A. Details of Pseudo Label Generation

A.1. Transferring 2D Labels to 3D

Open-vocabulary 2D segmentation models possess rich semantic knowledge and accurate pixel-level predictions, enabling them to localize arbitrary text queries. Since closed-set 2D perception models tailored to autonomous driving dataset classes are not available under an annotation-free condition, we employ open-vocabulary segmentation models as the teachers to generate 2D perception results. After evaluating multiple open-vocabulary models, we adopt FC-CLIP as the primary teacher to perform instance segmentation on the image set $\mathcal{I}$, producing masks $M_{\mathcal{I}}$ and corresponding labels $L_{\mathcal{I}}$. However, FC-CLIP exhibits poor performance on "barrier" and "traffic cone". To compensate for this, we introduce SAN as an auxiliary teacher to segment only these two classes, and directly overwrite the corresponding regions in $M_{\mathcal{I}}$ to obtain the refined mask $M'_{\mathcal{I}}$ and labels $L'_{\mathcal{I}}$.

To transfer 2D labels to 3D, we utilize known sensor calibration parameters to construct a transformation matrix Γ between LiDAR points and image pixels. As LiDAR and camera capture data asynchronously, each LiDAR point $(x_i, y_i, z_i) \in \mathbb{R}^3$ is projected to the pixel coordinates $pixel^i = (u_i, v_i) \in \mathbb{R}^2$ on the closest timestamp image set $\mathcal{I}$ via coordinate transformation. The point-to-pixel mapping is defined as follows:

$$
\begin{aligned}
&[u_i, v_i, 1]^{\mathbf{T}} = \Gamma_{c,in} \times \Gamma \times [x_i, y_i, z_i, 1]^{\mathbf{T}} / z_i \\
&\Gamma = \Gamma_{c \leftarrow ego_c} \times \Gamma_{ego_c \leftarrow global} \times \Gamma_{global \leftarrow ego_l} \times \Gamma_{ego_l \leftarrow l}, \\
&c = camera, \quad l = LiDAR,
\end{aligned}
\tag{4}
$$

where $\Gamma_{c,in}$ denotes the camera intrinsic matrix. $\Gamma_{c \leftarrow ego_c}$, $\Gamma_{ego_c \leftarrow global}$, $\Gamma_{global \leftarrow ego_l}$, and $\Gamma_{ego_l \leftarrow l}$ are extrinsic matrices used to transform coordinates from the LiDAR coordinate system to the camera coordinate system, all four matrices $\Gamma_{...}$ are provided.

Through Eq. 4, the single-frame point cloud $\mathcal{P}$ is projected onto the image $\mathcal{I}$. Based on the pixel grouping in $M'_{\mathcal{I}}$, we obtain point clusters $\mathcal{P}'_{\mathcal{I}}$ with corresponding labels $L'_{\mathcal{I}}$.

A.2. Pseudo Box Generation and Refinement

The pseudo box generation process in AnnofreeOD partially follows MODEST [63] and OYSTER [66], and consists of the following steps:

Ground Removal Since this task does not focus on background elements such as roads, ground points are first removed to better highlight objects and reduce noise. Assuming the ground is flat, RANSAC is used to fit a planar surface to each point with a 5 cm inlier threshold. Points located more than 15 cm above the fitted ground plane are retained, while others are treated as ground points and discarded.

Clustering The 2D-to-3D label transfer process in Appendix A.1 introduce noise due to two main factors: First, there may be segmentation errors from the 2D teacher model T, where the predicted mask $M'_{\mathcal{I}}$ does not perfectly align with the ground truth $M^{GT}_{\mathcal{I}}$. This is especially evident at object boundaries, where mask precision and granularity are limited. Second, inaccuracies in the transformation matrix Γ between the LiDAR and camera can also introduce errors, primarily due to small temporal misalignment between sensors. These two sources of error jointly lead to noisy labels when projected into 3D space. Empirically, distant objects tend to contain fewer points, sometimes fewer than noise clusters, making semantic-level clustering for many objects suboptimal. To address this, we adopt instance segmentation-based teachers and perform instance-level clustering for denoising. For each instance in $\mathcal{P}'_{\mathcal{I}}$, we apply clustering using HDBSCAN [32] to effectively filter out noisy points. In our implementation, the minimum cluster size and cluster selection epsilon are set to 5 points and 1 meter.

Box Fitting We apply a bounding box fitting algorithm [68] to assign a 3D box to each previously obtained object point cluster. The 3D box is denoted as $b = (x, y, z, l, w, h, \theta)$, where the parameters represent the centroid position (x, y, z), box dimensions (l, w, h), and orientation angle θ. In practice, since scene flow or motion information is unavailable, the velocity is set as $v = (0, 0, 0)$.

Box Refinement The refinement process primarily follows the commonsense-based method in [63]. First, bounding boxes that are located below the ground or floating are removed. The remaining boxes must satisfy the following conditions: the number of points inside the box is ≥ 5; the box volume is within $[0.5, 120] m^3$; the highest point in the box satisfies $h_{\text{max}} \geq h_{\text{ground}} + 0.5m$, and the lowest point satisfies $h_{\text{min}} < h_{\text{ground}} + 1m$. For each category c_j, a commonsense size prior is defined as $b^0_{c_j} = [l^0_{c_j}, w^0_{c_j}, h^0_{c_j}]$. Given a pseudo box $b = (x, y, z, l, w, h, \theta)$ of category c_j, if $l < 0.6 \times l^0_{c_j}$ or $l > 2 \times l^0_{c_j}$, then l is replaced with $l^0_{c_j}$; similarly for w. When l or w is modified, one of the four vertical edges that is closest to the LiDAR center remains fixed, which may result in changes to x and y. The bottom face of the box is aligned with the ground. This procedure yields boxes that are more consistent with commonsense constraints.

Ablation Target	Method	mAP↑	NDS↑
2D Teacher	(1)SEEM	16.4	20.0
	(2)FC-CLIP	20.5	25.4
	(3)SAN	18.0	23.3
	(4)FC-CLIP+SAN	**22.2**	**26.4**
Box Augmentation	(1)minBA	20.5	25.3
	(2)maxBA	21.7	25.7
	(3)FRBA	21.6	25.7
	(4)DRBA	21.9	25.9
	(5)DGBA	**22.2**	**26.4**
Loss	(1)L_1-loss	21.8	25.9
	(2)smooth L_1-loss	21.8	26.0
	(3)NLL-loss	**22.2**	**26.4**

Table 5. **Ablation study of different targets on nuScenes validation set.** The experiment is based on 10-class object detection (Tab. 1).

Method	ST	0-30m	30-50m	50-80m	0-80m
Supervised	-	34.5	10.0	2.9	18.2
MODEST [CVPR'22]	×	12.5	0.8	0.1	5.0
OYSTER [CVPR'23]	×	12.3	1.1	0.3	5.4
LiSe [ECCV'24]	×	4.7	0.2	0.2	1.8
AnnofreeOD (ours)	×	22.3	1.3	0.6	8.7
MODEST [63]	✓	17.1	1.4	0.3	6.6
OYSTER [66]	✓	19.3	1.8	0.4	8.0
LiSe [67]	✓	24.0	4.4	**1.3**	11.4
AnnofreeOD (ours)	✓	**28.1**	**4.6**	**1.3**	**12.1**

Table 6. **Class-agnostic object detection by converting nuScenes to KITTI format.** We report AP_{3D} at $IoU = 0.25$ for objects across various distances. ST denotes self-training.

B. Supplementary Experiments

B.1. Comparison Results

B.1.1. Converting nuScenes to KITTI Format for Class-Agnostic Object Detection

This experiment shares the same objective as Sec. 4.2.3, but it employs a different experimental setup. Specifically, the nuScenes dataset was converted into KITTI format, and scenes were sampled. As a result, the training set is significantly smaller than that in Sec. 4.2.3, rendering direct comparisons between the two infeasible. We present the results in Tab. 6. The baselines referenced for comparison include MODEST [63], OTSTER [66], and LiSe [67], all of which are based on scene flow. Slightly different, LiSe incorporates results from 2D detection. Given the limited data volume, we adopted LiSe's pseudo labels-based self-paced learning, conducting ten rounds of self-training. Overall, the evaluation demonstrates that AnnofreeOD outperforms other baselines across various metrics. Even without self-training, it surpasses many self-trained results.

2D Teacher	Car	Ped	Bar	Trailer	T.C.
FC-CLIP	43.1	54.5	10.6	10.1	3.4
SAN	50.9	51.0	32.9	10.0	33.5

Table 7. **Partial annotation-free segmentation results (% IoU) of FC-CLIP and SAN on nuScenes validation set.** The results demonstrate that different 2D teacher models exhibit varying capabilities in understanding different semantics.

B.2. Ablation Study

B.2.1. Differences Among Various Teacher Models in 2D-to-3D Knowledge Distillation

Most of the semantic and morphological knowledge required for AnnofreeOD comes from the 2D open-vocabulary segmentation model. The quality of the 2D teacher model determines the quality of the 3D pseudo-labels and, consequently, the performance of the annotation-free detector. In Tab. 5, we benchmark different 2D teacher models (or combinations): (1) SEEM [72] for 2D panoptic segmentation; (2) FC-CLIP for 2D panoptic segmentation; (3) SAN for 2D semantic segmentation; (4) FC-CLIP for 2D panoptic segmentation with SAN's semantic segmentation results for optimization, especially for "barrier" and "traffic cone". The comparison between (1) vs. (2) shows that using a better 2D teacher model brings better performance. The results of (2) vs. (3) indicate that panoptic (or instance) segmentation models are more suitable as teacher models, as they enable instance-level clustering for denoising, whereas semantic-level clustering struggles to filter out background noise. This is because a semantic-level mask may contain multiple objects, making it impossible to identify noise points through clustering. Comparing (4) and (2), we see an increase of +1.7% mAP, with notable gains of 10.8% AP and 5.8% AP for "barrier" and "traffic cone," respectively. In Tab. 7, we present the performance of the 3D segmentation models trained with FC-CLIP and SAN. Inspired by the mixture-of-experts (MoE), we fuse the labels from different teachers, achieving promising results.

B.2.2. Orientation Estimation

Tab. 8 presents the impact of different orientation estimation strategies. No estimation was used as a control group (1). Among the experimental groups, our method (4) using DINOv2 [36] for 2D patch feature matching, while (2) utilizes a common-sense-based approach, and (3) applies traditional 2D feature matching with SIFT [30] and Brute Force Matching (BFM) for feature point extraction and cor-

Ablation Target	Method	$AOE_{Car}\downarrow$	mAOE $\downarrow$
Orientation Estimation	(1)-	1.86	1.49
	(2)common-sense-base	1.30	1.37
	(3)SIFT+BFM	1.02	1.40
	(4)DINOv2	**0.62**	**1.31**

Table 8. **Ablation study for orientation estimation.** (m)AOE stands for (mean) Average Orientation Error.

respondence. Results shows that (4) improved 1.24% and 0.68% over (1) and (2) on AOE_{car}, respectively. The unsatisfactory result of (3) is attributed to its sensitivity to occlusion, background variations, and noise.

B.2.3. Box Augmentation

In Tab. 5, we compare various box augmentation (BA) methods, including MinBA, MaxBA, FRBA, DRBA, and DGBA. Experimental results show that DGBA improved +1.7% mAP and +1.1% NDS compared to no box augmentation (MinBA). The significance of BA is that it neutralizes some noise and eliminates bias in original pseudo boxes. Such bias can lead to fixed prediction patterns, which are unfavorable for training on large-scale datasets. Unlike conventional denoising techniques, BA aims to achieve better performance by increasing the model's tolerance to noise.

B.2.4. The Role of NRR

To verify the effectiveness of NRR in boundary regression, we designed an ablation experiment comparing traditional $L1$-loss, smooth $L1$-loss, and Gaussian NLL-loss. The result, reported in Table 5, shows that NLL-loss outperforms the other losses, demonstrating the advantage of probability optimization in handling noisy data. Compared to $L1$-loss, smooth $L1$-loss (Huber loss) does not demonstrate any clear advantage. This is because smooth $L1$-loss imposes more penalties for precise predictions but is less effective when dealing with high-noise data. In contrast, NLL-loss models the regression target as a probability distribution. Due to its high tolerance to uncertain data, NLL-loss can better adapt to the variations of BA targets.

B.3. Visualization

Fig. 4 visualizes the process of box generation, refinement, and augmentation. It can be observed that our method performs excellently in multi-class recognition, though there is still a gap compared to the ground truth.

C. Discussion

C.1. Distinction Between "Unsupervised" and "Annotation-Free":

In certain contexts, "unsupervised" [20, 40, 63, 66] refers to model generation without the involvement of implicit labels (including those from teacher models). "Annotation-free" indicates that no labels are used during the current training process [27, 45]. Consequently, employing an off-the-shelf pre-trained model aligns with the definition of "annotation-free". Given the potential ambiguity of these terms, we adopt "annotation-free" throughout this paper for clarity and rigor. However, using the expression "unsupervised" is also fine.

C.2. Comparison with weakly supervised methods:

We note that some weakly supervised 3D detection methods, *e.g.*, FGR [50], GAL [59], and ALPI [18], train 3D detectors using 2D annotations. These methods achieve impressive performance and outperform AnnofreeOD on certain baselines. We attribute this gap mainly stems from our method's inability to precisely localize all targets, whereas 2D boxes provide both the location and category information. These weakly supervised approaches are worth attention, and integrating them with AnnofreeOD could lead to more advanced label-efficient solutions.

C.3. Dataset Selection Criteria

Our experiments primarily rely on the nuScenes dataset. This choice is motivated by nuScenes' low frame rate and sparse point cloud density, which significantly impact the performance of scene flow-based methods. AnnofreeOD adapts effectively to these limitations. It also achieves multi-class detection (>3 classes), a capability absent in previous approaches. However, on datasets with higher frame rates, such as Waymo [46], scene flow-based methods may be more advantageous.

Additionally, Waymo comprises 1950 segments, each lasting 20 seconds and collected at 10 Hz (totaling 390000 frames). In comparison, nuScenes contains 1,000 scenes, each 20 seconds long, annotated at 2 Hz, totaling 40000 frames. Thus, Waymo offers ten times the number of samples compared to nuScenes, leading to a proportional increase in pseudo-label generation time. Given the computational cost and the target application environment of our method, selecting nuScenes as the primary dataset is a well-justified decision.

C.4. Limitations

First, AnnofreeOD is constrained by its reliance on 2D teacher models. Employing a more advanced 2D teacher model would yield superior results. Our approach integrates predictions from multiple teacher models. However, this increases computational demands.

Second, there is a noticeable disparity in detection accuracy across different classes. As discussed in Sec. 4.2.1, the 2D teacher model exhibits a limited understanding of certain semantics. Moreover, errors in class assignments within pseudo-labels further exacerbate this issue.

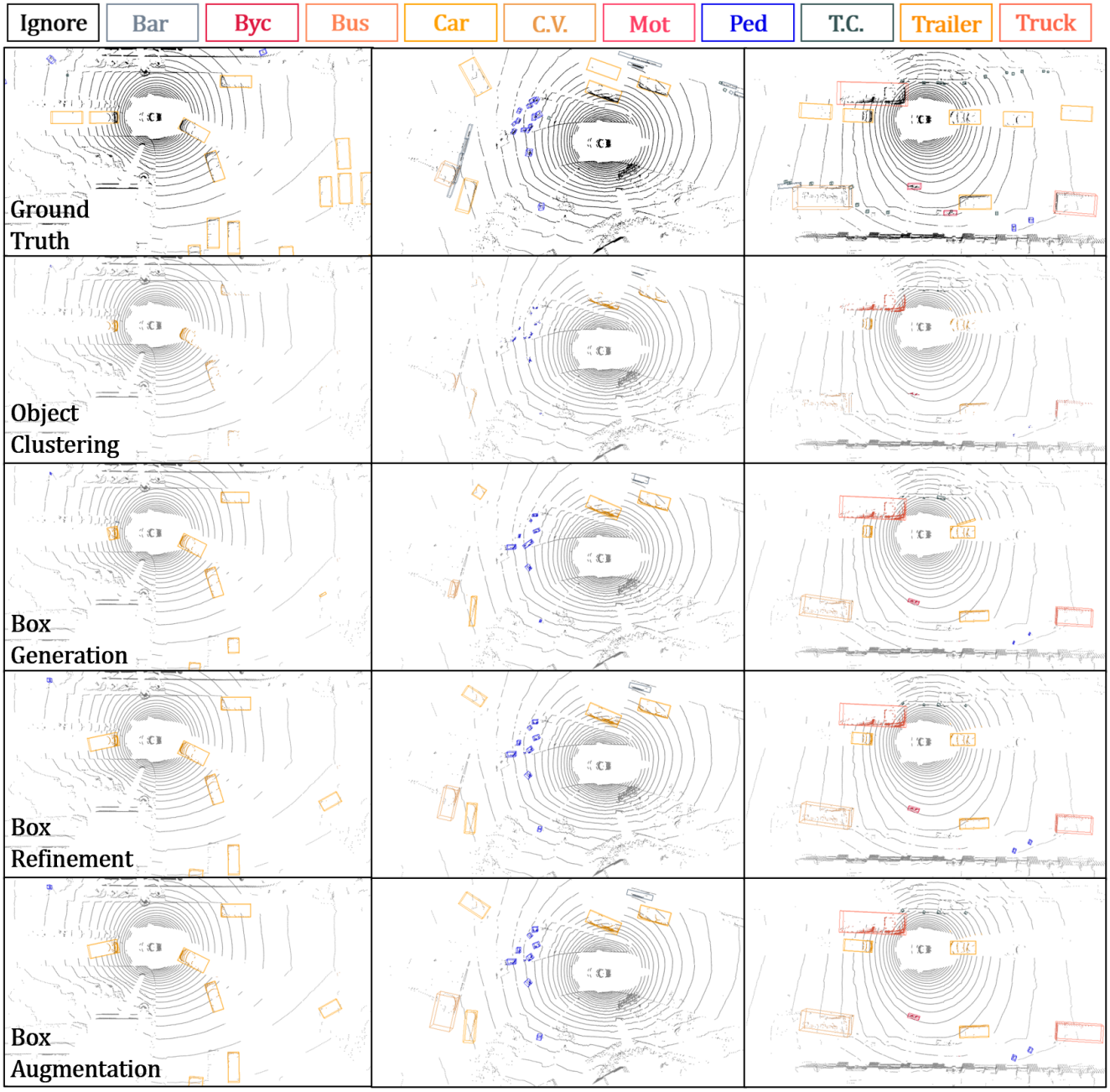

Figure 4. **Visualization of the process for generating pseudo boxes and box augmentation.**

AnnofreeOD outperforms scene flow-based methods in datasets with low frame rates and sparse point clouds. However, its effectiveness compared to scene flow-based approaches in high frame rate settings remains unverified. We plan to explore this topic in future work.