Towards Domain Generalization for Multi-view 3D Object Detection in Bird-Eye-View

Shuo Wang[1*] Xinhai Zhao[2*] Hai-Ming Xu[3] Zehui Chen[1] Dameng Yu[2],
Jiahao Chang [1], Zhen Yang[2], Feng Zhao[1†]
[1]University of Science and Technology of China [2]Huawei Noah's Ark Lab [3]University of Adelaide
{shuowang2323,lovesnow,changjh}@mail.ustc.edu.cn hai-ming.xu@adelaide.edu.au
{zhaoxinhai1,yudameng,yang.zhen}@huawei.com fzhao956@ustc.edu.cn

Appendix

A. Dataset Comparisons

We summarize the dataset information of nuScenes [1], Waymo [7] and Lyft [5] in detail in Tab. 1. To provide more intuitive comparisons among different datasets, we present images with projected ground-truth labels in Fig. 1. It is obvious that cameras utilized in these datasets are different, which is reflected in the image resolutions, object scale, *etc*.

Dataset	Size	Location	Shape	Number of Cameras	360°	Object	Night
nuScenes [1]	28130	Boston,SG.	(900,1600)	6	Yes	23	Yes
Waymo [7]	158081	USA	(1280, 1920),(886, 1920)	5	No	4	Yes
Lyft [5]	18900	Palo Alto	(1024, 1224)	6	Yes	9	No

Table 1. Dataset Overview. The size refers to the number of images used in the training stage and 360° indicates whether the cameras cover a 360° view. SG: Singapore.

B. Dynamic Perspective Augmentation

Suppose that two cameras shoot the plane P at different poses. As shown in the figure, the normal vector of the plane in frame 1 is N and the distance from P to frame 1 is d. For the point X_1 in Frame 1, we can get

$$N^T X_1 = d, \tag{1}$$

At the same time, it can be converted to frame 2 by the following equation

$$X_2 = RX_1 + T, \tag{2}$$

By using Eqs. (1) and (2), we can get that

$$X_2 = RX_1 + T\frac{1}{d}N^T X_1 = (R + T\frac{1}{d}N^T)X_1. \tag{3}$$

Then the 3D coordinate can be projected to the 2D image plane as:

$$x_1 = \frac{1}{z_1}KX_1, \tag{4}$$

*Shuo Wang and Xinhai Zhao contributed equally. This work was done when Shuo Wang was an intern at Huawei Noah's Ark Lab.
†Corresponding author.

Figure 1. Dataset visualizations with ground-truth labels.

where K is the intrinsic parameters of the camera and z is the depth.

By Eqs. (3) and (4), we can get that

$$x_2 = \frac{z_2}{z_1} K^{-1}(R + T\frac{1}{d}N^T)Kx_1 = \frac{z_2}{z_1} Hx_1, \tag{5}$$

where x_1 and x_2 are the homogeneous coordinates. Then we can get the homography matrix:

$$H = K^{-1}(R + T\frac{1}{d}N^T)K, \qquad H \in \mathbb{R}^{3\times 3}. \tag{6}$$

Since $\frac{z_2}{z_1}$ does not affect the validity of Eq. (5), we can determine one element in H to be 1 and the homography matrix has 8 degrees of freedom. So at least 4 corresponding point pairs are needed for recovering the matrix.

Algorithm 1 Dynamic Perspective Augmentation

Require: Multi-View Images $\mathbf{I} = \{I_1, I_2, \cdots, I_N\}$, Rotation Matrix from Ego Car to Camera $\mathbf{R} = \{R_1, R_2, \cdots, R_N\}$, Translation Matrix from ego car to Camera $\mathbf{T} = \{T_1, T_2, \cdots, T_N\}$, Intrinsic Parameters of Cameras $\mathbf{K} = \{K_1, K_2, \cdots, K_N\}$, N is the number of Surround Cameras, the Transformation of Rotation Matrix into Euler Angles ϕ, the Transformation of Euler Angles into Rotation Matrix Φ, Bottom Center and Bottom Corner Set of Ground Truth $\mathbf{O} = \{O_1, O_2, \cdots, O_M\}$, Yaw Range $[y_{min}, y_{max}]$, Pitch Range $[p_{min}, p_{max}]$, Roll Range $[r_{min}, r_{max}]$, Paired Matching Points Set L.

for n in range$(0, N)$ **do**
  for all m such that $O_m = (x_m, y_m, z_m) \in \mathbf{O}$ **do**
    $d \cdot (u, v, 1)^T = K_n(R_n \cdot O_m + T_n)$;
    // whether O_m is projected to the current image
    if $d > 0$ and (u, v) in $I_n.size()$ **then**
      $\Delta y = SAMPLE(y_{min}, y_{max})$;
      $\Delta p = SAMPLE(p_{min}, p_{max})$;
      $\Delta r = SAMPLE(r_{min}, r_{max})$;
      // get the origin camera pose
      $yaw, pitch, roll = \phi(R_n)$;
      // perturb the camera pose
      $R'_n = \Phi(yaw + \Delta y, pitch + \Delta p, roll + \Delta r)$;
      $d' \cdot (u', v', 1)^T = K_n(R'_n \cdot O_m + T_n)$;
      // whether O_m is projected to perturbed image
      if $d' > 0$ and (u', v') in $I_n.size()$ **then**
        $L.append([(u, v), (u', v')])$;
      end if
    end if
  end for
  // Whether over four pairs of points are matched
  if $len(L) >= 4$ **then**
    // LS denotes least squares method
    $H = LS(L)$;
    $I'_n = H \cdot I_n$;
  else
    $I'_n = I_n$;
  end if
end for

Output: Multi-View Perturbed Images $\mathbf{I'} = \{I'_1, \cdots, I'_N\}$

When the camera is only spinning ($T = 0$) or moving a small distance ($T \to 0$), the homography matrix is independent of d and N, which indicates Eq. (5) can be applied to the entire 3D space, rather than a plane. Therefore, we can leverage homography [2] to heuristically generate various perspective images for model learning. The implementation details are shown in Algorithm 1.

C. Detailed Training Settings

In this section, we introduce more detailed training settings. As for the model, we follow the basic config provided in [4]. Scale-invariant depth is determined by the metric depth and intrinsic parameters, so we employ [2, 90] for nuScenes, [1, 60] for Waymo and [1, 90] for Lyft. For the construction of the pseudo-domain categories, the settings of discretization thresholds are [500, 550, 600, 650, 700, 750] for nuScenes, [600, 650, 700, 750, 800, 850, 900] for Waymo and [500, 550, 600, 650] for Lyft.

D. Additional Experimental Results

Here, we provide some additional empirical results in the task of nuScenes→Waymo, building on top of BEVDet [4]. The results are shown in Tab. 2. The Source Only model cannot detect 3D objects where the mAP almost drops to 0 caused by the huge domain gap, just the same as BEVDepth [6]. We observe that CAM-Convs [3] hardly improves the performance of the

model on the target domain. By contrast, our approach greatly enhances the generalization ability of the model and achieves 64% NDS* of Oracle performance, which verifies the generalization of DG-BEV.

Nus → Waymo	Source Domain (nuScenes)					Target Domain (Waymo)				
Method	mAP↑	mATE↓	mASE↓	mAOE↓	NDS*↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS*↑
Oracle	-	-	-	-	-	0.487	0.582	0.147	0.078	0.609
Source Only	0.476	0.587	0.178	0.143	0.587	0.028	1.354	0.273	0.738	0.179
CAM-Convs [3]	0.477	0.580	0.177	0.136	0.590	0.034	1.346	0.273	0.721	0.185
DG-BEV (Ours)	**0.489**	**0.573**	**0.174**	**0.131**	**0.598**	**0.338**	**0.789**	**0.202**	**0.267**	**0.459**

Table 2. Performance of DG-BEV in nuScenes→Waymo task on top of BEVDet.

E. Broader Impacts

In autonomous driving, updating of cameras and iterations of vehicles often occur, which lead to a sharp decline in the performance of the previously trained model in real-world scenarios. In this paper, we use different domains to represent the above existing problems and propose DG-BEV, a domain generalization method for multi-view 3D object detection, which successfully alleviates the performance drop on the unseen target domain without impairing the accuracy of the source domain.

As for the limitation of our method, we do not take into account all the differences between domains and there are still numerous unsolved issues such as different color styles, various distributions of object dimensions, *etc*. We hope that our proposed DG-BEV can be a well-developed baseline for future research.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[2] Elan Dubrofsky. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5, 2009. 3

[3] Jose M Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. Cam-convs: Camera-aware multi-scale convolutions for single-view depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11826–11835, 2019. 3, 4

[4] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3

[5] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Level 5 perception dataset 2020. https://level-5.global/level5/data/, 2019. 1

[6] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 3

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