Supplemental Material for OccAM's Laser: Occlusion-based Attribution Maps for 3D Object Detectors on LiDAR Data

David Schinagl[1,2] Georg Krispel[1] Horst Possegger[1] Peter M. Roth[3,4] Horst Bischof[1,2]
{david.schinagl,georg.krispel,possegger,bischof}@icg.tugraz.at, peter.roth@tum.de
[1] Graz University of Technology [2] Christian Doppler Laboratory for Embedded Machine Learning
[3] Technical University of Munich [4] University of Veterinary Medicine, Vienna

This supplementary discusses additional insights and quantitative results, as well as potential limitations of our occlusion-based attribution maps.

1. Detector Comparison

Due to the page limit, we omitted the average attribution maps for pedestrians and cyclists, as well as Part-A^2 [5] (in favor of the also hybrid approach PV-RCNN [4]) and Voxel R-CNN [1] (in favor of the also voxel-based SECOND [7]) from the main manuscript (Section 4.6). Figure 2 shows these maps for all classes and all detectors.

Considering pedestrians, we see that independent of the model's architecture the head-shoulder silhouette is the most important feature for detection. In the case of cyclists, the voxel-based and hybrid methods show a stronger focus on the head than the pillar-based method. Furthermore, we can see that the two hybrid methods PV-RCNN [4] and Part-A^2 [5] lead to very similar average attribution maps.

2. Number of Iterations

Figure 3a shows how the number of iterations N influences the quality of individual (*i.e.* per-detection) attribution maps. While the attribution maps are very noisy for $N = 100$, the importance of individual regions can already be seen at $N = 300$ (which takes on average only 6 seconds for PointPillars [3]). Further iterations allow us to refine the details of the attribution maps notably. We found that the quality (regardless of the detector) saturates around $N = 3000$ and thus, use this setting for all experiments (as stated in the main manuscript, Section 4.1).

Considering the average attribution maps, shown in Figure 3b, we notice that the most important regions are already recognizable very early on. This is due to the fact that averaging the individual attribution maps is analogous to increasing N as above. Nevertheless, additional iterations further improve the precision of these maps.

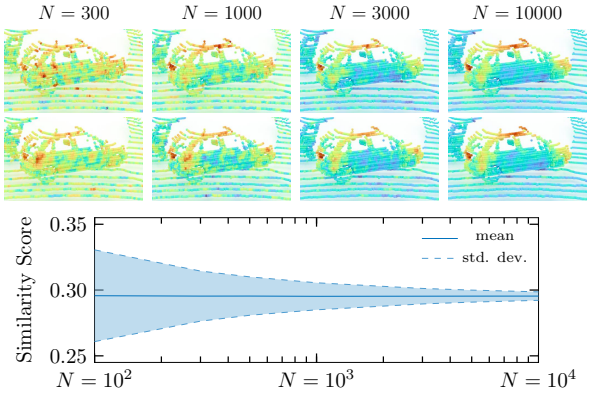


Figure 1. Convergence of the similarity score per point (for 500 detections over 20 runs with different random seeds). The visuals on top show the attribution map progression for a *car* over 2 runs.

3. Convergence

To demonstrate that our attribution maps converge towards the same result, we show in Fig. 1 (bottom) the mean similarity score and std. dev. for 500 detections over 20 runs with different random seeds (computed for all points within detections). In addition we provide an example of the attribution map progression for a car over 2 differently initialized runs.

4. Individual Similarity Sub-Metrics

Our similarity metric allows the generation of attribution maps for individual sub-metrics (*cf.* main manuscript Section 3.3). Figure 4 shows the average attribution maps w.r.t. orientation, translation, scale and confidence score compared to the overall similarity score. These show that (especially for cars) different structures are of varying importance for the detector's decision. For example, estimating the orientation focuses clearly on the roof, whereas for proper scaling the A-pillars are of higher importance.

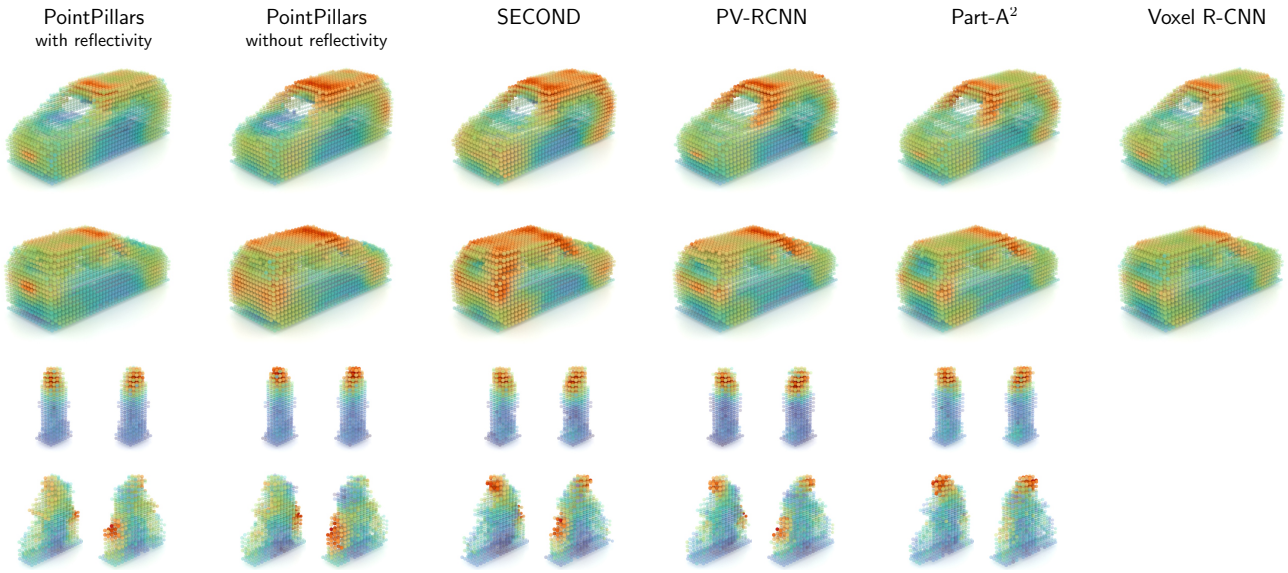


Figure 2. Comparison of average attribution maps for PointPillars [3] (trained w/o reflectivity), SECOND [7], PV-RCNN [4], Part-A² [5] and Voxel R-CNN [1] trained & evaluated on KITTI [2]. Note: Voxel R-CNN (from OpenPCDet [6] model zoo) is only available for cars.

5. Pointing Game

Another metric to assess saliency maps of image-based classification models is the pointing game [8]. For each saliency map the position of the max. value is determined. If this pixel is within the segmentation mask of the object, the example is considered a hit. The *pointing game score* is the number of hits divided by the number of evaluated image samples. We check for each correctly detected object whether the maximum attribution map value is within the ground truth 3D bounding box. For a PointPillars [3] model trained and evaluated on KITTI [2], we achieve a pointing game score of 0.9004. Note, however, that in KITTI the side mirrors of a car are not part of the annotated bounding box. If we thus increase the dimensions of the bounding boxes by only 10 %, the score increases to 0.9724.

6. Limitations

The first potential limitation is the runtime, which is primarily determined by the inference speed of the detector and by the number of iterations N. Note, however that creating very precise attribution maps is only needed for very few select examples, *e.g.* to analyze mis-detections. Creating the average attribution maps, on the other hand, can be done efficiently with a significantly smaller number of iterations. A second potential limitation can occur if an object is close to the LiDAR sensor, but still consists of only a few points. These edge cases can be caused by severe occlusions by other objects. In such cases the detections are more sensitive to the sub-sampling and the hyperparameter λ should be adapted individually, as we showed for the empty bounding boxes in Section 4.4 of the main manuscript.

References

[1] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection. In *Proc. AAAI*, 2021.

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. CVPR*, 2012.

[3] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection from Point Clouds. In *Proc. CVPR*, 2019.

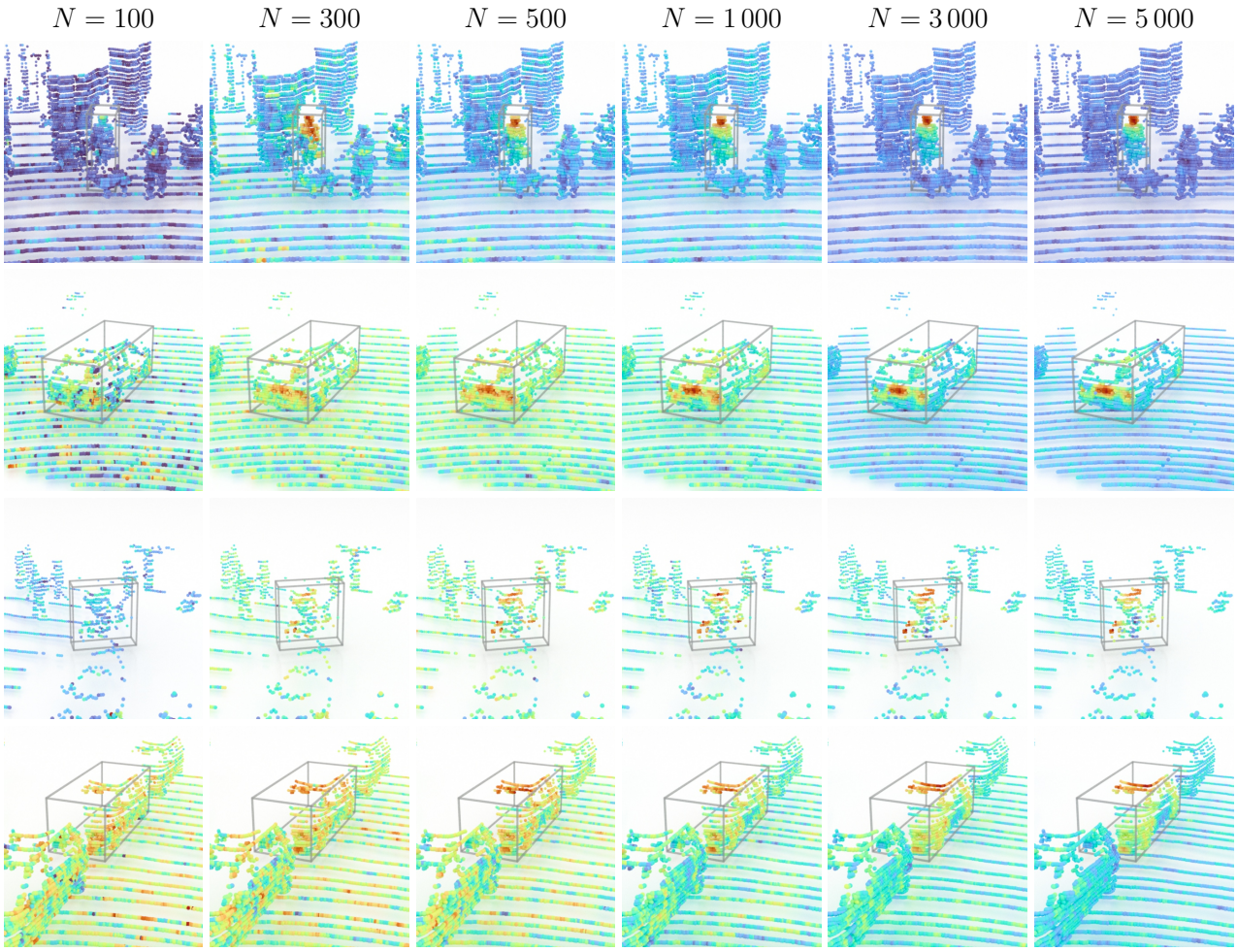
[4] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In *Proc. CVPR*, 2020.

[5] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From Points to Parts: 3D Object Detection from Point Cloud with Part-aware and Part-aggregation Network. *PAMI*, 43:2647–2664, 2021.

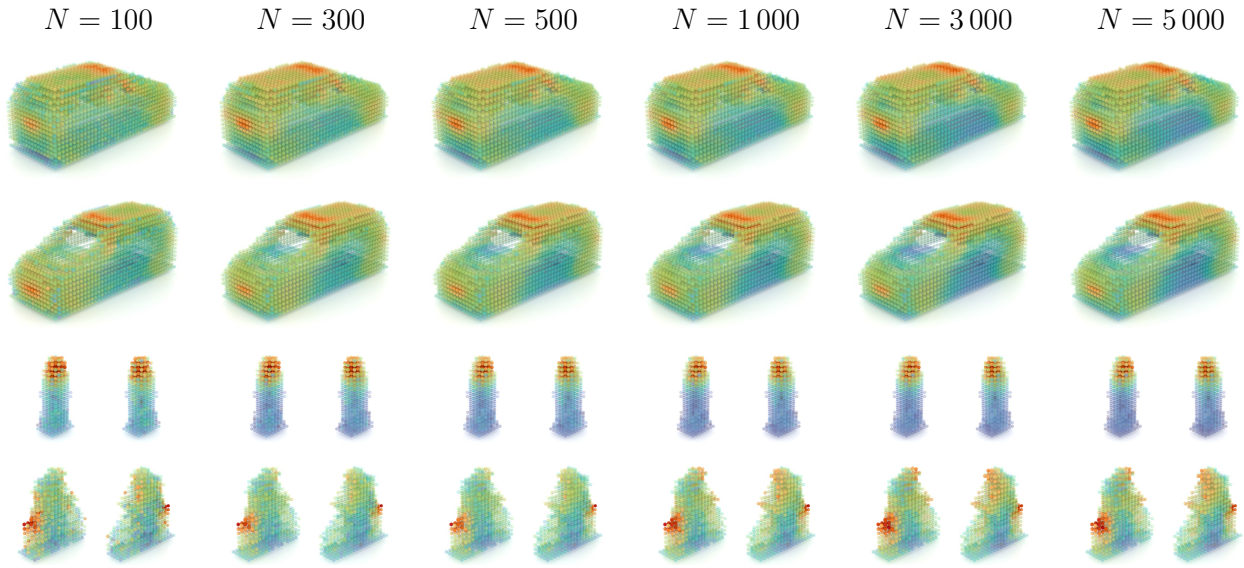
[6] OpenPCDet Development Team. OpenPCDet: An open-source toolbox for 3D object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020.

[7] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 18(10):3337, 2018.

[8] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. *IJCV*, 126(10):1084–1102, 2018.



(a) Detection-specific attribution maps from top-to-bottom: pedestrian, car, cyclist, car.



(b) Average attribution maps.

Figure 3. Influence of the number of iterations N on (a) individual (*i.e.* per-detection) and (b) average attribution maps for PointPillars [3] trained and evaluated on KITTI [2]. Best viewed on screen.

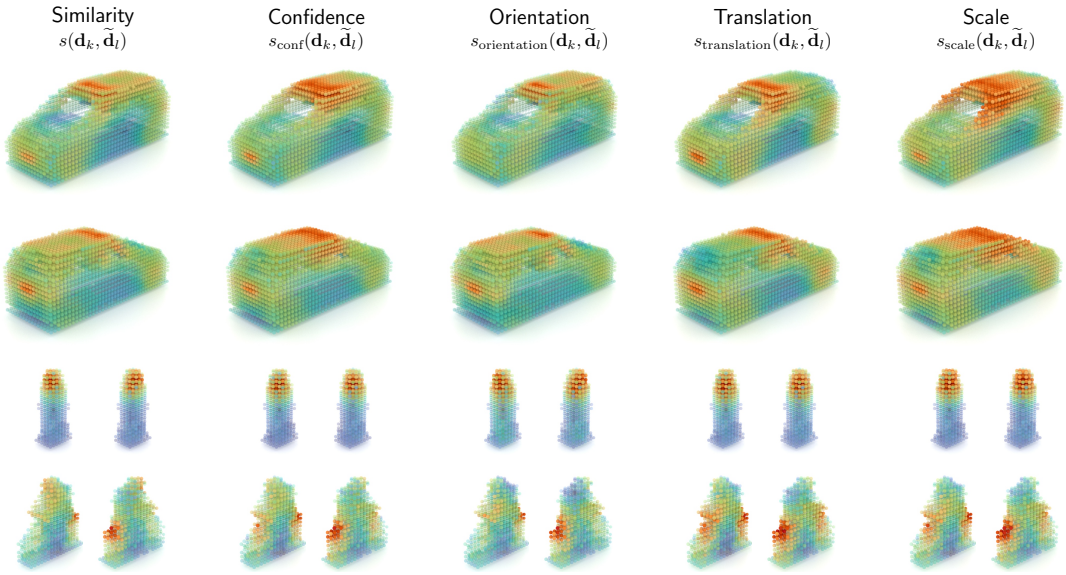


Figure 4. Average attribution maps w.r.t. the individual sub-metrics for a PointPillars [3] model trained and evaluated on KITTI [2].