A-Teacher: Asymmetric Network for 3D Semi-Supervised Object Detection

Supplementary Material

A. Overview

We first present the outline of this supplementary material. §B elaborates the implementation details of loss functions. §C offers the additional analyses of A-Teacher on quality and efficiency. §D indicates the potential of the proposed attention-based refinement model to apply to offline 3D object detection. §E contains the detailed data augmentation.

B. Details of Loss Function

In this section, we elaborate on the details of the loss functions in manuscript Sec. 3.4, including each component of the proposed attention-based refinement model.

Propagation-based Box Aggregation. For each candidate box, we first find its corresponding ground truth based on the intersection over union (IOU). As for the matched objects, we regard them as foreground and assign their $\mathbf{s}_{\mathrm{gt}}^{\mathrm{f}} = \mathbf{1}$. After that, we exploit the ground truth boxes denoted by superscript gt and candidate boxes denoted by superscript c to calculate the offset $\boldsymbol{\Delta}_{\mathrm{gt}} = \{\Delta_{\mathrm{x}}, \Delta_{\mathrm{y}}, \Delta_{\mathrm{z}}, \Delta_{\mathrm{l}}, \Delta_{\mathrm{w}}, \Delta_{\mathrm{h}}, \Delta_{\theta}\}$ follow [3], specifically,

$$\begin{aligned}
&\Delta_{\mathrm{x}} = \frac{\mathrm{x}^{\mathrm{gt}} - \mathrm{x}^{\mathrm{c}}}{\mathrm{d}^{\mathrm{c}}}, \Delta_{\mathrm{y}} = \frac{\mathrm{y}^{\mathrm{gt}} - \mathrm{y}^{\mathrm{c}}}{\mathrm{d}^{\mathrm{c}}}, \Delta_{\mathrm{z}} = \frac{\mathrm{z}^{\mathrm{gt}} - \mathrm{z}^{\mathrm{c}}}{\mathrm{h}^{\mathrm{c}}}, \\
&\Delta_{\mathrm{w}} = \log\frac{\mathrm{w}^{\mathrm{gt}}}{\mathrm{w}^{\mathrm{c}}}, \Delta_{\mathrm{l}} = \log\frac{\mathrm{l}^{\mathrm{gt}}}{\mathrm{l}^{\mathrm{c}}}, \Delta_{\mathrm{h}} = \log\frac{\mathrm{h}^{\mathrm{gt}}}{\mathrm{h}^{\mathrm{c}}}, \\
&\Delta_{\theta} = \sin\left(\theta^{\mathrm{gt}} - \theta^{\mathrm{c}}\right).
\end{aligned} \tag{1}$$

Conversely, for the unmatched objects, we regard them as background and assign their $\mathbf{s}_{\mathrm{gt}}^{\mathrm{b}} = \mathbf{0}$. After that, we merge the background and foreground to obtain the final label $\mathbf{s}$ for supervising the confidence. Eventually, we calculate the offset loss $\mathcal{L}_{\mathrm{pba}}^{\mathrm{o}}$ and the confidence loss $\mathcal{L}_{\mathrm{pba}}^{\mathrm{o}}$ as Eq. 13 in manuscript Sec. 3.4. As for the utilization of the predicted offset $\boldsymbol{\Delta}$, we use the reverse version of Eq. 1 to acquire the refined pseudo labels.

Dreaming-based Box Aggregation. First, we calculate the distance matrix $\mathbf{D}$ based on the categories and L1 norm between the predicted boxes and the ground truth. Then, conduct the Hungarian algorithm [2] to realize the assignment σ^* between the predicted boxes and the ground truth. After that, we obtain the loss $\mathcal{L}_{\mathrm{dba}}$ analogously to DETR3D [6]. Specifically, for matched objects, we calculate the class confidence loss for categories and regression loss for boxes. As for the unmatched objects, we only utilize the confidence loss.

Spatio-Temporal Deformable Aggregation. The $\mathcal{L}_{\mathrm{sta}}$ is composed by two parts, including the heatmap loss $\mathcal{L}_{\mathrm{hm}}$ and the regression loss $\mathcal{L}_{\mathrm{reg}}$, which is borrowed from CenterPoint [7]. The $\mathcal{L}_{\mathrm{hm}}$ is the focal loss [1] with K-channel, one channel for each of K classes. Specifically, we first convert the boxes in 3D space into grid space which is based on the resolution of the center-based detection head. Then, we generate the ground truth heatmap through the Gaussian function with boxes center and corresponding radius. After that, we transform the center of ground truth boxes into grid space. Eventually, we employ the L_1 norm to obtain the final loss, respectively.

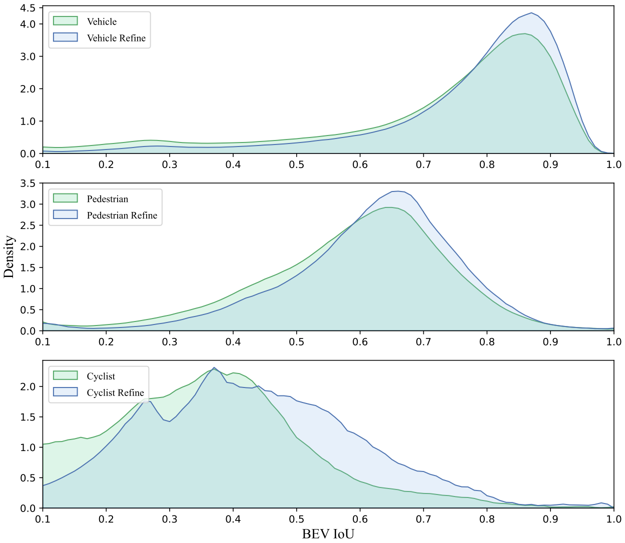

Figure 1. BEV IoU distribution in each category for pseudo labels generated by **vanilla teacher** and that after **refinement model**. The curves are smoothed for clear visualization.

C. Additional Analyses

Analysis about IoU distribution. The experiments in in manuscript Sec. 4.4 (Quality of Pseudo Labels) analyze the precision of pseudo labels, however, no specific analysis of how accurate it is. Therefore, as is shown in Fig 1, we draw the distribution in each category for pseudo labels directly generated by the vanilla teacher and after the refinement model. It is obvious that the IoU distribution after the refinement model is higher in each category, which indicates the pseudo labels generated by our refinement model are more precise. Above all, this experiment reveals the necessity of the utilization of multi-frame information from both the past and future to generate pseudo labels.

Precision based on remaining percentages. Since we can choose different thresholds for different models, solely referring to the analysis of the precision based on the thresholds may not be sufficient. Therefore, we also visualize

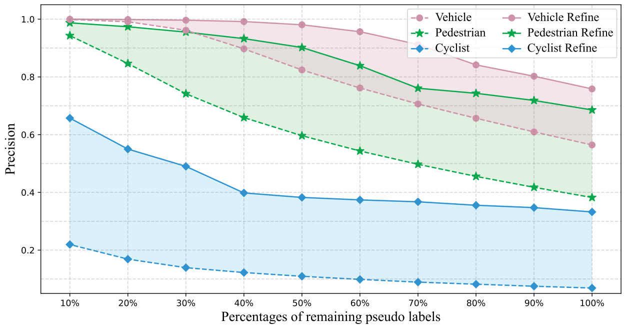

Figure 2. **Precision of pseudo labels based on percentages.** We compare the quality of pseudo labels generated by the vanilla teacher (dashed lines) and that after the refinement model (solid lines) under different remaining percentages applied to classification scores. The shadow region indicates the improvements brought by our refinement model.

the precision of pseudo labels based on different remaining percentages. It is self-evident to observe in Fig. 2 that the pseudo labels generated by the refinement model are more precise than those generated by the vanilla teacher. Combined with Fig. 4 in manuscript Sec. 4.4, the superiority of the proposed attention-based refinement model is proved.

Name	Train (V100/h)	Infer (V100/h)
Vanila Teacher (1f)	-	17.5
Refine Model (3f)	-	6.5
Threshold	60.3	-
HSSDA	84.6	-
A-Teacher	71.1	-

Table 1. **Speed comparison in solely inference and integrated in semi-supervised framework.** Time cost of one epoch on entire Waymo dataset with 1 × NVIDIA Tesla V100 GPU.

Efficient experiment of A-Teacher. Previous experiments exhibit the effectiveness of A-Teacher, we also want to prove the efficiency of our method, all of the speed tests conducted on V100 GPU. First, we compared the time consumption of the attention-based refinement model and the vanilla teacher. As is shown in Tab.1, our refinement model is extremely efficient, only needing 37% time of the vanilla teacher even using the three frames as input. To evaluate the practical efficiency, we record the time consumption when training, compared with vanilla teacher, our approach used an additional 9.8 hours but got a remarkable improvement. Besides, compared with HSSDA [4] which exploits test-time-augmentation (TTA), our approach can save 13.5 hours (15.9%) and achieve better performance.

D. More Experiments

Extension to offline detection. A-Teacher is capable of extending to offline 3D detection tasks. As is shown in Tab. 2, we conduct experiments on 5% (40 sequences) of Waymo dataset [5], compared with directly concatenating the multi-frame point clouds in the past and future, our attention-based refinement model can substantially promote the detection performance, especially for *Pedestrian* and *Cyclist*, which demonstrates the great potential of our approach.

Model	Veh. (mAP)		Ped. (mAP)		Cyc. (mAP)	
	L1	L2	L1	L2	L1	L2
PV(1f)	67.7	59.4	66.4	57.6	43.5	41.9
PV(3f-offline)	67.8	59.6	68.8	60.0	49.7	48.0
PV(1f+Re3f)	69.2	60.8	70.0	61.1	56.9	54.9

Table 2. **Implementation on offline object detection.** Compared with vanilla offline method (point clouds concatenation).

E. Data Augmentation for Refinement Model

In order to exhaustively utilize the labeled data for training, we exploit a series of meticulously designed data augmentation approaches. Our augmentation towards the input point clouds and candidate boxes respectively.

Random box-based perturbation and dropout. Since the vanilla teacher that generates the candidate boxes has already been trained on the labeled data. Thus, the candidate boxes generated by the vanilla teacher may be overfitting for labeled data. Therefore, to enhance the robustness of the refinement model and ameliorate the overfitting issue, the predicted boxes are added with slight perturbation and randomly discarded. Meanwhile, to increase the diversity of the original point clouds, we randomly drop parts of the point clouds in the candidate boxes.

Multi-stage candidate boxes injection. Since the vanilla teacher has different learning levels for different scenarios. Therefore, the proposed refinement model should handle inputs with inconsistent deviations. In order to further increase the diversity of training data, we utilize the candidate boxes generated by the vanilla teacher in different training stages to enrich the bias situations of candidate boxes.

Sequence GT boxes and point clouds paste. The class imbalance is considerable in the Waymo dataset, since the number of *Cyclist* is the minority compared with *Vehicle* and *Pedestrian*. Therefore, it is critical to conduct object-based paste to alleviate class imbalance. In the practical implementation, we paste the point sequences of the identical objects to the original point clouds, meanwhile, removing the points at the original location. Moreover, as for the candidate boxes, we implement augmentations that mimic the above-mentioned box-based perturbation and dropout.

References

[1] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object

detection. In *ICCV*, pages 6569–6578, 2019. 1

[2] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 1

[3] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019. 1

[4] 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 *CVPR*, pages 23819–23828, 2023. 2

[5] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020. 2

[6] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 1

[7] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 1