Multimodal 3D Object Detection on Unseen Domains
Supplementary Material

Deepti Hegde[1] Suhas Lohit[2] Kuan-Chuan Peng[2] Michael J. Jones[2] Vishal M. Patel[1]

[1]Johns Hopkins University [2]Mitsubishi Electric Research Laboratories

{dhegde1, vpatel36}@jhu.edu {slohit, kpeng, mjones}@merl.com

1. Datasets for experiments

For our experiments, we choose four popular autonomous driving LiDAR-image datasets for 3D object detection: Lyft [3], KITTI [2], Waymo Open Dataset [5], and nuScenes [1]. In Table 1, we compare various properties of these datasets. This includes the conditions of data capture, such as sensor specifications, location, and weather as well as the properties of the data itself such as the size of the scene and the average dimensions of objects.

2. Qualitative evaluation of 3D object detection

We provide a qualitative comparison of the detection results of our proposed method **CLIX$^{\mathbf{3D}}$** against those of the single and multi-source direct transfer (DT) baselines for the Part-A^2 network for the domain shift scenario of Waymo, nuScenes $\rightarrow$ KITTI. This comparison is shown in Figure 1, where the columns correspond to the results from each method, while each row corresponds to the samples from the KITTI validation dataset. We visualize the bounding boxes that are predicted with a confidence score greater that 0.3. Our method **CLIX$^{\mathbf{3D}}$** addresses the problem of missed detections (false negatives) as well as superfluous predictions (false positive) faced by the baseline approaches that affect the precision score. The DT Waymo$\rightarrow$KITTI method in particular predicts numerous false positives with high confidence. The DT nuScenes$\rightarrow$KITTI model does not suffer from false positives, but fails to predict most instance of the "Cyclist" class (see column 1, rows 2 and 3). Multi-source DT (column 3) addresses some of these problems but still fails to detect some instance of "Car" and "Pedestrian". Column 4 shows the qualitative improvement our method, which predicts more instance of "Pedestrian" with fewer false positives of the "Car" category.

3. Additional implementation details

Evaluation metrics We report the 3D mean average precision of the "Car," "Pedestrian," and "Cyclist" categories at the medium difficulty, following the KITTI evaluation metric [2]. Since all networks are converted to the uniform format of the KITTI dataset, we use this same evaluation metric across all datasets, and consider only the image field-of-view for all lidar scenes. In the case of Part-A^2 evaluation on the Waymo [5] dataset, we report performance at 3D IoU thresholds $0.5, 0.25, 0.25$ for the "Car," "Pedestrian," and "Cyclist" categories respectively. This is done to perform a fair comparison with 3D-Vfield [4], which uses the same metric specification, and to be consistent for model selection.

When performing domain transfer to the Waymo dataset, we lower the target point clouds and ground truth bounding boxes by 1.6m to align them with the ground planes of the source datasets of Lyft and KITTI. This is done during the evaluation step only, and is consistent with the procedure followed by Lehner *et al*. [4] in 3D-Vfield.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Q. Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. 1

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[3] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*, 2020. 1

[4] Alexander Lehner, Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Mohammad-Ali Nikouei Mahani, Nassir Navab, Benjamin Busam, and Federico Tombari. 3D-VField: Adversarial augmentation of point clouds for domain generalization in 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17295–17304, 2022. 1

[5] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou,

	KITTI	Waymo	nuScenes	Lyft
LiDAR sensor	Velodyne HDL-64	$1\times360°$, $4\times$HoneyComb	Velodyne HDL-32	$1\times$64-beam, $2\times$40-beam
Point cloud size	100K	150K	70K	40K
Point cloud range	$[0,-40,-3,70.4,40,1]$	$[-75.2,-75.2,-2,75.2,75.2,4]$	$[-51.2,-51.2,-5.0,51.2,51.2,3.0]$	$[-80.0,-80.0,-5.0,80.0,80.0,3.0]$
LiDAR height	1.73	3.33	1.8	1.45
"Car" anchor	[3.90,1.60,1.56]	[4.70,2.10,1.70]	[4.63,1.97,1.74]	[4.75,1.92,1.71]
"Cyclist" anchor	[1.76,0.60,1.73]	[1.78,0.84,1.78]	[1.70,0.60,1.28]	[1.76,0.63,1.44]
"Pedestrian" anchor	[0.80,0.60,1.73]	[0.91,0.86,1.73]	[0.73,0.67,1.77]	[0.80,0.76,1.76]
# Annotated 3D bounding box	200K	12M	1.4M	1.3M
Location of capture	Germany	USA	USA, Singapore	USA
Weather conditions	sunny	variety	variety	variety

Table 1. Comparison between the autonomous driving datasets used in our experiments. All distances are in meters.

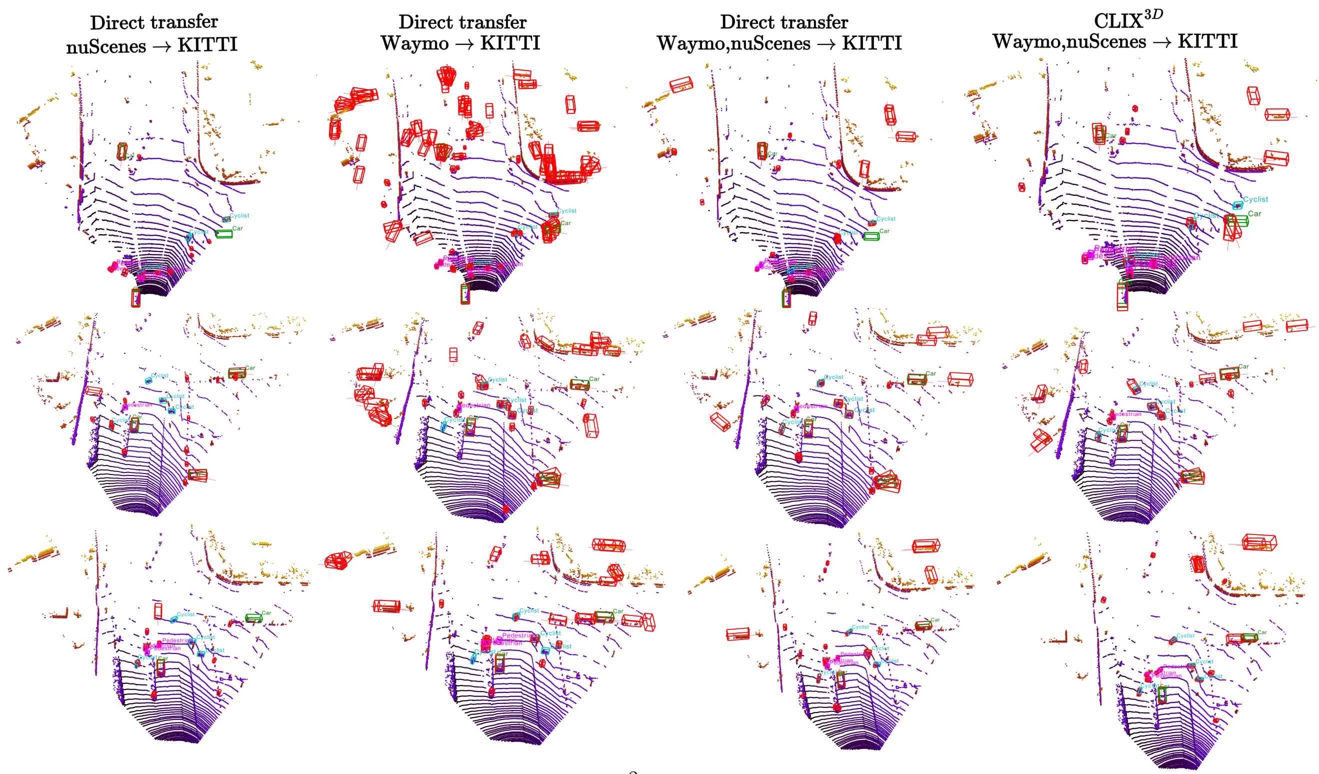

Figure 1. A qualitative comparison of the detection results of Part-A^2 trained for the domain shift scenario Waymo, nuScenes$\rightarrow$KITTI. Ground truth bounding boxes for the "Car" category are in green, in magenta for the "Pedestrian" category, and in cyan for the "Cyclist" category. Predictions are in red. (Best viewed zoomed in and in color).

Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 1