CorrBEV: Multi-View 3D Object Detection by Correlation Learning with Multi-modal Prototypes

Supplementary Material

The supplementary material presents additional details and analysis of our model.

- **Patch Size for Random Masking in Pseudo-occlusion Processor**
  We explore the influence of masking operator with different sizes in the proposed pseudo-occlusion processor.
- **Discriminative Correlation Heatmap**
  We analyze the strong discriminative semantic property within the correlation heatmap for perceiving occluded objects.
- **More Visualization**
  We provide more visualization results to prove the effectiveness of our method.
- **Visualizations on Truncated Objects**
  We provide an additional qualitative comparison of truncated objects to prove the robustness of our method.

1.1. Patch Size in Pseudo-occlusion Processor (P2)

The proposed P2 randomly replaces a patch (*e.g.,* 3×3 pixels) in the 2D bounding boxes of non-occluded objects (*i.e.,* visibility level of 3 or 4) with the mean value of the template image to balance the distribution of different visibility levels. Tab. 1 shows the influence of basic patch size for masking. It's observed that a moderate patch size, *i.e.,* 3×3, obtains relatively better average performance. The possible reason is that bigger patch sizes unconsciously lose some important context, while an overly small patch size, *i.e.,* 1×1, doesn't effectively mimic occlusion (more like noise).

Table 1. Ablation on the patch size in pseudo-occlusion processor.

#	Patch Size	mAP↑	NDS↑	Vis_1 † Recall↑	Vis_2 † Recall↑	Vis_3 † Recall↑	Vis_4 † Recall↑
①	1×1	47.2	57.0	68.5	75.9	82.4	86.7
②	3×3	**47.5**	**57.4**	**69.1**	**77.6**	**83.0**	**87.4**
③	5×5	47.0	56.9	68.7	76.5	82.8	87.2
④	7×7	46.4	56.3	67.9	74.7	82.7	87.0

1.2. Discriminative Correlation Heatmap

We analyze the strong discriminative semantic property within the correlation heatmap for perceiving occluded objects. As mentioned in the the manuscript, we observe that perceiving partially occluded objects in 2D correlation heatmaps is easier than that in 3D BEV space (*e.g.,* the missed targets of the baseline in qualitative results). Therefore, we exploit the heatmap to generate target-aware object queries to improve the capability of detecting occluded objects. Here we show the correlation heatmaps in Fig. 1. It is observed that for the partially occluded targets (*e.g.,* pedestrians in the rectangle), the predicted heatmap can generate strong responses, proving our claim. Notably, in the first image, the cars that are far from the ego vehicle (beyond the pre-defined detection range), are also highlighted in our correlation heatmap. This shows the potential of our method for improving the detection quality of far and small objects.

1.3. More Visualizations.

We present more qualitative results to illustrate the superior occlusion-perception capabilities of our method. As shown in Fig. 2, our approach successfully identifies the occluded objects that are ignored by baseline method SparseBEV [3] in different scenarios, *e.g.,* the pedestrian behind a barrier, the car that is heavily covered by nearby vehicles.

1.4. Visualizations on Truncated Objects.

In KITTI 3D object detection benchmark [2], truncation label is provided, which refers to the object leaving image boundaries. Inspired by this, we provide qualitative comparison of truncated objects, which can be considered a form of "occlusion" caused by the limited camera range. The results in Fig. 3 illustrate the robustness of our method in handling such truncated cases.

1.5. Class-wise Performance Improvement.

We provide per-class improvement compared to the baseline SparseBEV [3], and the results are presented in Fig. 4 (left) of the attached PDF. We observe that the relative improvement ratio for categories with a higher number of samples (e.g. car and pedestrian) tends to saturate. This could be attributed to the abundant training data available for these categories, making them relatively easier to learn. Surprisingly, we are pleased to note a significant improvement in categories with fewer samples (e.g. truck, bicycle and motorcycle). This result indicates that our method is effective in enhancing performance, particularly in categories where data is limited, which is often a challenging scenario.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the*

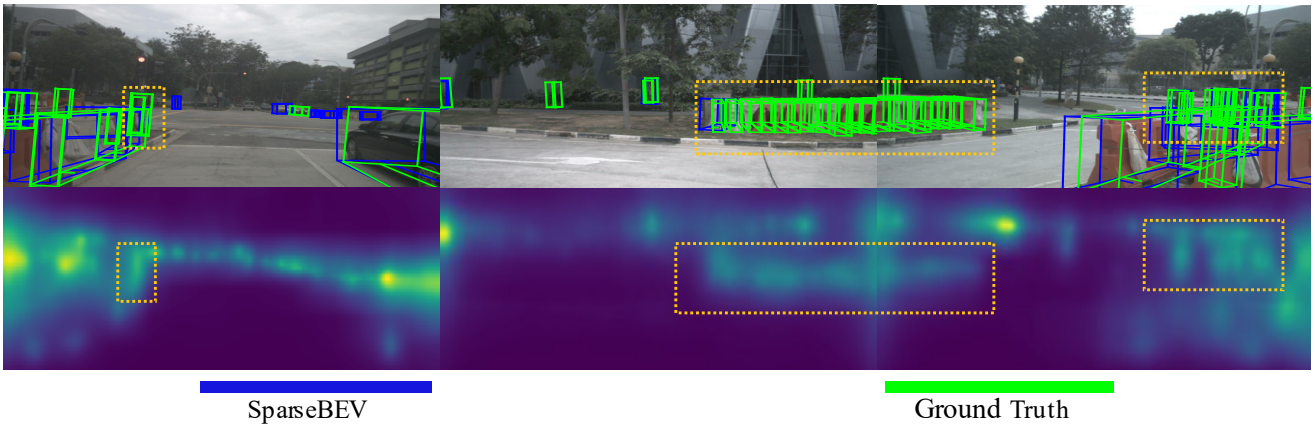

Figure 1. Visualization of the correlation heatmap. The introduced multi-modal prototypes help to enhance the responses of occluded targets.

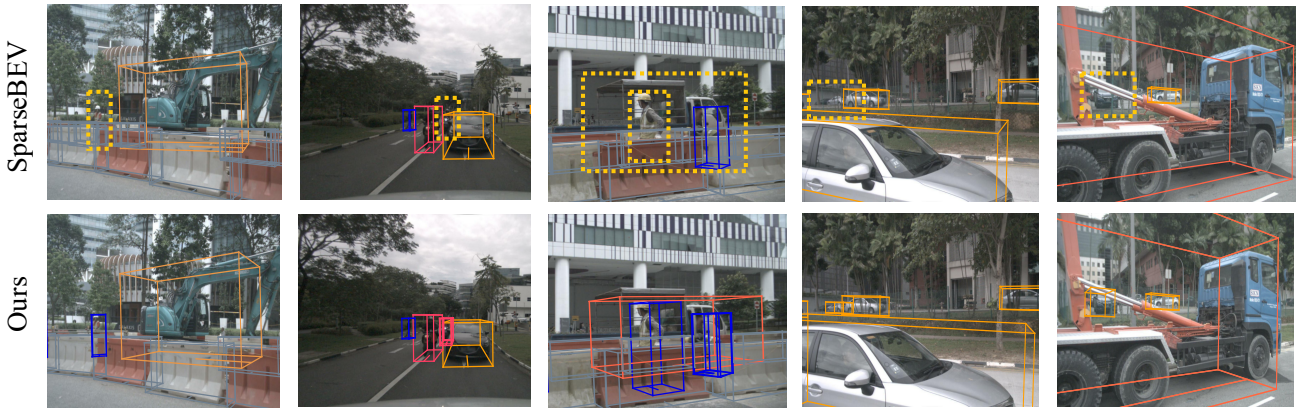

Figure 2. More qualitative comparison on occlusion perception. Our method shows superior capabilities of detecting occluded objects compared with the baseline (see dotted yellow rectangles).

Figure 3. The qualitative truncation comparison between SparseBEV and our CorrBEV$_{sp}$.

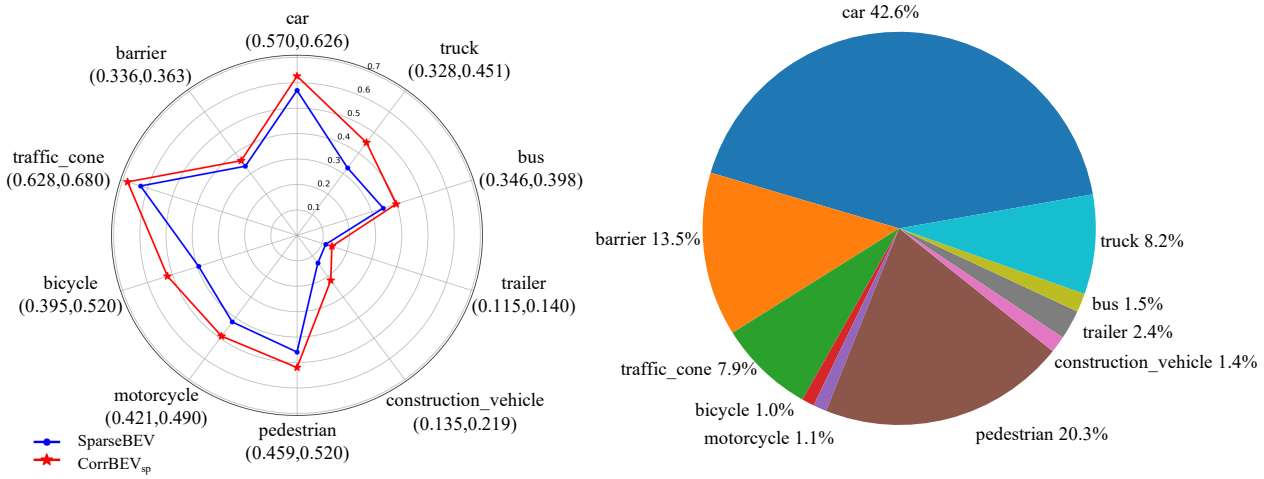

Figure 4. The per-class result comparison between our CorrBEV$_{sp}$ and SparseBEV (left) and per-class number in nuScenes [1] Train split(right).

IEEE/CVF conference on computer vision and pattern recognition, 2020. 3

[2] Andreas Geiger, Philip Lenz, and Raquel 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

[3] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1