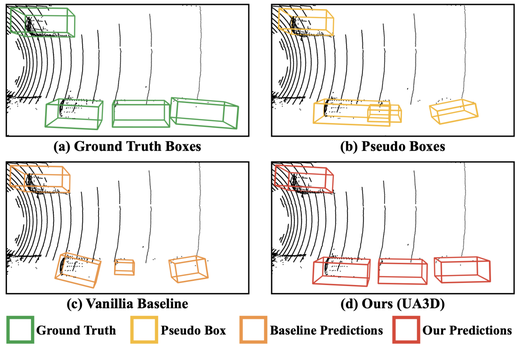
Harnessing Uncertainty-aware Bounding Boxes for Unsupervised 3D Object Detection

Ruiyang Zhang[1] Hu Zhang[2] Zhedong Zheng[1]

[1]FST and ICI, University of Macau, China [2]CSIRO Data61, Australia

Download Code

1. Motivation



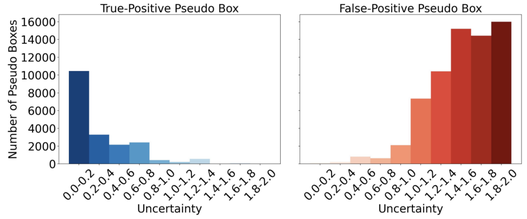
- As shown in the Figure, we notice that pseudo boxes generated by clustering-based algorithms often contain noise (comparing (a) and (b)). Previous methods (c) directly utilize those noisy pseudo boxes to train detection model, leading to suboptimal performance.

2. Contributions

- We introduce fine-grained uncertainty estimation to assess the quality of pseudo boxes in a learnable manner. Following this, we leverage the estimated uncertainty to regularize the iterative training process, realizing the coordinate-level adjustment in optimization.
- Quantitative experiment and Qualitative analysis on nuScenes and Lyft validate the efficacy of our uncertainty-aware framework.

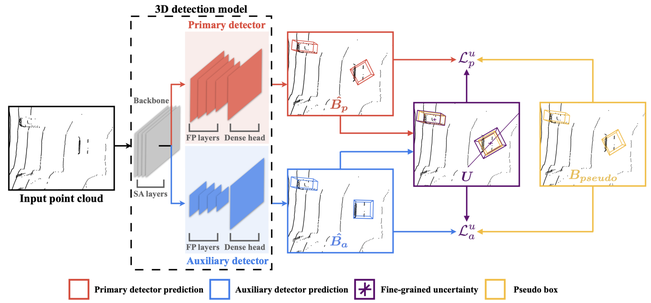
3. Method

- Relation between Uncertainty and Pseudo Boxes



Generally, Our UA3D reliably assigns low uncertainty to accurate pseudo boxes and high uncertainty for noisy ones. For illustration, we average coordinate-level uncertainty to box level.

- Overview Pipeline



- Objectives

Coordinate-level Uncertainty

$$\Delta_x = |x_p - x_a|, \Delta_y = |y_p - y_a|, \Delta_z = |z_p - z_a|,$$
$$\Delta_l = |l_p - l_a|, \Delta_w = |w_p - w_a|, \Delta_h = |h_p - h_a|,$$
$$\Delta_\theta = |\theta_p - \theta_a|,$$

$$U = [\Delta_x, \Delta_y, \Delta_z, \Delta_l, \Delta_w, \Delta_h, \Delta_\theta] \in \mathbb{R}^{n\times 7}$$

Adaptive Uncertainty Regularization

$$\mathcal{L}_p^u = \sum_{i=1}^{7}\left(\frac{\mathcal{L}_{p,i}}{\exp(U_i)} + \lambda \cdot U_i\right), \mathcal{L}_a^u = \sum_{i=1}^{7}\left(\frac{\mathcal{L}_{a,i}}{\exp(U_i)} + \lambda \cdot U_i\right) \qquad \mathcal{L}_{total} = \mathcal{L}_p^u + \mu \cdot \mathcal{L}_a^u$$

4. Experiments

- Quantitative Results

Method	Conference	Data	Round	0m-30m AP_{BEV}	0m-30m AP_{3D}	30m-50m AP_{BEV}	30m-50m AP_{3D}	50m-80m AP_{BEV}	50m-80m AP_{3D}	0m-80m AP_{BEV}	0m-80m AP_{3D}
Supervised [50]	-	-	-	39.8	34.5	12.9	10.0	4.4	2.9	22.2	18.2
LiDAR-Based											
MODEST [50]	CVPR'22	L	0	16.5	12.5	1.3	0.8	0.3	0.1	7.0	5.0
MODEST [50]	CVPR'22	L	10	24.8	17.1	5.5	1.4	1.5	0.5	11.8	6.6
OYSTER [53]	CVPR'23	L	0	14.7	12.3	1.5	1.1	0.5	0.3	6.2	5.4
OYSTER [53]	CVPR'23	L	2	26.6	21.3	4.4	1.8	1.7	0.4	12.7	8.0
LiSe [54]	ECCV'24	L	0	14.8	12.3	1.5	0.4	0.4	0.2	6.1	4.2
LiSe [54]	ECCV'24	L	10	31.4	21.1	7.0	2.5	2.6	0.5	15.7	9.0
UA3D (ours)	-	L	0	13.7	11.5	0.9	0.6	0.5	0.2	5.4	4.9
UA3D (ours)	-	L	2	30.1	19.8	7.8	2.9	3.1	0.5	15.1	9.1
UA3D (ours)	-	L	10	**38.3**	**23.8**	**10.1**	**3.5**	**4.3**	**0.7**	**19.6**	**10.5**
LiDAR-Image Fusion											
LiSe [54]	ECCV'24	L & I	0	5.8	4.7	0.6	0.2	0.3	0.2	2.1	1.8
LiSe [54]	ECCV'24	L & I	10	35.0	24.0	11.4	4.4	4.8	1.3	19.8	11.4
UA3D (ours)	-	L & I	0	8.4	7.3	0.8	0.5	0.4	0.8	3.5	2.4
UA3D (ours)	-	L & I	10	**38.2**	**24.7**	**12.5**	**4.9**	**5.0**	**1.7**	**21.3**	**12.1**

Table 1. **Quantitative Results on nuScenes [2].** UA3D significantly surpasses the state-of-the-art LiSe [54] across all evaluated metrics. This validates the efficacy of proposed coordinate-level uncertainty estimation and regularization in mitigating negative impacts of noisy pseudo boxes, thereby enhancing detection performance. We report AP_{BEV} / AP_{3D} at IoU=0.25. 'L' for LiDAR data and 'I' for image data. Round refers to the number of self-training round. The best results are in **bold**, and the second-best results are underlined.

Method	Conference	Data	Round	0m-30m AP_{BEV}	0m-30m AP_{3D}	30m-50m AP_{BEV}	30m-50m AP_{3D}	50m-80m AP_{BEV}	50m-80m AP_{3D}	0m-80m AP_{BEV}	0m-80m AP_{3D}
Supervised [50]	-	-	-	82.8	82.6	70.8	70.3	50.2	49.6	69.5	69.1
LiDAR-Based											
MODEST-PP [50]	CVPR'22	L	0	46.4	45.4	16.5	10.8	0.9	0.4	21.8	18.0
MODEST-PP [50]	CVPR'22	L	10	49.9	49.3	32.3	27.0	3.5	1.4	30.9	27.3
MODEST [50]	CVPR'22	L	0	65.7	63.0	41.4	36.0	8.9	5.7	42.5	37.9
MODEST [50]	CVPR'22	L	10	73.8	71.3	62.8	60.3	27.0	24.8	57.3	55.1
LiSe [54]	ECCV'24	L	0	42.9	42.6	11.0	10.7	0.5	0.4	20.0	19.6
LiSe [54]	ECCV'24	L	10	76.0	73.4	64.7	61.8	28.5	24.9	59.8	56.1
UA3D (ours)	-	L	0	66.0	63.3	43.8	36.3	8.9	5.1	43.2	38.0
UA3D (ours)	-	L	10	**76.5**	**73.6**	**64.6**	**62.0**	**36.8**	**29.0**	**62.1**	**57.9**
LiDAR-Image Fusion											
LiSe [54]	ECCV'24	L & I	0	54.5	54.0	24.2	22.8	1.4	1.2	29.2	27.5
LiSe [54]	ECCV'24	L & I	10	76.7	74.0	66.1	64.4	46.6	43.7	65.6	62.5
UA3D (ours)	-	L & I	0	60.3	57.4	35.5	28.6	2.4	2.5	35.8	31.1
UA3D (ours)	-	L & I	10	**78.2**	**74.6**	**67.3**	**65.1**	**49.2**	**46.0**	**68.1**	**64.2**

Table 2. **Quantitative Results on Lyft [11].** UA3D outperforms LiSe [54] by a clear margin, under both LiDAR-based and LiDAR-image fusion settings. Notably, we employ same hyper-parameters as those in nuScenes, validating robustness of UA3D across different datasets.

- Ablation Studies

Method	0m-30m BEV	0m-30m 3D	30m-50m BEV	30m-50m 3D	50m-80m BEV	50m-80m 3D	0m-80m BEV	0m-80m 3D
Rule-Based								
Distance Rule	29.6	19.6	7.2	2.2	3.2	0.5	14.8	8.1
Volume Rule	25.7	17.7	5.6	2.2	2.5	0.4	12.3	7.4
Num. Point Rule	27.3	17.6	7.3	2.8	2.3	0.3	13.7	7.5
Regression-Based								
Additional Channel	26.3	18.8	4.9	2.2	2.0	0.3	12.1	7.7
Additional FC	27.2	19.7	4.0	1.9	1.2	0.1	12.5	8.1
Ensemble-Based								
10 Members	32.5	20.7	5.5	2.3	3.1	0.4	15.0	8.6
20 Members	32.1	23.8	10.1	3.5	3.6	0.7	15.3	9.1
Monte Carlo Dropout-Based								
$p = 0.1, N = 10$	29.6	19.6	7.2	2.2	3.2	0.2	14.8	8.1
$p = 0.2, N = 20$	28.1	20.3	8.0	3.3	3.9	0.5	15.7	9.2
UA3D (ours)	**38.3**	**23.8**	**10.1**	**3.5**	**4.3**	**0.7**	**19.6**	**10.5**

Table 3. **Comparison with Other Uncertainty.** Our learnable uncertainty surpasses all other types of uncertainty, validating its superiority in handling complex cases. Results are from nuScenes. BEV is short for AP_{BEV}, and 3D for AP_{3D}.

- Qualitative Comparison