MonoSAOD: Monocular 3D Object Detection with Sparsely Annotated label
– *Supplementary Material* –

Algorithm 1 Road-Aware Patch Augmentation (RAPA)

Input: Source patch P_s with label $l_s = (x_s, y_s, z_s, h, w, l, r_y)$, target image I_t, camera extrinsics $[R_s|T_s]$, $[R_t|T_t]$, road mask M_{road}, existing boxes $\mathcal{B}_{\text{existing}}$
Output: Augmented image I'_t and label l'_t

$(x_t, y_t, z_t)^\top = [R_t|T_t][R_s|T_s]^{-1}(x_s, y_s, z_s)^\top$

for $n = 1$ to $N_{\max}$ **do**
  Sample $x_{\text{offset}} \in [-\delta, \delta]$
  $(x'_t, y'_t, z'_t)^\top = (x_t, y_t, z_t)^\top + (x_{\text{offset}}, 0, 0)^\top$
  $\theta_{\text{new}} = \arctan 2(x'_t, z'_t)$
  $r'_y = \alpha + \theta_{\text{new}}$
  Project $(x'_t, y'_t, z'_t, h, w, l, r'_y)$ to 2D box $\mathbf{b}$
  if $\frac{\sum_{(i,j)\in\mathbf{b}} M_{\text{road}}(i,j)}{|\mathbf{b}|} < \tau_{\text{road}}$ **then**
    continue
  end if
  if $\exists \mathbf{b}_e \in \mathcal{B}_{\text{existing}} : \text{IoU}(\mathbf{b}, \mathbf{b}_e) \geq \tau_{\text{overlap}}$ **then**
    continue
  end if
  break
end for

Resize P_s to match $\mathbf{b}$ and paste onto $I_t \rightarrow I'_t$
$l'_t = (x'_t, y'_t, z'_t, h, w, l, r'_y)$
return (I'_t, l'_t)

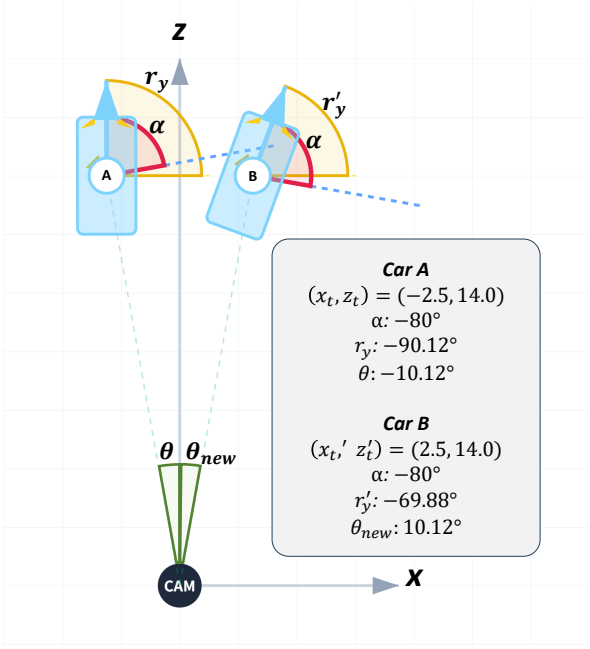

Figure S.1. Conceptual visualization of 3D geometric translation of RAPA. The coordinates (x'_t, z'_t) represent the new 3D position of the object after applying the horizontal translation. Car A is translated to position B while preserving the observation angle α to maintain consistent patch appearance. The global orientation r_y is updated to r'_y based on the new camera viewing angle (θ to θ_{new}), ensuring 3D geometric consistency.

1. Additional Implementation Details

In Table 3 of the main paper, we set the depth reliability threshold to $\tau_{\text{depth}} = 0.7$ for the setting without RAPA, while all other experiments use $\tau_{\text{depth}} = 1.0$. This adjustment is necessary because Table 3 of the main paper evaluates the model without RAPA, resulting in a different depth–uncertainty distribution from the full framework. Since RAPA alleviates geometric ambiguity by enforcing more consistent object appearances, the model produces more reliable depth estimates when RAPA is applied. To ensure a fair and controlled comparison in the no-RAPA ablation, we lower τ_{depth} accordingly.

Additional Implementation Details of RAPA. During offline patch extraction, we apply strict filtering to ensure high-quality patches. From the sparse ground-truth annotations, we select only objects with truncation level 0 and occlusion level 0 to ensure full visibility without boundary cuts. We also enforce depth constraints of $2.0 \leq z < 65.0$ meters, as MonoDETR [8] and MonoDGP [2] internally filter out labels outside this range; this prevents augmented patches from appearing without valid labels.

For segmentation, SAM is prompted with 2D bounding boxes to extract clean car-only patches with transparent backgrounds, saved as RGBA images. We additionally generate road segmentation masks M_{road} for each training scene using SAM with manually provided point prompts on drivable areas. Each mask is a binary image where non-zero pixels denote road regions. These masks are created once before training and cached for fast lookup.

For valid placement search (detailed in Algorithm 1), we set the horizontal search range to $\delta = 5.0$ meters and uniformly sample $m = 10$ candidate offsets, exploring positions within $[-5.0, 5.0]$ meters from the transformed 3D location. Each candidate must satisfy two constraints. (1) A road-overlap threshold of $\tau_{\text{road}} = 0.7$, requiring at least 70% of the projected 2D box to lie on drivable regions. (2) An overlap threshold of $\tau_{\text{overlap}} = 0.1$, limiting IoU with existing labeled objects to at most 10% to avoid unrealistic

Table S.1. Detection comparison on the Clear and Foggy KITTI images under the 30% annotation ratio. We report both AP_{3D} and AP_{BEV}.

Method	Clear Image (30%)						Foggy Image (30%)					
	AP_{3D}			AP_{BEV}			AP_{3D}			AP_{BEV}		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Baseline	11.17	8.73	7.56	17.24	13.62	11.49	11.10	7.43	5.81	16.78	11.99	9.91
Co-mining	16.01	12.62	10.38	24.81	18.31	15.28	11.22	7.91	6.20	17.04	12.09	9.76
SparseDet	16.95	13.30	10.97	24.78	19.41	16.52	11.40	7.93	6.52	17.64	12.77	10.41
Calibrated Teacher	17.14	12.96	10.58	24.35	18.04	15.01	11.97	8.65	7.26	18.99	13.86	11.53
Co-student	15.99	12.67	10.38	24.76	18.27	15.20	11.57	7.97	6.20	17.58	12.36	9.97
Proposed	**21.28**	**15.60**	**12.79**	**28.45**	**20.40**	**17.04**	**19.11**	**13.72**	**10.35**	**28.28**	**19.49**	**15.16**

Table S.2. Detection results of car category on the KITTI validation set under the lower annotation ratios of 10%, 20%, and 30%. We compare our method with existing SAOD approaches reproduced using their official implementations for fair comparison. **Bold**/underline fonts indicate the best/second-best results.

Method	10%			20%			30%		
	Easy	**Mod.**	**Hard**	**Easy**	**Mod.**	**Hard**	**Easy**	**Mod.**	**Hard**
Baseline [8]	1.04	1.15	0.15	5.38	4.70	3.70	11.17	8.73	7.56
Co-mining [5]	0.00	2.50	2.50	3.88	3.90	3.26	16.01	12.62	10.38
SparseDet [3]	1.88	1.67	1.67	8.83	7.07	5.59	16.95	13.30	10.97
Calibrated Teacher [4]	1.62	1.55	1.05	6.04	4.88	4.17	17.14	12.96	10.58
Co-student [6]	0.00	2.50	2.50	1.25	1.66	1.66	15.99	12.67	10.38
Proposed Method	**14.18**	**10.12**	**7.58**	**19.48**	**13.78**	**10.93**	**21.28**	**15.60**	**12.79**

occlusions.

Once a valid placement is found, we update the rotation by recomputing the camera-viewing angle to preserve geometric consistency. This viewpoint-driven rotation adjustment is shown in Figure S.1, where the object orientation changes depending on its horizontal position relative to the camera.

We apply RAPA consistently across the entire training pipeline, including the teacher initialization stage (195 epochs). Patch-augmented images are regenerated at every epoch. To avoid duplicating objects within the same scene, we exclude patches originating from the target scene. For each image, we set the maximum number of placement trials to $N_{\text{max}} = 40$, meaning the algorithm attempts up to 40 candidate locations to find a valid road-consistent placement before moving to the next sample.

2. Additional Experiments

Additional BEV Results. We report both AP_{3D} and AP_{BEV} results under 30% annotation settings for clear and foggy images in Table S.1. Our method consistently outperforms the existing Sparsely-Annotated Object Detection (SAOD) methods in both metrics.

Results under Extremely Sparse Annotation Settings. We further evaluate our method under extremely sparse annotation settings of 10% and 20% in Table S.2. Our method achieves 10.12 AP (Moderate) at 10% annotation, significantly outperforming the best baseline SparseDet (1.67 AP_{3D}) by +8.45. This large improvement demonstrates that RAPA effectively leverages the limited training data, while PBF provides more reliable pseudo-label selection than confidence-based approaches. At 20% annotation, our method maintains strong performance with 13.78 AP (Moderate), showing +6.71 improvement over SparseDet. The results validate that our approach is particularly effective in extreme low-data regimes.

PBF Module Ablation Study. Our Prototype-Based Filtering (PBF) module uses two scoring functions: S_{proto} for prototype similarity and S_{depth} for depth uncertainty. We ablate these components with and without Road-Aware Patch Augmentation (RAPA) in Table S.3. The results show that both scoring functions contribute to performance gain compared to the baseline, and their effectiveness is further amplified when combined with RAPA.

Hyperparameter Sensitivity. Table S.4 shows that our method is relatively robust to variations in τ_{depth} and τ_{proto}, consistently outperforming existing methods across different settings. However, when τ_{proto} is set to an extreme value (*i.e.*, $\tau_{\text{proto}} \geq 0.95$), overly strict filtering results in too few pseudo-labels being accepted, reducing the effective supervision and leading to a slight performance drop.

Computational Complexity and Overhead. Our frame-

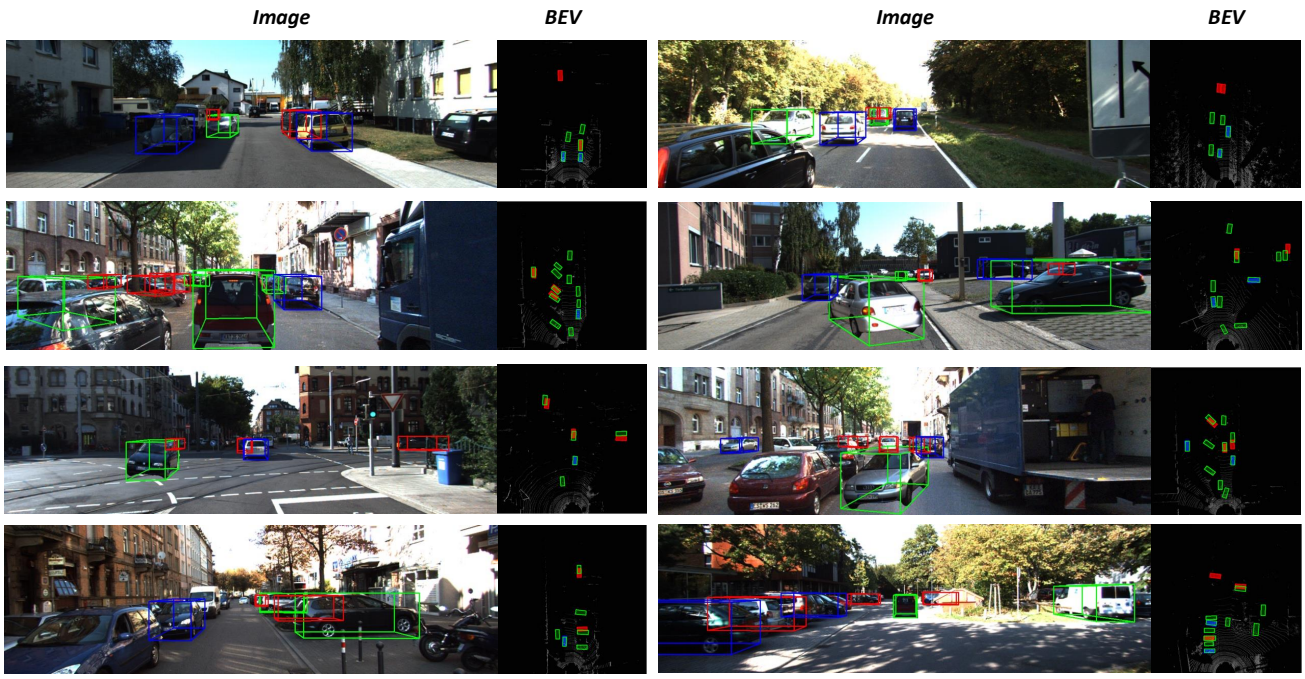

Figure S.2. Visualization of PBF filtering results. In the image view, green boxes indicate the sparsely annotated ground-truth labels, while blue boxes denote the selected pseudo-labels and red boxes indicate predictions filtered out by PBF. In the BEV view (right), green boxes represent the original full ground-truth annotations, and blue boxes correspond to the retained pseudo-labels. The BEV visualization highlights how well the selected pseudo-labels align with the original ground-truth geometry, demonstrating the effectiveness of PBF in filtering out predictions.

Table S.3. Ablation study of PBF modules on the KITTI validation set for the car category under 30% annotation ratio.

RAPA	S_{proto}	S_{depth}	Easy	Mod.	Hard
-	-	-	11.17	8.73	7.56
-	✓	-	16.07	12.26	9.90
-	✓	✓	**16.49**	**12.65**	**10.32**
✓	✓	-	18.73	14.85	12.13
✓	✓	✓	**21.28**	**15.60**	**12.79**

Table S.4. Hyperparameter sensitivity analysis on KITTI (30%).

τ_{depth}	τ_{proto}	Easy	Mod.	Hard
0.8		21.75	15.45	12.48
1.0	0.85	21.28	15.60	12.79
1.2		22.12	15.47	12.32
	0.75	21.56	15.57	12.62
1.0	0.85	21.28	15.60	12.79
	0.95	20.46	14.58	11.80

work introduces additional training cost compared with the baseline due to the teacher–student architecture. However, such a design is commonly adopted in sparse annotation learning frameworks and is used here to ensure fair comparison with prior SAOD methods. Compared with Co-Student [6], our method shows more efficient computation, requiring 1.92 s/iter versus 2.58 s/iter. In addition, the memory overhead is modest, increasing GPU memory usage from 20.45 GB to 21.87 GB (+1.42 GB, 6.9%). The prototype module itself introduces negligible computational overhead, requiring only 0.38 ms/iter and 31 MB of memory.

Robustness against segmentation noise. To evaluate whether our method depends on precise segmentation quality, we simulate common segmentation errors following prior protocols [1], by injecting boundary perturbations to both road and object masks. Specifically, we apply three types of noise: dilation, erosion (5px), and polygonal boundary approximation, as illustrated in Figure. S.3 Table S.6 reports the results under these corrupted masks on KITTI with 30% annotations. Despite mask distortions, our method maintains stable performance across all settings. This robustness arises because segmentation masks are used only for coarse region extraction and placement zone identification. The final object placement and geometry are governed by the 3D geometric constraints in Eq. (1–2, 5–6), which remain unaffected by small boundary inaccuracies. Consequently, approximate masks are sufficient, indicating that our method does not rely on precise segmentation quality or SAM-specific accuracy.

Table S.5. Comparison with other patch augmentation methods used in monocular 3D object detection tasks, on the KITTI validation set under the 30% annotation ratio. Results are reported with and without PBF to show the isolated contribution of each augmentation approach.

Method	Easy	Mod.	Hard
Mix-Teaching [7]	15.42	13.26	11.17
CMAug [9]	17.15	13.30	11.44
RAPA	**20.31**	**14.51**	**11.72**
Mix-Teaching + PBF	18.96	14.77	12.22
CMAug + PBF	18.93	14.79	12.61
RAPA + PBF	**21.28**	**15.60**	**12.79**

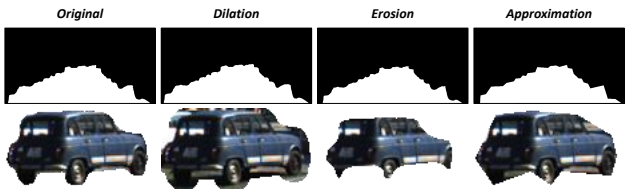

Figure S.3. Examples of simulated segmentation errors: dilation, erosion, and boundary approximation.

Comparison with other Patch Augmentation Methods. We compare RAPA with existing patch augmentation methods Mix-Teaching [7] and CMAug [9] in Table S.5. To isolate the contribution of augmentation from pseudo-label filtering, we report results both with and without our PBF module. RAPA outperforms both baselines even without PBF (14.51 vs. 13.30 and 13.26 AP_{3D} on Moderate), demonstrating the effectiveness of our road-aware placement strategy and 3D geometric consistency. When combined with PBF, RAPA achieves the best performance of 15.60 AP_{3D} on Moderate.

3. Additional Visualization Results

Effect of the PBF Module. Figure S.2 visualizes the filtering behavior of PBF. In the image view, green boxes represent the sparsely annotated ground-truth labels, while blue boxes denote the selected pseudo-labels and red boxes indicate predictions discarded by PBF due to low prototype similarity or high depth uncertainty. In the BEV view, green boxes correspond to the full ground-truth annotations (not the sparse subset), enabling a direct comparison between the retained pseudo-labels (blue) and the true object geometry. To provide a broader set of qualitative examples, we additionally include predictions with confidence scores of at least 0.2. As shown in the figure, the discarded predictions often exhibit incorrect depth estimates, whereas the selected pseudo-labels align well with the original ground truth, demonstrating the effectiveness of PBF in distinguishing reliable predictions from erroneous ones.

Table S.6. Robustness to segmentation noise on KITTI with 30% annotations.

Method	Easy	Mod.	Hard
Baseline (No Aug)	11.17	8.73	7.56
Scale (Dilation)	18.47	14.07	11.38
Scale (Erosion)	18.04	13.86	11.19
Boundary Approx.	20.04	14.30	11.54
SAM (Ours)	**20.31**	**14.51**	**11.72**

Visualization of the GT Bank Updating Process. Figure S.4 provides additional details on how pseudo-labels are accumulated throughout training. In (a), only the sparsely annotated ground-truth boxes are available. As training progresses, PBF selects reliable predictions based on prototype similarity and depth uncertainty, and (b) shows the combined set of sparse ground truth and newly added pseudo-labels, where the latter are highlighted in blue. The green boxes in (b) indicate the current GT Bank, which contains both the original sparse annotations and the pseudo-labels accumulated from previous iterations.

To assess the geometric correctness of the added pseudo-labels, (c) compares them with the original full ground-truth annotations in the BEV view: green boxes represent the original ground truth, while blue boxes denote the pseudo-labels added to the GT Bank. The strong alignment between the two sets demonstrates that PBF effectively selects and stores only geometrically reliable predictions in the GT Bank.

Example of Pseudo-labeling Progression in Foggy Images. We visualize the progression of our method on foggy images in Figure S.5. Similar to clear images shown in the main paper (Figure 3), our framework progressively generates high-quality pseudo-labels even under adverse weather conditions. As training progresses, low-quality predictions are filtered out by PBF, while reliable predictions that satisfy both prototype similarity and depth uncertainty criteria are retained as pseudo-labels (shown in red). The results demonstrate that our method maintains robust pseudo-label generation capabilities on foggy images, validating the effectiveness of feature-level filtering over confidence-based approaches in challenging conditions.

Visual Comparison with Other Patch Augmentation Methods. We visually compare RAPA with other augmentation methods in Figure S.6. Mix-Teaching [7] applies random border cut, color-padding, and mix-up for patch-level augmentation, but maintains objects at their original locations. CMAug [9] considers 3D transformations when placing patches, but uses classic copy-paste augmentation that includes background context, creating unrealistic composite

images. In contrast, RAPA extracts clean object patches using SAM and places them at geometrically valid road locations. This generates more realistic training samples while preserving 3D geometric consistency.

References

[1] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 3

[2] Fanqi Pu, Yifan Wang, Jiru Deng, and Wenming Yang. Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors. In *CVPR*, 2025. 1

[3] Saksham Suri, Saketh Rambhatla, Rama Chellappa, and Abhinav Shrivastava. Sparsedet: Improving sparsely annotated object detection with pseudo-positive mining. In *ICCV*, 2023. 2

[4] Haohan Wang, Liang Liu, Boshen Zhang, Jiangning Zhang, Wuhao Zhang, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Haoqian Wang. Calibrated teacher for sparsely annotated object detection. In *AAAI*, 2023. 2

[5] Tiancai Wang, Tong Yang, Jiale Cao, and Xiangyu Zhang. Co-mining: Self-supervised learning for sparsely annotated object detection. In *AAAI*, 2021. 2

[6] Lianjun Wu, Jiangxiao Han, Zengqiang Zheng, and Xinggang Wang. Co-student: Collaborating strong and weak students for sparsely annotated object detection. In *ECCV*, 2024. 2, 3

[7] Lei Yang, Xinyu Zhang, Jun Li, Li Wang, Minghan Zhu, Chuang Zhang, and Huaping Liu. Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection. *TCSVT*, 2023. 4

[8] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *ICCV*, 2023. 1, 2

[9] Weijia Zhang, Dongnan Liu, Chao Ma, and Weidong Cai. Alleviating foreground sparsity for semi-supervised monocular 3d object detection. In *WACV*, 2024. 4

Figure S.4. Visualization of pseudo-labels added during training. (a) shows the sparsely annotated ground-truth labels. (b) presents the combined set of ground-truth and newly added pseudo-labels, where blue boxes denote the newly added pseudo-labels and green boxes represent all available labels (sparse ground-truth + accumulated pseudo-labels). (c) visualizes the BEV alignment between the newly added pseudo-labels (blue) and the original full ground-truth annotations (green), illustrating that the generated pseudo-labels are geometrically consistent with the true object locations.

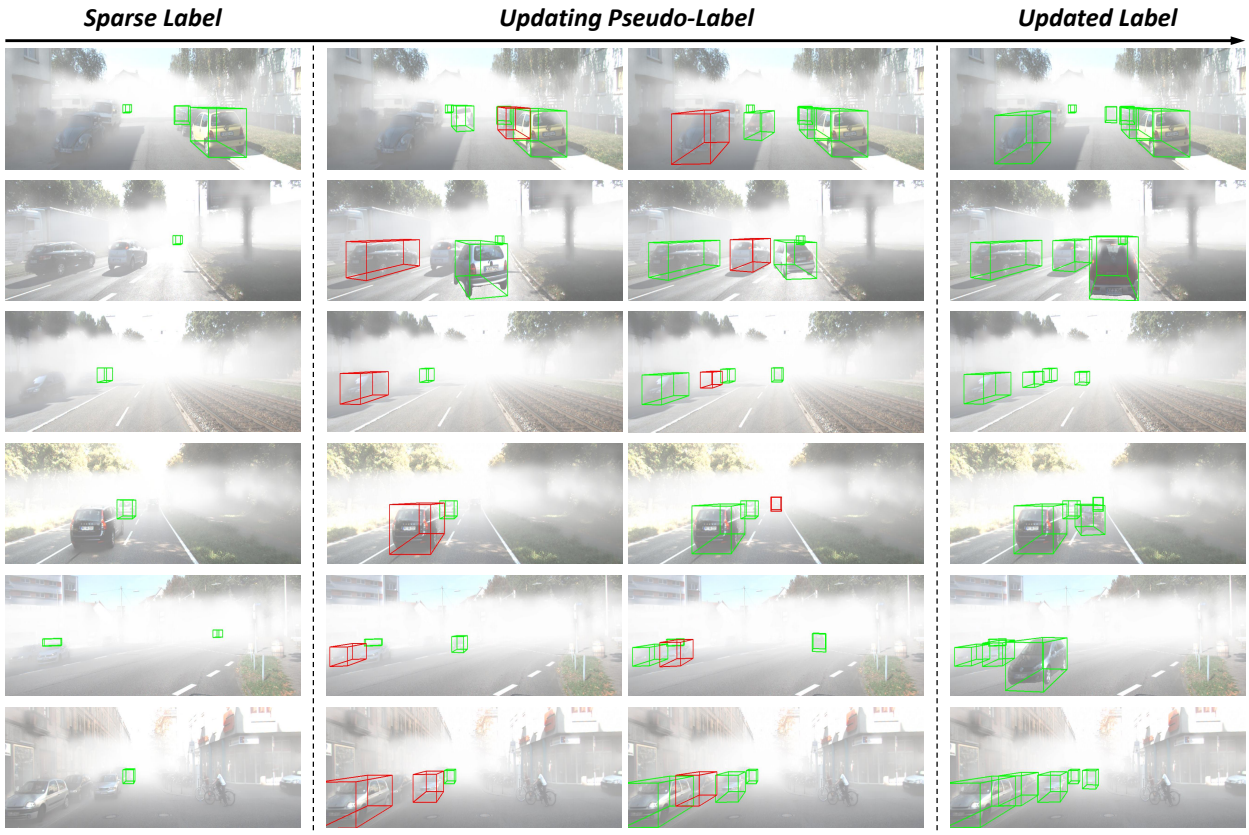

Figure S.5. Progression of pseudo-labels selected by the proposed PBF module for GT Bank enrichment on foggy images. Green boxes denote sparse ground truths and previously accumulated pseudo-labels, while red boxes indicate high-quality pseudo-labels newly selected at the current step. The consistent selection of geometrically and semantically reliable pseudo-labels highlights the effectiveness of the PBF module.

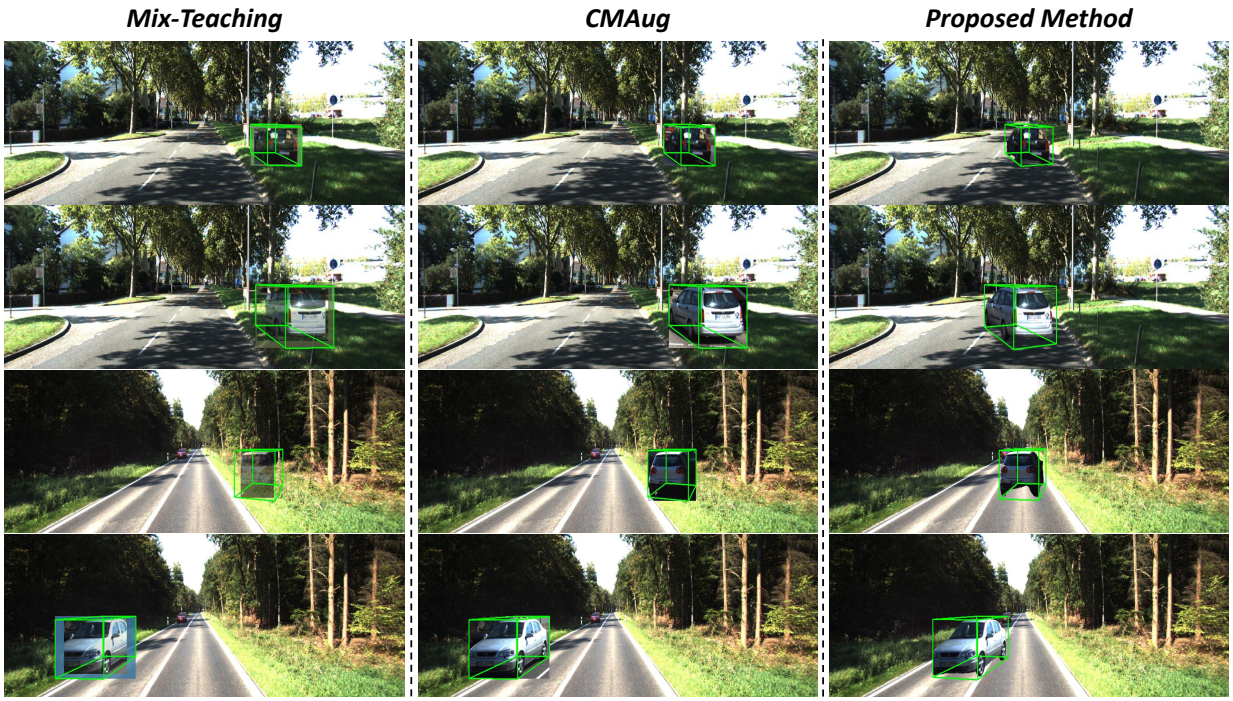

Figure S.6. Qualitative comparison of augmentation methods. Proposed method (RAPA) produces realistic augmentations by placing objects at valid locations with geometrically consistent 3D bounding boxes.