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

Shuo Wang1*  Xinhai Zhao2*  Hai-Ming Xu3  Zehui Chen1  Dameng Yu2,  Jiahao Chang1,  Zhen Yang2,  Feng Zhao1†
1University of Science and Technology of China  2Huawei Noah’s Ark Lab  3University 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.

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

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

$$N^T X_1 = d,$$  \hspace{1cm} (1)

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

$$X_2 = RX_1 + T,$$  \hspace{1cm} (2)

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

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

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

$$x_1 = \frac{1}{z_1} K X_1,$$  \hspace{1cm} (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.
where $K$ is the intrinsic parameters of the camera and $z$ is the depth.

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

$$x_2 = \frac{z_2}{z_1} K^{-1} (R + T \frac{1}{d} N^T) K x_1 = \frac{z_2}{z_1} H x_1,$$

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

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

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.

Figure 1. Dataset visualizations with ground-truth labels.
Algorithm 1 Dynamic Perspective Augmentation

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

```
1: for n in range(0, N) do
2:   for all m such that \( O_m = (x_m, y_m, z_m) \in O \) do
3:     d \cdot (u, v, 1)^T = K_n(R_n \cdot O_m + T_n);
4:     // whether \( O_m \) is projected to the current image
5:     if \( d > 0 \) and \((u, v)\) in \( I_n, \text{size}()\) then
6:       \( \Delta y = \text{SAMPLE}(y_{\text{min}}, y_{\text{max}})\);
7:       \( \Delta p = \text{SAMPLE}(p_{\text{min}}, p_{\text{max}})\);
8:       \( \Delta r = \text{SAMPLE}(r_{\text{min}}, r_{\text{max}})\);
9:       // get the origin camera pose
10:      yaw, pitch, roll = \( \Phi(R_n)\);
11:     // perturb the camera pose
12:      \( R'_n = \Phi(yaw + \Delta y, pitch + \Delta p, roll + \Delta r)\);
13:      \( d' \cdot (u', v', 1)^T = K_n(R'_n \cdot O_m + T_n)\);
14:     // whether \( O_m \) is projected to perturbed image
15:     if \( d' > 0 \) and \((u', v')\) in \( I_n, \text{size}()\) then
16:       \( L.append([\{(u, v), (u', v')\}]\);
17:     end if
18:   end if
19: end for
20: // Whether over four pairs of points are matched
21: if \( len(L) >= 4 \) then
22:   // LS denotes least squares method
23:   H = LS(L);
24:   \( I'_n = H \cdot I_n; \)
25: else
26:   \( I'_n = I_n; \)
27: end if
28: end for
```

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

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.

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

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


