad \triangleq \quad IoU(b_i, b_i^{GT}) > \tau_{iou} \tag{3}$$

The threshold determination function $\mathcal{T}(\cdot)$ is implemented using a neural network. CTE takes the context inputs, represented as $x_i^c = [c_i, d_i]$, and includes an embedding layer for each context. It is followed by a MLP module and a sigmoid function, predicting the context-aware threshold. To train the CTE, we introduce a threshold error to evaluate the accuracy of the determined threshold, which serves as a loss function. The threshold error of the score s and threshold τ is quantified as:

$$\mathcal{L}_{thr}(\tau, s, b, b^{GT}) = \begin{cases} ||\tau - s||_2^2 & \left[\begin{array}{l}(IoU(b, b^{GT}) \geq \tau_{iou} \wedge s \leq \tau) \vee \\ (IoU(b, b^{GT}) < \tau_{iou} \wedge s > \tau)\end{array}\right] \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

We assign the L2 loss between the predicted pseudo-label quality s and the threshold τ for the false cases in Eq. (3). Specifically, when a pseudo-label is correct ($IoU(b, b^{GT}) \geq \tau_{iou}$) but τ is higher than s (False Negative), a loss is applied. Conversely, when a pseudo-label is incorrect ($IoU(b, b^{GT}) < \tau_{iou}$) but τ is lower than s (False Positive), it also contributes to the loss. A lower threshold error indicates a more optimal threshold in a global view, according to Eq. (3). Through learning with the threshold errors of instances in a batch, the model progressively learns the threshold determination function $\mathcal{T}(\cdot)$. The training objective of the CTE is as follows:

$$\mathcal{L}_{CTE} = \frac{1}{N_l} \sum_i \mathcal{L}_{thr}(\mathcal{T}(x_i^c), \overline{\mathcal{Q}}(x_i^s), b_i, b_i^{GT}) \tag{5}$$

$\overline{\mathcal{Q}}(x_i^s)$ is the predicted pseudo-label quality that is stop-gradient, preventing gradient flow from the CTE to the PQE to avoid interference. Using $\mathcal{L}_{CTE}$, the model learns the appropriate context-specific threshold $\mathcal{T}(x_i^c)$ for $\mathcal{Q}(x_i^s)$.

D. Soft Supervision

Despite the proposed pseudo-labeling, unavoidable noises in pseudo-labels occur. To mitigate the impact of this noise, we propose a Soft Supervision that helps robust learning against pseudo-label noise. In the previous work HSSDA [8], the hierarchical supervision categorized pseudo-labels into a high-level and ambiguous-level. The loss for ambiguous-level pseudo-labels was softened, while high-level pseudo-labels were utilized for GT Sampling augmentation [16]. This approach amplifies the influence of clean pseudo-labels and reduces the impact of noisy ones. Note that the pseudo-labels generated by PSM achieve a high precision and recall (see Figure 6), making single-level pseudo-labels sufficient. We integrated and modified operations for both high-level (GT sampling augmentation) and ambiguous-level pseudo-labels (softened loss). Consequently, our supervision process is simplified yet reducing the effects of pseudo-label noise. The Soft Supervision includes Soft GT Sampling and Loss re-weighting.

Soft GT Sampling augmentation. GT Sampling augmentation counteracts foreground sparsity by sampling GT from the labeled dataset and placing it into different frames. However, directly applying GT Sampling augmentation to inaccurate pseudo-labels results in excessive supervision signals containing noises, increasing the risk of overfitting to the noise. Therefore, we sample both the GT and their joint confidence score $w = s^{obj} * \max(p^{cls})$, as in HSSDA [8]. The joint confidence score is then used for the loss re-weighting to reduce the influence of noise. During SSL, we accumulate pseudo-labels in the Soft GT Database.

Loss re-weighting. We soften the impact of noisy pseudo-labels using their associated joint confidence score w. These pseudo-labels are sourced from scene-generated pseudo-labels and samples from the Soft GT Database. This ensures the student focuses more on high-confidence pseudo-labels than noisy ones.

Soft Supervision simplifies and generalizes the hierarchical supervision [8], effectively addressing pseudo-label noise while maintaining the benefits of high-precision and high-recall pseudo-labels generated by PSM.

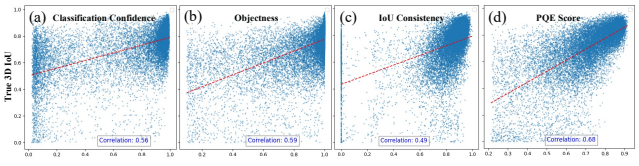


Fig. 4. The correlation between GT-IoU and each score for KITTI 1% split. (a) Classification confidence, (b) Objectness, (c) IoU consistency [8], and (d) the output score of PQE.

E. Training Strategy

During the burn-in stage, both the teacher and student networks are initialized after training the detector. The PSM is then trained using the teacher network's output. Since CTE takes PQE as input, PQE's learning states influence CTE. Both networks are trained together with a single optimizer, where PQE converges first and then CTE. The gradient of PSM does not backpropagate to the teacher network to avoid interfering with the detector training. In the semi-supervision stage, the student network is trained on both unlabeled and labeled datasets, while the PSM is trained exclusively on the labeled dataset. The total loss function incorporates three components: the labeled loss $\mathcal{L}_l$ and unlabeled loss $\mathcal{L}_u$ for the student network, along with $\mathcal{L}_{PSM}$ for the PSM network.

$$\mathcal{L} = \underbrace{\frac{1}{N_l}\sum_i (\mathcal{L}_i^{cls} + \mathcal{L}_i^{reg})}_{\mathcal{L}_l} + \underbrace{\frac{1}{N_u}\sum_i w_i(\mathcal{L}_i^{cls*} + \mathcal{L}_i^{reg*})}_{\mathcal{L}_u} + \mathcal{L}_{PSM} \tag{6}$$

Where $\mathcal{L}^{cls}$ and $\mathcal{L}^{reg}$ represent the classification and regression losses using ground truth labels, respectively. For the unlabeled data, $\mathcal{L}^{cls*}$ and $\mathcal{L}^{reg*}$ are measured using the pseudo-labels by PSM and w_i serves as the joint confidence score for Soft Supervision. The PSM's training loss $\mathcal{L}_{PSM}$ includes both PQE and CTE losses, as detailed in Section III-C, where PSM is trained using teacher's pseudo-label candidates and ground truth labels for labeled scenes.

$$\mathcal{L}_{PSM} = \mathcal{L}_{PQE} + \mathcal{L}_{CTE} \tag{7}$$

By minimizing the PSM loss, PSM evolves along with the learning state θ_t through the joint training. Pseudo-labels selected by CTE are used to train the student network, and the teacher network is updated via EMA of the student. The teacher's predictions are then used to train CTE and PQE. This process repeats during training, establishing interactions between the student and PSM.

IV. EXPERIMENTS

A. Dataset and Evaluation Metric

KITTI. To evaluate the proposed framework, we utilize the KITTI 3D object detection benchmark [39], comprising 3,712 training scenes and 3,769 validation scenes. Following prior works [3], [8], we randomly select 1% and 2% of the labeled data for the semi-supervised setting. For each ratio, we sample three distinct labeled sets and average the results to measure generalized performance independent of specific labeled sets. We evaluate three classes: Car, Pedestrian, and Cyclist—using Average Precision (AP) at 40 recall positions, applying Intersection over Union (IoU) thresholds of 0.7, 0.5, and 0.5 for each class, respectively.

Waymo. We additionally evaluate our framework on the Waymo Open Dataset [40], which is the largest autonomous driving dataset containing 1,000 sequences. It includes 798 training sequences with approximately 150K point cloud samples and 202 validation sequences with about 40K samples. We sample 1% of the training sequences (approximately 1.4K frames) for the semi-supervised setting. Due to the large scale of the Waymo dataset, we evaluate the results using a single split instead of averaging over three splits. We present AP and APH results at LEVEL 1 and LEVEL 2 difficulties for Vehicle, Pedestrian, and Cyclist classes.

B. Implementation Details

Network Architecture. In PSM, CTE and PQE are lightweight 4-layer MLPs with channel dimensions $D_{MLP} = [16, 32, 32, 1]$. For PQE, the score inputs are concatenated and then fed into the MLP. For CTE, the classes are linearly embedded to $D_{class} = 8$, and distances are embedded to $D_{distance} = 8$ dimensions using Fourier embedding [41] after normalization. The embedded contexts are concatenated and then passed into the MLP.

Training Details. Following prior works [8], we adopt PV-RCNN [22] and Voxel-RCNN [17] as our baseline 3D detectors. During the semi-supervision stage, PSM and detector are jointly optimized using a single ADAM optimizer. The PSM is trained for 60 epochs with batch size 16, and we set the GT-IoU threshold $\tau_{iou} = 0.8$. See Table VI for the effect of using different values. We use the same augmentation and experimental settings as in [8].

C. Main Results

KITTI. We compare our method with state-of-the-art methods on the KITTI val set. Table I presents the results based on the PV-RCNN. Compared to previous methods using PV-RCNN, our approach achieves the highest mAP, with absolute improvements of 20.2 and 15.0 at 1% and 2%, respectively. Notably, in the Cyclist class, we observe significant performance gains of 17.2 and 3.2 compared to the previous state-of-the-art at 1% and 2% settings. Table III shows the performance based on Voxel-RCNN. We observe similar behaviors of performance improvement. Under the setting of minimum labeled datasets 1%, our method demonstrated substantial performance gains. Note also these results are obtained by the simpler pipeline that eliminates the dual-threshold based pseudo-label hierarchization and complex supervision strategies required by HSSDA [8]. Moreover, learning PSM during SSL removes the need for iterative threshold recalculation.

Waymo. Table II presents the experimental results on the large-scale Waymo dataset. For comparison, we focus on works specifically addressing pseudo-label selection.Our framework shows significant improvements for the Vehicle and Cyclist classes. These gains mainly owe to the rich supervision signal provided by the pseudo-labels from PSM

TABLE I

PERFORMANCE COMPARISON ON KITTI VAL SET BY PV-RCNN. ALL COMPARED METHODS USE PV-RCNN AS THE BASE DETECTOR. THE TOP ROW SHOWS THE RESULT OF THE DETECTOR TRAINED ON THE LABELED-ONLY DATASET.

Model	Threshold	1%				2%			
		Car	Ped.	Cyc.	mAP	Car	Ped.	Cyc.	mAP
PV-RCNN (in [1])	Detector	73.5	28.7	28.4	43.5	76.6	40.8	45.5	54.3
3DIoUMatch [1]	Manual	76.0	31.7	36.4	48.0	78.7	48.2	56.2	61.0
DDS3D [2]	Manual	76.0	34.8	38.5	49.8	78.9	49.4	53.9	60.7
Reliable Student [5]	Manual	77.0	41.9	35.4	51.4	79.5	53.0	59.0	63.8
DetMatch [3]	Manual	77.5	**57.3**	42.3	59.0	78.2	54.1	64.7	65.6
HSSDA [8]	Auto	80.9	51.9	45.7	59.5	81.9	**58.2**	65.8	68.6
Ours	Auto	**81.3**	47.0	**62.9**	**63.7**	**82.0**	56.8	**69.0**	**69.3**

TABLE II

EXPERIMENTAL RESULTS ON THE WAYMO VALIDATION SET. * DENOTES OUR REPRODUCED RESULTS.

1%	Veh. (L1)		Veh. (L2)		Ped. (L1)		Ped. (L2)		Cyc. (L1)		Cyc. (L2)	
	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH
PV-RCNN (in [3])	47.3	45.6	43.6	42.0	28.9	15.6	26.2	14.1	-	-	-	-
DetMatch [3] (using PV-RCNN)	52.2	51.1	48.1	47.2	39.5	18.9	35.8	17.1	-	-	-	-
PV-RCNN (in [8])	48.5	46.2	45.5	43.3	30.1	15.7	27.3	15.9	4.5	3.0	4.3	2.9
HSSDA [8] (using PV-RCNN)	56.4	53.8	49.7	47.3	40.1	20.9	33.5	17.5	29.1	20.9	27.9	20.0
Voxel-RCNN (in [8])	49.0	48.0	42.4	41.5	41.2	32.8	34.7	27.7	5.8	5.6	5.6	5.4
HSSDA [8] (using Voxel-RCNN)	54.9	54.1	48.3	47.5	**43.9**	**37.8**	**36.6**	**31.6**	17.5	16.7	16.7	16.0
*Voxel-RCNN [17]	53.9	51.2	46.8	45.0	41.4	21.7	34.9	18.3	5.8	3.4	5.6	3.3
Ours (using Voxel-RCNN)	**58.8**	**57.3**	**51.1**	**49.8**	30.6	16.5	25.5	13.8	**34.8**	**22.3**	**33.5**	**21.4**

TABLE III

EXPERIMENTAL RESULTS ON KITTI VAL SET USING VOXEL-RCNN AS THE BASE DETECTOR FOR ALL METHODS.

Model	1%				2%			
	Car	Ped.	Cyc.	mAP	Car	Ped.	Cyc.	mAP
Voxel-RCNN (in [8])	74.0	19.0	37.0	43.3	76.5	40.2	39.9	52.2
HSSDA [8]	**81.7**	43.9	48.3	58.0	**82.0**	58.3	65.7	68.7
Ours	81.4	**52.2**	**61.5**	**65.0**	81.8	**58.6**	**70.6**	**70.3**

TABLE IV

ABLATION STUDIES OF EACH COMPONENT ON KITTI VAL.

Exp	PSM		Soft Supervision	Car	Ped.	Cyc.	mAP
	PQE	CTE					
1	-	-	-	79.3	49.3	43.8	57.5
2	✓	-	-	80.6	43.8	59.4	61.2
3	✓	-	✓	81.1	47.0	60.3	62.8
4	✓	✓	-	81.3	50.9	**64.4**	**65.5**
5	✓	✓	✓	**81.4**	**52.2**	61.5	65.0

TABLE V

ABLATION STUDIES ON CONTEXTS CONSIDERED IN CTE

Exp.	Context		Car	Ped.	Cyc.	mAP
	Distance	Class				
1	-	✓	80.8	59.3	67.7	69.3
2	✓	-	**82.0**	50.4	68.5	67.0
3	✓	✓	81.8	**58.6**	**70.6**	**70.3**

TABLE VI

EFFECT OF GT-IOU THRESHOLD τ_{iou}

GT-IoU Threshold	Car	Ped.	Cyc.	mAP
0.70	80.3	41.1	67.1	62.8
0.75	80.8	47.0	**67.4**	**65.1**
0.80	**81.4**	**52.2**	61.5	65.0
0.85	80.8	51.7	47.6	60.0

and the Soft GT Sampling augmentation. However, the performance of the Pedestrian class remains sensitive to noise from excessive GT samples despite the effort to mitigate this with Soft Supervision. The Pedestrian class exhibits particularly noisy patterns compared to other classes. The issues with the performance of the Pedestrian are known to the community. According to the official implementation of HSSDA [8], a different pseudo-label selection policy specific to Pedestrian is applied, whereas our method applies the same setting to all classes.

D. Ablation Studies and Analyses

In this section, we present experimental analyses to demonstrate the effect of our proposed framework. All results in this section are obtained using the KITTI 1% split.

Contribution of Each Component. Table IV shows the results of each proposed component. Exp 1 presents the baseline from HSSDA [8]. Exp 2 demonstrates the effect of Pseudo-label Quality Estimation (PQE). The threshold generated by the Dual Threshold Generation [8] for PQE is used for pseudo-label selection. We apply supervision without distinguishing between high-level and ambiguous-level pseudo-labels, and PQE alone outperforms the baseline. Exp 4 shows the effect of Context-aware Threshold Estimator (CTE). It demonstrates a significant performance gain with 4.2 mAP improvement over handcrafted thresholds. Exp 3 and Exp 5 illustrate the impact of Soft Supervision. They show meaningful performance improvements for the Pedestrian class, which has more noise in pseudo-labels compared to other classes.

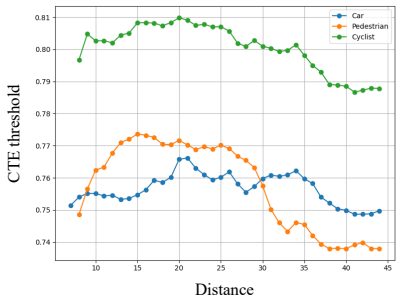


Fig. 5. CTE thresholds by classes and distances.

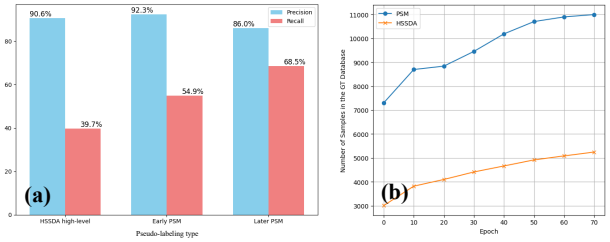


Fig. 6. Quantitative comparisons of pseudo-label qualities on KITTI. PSM is pre-trained with the 1% split.

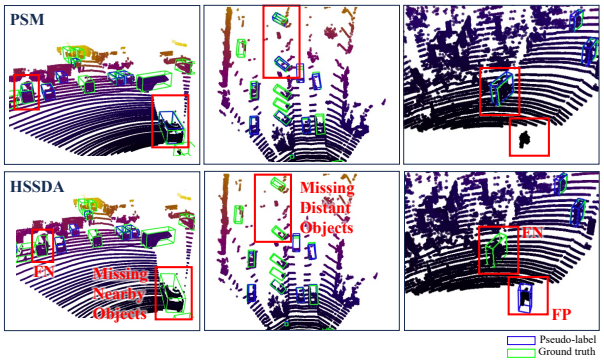


Fig. 7. Qualitative comparisons of pseudo-labels on KITTI. PSM is pre-trained with the 1% split.

Impact of Contexts in CTE. We conducted an ablation study on different contexts in CTE. Using both class and distance contexts achieved the best accuracy, as in Table V. We observed that using the distance improved the recall rate of pseudo-labels, which contributed to the performance gain for Car and Cyclist. Figure 5 shows how the CTE thresholds vary across different classes and distances, similar to the score variations for distances as in Figure 2b while exhibiting class-specific characteristics. Furthermore, unlike ATF-3D [7] and HSSDA [8] where contexts are discretized, CTE operates in continuous context space, enabling a more flexible threshold determination mechanism without overfitting.

GT-IoU Threshold. We define the pseudo-labels among teacher predictions where the GT-IoU exceeds the threshold $\tau_{iou} = 0.8$ for PSM training. While this is considered a hyperparameter, it is more general and interpretable than multiple score-level thresholds, which involve multiple dynamic scores ($s^{obj}, \tilde{s}^{obj}, s^{cls}, v$) and are sensitive and computationally complex. In contrast, the GT-IoU threshold is easier to set thanks to its geometrical and statistical intuitions. Choosing this value as accurate labels is straightforward from visual overlaps and prior studies [8]. Existing automatic thresholding methods i.e. HSSDA, also involve a few hyperparameters to tune (e.g., the negative/positive sample ratios [7] and matching IoU threshold [8]). Given the value of τ_{iou}, the CTE automatically and adaptively determines the score-level threshold while accounting for contextual factors. Table VI shows the ablation study on the GT-IoU threshold τ_{iou}. There is little performance change for the Car class with different values of τ_{iou}. In contrast, the Pedestrian class exhibits a decline in performance as τ_{iou} decreases, while the Cyclist class performs poorly at higher τ_{iou} values. We set τ_{iou} as 0.8 which yields the most balanced performance among classes, and fixed for all datasets. Note the mAP is not sensitive to τ_{iou}.

Quality of Pseudo-Labels. The quality and coverage of pseudo-labels can be quantified using precision and recall. As shown in Figure 6a, PSM's pseudo-labels show 1.7 higher precision and 15.2 higher recall than HSSDA's high-level pseudo-labels. Furthermore, after 80 epochs of SSL, PSM's pseudo-labels show only a 6.3 decrease in precision while demonstrating a notable 13.6 increase in recall compared to HSSDA. Consequently, PSM selects more precise and diverse pseudo-labels through context-aware pseudo-labeling, as illustrated in Figure 7. Figure 6b displays the number of pseudo-labels stored in the GT Database and Soft GT Database over training epochs. Our framework stores a substantially larger amount of pseudo-labels, providing the student with rich supervision signals.

V. CONCLUSION

In this paper, we propose a novel learning-based pseudo-labeling method that predicts pseudo-label quality and determines context-aware thresholds within the SSL framework. This approach enables the generation of a large volume of high-quality pseudo-labels. We also introduce Soft Supervision to prevent the student model from overfitting to pseudo-label noise. The extensive experiments and ablation studies support the effectiveness of our framework. In the future, we plan to extend the proposed pseudo-labeling to more complex SSL scenarios that involve richer pseudo-label contexts, such as multi-modal settings.

VI. ACKNOWLEDGEMENT

This work was supported in part by NST grant (CRC 21015, MSIT), IITP grants (RS-2023-00228996, RS-2024-00459749, RS-2025-25441313, RS-2025-25443318, MSIT), KOCCA grant (RS-2024-00442308, MCST), and LG Electronics Co., Ltd.

REFERENCES

[1] H. Wang, Y. Cong, O. Litany, Y. Gao, and L. J. Guibas, "3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 615–14 624. 1, 2, 3, 6

[2] J. Li, Z. Liu, J. Hou, and D. Liang, "Dds3d: Dense pseudo-labels with dynamic threshold for semi-supervised 3d object detection," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9245–9252. 1, 2, 6

[3] J. Park, C. Xu, Y. Zhou, M. Tomizuka, and W. Zhan, "Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 370–389. 1, 2, 5, 6

[4] H. Wang, Z. Zhang, J. Gao, and W. Hu, "A-teacher: Asymmetric network for 3d semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 978–14 987. 1, 2

[5] F. Nozarian, S. Agarwal, F. Rezaeianaran, D. Shahzad, A. Poibrenski, C. Müller, and P. Slusallek, "Reliable student: Addressing noise in semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4981–4990. 1, 2, 6

[6] X. Wu, L. Peng, L. Xie, Y. Hou, B. Lin, X. Huang, H. Liu, D. Cai, and W. Ouyang, "Semi-supervised 3d object detection with patchteacher and pillarmix," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6153–6161. 1, 2

[7] Z. Zhang, Y. Ji, W. Cui, Y. Wang, H. Li, X. Zhao, D. Li, S. Tang, M. Yang, W. Tan *et al.*, "Atf-3d: Semi-supervised 3d object detection with adaptive thresholds filtering based on confidence and distance," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 573–10 580, 2022. 1, 2, 7

[8] C. Liu, C. Gao, F. Liu, P. Li, D. Meng, and X. 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*, 2023, pp. 23 819–23 828. 1, 2, 3, 4, 5, 6, 7

[9] X. Shi, Z. Chen, and T.-K. Kim, "Distance-normalized unified representation for monocular 3d object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 91–107. 2

[10] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, and T.-K. Kim, "Geometry-based distance decomposition for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 172–15 181. 2

[11] Y. Hong, H. Dai, and Y. Ding, "Cross-modality knowledge distillation network for monocular 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 87–104. 2

[12] X. Shi, Z. Chen, and T.-K. Kim, "Multivariate probabilistic monocular 3d object detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 4281–4290. 2

[13] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 770–779. 2

[14] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1951–1960. 2

[15] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3dssd: Point-based 3d single stage object detector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 040–11 048. 2

[16] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018. 2, 3, 4

[17] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1201–1209. 2, 5, 6

[18] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705. 2

[19] Y. Wang, A. Fathi, A. Kundu, D. A. Ross, C. Pantofaru, T. Funkhouser, and J. Solomon, "Pillar-based object detection for autonomous driving," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 2020, pp. 18–34. 2

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017. 2

[21] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499. 2

[22] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 529–10 538. 2, 5

[23] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection," *International Journal of Computer Vision*, vol. 131, no. 2, pp. 531–551, 2023. 2

[24] L. Samuli and A. Timo, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, vol. 4, no. 5, 2017, p. 6. 2

[25] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017. 2

[26] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in neural information processing systems*, vol. 33, pp. 6256–6268, 2020. 2

[27] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896. 2

[28] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5070–5079. 2

[29] H. Choi, Z. Chen, X. Shi, and T.-K. Kim, "Semi-supervised object detection with object-wise contrastive learning and regression uncertainty," *arXiv preprint arXiv:2212.02747*, 2022. 2

[30] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020. 2

[31] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021. 2

[32] S. Li, W. Jin, Z. Wang, F. Wu, Z. Liu, C. Tan, and S. Z. Li, "Semireward: A general reward model for semi-supervised learning," *arXiv preprint arXiv:2310.03013*, 2023. 2

[33] N. Zhao, T.-S. Chua, and G. H. Lee, "Sess: Self-ensembling semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 079–11 087. 2

[34] J. Yin, J. Fang, D. Zhou, L. Zhang, C.-Z. Xu, J. Shen, and W. Wang, "Semi-supervised 3d object detection with proficient teachers," in *European Conference on Computer Vision*. Springer, 2022, pp. 727–743. 2

[35] M. Kang, T. Kong, and T.-K. Kim, "Semi-supervised 3d object detection with channel augmentation using transformation equivariance," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 638–644. 2

[36] J. Zhan, T. Liu, R. Li, Z. Zhang, and Y. Chen, "Csot: Cross-scan object transfer for semi-supervised lidar object detection," in *European Conference on Computer Vision*, 2024. 2

[37] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 386–10 393. 3

[38] ——, "Fast-CLOCs: Fast camera-lidar object candidates fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 187–196. 3

[39] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361. 5

[40] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454. 5

[41] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in neural information processing systems*, vol. 33, pp. 7537–7547, 2020. 5