BEVSpread: Spread Voxel Pooling for Bird's-Eye-View Representation in Vision-based Roadside 3D Object Detection

Supplementary Material

A. Appendix

A.1. Potential Impacts

We propose a novel spread voxel pooling approach, named BEVSpread, which is a plug-in and can enhance the performance of existing frustum-based BEV methods in roadsize perception. However, it may produce inaccurate predictions for autonomous vehicles, causing wrong decision-making and potential traffic accidents, and it may help tracking someone else, making privacy invasion happens. Compared with roadsize scenarios, vehicle-side perception is quite different, we think it is worth further research on how to apply spread voxel pooling on vehicle-side methods.

A.2. Analysis on BEV grid size

As shown in Fig. 1a, the predicted point is usually not located in a BEV grid center, previous work simply accumulates this point feature into its corresponding BEV grid, which causes a approximation error. As shown in Tab. 7, Augmenting the density of BEV grids can alleviate this error, and as BEV grid size decreases , the performance gradually improves. When the grid size is set to 0.2m, results of three categories are better than others. However, augmenting the density of BEV grids results in a notable increase in computational workload and memory overhead, especially because of the long perception range in roadside scenarios. Therefore, when keep BEV gird a certain size, spread voxel pooling module can enhance the performance of existing frustum-based BEV methods without causing increased memory consumption.

A.3. Analysis on Weight Function

In order to achieve better performance, we attempt a variety of functions, including L2, Linear and Gaussian distributions, and their function curves are shown in Fig. 7. We compare the mAP of cyclist on DAIR-V2X-I [46] dataset with different weight functions, as shown in Fig. 8, BEVSpread is significantly better than baseline (BEVHeight [44]) and Gaussian function outperforms other counterparts. We believe that this is because Gaussian function's curve is smoother than others near original point, which makes it retaining more location information. And with distant increases, the weight decreases faster to 0, which prevents assigning information to wrong positions. The results of three categories with different weight functions on DAIR-V2X-I dataset can be found in Tab. 8.

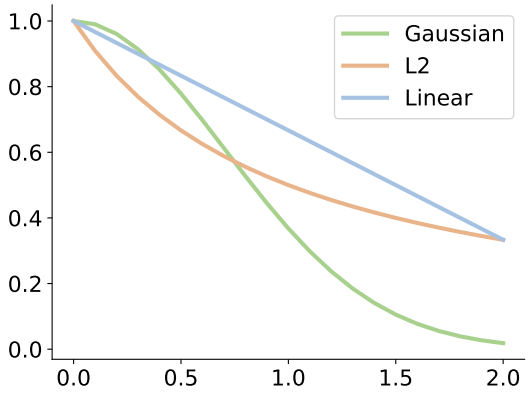

Figure 7. **Different weight functions.** We attempt a variety of functions, including L2, Linear and Gaussian distributions.

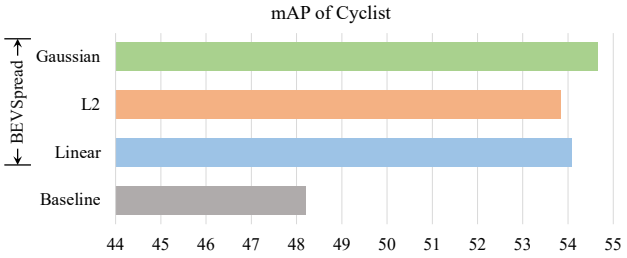

Figure 8. **Analysis on weight function.** BEVSpread is significantly better than baseline (BEVHeight [44]). And Gaussian function outperforms other counterparts.

A.4. Results on Neighbors Number

To investigate the effect of the hyperparameter neighbors number on the performance of BEVSpread, We repect 3 times for each neighbors number k selection. Fig. 6 shows how the mAP of three categoties changes with neighbors number k. The light-blue area indicates the error range and it can be observed that the performance of $k \geq 2$ is significantly better than $k = 1$ (baseline). As k increases, the performance gradually improves and becomes stable. The specific experiment results are presented in Tab. 9.

A.5. Ablation Study on Rope3D

The proposed spread voxel pooling strategy, as a plug-in, can significantly improve the performance of existing frustum-based BEV methods. We further conduct ablation

Table 7. **Analysis on the BEV grid size.** Here we conduct experimets with different BEV grid size on DAIR-V2X-I dataset. BEVHeight [44] is used as base model, resnet-50 is used as image encoder, the BEV grid size is set to 0.8 meters, and the detection range is set to $0\sim100m$

Grid Size	Vehicle (*IoU=0.5*)			Pedestrian (*IoU=0.25*)			Cyclist (*IoU=0.25*)		
	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
0.8m	76.59	64.69	64.76	27.08	25.79	25.29	49.39	52.30	52.84
0.4m	78.07	65.90	65.96	40.94	39.02	39.15	56.69	59.44	59.88
0.2m	**79.05**	**66.86**	**66.91**	**46.06**	**44.02**	**44.30**	**58.10**	**60.37**	**60.76**

Table 8. **Analysis on Weight Function.** We attempt a variety of functions, including L2, Linear and Gaussian distributions. ResNet-50 is used as image encoder, the BEV grid size is set to 0.8 meters, and the detection range is set to $0\sim100m$

Function	Vehicle (*IoU=0.5*)			Pedestrian (*IoU=0.25*)			Cyclist (*IoU=0.25*)		
	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
Base	76.24	64.54	64.13	26.47	25.79	25.72	48.55	48.21	47.96
Linear	77.44	65.43	65.51	31.31	29.84	30.08	52.60	54.10	54.67
L2	77.49	65.46	65.54	30.64	29.29	29.43	52.22	53.85	54.42
Gaussian	**77.67**	**65.61**	**65.69**	**31.34**	**29.94**	**30.08**	**53.53**	**54.65**	**55.17**

Table 9. **Analysis on neighbors number.** For each neightbors number, we repeat 3 times. ResNet-101 is used as image encoder, the BEV grid size is set to 0.4 meters, and the detection range is set to $0\sim100m$

Neighbors Num	Vehicle (*IoU=0.5*)			Pedestrian (*IoU=0.25*)			Cyclist (*IoU=0.25*)			mAP
	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard	
2	78.57	66.34	66.43	45.80	43.74	43.89	58.20	61.09	61.48	58.68
	78.59	66.21	66.39	45.22	43.31	43.46	59.93	61.68	62.14	
	78.73	66.47	66.54	44.38	42.54	42.76	62.89	63.54	63.96	
3	79.00	66.69	66.76	45.86	43.90	44.01	62.26	62.88	63.19	59.01
	78.88	66.65	66.72	45.10	43.13	43.38	60.16	61.65	62.05	
	78.92	66.68	66.76	45.40	43.52	43.66	61.21	62.22	62.56	
4	79.01	66.77	66.83	44.56	42.63	42.76	63.16	63.40	63.73	59.26
	78.70	66.47	66.53	45.67	43.63	43.84	62.65	63.03	63.40	
	78.78	66.63	66.72	45.25	43.30	43.47	62.82	63.17	63.26	
5	78.60	66.38	66.44	45.21	43.14	43.36	62.76	63.41	63.78	59.65
	79.05	66.74	66.80	46.70	44.61	44.89	**63.55**	**63.87**	**64.21**	
	78.80	66.65	66.59	46.05	44.00	44.30	63.07	63.70	63.97	
6	78.61	66.41	66.46	**47.05**	**45.05**	**45.29**	62.05	62.73	63.18	**59.71**
	78.80	66.56	66.53	46.53	44.39	44.70	62.53	63.26	63.67	
	79.07	**66.82**	**66.88**	46.54	44.51	44.71	62.64	63.50	63.75	

Table 10. **Ablation study of spread voxel pooling on the Rope3D [45].** ResNet-50 is used as image encoder, the BEV grid size is set to 0.8 meters, and the detection range is set to $0\sim100m$, and top-k (k=2) nearest BEV grid centers are selected as neighbors.

Method	Vehicle (*IoU=0.5*)			Pedestrian (*IoU=0.25*)			Cyclist (*IoU=0.25*)		
	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
BEVDepth [16]	75.90	65.14	65.10	16.82	16.14	16.31	52.53	51.70	49.81
+ spread voxel pooling	**79.27**	**68.19**	**68.17**	**21.97**	**21.09**	**21.19**	**54.95**	**54.20**	**54.13**
w.r.t. BEVDepth	+3.37	+3.05	+3.07	+5.15	+4.95	+4.88	+2.43	+2.49	+4.32
BEVHeight [44]	76.42	67.24	67.07	21.57	19.79	19.98	56.57	54.80	54.68
+ spread voxel pooling	**80.16**	**70.79**	**70.72**	**23.75**	**21.70**	**21.00**	**59.34**	**57.34**	**57.23**
w.r.t. BEVHeight	+3.74	+3.54	+3.66	+2.18	+1.90	+1.03	+2.77	+2.54	+2.56

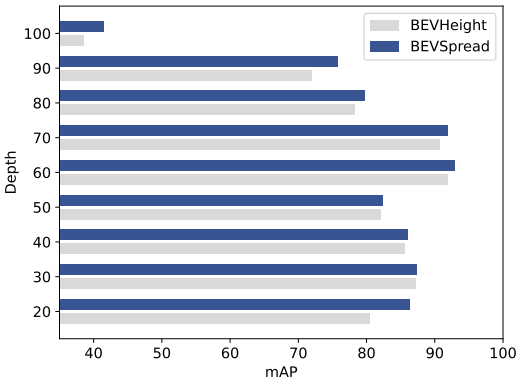

Figure 9. Range-wise evaluation on the DAIR-V2X-I validation set. Metric is $AP_{3D|R40}$ of the **Vehicle** category under moderate setting. The sample interval is 10m, e.g., the value at vertical axis 50 indicates the overall performance of all samples between 45m and 55m.

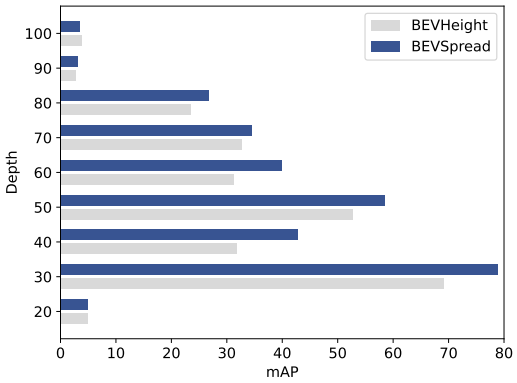

Figure 10. Range-wise evaluation on the DAIR-V2X-I validation set. Metric is $AP_{3D|R40}$ of the **Pedestrian** category under moderate setting. The sample interval is 10m, e.g., the value at vertical axis 50 indicates the overall performance of all samples between 45m and 55m.

study on Rope3D dataset. As shown in Tab. 10, after being deployed to BEVDepth [16], the detection performance has been significantly improved by a margin of (3.16, 4.99 and 3.08) AP in three categories. After being deployed to BEVHeight [44], the detection performance has been improved by a margin of (3.65, 1.70 and 2.62) AP in three categories.

A.6. Robust of BEVSpread

In real-world scenarios, roadside cameras are mounted on poles a few meters above the ground, and are often subjected to variations in extrinsic parameters caused by factors such as wind, vibrations, human adjustments, and other environmental conditions. Additionally, the intrinsic parameters also change between different cameras. So we investigate the robustness of BEVHeight and BEVSpread in the context of fluctuations in camera parameters. We introduce offset noise with a $N(0, 1.67)$ distribution to $roll$ and $pitch$ angles associated with the extrinsic matrix. For the camera $focal$ length, we introduce scale noise, with the scaling coefficient following a $N(1, 0.2)$ distribution. As shown in **??**, BEVSpread maintains the best accuracy across all test-time scenarios involving noisy camera parameters. When only the $pitch$ angle is disturbed, BEVSpread exhibits significantly enhanced robustness compared to BEVHeight, with an improvement of (3.29, 8.75 and 7.71) AP in three categories. These results reveal BEVSpread's excellent robustness and resistance to interference.

A.7. Range-wise Evaluation

We present the accuracy distributions of BEVHeight [44] and the proposed BEVSpread within different range intervals. As shown in Fig. 9, we can observe that BEVSpread exhibit a notable advantage on vehicle category in long-range scenarios, particularly at distances of more than 50 meters. We believe that this advantage stems from the fact that BEVSpread assigns larger weights to the surrounding BEV grids for distant targets, which results in distant objects containing more image features, leading to a better performance in long-range scenarios. As shown in Fig. 10, BEVSpread outperforms BEVHeight at all distances on pedestrian category. We believe that this is because BEVSpread can reduce the approximation error in voxel pooling, which significantly influences the detection of small scale objects like pedestrian.

A.8. Customized CUDA Parallel Acceleration

We refined the cuda acceleration operator of origin voxel pooling. Except for parallel processing of all points, we further parallelize the addition of the 80 channels of context features, which makes voxel pooling faster. BEVHeight takes 74.3ms for one inference and under same configuration, BEVSpread takes 69.8ms when neighbors number is set to 1 and takes 73.9ms when neighbors number is set to 2. As shown in Fig. 6, when neightbors number $k = 2$, the performance of BEVSpread is better than the baseline. So we can make a balance between accuracy and inference time, and BEVSpread can achieve comparable inference time as BEVHeight while significantly improves its performance.

A.9. More Visualizations

In Fig. 11, we present more visuaization results on the DAIR-V2X-I [46] dataset. We can see that BEVSpread has advantages in detection of long-range objects and small scale objects.

Table 11. **Robustness analysis on the DAIR-V2X-I validation set.** Three disturbed factors of roadside cameras are investigated, including focal length, roll angle, and pitch angle.

Method	Disturbed			Vehicle ($IoU = 0.5$)			Pedestrian ($IoU = 0.25$)			Cyclist ($IoU = 0.25$)		
	focal	roll	pitch	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
BEVDepth[16]	-	-	-	75.31	65.24	65.32	32.68	31.01	31.33	46.96	50.88	51.44
	✓	-	-	72.17	60.19	60.20	25.75	25.16	24.35	40.65	47.09	47.21
	-	✓	-	74.78	62.72	62.81	30.80	30.20	30.43	45.58	50.07	50.72
	-	-	✓	74.83	62.76	62.85	30.21	28.62	28.91	46.07	50.15	50.85
	-	✓	✓	74.62	62.57	62.66	30.38	28.87	29.13	45.96	50.15	50.79
	✓	✓	✓	71.91	59.94	59.96	26.61	25.18	25.22	39.79	46.11	46.13
BEVHeight[44]	-	-	-	78.08	65.97	66.04	40.01	38.21	38.38	58.01	60.46	60.95
	✓	-	-	72.30	60.45	60.47	32.18	30.65	29.65	50.06	55.04	55.14
	-	✓	-	77.65	65.57	65.65	38.38	36.60	36.72	56.15	59.11	59.52
	-	-	✓	75.37	63.31	63.38	33.13	31.47	31.63	52.88	56.07	56.44
	-	✓	✓	75.06	63.08	63.16	33.67	31.19	31.30	51.65	54.93	56.83
	✓	✓	✓	71.71	59.92	59.96	27.81	26.43	26.36	47.42	51.19	51.26
BEVSpread (Ours)	-	-	-	79.15	66.86	66.92	46.64	44.61	44.73	63.15	63.55	63.94
	✓	-	-	75.41	63.37	63.40	35.09	35.09	33.33	53.61	56.98	56.90
	-	✓	-	78.44	66.20	66.29	42.19	39.27	40.33	59.93	62.75	63.15
	-	-	✓	78.84	66.51	66.59	42.03	40.14	40.30	61.22	63.47	63.84
	-	✓	✓	77.96	65.77	65.87	39.69	37.87	37.96	59.90	62.51	62.86
	✓	✓	✓	**72.66**	**60.70**	**60.70**	**32.41**	**30.73**	**30.71**	**51.52**	**56.69**	**56.67**
	w.r.t. BEVHeight			+0.95	+0.78	+0.78	+4.60	+4.30	+4.35	+4.10	+5.50	+5.41

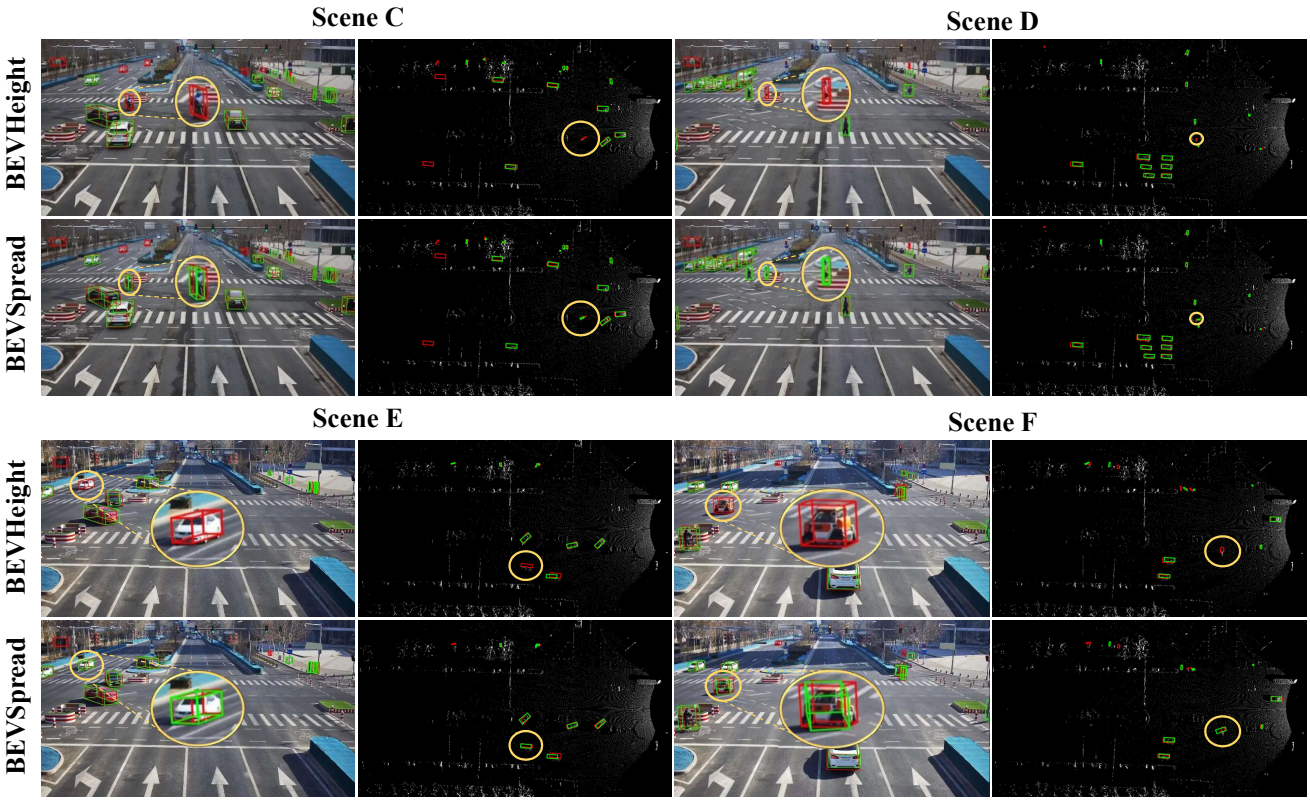

Figure 11. More Visualization results of BEVHeight and proposed BEVSpread in image and BEV view.