Multi-View Attentive Contextualization for Multi-View 3D Object Detection

Supplementary Material

Overview

In this supplementary material, we provide more details on the following aspects that are not presented in the main paper due to space limit:

- *Computation and memory cost* are provided in Sec. 1.
- *Supplementary qualitative results on NuScenes validation split* are provided in Sec. 2.

1. Computation and Memory Cost

Method	Speed (FPS)	GPU Mem (MB)	#Param (M)	NDS↑
PETR-VovNet-99 [29]	9.8	3638	83.07	42.6
PETR-VovNet-99-MvACon	9.6	3638	84.75	**43.4 (+0.8)**
BEVFormer-b [25]	3.9	6928	69.14	51.7
BEVFormer-b-MvACon-lite	3.2	6936	70.75	**52.5 (+0.8)**
BEVFormer-b-MvACon	3.0	11452	70.75	**52.8 (+1.1)**

Table 7. Efficiency and resource consumption of MvACon on PETR and BEVFormer. MvACon-lite refers to the model without using the concatenation of cluster contexts from all feature pyramids. This will greatly reduce extra GPU memory consumption, with only 0.3 NDS droppped compared with our full model.

Computation and memory cost of our MvACon is provided in Tab. 7. We use the parameter calculation script provided by BEVFormer's open source codebase: https://github.com/fundamentalvision/BEVFormer.

Our MvACon is able to improve PETR with negligible computation cost. We tested two versions of BEVFormer-b-MvACon: a lite version and a full model. In the lite version, we enforce cluster attention within each feature pyramid level instead of using clusters from all levels. This will largely reduce the computation cost. It shows that our lite version is able to improve the baseline with only 8 MB extra GPU memory cost with 0.8 NDS improvement. Using our full model, we are able to improve the baseline with 1.1 NDS improvement. These results clearly demonstrates the effectiveness and necessity of incorporating useful contexts before feature lifting.

2. More Qualitative Results on NuScenes

Qualitative results for deformable points across consecutive frames. We visualize the deformable points across 3 consecutive frames in Fig. 6. We observe that our MvACon is able to learn stable and meaningful high-response deformable points on both cars and surrounding buildings.

Supplementary qualitative results for deformable points in different scenes. We visualize the deformable points in different scenes in Fig. 7 and Fig. 8. We observe that our MvACon is able to learn meaningful high-response deformable points on cars and surrounding references, which could be helpful in improving the prediction of object location, orientation and velocity.

Supplementary qualitative comparison for detection results on NuScenes validation set. We visualize prediction results on NuScenes validation set and compare it with BEVFormer in Fig. 9, Fig. 10 and Fig. 11. We observe that our MvACon performs better in dense scenes.

Supplementary visualization for learned cluster heatmap We provide the detailed visualization results of learned cluster contexts with raw images in Fig. 12. This uses the same scene shown in Fig. 3 of our main paper. The only difference is that we include raw images in supplementary materials. We observe that the learned cluster heatmap has high response on foreground contexts.

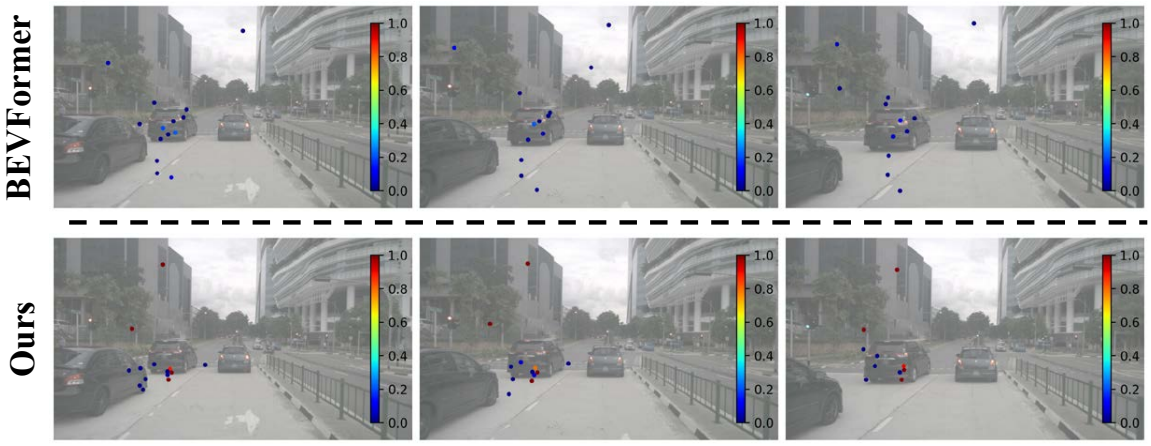

Figure 6. Visualization results of the deformable points originating from a 2D reference point across 3 consecutive frames on NuScenes validation set. This 2D reference point is projected from a 3D BEV (Bird's Eye View) anchor point in the BEVFormer encoder. We use the same BEV anchor point as the one presented in our main paper. From left to right, we exhibit the deformable points outputted from the encoder's final layer, arranged in chronological order (t-1, t, t+1).

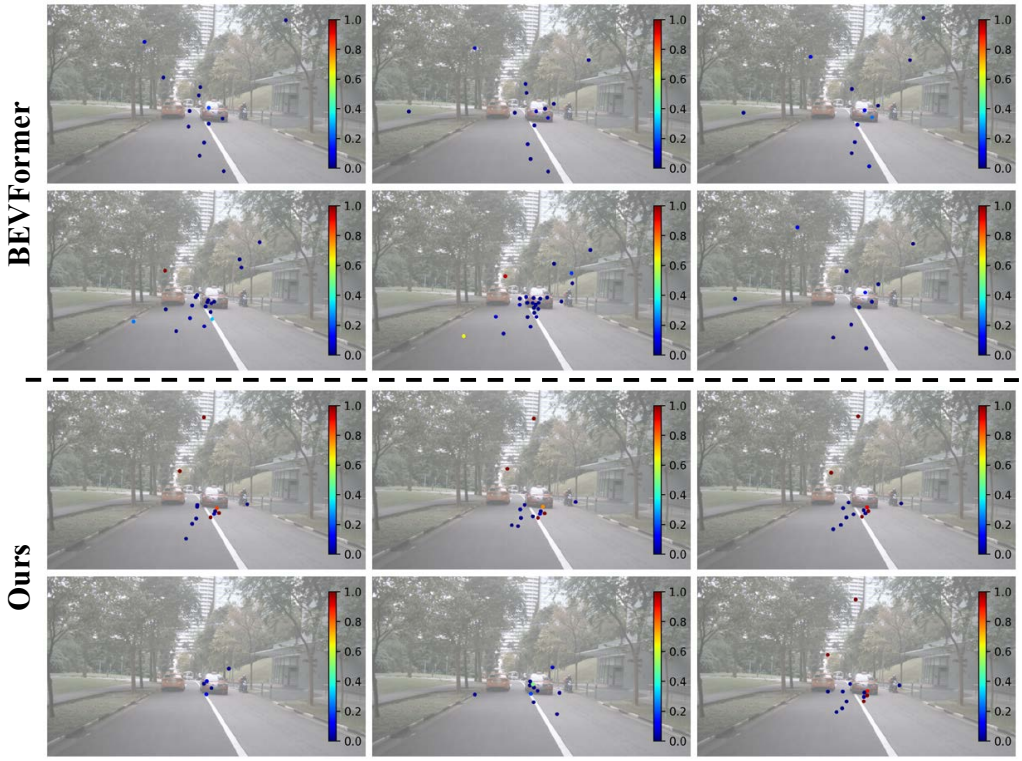

Figure 7. Visualization results of the deformable points originating from a 2D reference point, which is projected from a 3D BEV anchor point in the BEVFormer encoder, on NuScenes validation set. We utilize the a BEV anchor point one the right car. From left to right and up to bottom, we display the deformable points output from each layer (#1-#6) in the encoder, respectively.

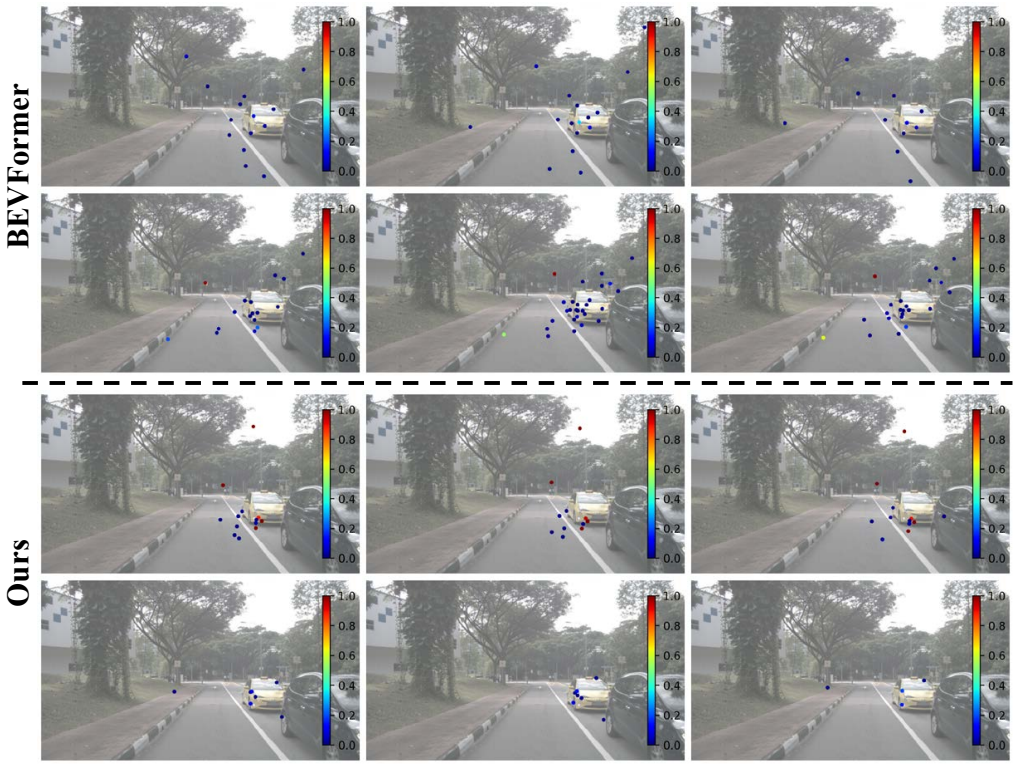

Figure 8. Visualization results of the deformable points originating from a 2D reference point, which is projected from a 3D BEV anchor point in the BEVFormer encoder, on NuScenes validation set. We utilize the a BEV anchor point one the yellow car. From left to right and up to bottom, we display the deformable points output from each layer (#1-#6) in the encoder, respectively.

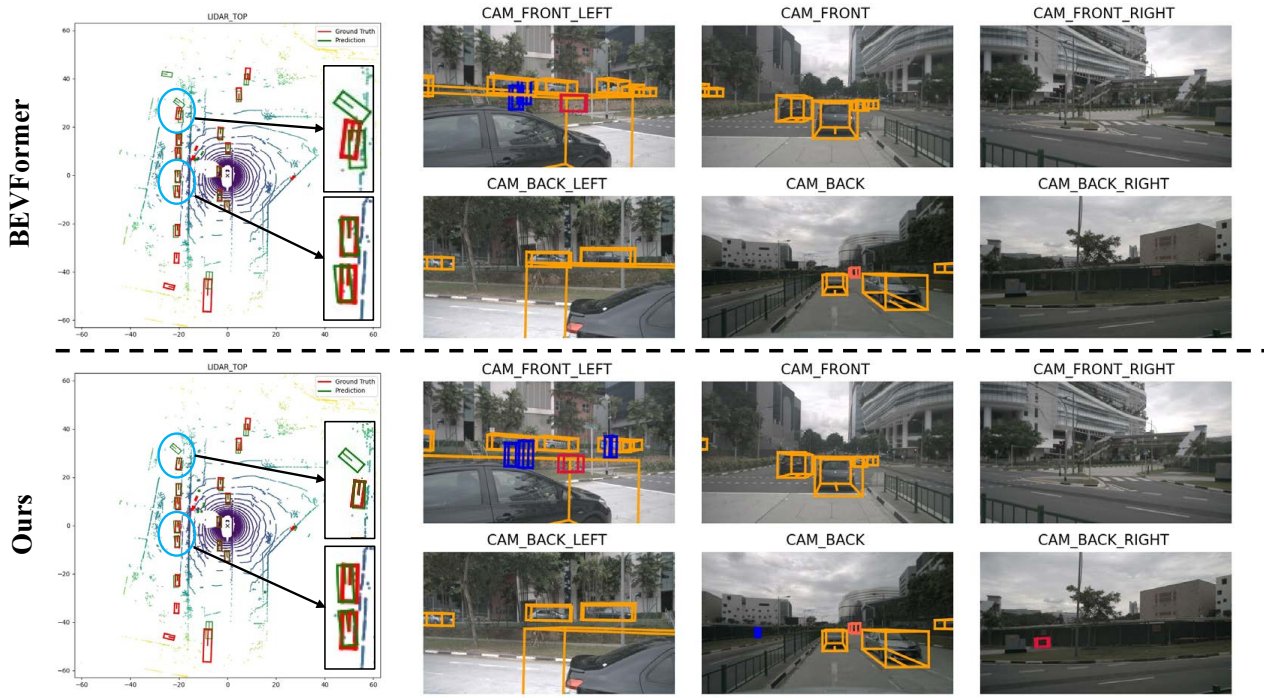

Figure 9. Qualitative comparisons between BEVFormer and our MvACon method on NuScenes validation set.

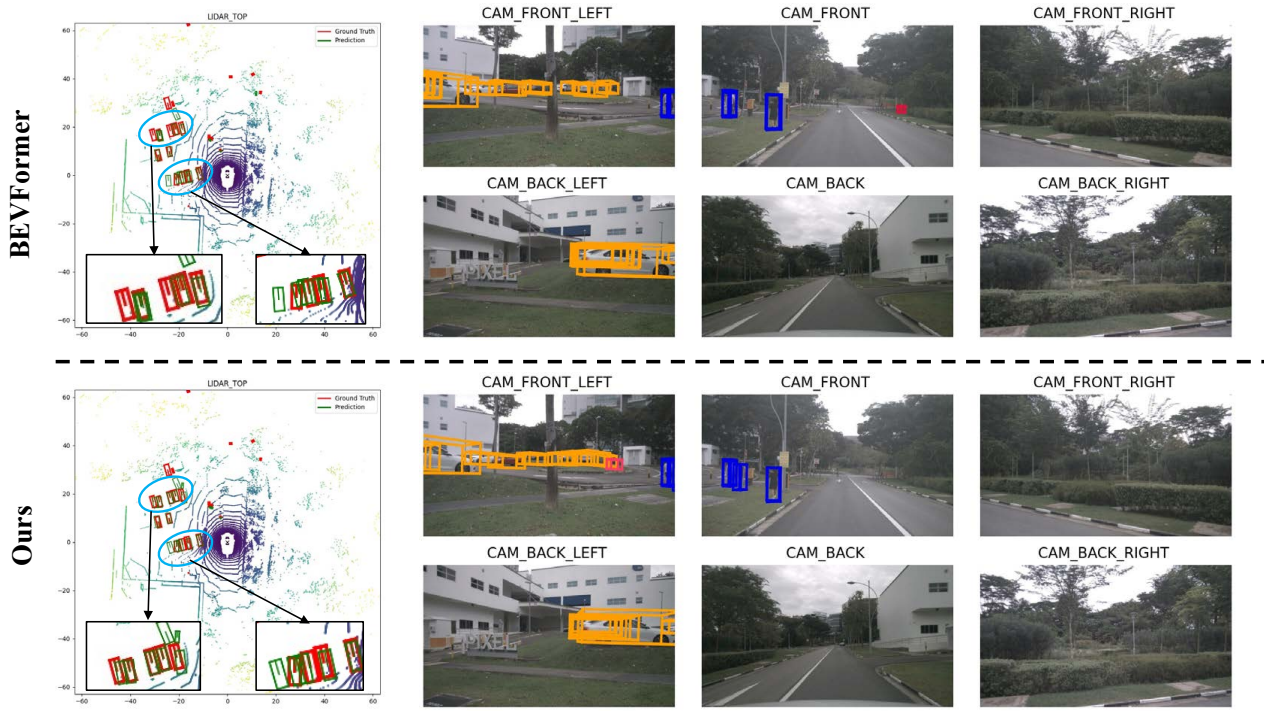

Figure 10. Qualitative comparisons between BEVFormer and our MvACon method on NuScenes validation set.

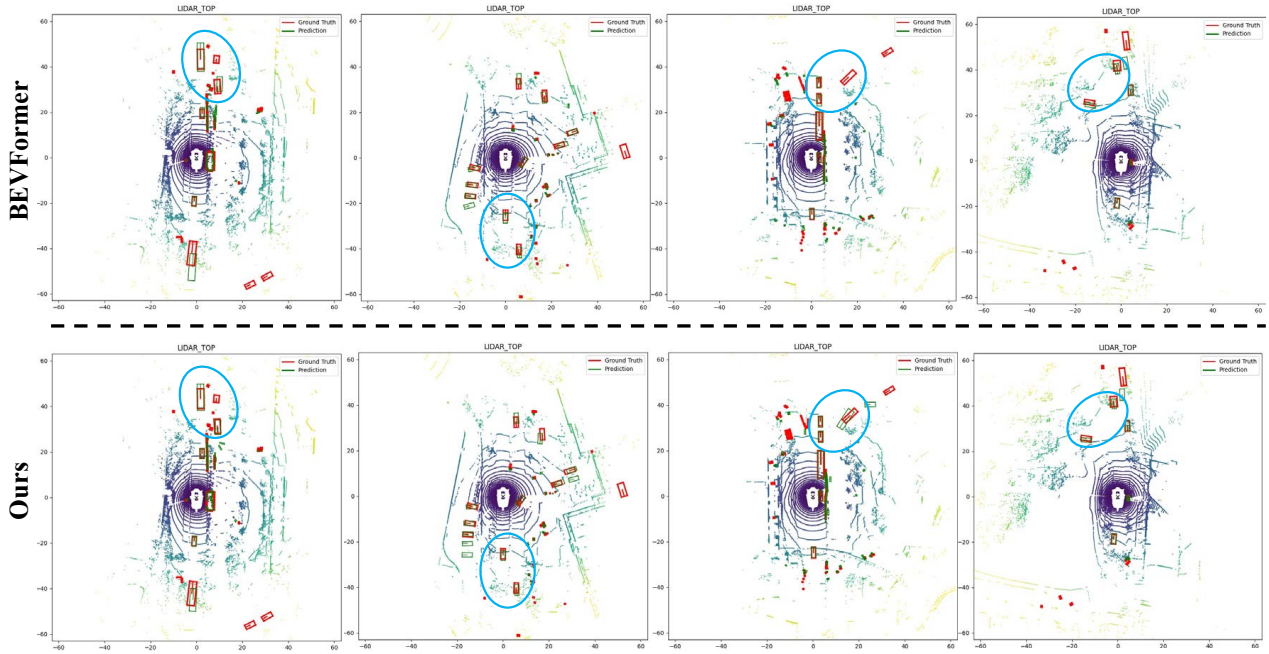

Figure 11. Qualitative comparisons between BEVFormer and our MvACon method on NuScenes validation set.

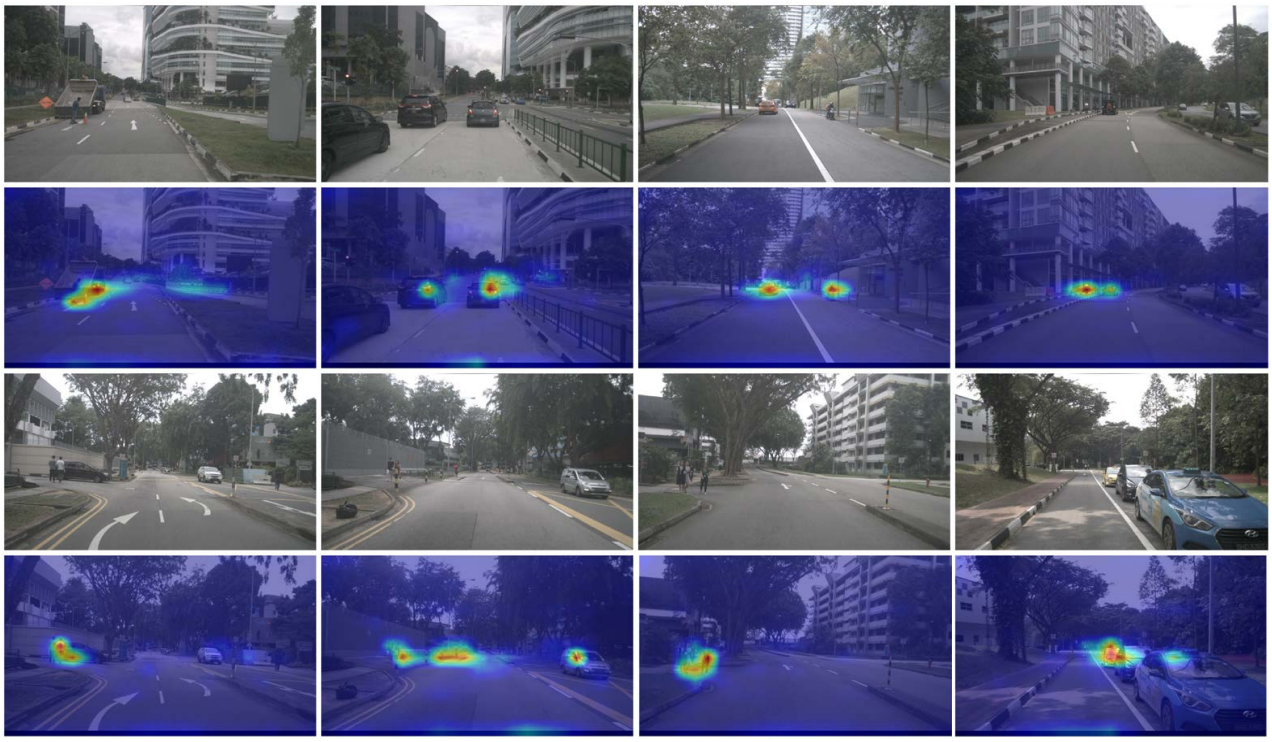

Figure 12. Visualization results of learned cluster contexts with raw images in our proposed attentive contextualization module on NuScenes validation set. We sum all the learned clusters along the channel and upsample it to the original image resolution through bilinear interpolation. We observed that the learned cluster context encodes abundant context information in the scene.