TCC-Det: Temporarily consistent cues for weakly-supervised 3D detection Supplementary material

Jan Skvrna and Lukas Neumann

Visual Recognition Group, Department of Cybernetics
FEE, Czech Technical University, Prague, Czech Republic

1 Waymo Open Dataset implementation

Due to the differences between Waymo Open Perception Dataset [4] (WOD) and KITTI [2], we had to modify hyper-parameters in the following way. The Voxel-RCNN [1] was pre-trained on the pseudo-ground truth for 20 epochs with batch size equal to 48 on 4x NVIDIA A100. It took approximately 20 hours to pretrain the network, while creating the pseudo ground truth labels took approximately 60 hours. The full training of the network with our loss function was for 1 epoch with batch size 8 on 4x NVIDIA A100. It took approximately 100 hours to train the network.

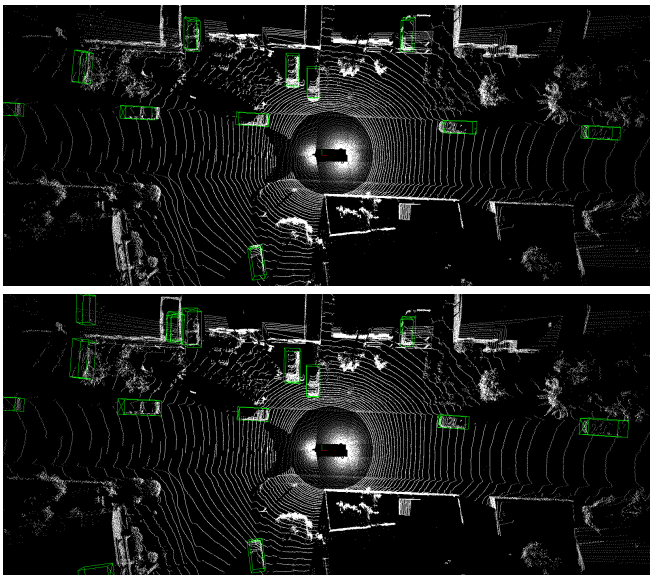

Fig. 1: Qualitative examples of 3D object detections on the Waymo Open Perception dataset [4]. The upper figure shows detections of weakly-supervised Voxel-RCNN [1], while the bottom figure shows detections of the fully-supervised one.

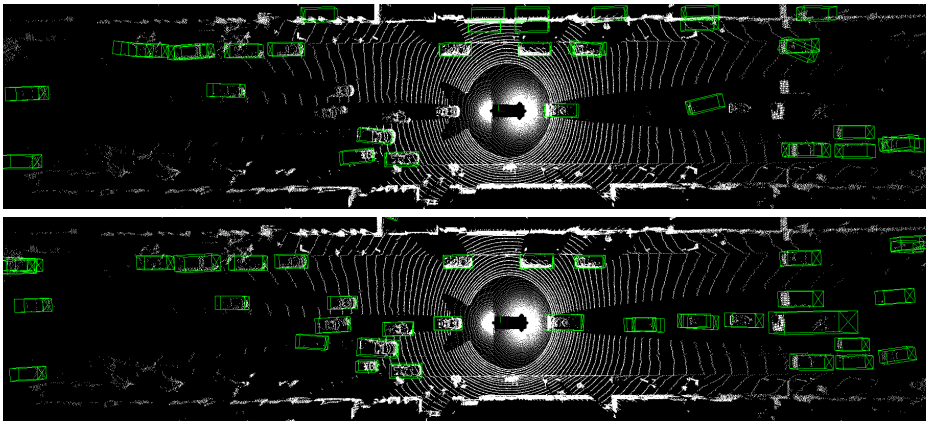

Fig. 2: Typical failure modes of our method on the Waymo Open Perception dataset [4]. The upper figure shows detections of weakly-supervised Voxel-RCNN [1], while the bottom figure shows detections of the fully-supervised one.

Further, we show qualitative analysis in Fig. 1 and Fig. 2. The difference between fully and weakly supervised detectors is mostly the inability to detect cars moving behind the ego-vehicle and fine refinement of the 3D bounding boxes.

2 Ablations

To further support the design choices we made, we provide three additional ablation studies regarding the hyper-parameters of our method. All ablations are evaluated on the KITTI dataset.

Table 1: The effect of the number of frames used in aggregating the point clouds

Frames	BEV			3D		
	Easy @0.7	Moderate @0.7	Hard @0.7	Easy @0.7	Moderate @0.7	Hard @0.7
± 0	88.83	84.7	78.11	53.70	45.17	41.26
± 10	89.70	87.71	86.75	83.77	68.05	67.13
± 30	90.09	88.25	86.95	85.92	75.33	73.74

Number of frames. In Table 1 we demonstrate the accuracy of our method depending on how many frames of LiDAR point clouds are aggregated and therefore considered in the temporal consistency. Using only 10 frames before and after the reference frame, which translates into 1 second before and after, shows great performance improvement, proving that exploiting temporary consistent frames

is key to achieving good performance. As expected, using longer periods of time to aggregate LiDAR points and track cars provides an additional performance boost, especially in 3D. It is worth noting that when the frame count is 0, we cannot perform tracking of moving cars, so the motion model is not applied.

Iterative Closest Point. As mentioned in the main text, data from the IMU of the ego-vehicle does not provide perfect alignment of the frames. In Table 2, we demonstrate the effect by removing the ICP step [3] and show that ICP is indeed necessary to achieve good alignment in aggregating LiDAR point clouds.

Table 2: The effect of Iterative Closest Point for better frame alignment

ICP	BEV			3D		
	Easy @0.7	Moderate @0.7	Hard @0.7	Easy @0.7	Moderate @0.7	Hard @0.7
✗	89.80	87.56	86.34	74.78	70.04	64.61
✓	90.09	88.25	86.95	85.92	75.33	73.74

Template Fitting Loss steepness. Template Fitting Loss relies on a single parameter – k, which defines sigmoid steepness. As the steepness increases, the smaller distances between two points achieve saturation, and the gradient becomes negligible. In other words, as the steepness increases, the distance threshold, which decides if the correspondence of two points is worth optimizing, decreases. Table 3 shows that steepness equal to 10 achieves the best accuracy.

Table 3: The effect of the steepness parameter value in the Template Fitting Loss

k	BEV			3D		
	Easy @0.5	Moderate @0.5	Hard @0.5	Easy @0.5	Moderate @0.5	Hard @0.5
5	78.78	80.23	72.76	75.81	70.40	69.45
10	78.98	80.33	72.91	75.93	70.29	69.44
15	77.93	73.09	72.16	75.00	69.47	62.12
25	75.88	71.57	70.48	72.84	67.65	60.49

References

1. Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., Li, H.: Voxel r-cnn: Towards high performance voxel-based 3d object detection. vol. 35, pp. 1201–1209 (May 2021). https://doi.org/10.1609/aaai.v35i2.16207, https://ojs.aaai.org/index.php/AAAI/article/view/16207
2. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)

3. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings Third International Conference on 3-D Digital Imaging and Modeling. pp. 145–152 (2001). https://doi.org/10.1109/IM.2001.924423
4. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)