ObjectFusion: Multi-modal 3D Object Detection with Object-Centric Fusion — ICCV 2023 Supplementary Material

Qi Cai, Yingwei Pan, Ting Yao, Chong-Wah Ngo and Tao Mei
University of Science and Technology of China
HiDream.ai Inc
Singapore Management University
{cqcaiqi, panyw.ustc, tingyao.ustc}@gmail.com, cwngo@smu.edu.sg, tmei@hidream.ai

This supplementary material contains 1) more ablation studies on voxel sizes, image sizes and data augmentations. 2) an analysis of robustness to corrupted images. 3) qualitative comparisons with other methods.

1. More Ablations on Design Choices

Here we conduct more ablation studies to study the effect of voxel size, which corresponds to the resolution of LiDAR branch. A smaller voxel size results in higher-resolution inputs of LiDAR branch, while increasing the number of voxels and computational cost. As shown in Table 9, the use of a smaller voxel size (0.0075m) leads to a 0.9% improvement in mAP compared to voxel size of 0.1m. Moreover, we also perform ablation study on image size, which affects the resolution of the camera branch. As shown in Table 10, the use of a larger image size (384 * 1056) leads to 0.4% improvement in mAP compared to the image size of 256 * 704. In an effort to seek a better balance between performance and computational cost, we select 0.075m for voxel size and 256 * 704 for image size.

In addition, we perform ablation experiments on different data augmentation strategies: L and C denote the independent augmentation on LiDAR and camera branch respectively, L + C represents joint augmentation on both branches. Table 11 shows that the use of joint augmentation (L+C) leads to a 0.8% boost in mAP compared to only LiDAR augmentation (L). Furthermore, augmentation in the LiDAR exhibits better performance than that in the camera since the LiDAR conveys richer geometry information.

2. Robustness to Corrupted Images

Our ObjectFusion inevitably relies on projected features from the camera branch for multi-modal fusion, which might be vulnerable to corruption in images. To assess the robustness of ObjectFusion to corrupted images, we conduct an additional experiment by randomly dropping one view image during inference. This results in a slight decrease in mAP score from 69.8% to 69.6%, demonstrating the robustness of ObjectFusion with corrupted images.

Table 9: Performance comparisons on nuScenes validation set with different voxel sizes.

Voxel Size (m)	0.0075	0.1	0.125
mAP (%)	69.8	68.9	67.3

Table 10: Performance comparisons on nuScenes validation set with different image sizes.

Image Size	128 * 352	256 * 704	384 * 1056
mAP (%)	67.2	69.8	70.2

Table 11: Performance comparisons on nuScenes validation set with different augmentation strategies.

Augmentation	L	C	L + C
mAP (%)	69.0	67.5	69.8

3. Qualitative Comparisons

Figure 4 further illustrates more examples of the 3D object detection results achieved by BEVFusion and our ObjectFusion. Due to the incorporation of enhanced object-level features from multiple modalities, ObjectFusion demonstrates improved recall of targeted objects.

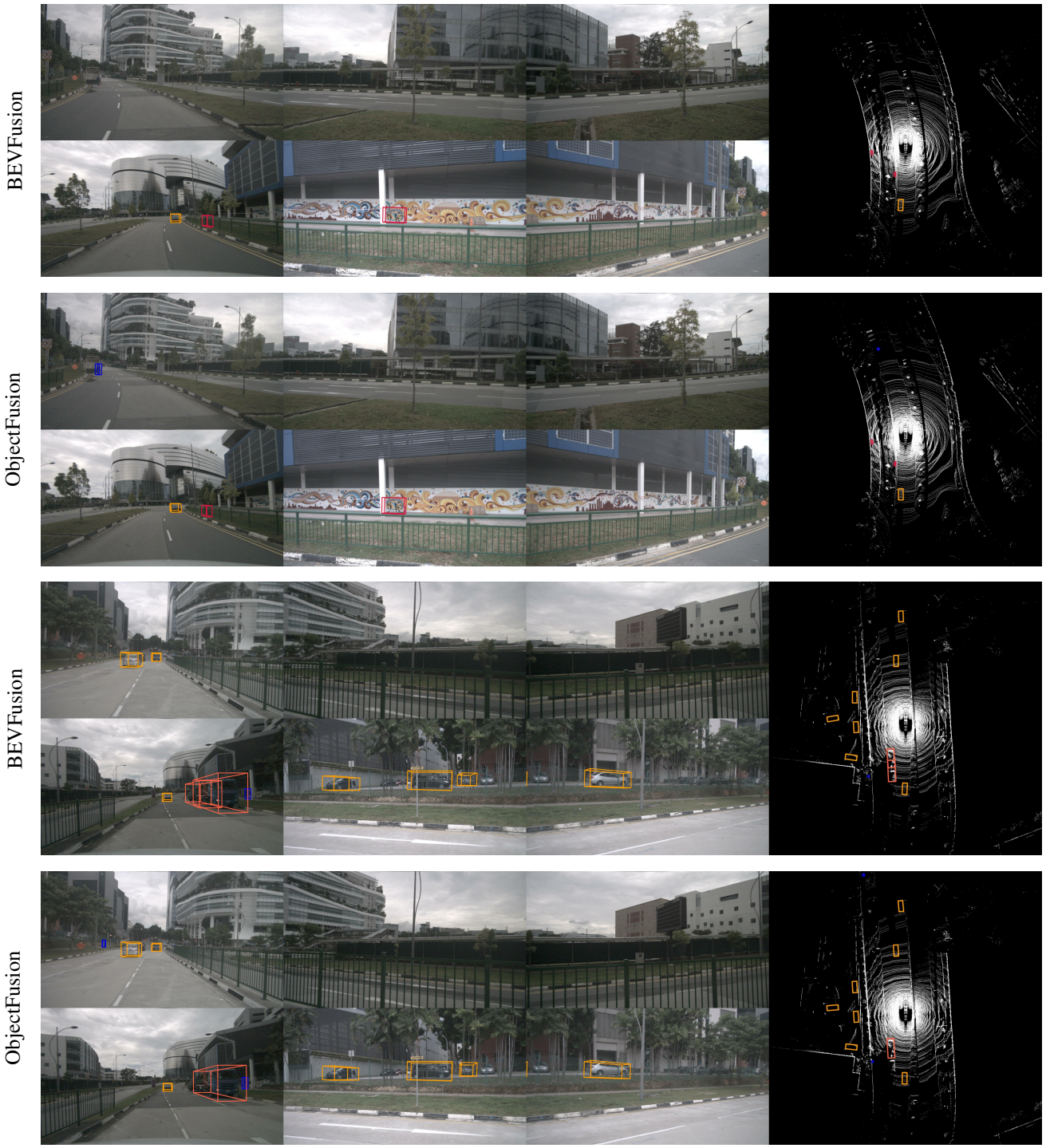

Figure 4: Qualitative results of BEVFusion and ObjectFusion on nuScenes validation set. The first row displays images from the CAM_FRONT, CAM_FRONT_RIGHT, and CAM_BACK_RIGHT cameras. The second row displays images from the CAM_BACK, CAM_BACK_LEFT, and CAM_FRONT_LEFT cameras. The final image presents the LiDAR view projected on a bird's-eye view (BEV). The bounding box colors are defined in the following way: car=, truck=, construction_vehicle=, bus=, trailer=, barrier=, motorcycle=, bicycle=, pedestrian=, traffic_cone=. Best viewed with color and zoom-in. (Part 1/3)

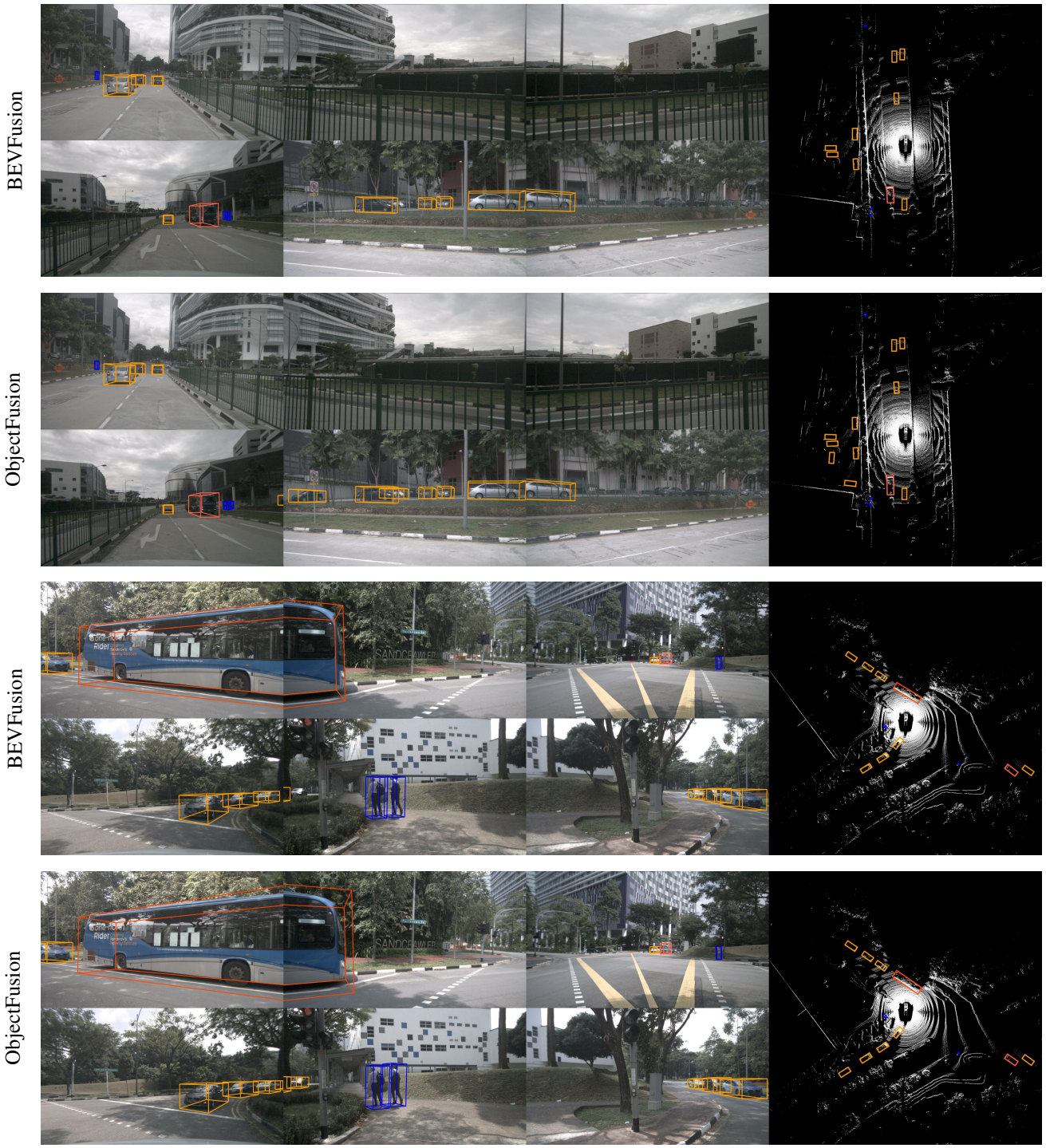

Figure 4: Qualitative results of BEVFusion and ObjectFusion on nuScenes validation set. The first row displays images from the CAM_FRONT, CAM_FRONT_RIGHT, and CAM_BACK_RIGHT cameras. The second row displays images from the CAM_BACK, CAM_BACK_LEFT, and CAM_FRONT_LEFT cameras. The final image presents the LiDAR view projected on a bird's-eye view (BEV). The bounding box colors are defined in the following way: car=, truck=, construction_vehicle=, bus=, trailer=, barrier=, motorcycle=, bicycle=, pedestrian=, traffic_cone=. Best viewed with color and zoom-in. (Part 2/3)

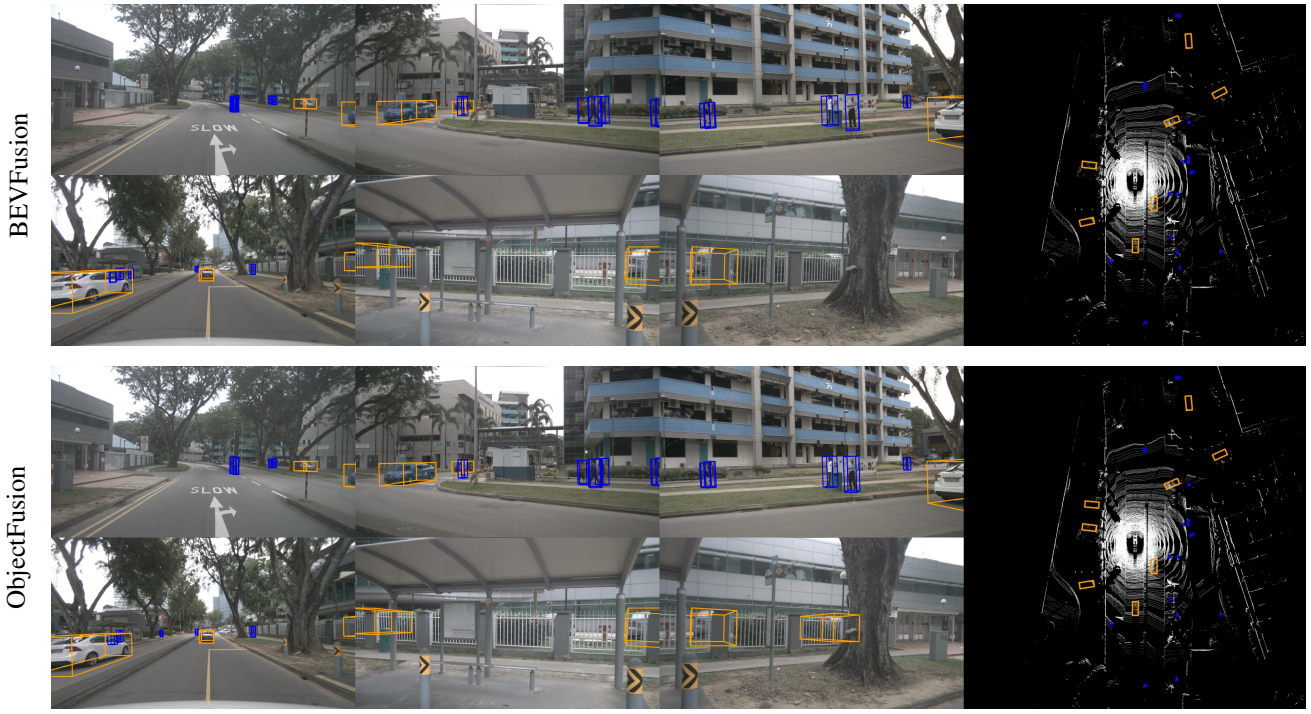

Figure 4: Qualitative results of BEVFusion and ObjectFusion on nuScenes validation set. The first row displays images from the CAM_FRONT, CAM_FRONT_RIGHT, and CAM_BACK_RIGHT cameras. The second row displays images from the CAM_BACK, CAM_BACK_LEFT, and CAM_FRONT_LEFT cameras. The final image presents the LiDAR view projected on a bird's-eye view (BEV). The bounding box colors are defined in the following way: car=, truck=, construction_vehicle=, bus=, trailer=, barrier=, motorcycle=, bicycle=, pedestrian=, traffic_cone=. Best viewed with color and zoom-in. (Part 3/3)