Fast-CLOCs: Fast Camera-LiDAR Object Candidates Fusion for 3D Object Detection – Supplementary Materials

1. Qualitative Results

Fig 1 shows some qualitative results of our proposed fusion method on KITTI test set. Red bounding boxes represent false positive detections from PV-RCNN [4] that are deleted by our Fast-CLOCs, green bounding boxes are true positive detections that are confirmed by Fast-CLOCs.

Incorporating image visual information can help remove LiDAR false positive detections and confirm LiDAR true positive detections, as shown in Fig 1. The image projection region of a false positive detection in the LiDAR point cloud, will be classified and usually rejected as a detection through our 3D-Q-2D detector. Then Fast-CLOCs can leverage this inconsistency to remove the false positive. Objects that are double confirmed by LiDAR detector and 3D-Q-2D image detector will be kept by Fast-CLOCs.

2. Failure Cases Analysis

There are mainly two types of failure cases for Fast-CLOCs. One is mistakenly removing the true positives, the other is fail to delete false positives. These failure cases happen in the scenarios in which objects are heavily occluded, at long distance or in poor lighting conditions. Fig 2 shows some failure examples. In fail case#1 and case#2, LiDAR-only detector detects the true positive cars. But the true positives are suppressed by our 3D-Q-2D detector due to high level of occlusion (case#1) and poor lighting condition (case#2) in the image plane. So Fast-CLOCs fusion removes these true positives. In fail case#3 and case#4, LiDAR-only detector detects the cars but with wrong poses, so they are false positive detections. But these false positive detections are not rejected by our 3D-Q-2D detector. Because parts of the cars are visible in the image plane, the 3D-Q-2D confirms them as cars but fails to provide the full sizes of the cars in the image plane. Fast-CLOCs fusion therefore keeps these false positive.

Detecting objects under occlusion is an open problem in computer vision community. Designing better detection networks and collecting more training data with these corner cases could be one direction to resolve this issue. But it would require more computing resources. We believe adding temporal information from multiple frames would be a simpler direction.

3. Ablation Study on Minor Modifications of the Fusion Network

We made two minor modifications on the CLOCs [3] fusion network. One is adding another $flag$ channel to highlight whether a 3D detection overlaps with at least one 2D detection. The motivation is that we want to keep the 3D detection candidate that has no 2D detections overlap with it. Adding this channel also helps the network distinguish this case from other examples with very small IoU and 2D confidence score. The other modification is we apply Residual blocks [2] instead of standard 1×1 convolutional layers to slightly improve the performance. The ablation studies regarding these changes are present in Table 1. The first row in Table 1 without Flag channel and residual blocks represent the performance for original CLOCs. As shown in the Table 1, adding flag channel and using residual blocks slightly improves the detection performance.

Flag channel	Residual blocks	3D AP (%)		
		easy	moderate	hard
		92.52	85.55	82.64
✓		92.93	85.77	82.97
	✓	92.83	85.82	82.98
✓	✓	**93.18**	**86.01**	**83.07**

Table 1: Ablation studies of modifications in the fusion network on KITTI validation set. PV-RCNN is applied as the 3D detector in this experiments. We add a flag channel to highlight whether a 3D detection overlaps with at least one 2D detection. Residual blocks instead of standard 1×1 convolutional layers to slightly improve the performance. The first row represents the performance for original CLOCs fusion network. As shown in the table, adding flag channel and using residual blocks slightly improve the detection performance.

References

[1] 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.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Su Pang, Daniel Morris, and Hayder Radha. Clocs: Camera-lidar object candidates fusion for 3d object detection. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[4] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-

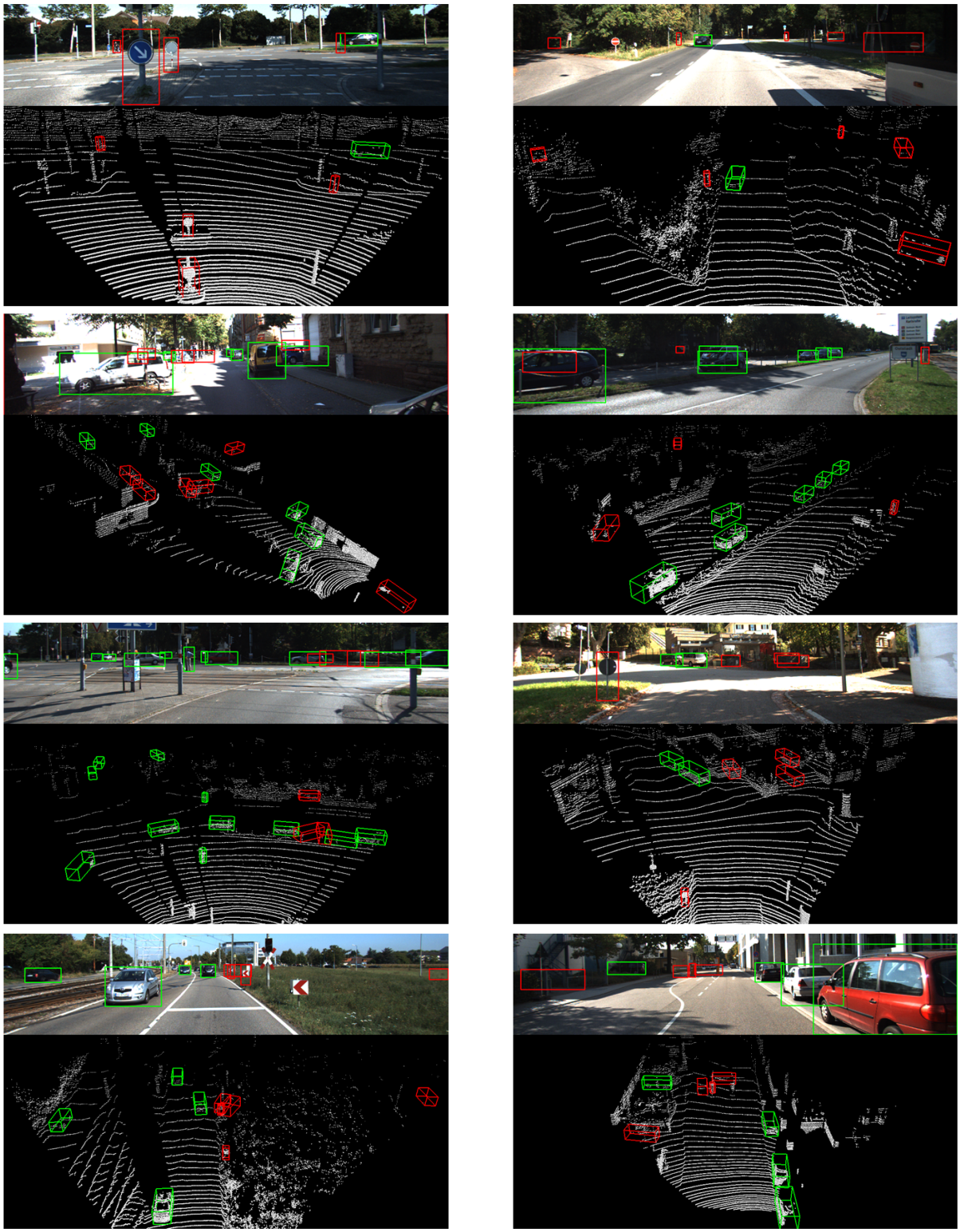
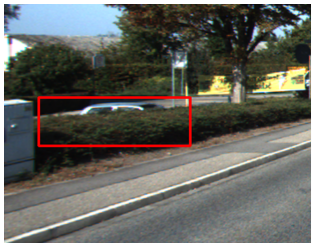
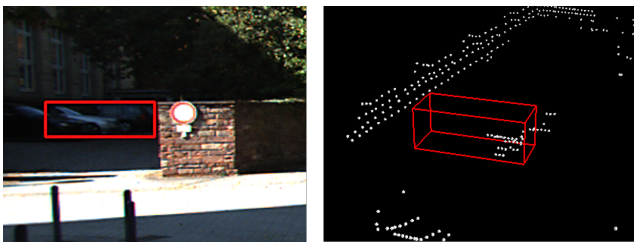


Figure 1: Qualitative results of our Fast-CLOCs on KITTI [1] test set compared to PV-RCNN [4]. Red bounding boxes are false positive detections from PV-RCNN that are removed by our Fast-CLOCs. Green bounding boxes are confirmed true positive detections. The upper row in each image is the 3D detection projected to image, the others are 3D detections in LiDAR point clouds.

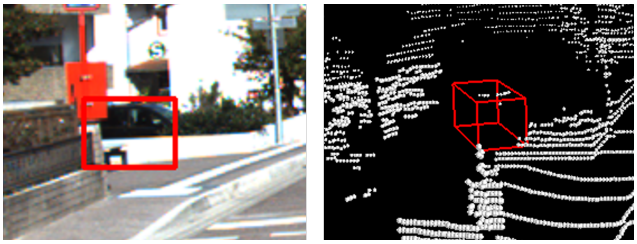


(a) Fail Case#1: Fast-CLOCs removes a true positive from LiDAR 3D detector.



(b) Fail Case#2: Fast-CLOCs removes a true positive from LiDAR 3D detector.

(c) Fail Case#3: Fast-CLOCs fails to remove a false positive from LiDAR 3D detector.



(d) Fail Case#4: Fast-CLOCs fails to remove a false positive from LiDAR 3D detector.

Figure 2: Some failure cases. (a) and (b): LiDAR-only detector detects the true positives (car). But the cars are suppressed by 3D-Q-2D detector due to high level of occlusion (case#1) and poor lighting conditions (case#2) in image plane. Therefore, Fast-CLOCs fusion removes these true positives. (c) and (d): LiDAR-only detector detects the cars but with wrong poses, so they are false positives. But these false positives are not rejected by 3D-Q-2D detector because parts of the cars are visible in the image plane. Therefore, Fast-CLOCs fusion keeps these false positives.

voxel feature set abstraction for 3d object detection. In *CVPR*, 2020.