Pseudo Label Refinery for Unsupervised Domain Adaptation on Cross-dataset 3D Object Detection

Supplementary Material

Considering the space limitation of the main text, we provided more results and discussion in this supplementary material, which is organized as follows:

A. Explanations and Discussions.

A.1. Additional Implementation Details

Throughout both the pre-training and self-training processes, we follow [6, 7] to employ commonly used data augmentation techniques, which encompass random flipping (along the X and Y axes), random global scaling, random global rotation, random object scaling, and random object rotation. We utilize the entire training and validation sets of the NuScenes[1] and KITTI [2] datasets, while for the Waymo [4] dataset, we randomly sample 50% of the training set along with the entire validation set.

A.2. Detailed Generation Process of Pseudo Labels

In a two-stage detector, the generation of pseudo labels, *i.e.* 3D boxes, involves two stages. In the first stage, a set of basic proposals are generated along with their corresponding IoU confidence scores. Each proposal represents a preliminary candidate pseudo box that may contain an object. In the second stage, the extracted Region-of-Interest (RoI) features from each proposal are utilized to refine pseudo boxes. Subsequently, a Non-Maximum Suppression (NMS) is applied to select the most confident pseudo boxes, while suppressing others that exhibit high overlap. Throughout this process, the presence of IPNI introduces imprecise proposals and confuses the RoI features across different object categories, as discussed in the manuscript. Considering that pseudo labels are derived from the basic proposals and their corresponding RoI features, it is evident that IPNI significantly degrades the overall quality of pseudo labels, which is also supported by Fig. 6 and Fig. 7.

A.3. Random Object Scaling *vs.* Complementary Augmentation

Since both Random Object Scaling (ROS) in ST3D [6] and our Complementary Augmentation (CA) adopt the same coordinate conversion method in [5], it is important to note the differences between CA and ROS. Firstly, in CA, we replace the box b with a high-quality pseudo box b_h instead of solely scaling the box itself. Secondly, the scale sizes used in CA are determined by the ratios $(\frac{l_b}{l_h}, \frac{w_b}{w_h}, \frac{h_b}{h_h})$ rather than random values. In this way, we can retain valuable localization and categorization information from the original box b within $\hat{b_h}$. Thirdly, CA is specifically applied in the target domain to enhance the reliability of pseudo labels, whereas ROS is implemented in the source domain to address the domain shift resulting from object size bias.

A.4. Limitations and Future Work

While our proposed method, PERE, demonstrates promising results in cross-dataset 3D object detection, its generalizability to unseen datasets or domains may be limited. The performance can degrade when faced with novel object categories not present in the training data. Addressing this limitation requires exploring techniques to ensure robust performance. In future work, we plan to enhance the current datasets by incorporating synthetic data, which can be can be easily annotated. This augmentation aims to make the datasets more suitable for accommodating novel categories and to enhance the overall training process.

B. Additional Ablation Studies

To gain a deeper understanding of the influence of individual modules within PERE, we conducted additional ablation studies. These experiments are based on PVRCNN [3], conducted on the N $\rightarrow$ K task, and evaluated for the *Car* category, unless otherwise stated.

B.1. Sensitivity Analysis of Key Hyper-parameters

In this section, we perform a sensitivity analysis on several key hyper-parameters, including the deviation level λ,

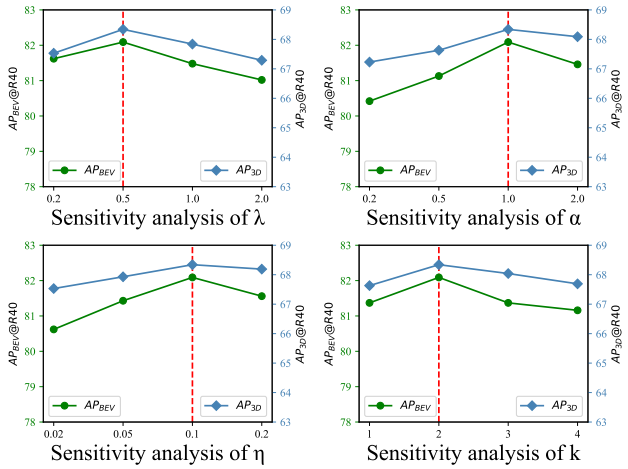

Figure S1. Ablation studies for the deviation level λ, the margin α in the triplet loss, the trade-off parameter η and the update frequency k. The red dotted lines represent the optimal parameters adopted in our experiments.

$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	$\mathrm{AP}_{\mathrm{BEV}}$ / Closed Gap	$\mathrm{AP}_{\mathrm{3D}}$ / Closed Gap
		79.56 / 57.57%	61.35 / 49.76%
✓		79.83 / 58.98%	63.70 / 55.97%
	✓	81.51 / 67.74%	64.46 / 57.98%
✓	✓	**82.09 / 70.77%**	**68.34 / 68.23%**

Table S1. Ablation studies of cross-domain triplet loss

the margin α in the triplet loss, the trade-off parameter η and the update frequency k. Fig. S1 illustrates that our proposed PERE exhibits low sensitivity to these parameters. It consistently achieves remarkable generalization performance across a wide range of parameter values. This finding demonstrates the robustness of our PERE under diverse hyper-parameter settings.

B.2. Analysis of Cross-domain Triplet Loss

In this part, we assess the individual importance within $\mathcal{L}_{triplet}$ (Sec. 3.5). As shown in Table S1, the absence of $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ results in the poorest performance. Furthermore, $\mathcal{L}_{inter}$ contributes more significantly to performance improvements compared to $\mathcal{L}_{intra}$, resulting in respective increases of 5.07% and 3.83% in $\mathrm{AP}_{\mathrm{3D}}$. The inclusion of both $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ yields the best performance, confirming the effectiveness of leveraging the intra-domain and the inter-domain triplet losses to align RoI features.

B.3. Analysis of BoxReplace

In this section, we design two variants of BoxReplace (Sec. 3.3). As shown in Table S2, *w/o scaling* and *random-category* result in a performance decrease of 2.4% and 3.5% (0.4%and 0.9%) in terms of $\mathrm{AP}_{\mathrm{3D}}$ ($\mathrm{AP}_{\mathrm{BEV}}$). These results indicate that both the scaling operation and maintaining category consistency in our BoxReplace approach contribute to performance enhancements.

Method	$\mathrm{AP}_{\mathrm{BEV}}$ / Closed Gap	$\mathrm{AP}_{\mathrm{3D}}$ / Closed Gap
w/o scaling	81.80 / 69.26%	66.72 / 63.95%
random-category	81.38 / 67.07%	65.95 / 61.92%
BoxReplace (ours)	**82.09 / 70.77%**	**68.34 / 68.23%**

Table S2. Effectiveness analysis of each module in BoxReplace. Unlike our approach, which replaces the unreliable box b with a scaled high-confidence box b_h under the same category, *w/o scaling* indicates that box b_h is not scaled, and *random-category* means randomly selecting b_h without maintaining category consistency.

Method	$\mathrm{AP}_{\mathrm{BEV}}$ / Closed Gap	$\mathrm{AP}_{\mathrm{3D}}$ / Closed Gap
V-$\hat{j}$	79.54 / 57.46%	61.42 / 49.95%
V-average	81.57 / 68.06%	63.66 / 55.87%
V-i (ours)	**82.09 / 70.77%**	**68.34 / 68.23%**

Table S3. Effectiveness analysis of proposal generation strategy. V-i, V-$\hat{j}$ and V-average denotes that proposals $\mathcal{I}$ and $\mathcal{E}$ inherit the values of (l, w, h, θ) from proposal i, proposal $\hat{j}$, and the average values of proposals i and $\hat{j}$, respectively.

B.4. Analysis of Proposal Generation Strategy

In this section, we formulate two variants to generate additional proposals (Sec. 3.4). As shown in Table S3, the performance of proposals $\mathcal{I}$ and $\mathcal{E}$ inheriting the values of (l, w, h, θ) from proposal i exhibits the highest precision. Additionally, the V-average variant achieves the second-best performance. These results emphasize that inheriting proposal $i \in \mathsf{P}^h$ yields superior precision compared to proposals $j \in \mathsf{P}^r$ and their corresponding average values.

C. Qualitative Results

C.1. Detection Performance

To provide a more intuitive illustration of detection performance, we randomly selected four point cloud samples from the validation set of KITTI [2]. The qualitative results are presented in Fig. S2. Compared to the source only model, our PERE significantly enhances the generalization ability through pseudo label refinement. Specifically, the results demonstrate that the source only model generates a higher number of redundant false positive boxes while overlooking more true positive boxes. In contrast, PERE achieves a more reasonable performance by producing fewer false boxes. However, despite the improvements, as shown in the lower part of Fig. S2, PERE still exhibits some failure cases due to the inevitable domain gaps.

C.2. Direct Removal *vs.* Complementary Augmentation

As mentioned in our paper, another naive solution is to directly remove all unreliable boxes and their internal points (Direct Removal). To facilitate a more intuitive comparison between Direct Removal and our complementary augmentation, we randomly selected four point cloud samples from the KITTI validation set [2]. Fig. S3 illustrates that Direct Removal overlooks more true positive boxes. This occurs because, during testing, points within unreliable boxes are misclassified as background points, leading to the model becoming trapped in local minima. In contrast, our complementary augmentation, which incorporates BoxReplace, utilizes points within high-confidence boxes as foreground points at unreliable locations. This operation effectively prevents the model from getting stuck in local minima.

C.3. Interpolation and Extrapolation

In this section, to present the interpolation and extrapolation operations in 3D space more intuitively, we randomly selected four instances along with their corresponding proposals. As shown in Fig. S4, the interpolation and extrapolation operations exhibit the ability to generate more precise proposals compared to the basic proposals. These operations go beyond solely focusing on regions with similar point numbers as instances in D_s, resulting in an effective enhancement of proposal precision.

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[2] 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*, pages 3354–3361. IEEE, 2012. 1, 2, 3

[3] 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*, pages 10529–10538, 2020. 1, 4, 5

[4] 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*, pages 2446–2454, 2020. 1

[5] Y. Yan, Y. Mao, and B. Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1

[6] J. Yang, S. Shi, Z. Wang, H. Li, and X. Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10368–10378, 2021. 1

[7] Z. Yihan, C. Wang, Y. Wang, H. Xu, C. Ye, Z. Yang, and C. Ma. Learning transferable features for point cloud detection via 3d contrastive co-training. *Advances in Neural Information Processing Systems*, 34:21493–21504, 2021. 1

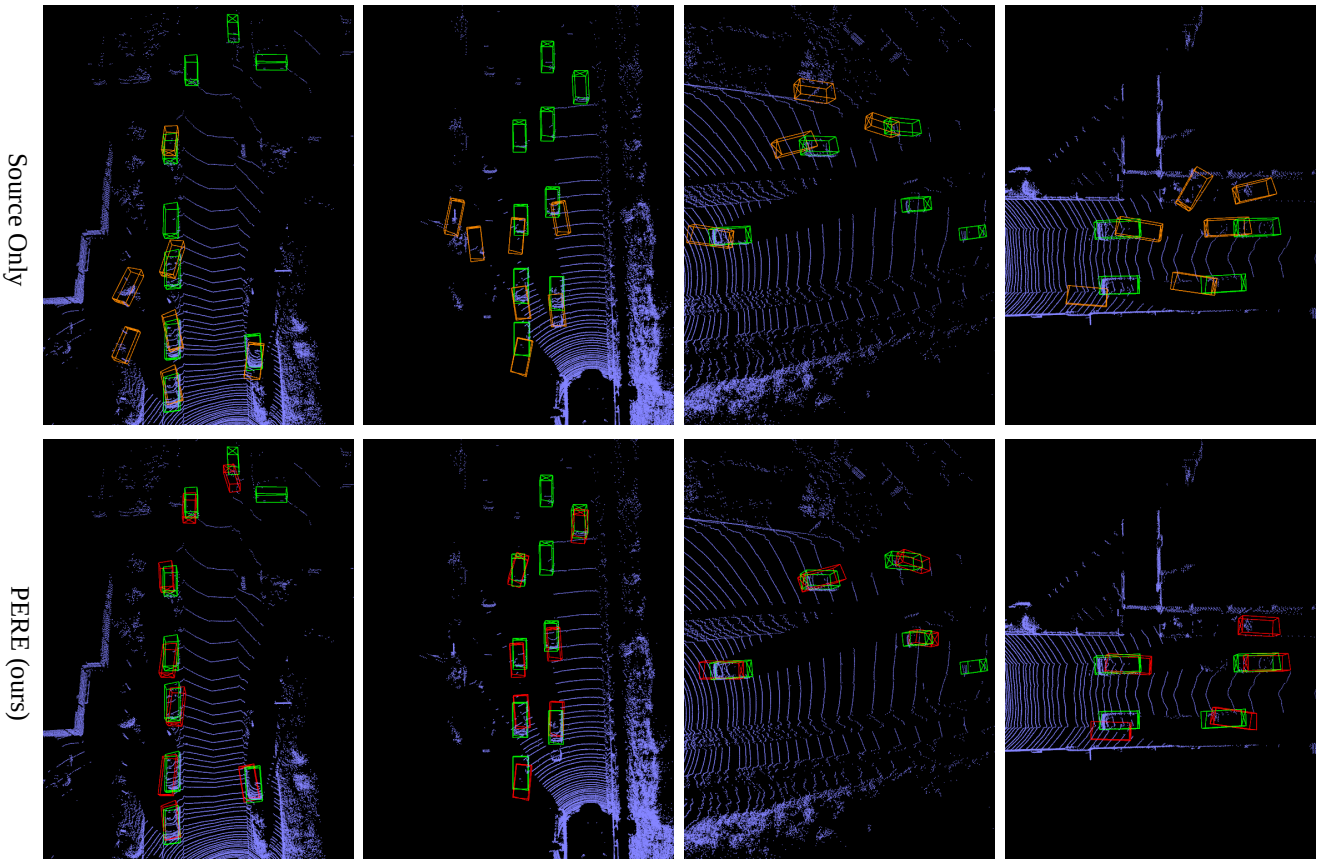

Figure S2. Qualitative results of detection performance on the KITTI val set. The ground-truth boxes are highlighted in green, the detection results obtained from our PERE are highlighted in red, and the detection results derived from the source only model are highlighted in orange. All qualitative results are generated based on PVRCNN [3] and conducted on the N → K task.

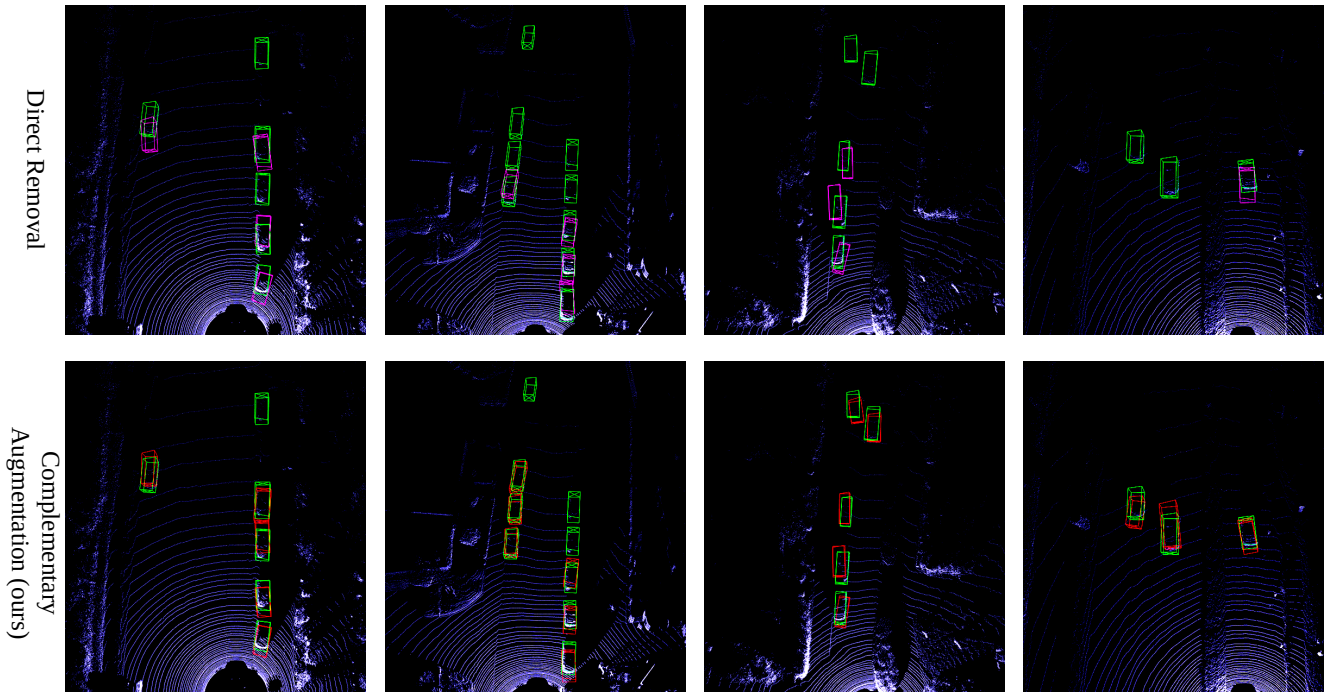

Figure S3. Qualitative results of the comparison between Direct Removal and our complementary augmentation. The ground-truth boxes are highlighted in green, the detection results obtained from our PERE are highlighted in red, and the detection results derived from the Direct Removal model are highlighted in purple. All qualitative results are evaluated on the KITTI val set, based on PVRCNN [3] and conducted on the N → K task.

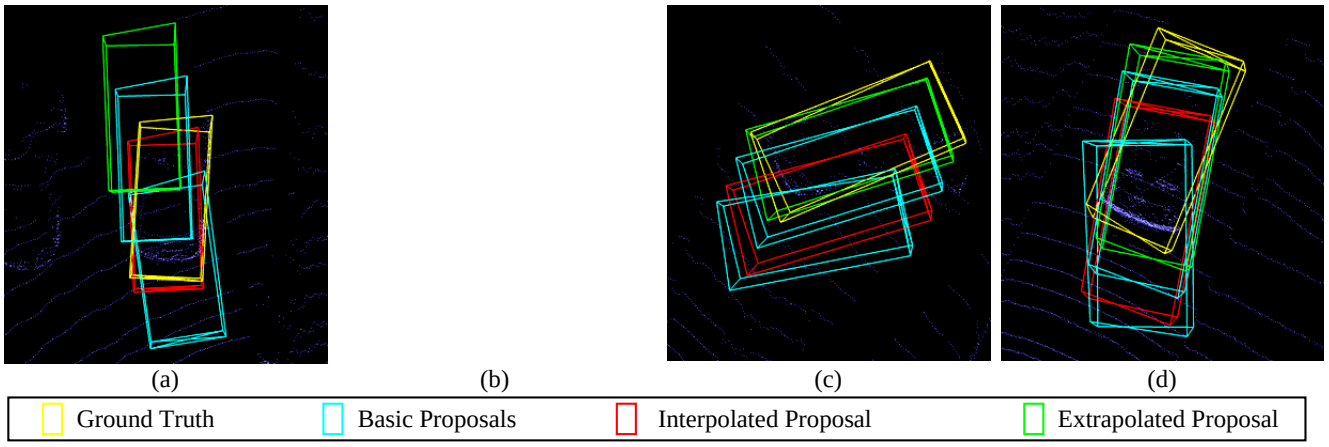

Figure S4. Qualitative results of proposals. For the sake of brevity, we omit other basic low-confidence proposals to present the interpolation and extrapolation operations in 3D space. (a) and (b) demonstrate that the interpolated proposals exhibit the closest alignment with their corresponding instances, while (c) and (d) depict that the extrapolated proposals align closest with their corresponding instances. All qualitative results are generated based on PVRCNN [3] and conducted on the N → K task.