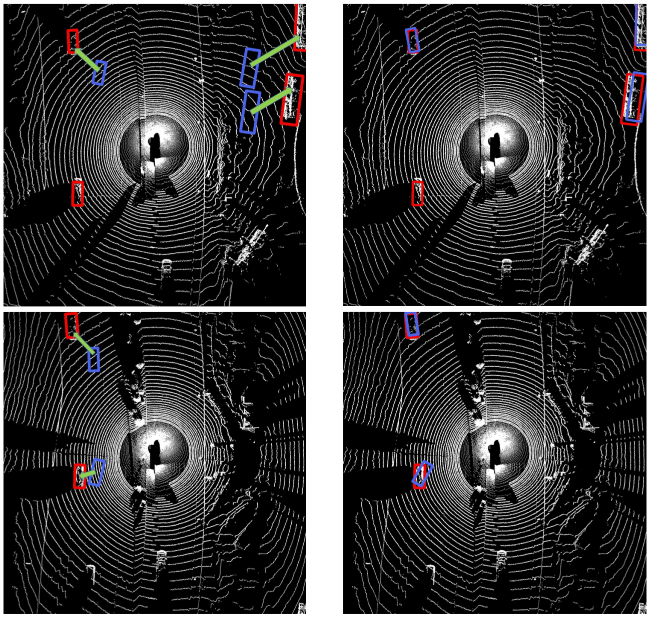
Towards Domain Generalization for Multi-view 3D Object Detection in Bird-Eye-View

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

Abstract

Multi-view 3D object detection (MV3D-Det) in Bird-Eye-View (BEV) has drawn extensive attention due to its low cost and high efficiency. Although new algorithms for camera-only 3D object detection have been continuously proposed, most of them may risk drastic performance degradation when the domain of input images differs from that of training. In this paper, we first analyze the causes of the domain gap for the MV3D-Det task. Based on the covariate shift assumption, we find that the gap mainly attributes to the feature distribution of BEV, which is determined by the quality of both depth estimation and 2D image's feature representation. To acquire a robust depth prediction, we propose to decouple the depth estimation from the intrinsic parameters of the camera (i.e. the focal length) through converting the prediction of metric depth to that of scale-invariant depth and perform dynamic perspective augmentation to increase the diversity of the extrinsic parameters (i.e. the camera poses) by utilizing homography. Moreover, we modify the focal length values to create multiple pseudo-domains and construct an adversarial training loss to encourage the feature representation to be more domain-agnostic. Without bells and whistles, our approach, namely DG-BEV, successfully alleviates the performance drop on the unseen target domain without impairing the accuracy of the source domain. Extensive experiments on various public datasets, including Waymo, nuScenes, and Lyft, demonstrate the generalization and effectiveness of our approach. To the best of our knowledge, this is the first systematic study to explore a domain generalization method for MV3D-Det.



(a) Baseline (b) DG-BEV

Figure 1. Qualitative comparisons between BEVDepth and the proposed DG-BEV. The red and blue bounding boxes represent ground truth and detected results on the target domain respectively. Depth-shift is shown in green arrows. Our approach can detect correct 3D results on unknown domains.

1. Introduction

3D object detection, aiming at localizing objects in the 3D space, is critical for various applications such as autonomous driving [6, 38], robotic navigation [2], and virtual reality [32], *etc.* Despite the remarkable progress of LiDAR-based methods [17,30,33], camera-based 3D object detection in Bird-Eye-View (BEV) [14, 19, 21] has drawn increasing attention in recent years due to its rich semantic information and low cost for deployment.

[*] 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.

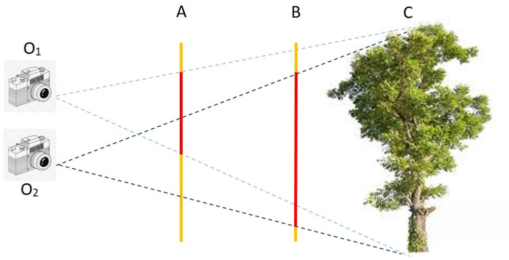


Figure 2. Illustration of the difficulty in estimating depth based on cameras with different focal length. O_1 and O_2 are the optical centers of two cameras and C is the object being photographed. A and B denote the imaging planes of the two cameras respectively and the red parts show the size of the same object in their corresponding image planes.

However, most of the detectors assume that the training and testing data are obtained in the same domain which may be hardly guaranteed in realistic scenarios. Thus, tremendous performance degradation will appear when the domain of the input image shifts. For example, nuScenes [3] and Waymo [34] are two popular benchmarks for 3D object detection and their data collection devices are not identical, *i.e*., both of the intrinsic and extrinsic parameters are different. Empirical results presented in Fig. 1 show that detectors trained on nuScenes have location bias when predicting objects on the Waymo dataset.

Domain Generalization (DG) [8, 18, 26], aiming to learn a model that generalizes well on unseen target domains, can be a plausible solution to alleviate the bias mentioned above. In the literature, DG has been widely explored for 2D vision tasks, *e.g*., image recognition [7, 16], object detection [31, 44], and semantic segmentation [27, 41]. However, most of these works are designed for the case where there are multiple source domains available which are obviously infeasible due to the diversity of the real world in autonomous driving scenarios. Alternatively, one recent work [39] proposed to study the single-domain generalization for LiDAR-based detection. However, it is not tractable to directly adapt this method to solve the camera-based detection task due to the fundamental differences between the characteristics of points and images. Therefore, developing a general domain generalization framework for MV3D-Det is still highly desirable.

In this paper, we theoretically analyze the causes of the domain gap for MV3D-Det. Based on the covariate shift assumption [4], we find that such a gap mainly attributes to the feature distribution of BEV, which is determined by the depth estimation and 2D image feature jointly. Based on this, we propose DG-BEV, a domain generalization method for MV3D-Det in BEV. Specifically, we first conduct a thorough analysis of why the estimated depth becomes inaccurate when the domain shifts and find the key factor lies in that intrinsic parameters of cameras used in various domains are hardly guaranteed to be identical (please refer to Fig. 2 for a better understanding). To alleviate this issue, we propose to decouple the depth estimation from the intrinsic parameters by converting the prediction of metric depth to that of scale-invariant depth. On the other hand, extrinsic parameters of cameras (*e.g*. camera poses) also play an important role in camera-based depth estimation, which is often ignored in previous works. Instead, we introduce homography learning to dynamically augment the image perspectives by simultaneously adjusting the imagery data and the camera pose.

Moreover, since domain-agnostic feature representations are favored for better generalization, we propose to build up multiple pseudo-domains by modifying the focal length values of camera intrinsic parameters in the source domain and construct an adversarial training loss to further enhance the quality of feature representations. In summary, the main contributions of this paper are:

- We present a theoretical analysis on the causes of the domain gap in MV3D-Det. Based on the covariate shift assumption, we find the gap lies in the feature distribution of BEV, which is determined by the depth estimation and 2D image feature jointly.
- We propose DG-BEV, a domain generalization method to alleviate the domain gap from both of the two perspectives mentioned above.
- Extensive experiments on various public datasets, including Waymo, nuScenes, and Lyft, demonstrate the generalization and effectiveness of our approach.
- To the best of our knowledge, this is the first systematic study to explore a domain generalization method for multi-view 3D object detectors.

2. Related Works

2.1. Vision-based 3D object detection

Vision-based 3D object detection [25] is gaining more and more attention from researchers due to rich semantic information and low cost for deployment. In the last few years, many efforts have been made on predicting objects directly from a single image. For example, inspired by FCOS [35], FCOS3D [36] extends this paradigm to 3D object detection and achieves great performance. Since single view-based prediction does not integrate information from multiple cameras well, there is a growing interest in MV3D-Det. LSS [29] is the first to explore the mapping of multi-view features to BEV space. Based on LSS, BEVDet [14] enables this paradigm to perform competitively. BEVDepth [19] regard LiDAR as supervisory information for depth and enhance the model's ability of depth perception. DETR3D [37] integrates information from multiple perspectives in an attention pattern and GraphDETR3D [5] improves performance further by utilizing graph neural networks. Moreover, PETR [23] proposes 3D position-aware encoding, which greatly improves the performance of DET3D.

2.2. Domain Adaption and Domain Generalization

Domain adaptation (DA) aims to improve models' performance on a known target domain. Many approaches have been designed for 2D detection. Particularly, [4] proposed to align both feature-level and instance-level distributions through an adversarial mechanism [12]. Subsequent work [1, 13, 40, 42] expands on this foundation. Since sometimes we can not get access to the target domain, some studies have started to focus on domain generalization (DG) [8, 10, 18, 26], which targets generalizing a model trained on source domains to many unseen target domains.

However, the aforementioned methods based on 2D detection mainly focus on handling lighting, color, and texture variations. Obviously, they can not be directly applied to 3D detection, the focus of which is to accurately estimate the spatial information of objects. Thus, some domain adaption methods specific to 3D perception have been explored. CAM-Convs [10] is a new type of convolution that improves the generalization capabilities of depth prediction networks considerably. For LiDAR-based detection [39], differences in data structures and network architectures make it impossible to apply to multi-view 3D detection. STMono3D [20] only explores single-view 3D detection instead of multi-view.

3. Method

3.1. Problem Definition

Under the domain generalization setting, we can access the labeled images from the source domain $D_S = \{x_s^i, y_s^i, K_s^i, E_s^i\}_{i=1}^{N_S}$ but the target domain $D_T = \{x_t^i, y_t^i, K_t^i, E_t^i\}_{i=1}^{N_T}$ is not available, of which N_s and N_t are the numbers of samples from the source and target domains, respectively. Each 2D image x^i comes with the camera intrinsic parameter K^i and the extrinsic parameter E^i. K^i is responsible for projecting the points in 3D space to the 2D image plane and E^i indicates the camera pose, which is composed of yaw, pitch, and roll. Label y^i consists of object class k, location (c_x, c_y, c_z), size in each dimension (d_x, d_y, d_z), and orientation θ. We aim to train models with D_S and achieve as good results as possible when inferring in any other target domain D_T. At the same time, the process described above will not impair the accuracy of the source domain.

3.2. A Probabilistic View of the Domain Gap

MV3D-Det in Bird-Eye-View (BEV) can be viewed as a component of two parts, one is a mapping that projects a 2D image into the feature map of 3D space (*i.e.* BEV), and the other one is to learn the posterior $P(Y|X, K, E)$, where Y is the ground truth consisting of category, location, dimension and orientation, X is the image representation, K is the intrinsic parameter and E is the extrinsic parameter. Let $P_S(Y, X, K, E)$ and $P_T(Y, X, K, E)$ represent the joint distribution of training samples in the source domain and the target domain, respectively. When there exists no domain shift in theory, it means that $P_S(Y, X, K, E) = P_T(Y, X, K, E)$. According to Bayes's Formula, we decompose the joint distribution as:

$$P(Y, X, K, E) = P(Y|X, K, E)P(X, K, E). \quad (1)$$

Similar to [4], we make the covariate shift assumption for $P(Y|X, K, E)$, *i.e.*, different domains naturally have the same conditional probability, and the domain distribution shift results from the inconsistent marginal distribution $P(X, K, E)$. In MV3D-Det, $P(X, K, E)$ indicates the feature distribution of 2D images projected into 3D space, which is determined by the depth estimation and 2D image feature jointly. Hence, we try to improve the existing domain shift from the above two aspects.

3.3. DG-BEV

In this section, we introduce our domain generation framework for multi-view 3D object detection, DG-BEV. Building on top of BEVDepth, we designate three simple approaches: (i) intrinsic-decoupled depth prediction, (ii) dynamic perspective augmentation, and (iii) domain-agnostic feature learning. Fig. 3 illustrates the overall framework of our approach.

3.3.1 Intrinsics-Decoupled Depth Prediction

As shown in Fig. 2, when two cameras with various intrinsic parameters (*i.e.* focal lengths) shoot the same object at the same distance, the imaging size of the object, which is determined by the intrinsic parameters, can be quite different. If a model is only optimized on the dataset collected from a specific camera, it can be difficult for the model to predict an identical depth for the object pictured from another camera, and thus it is the cause of inaccurate depth prediction when domain shifts, similar in [20]. Furthermore, we empirically find that the estimated depth has been entangled with the intrinsic parameters of the camera, and results in non-compliance with the intuition of "Everything looks small in the distance and big on the contrary". Hence, we attempt to decouple the estimated depth from the intrinsic parameters.

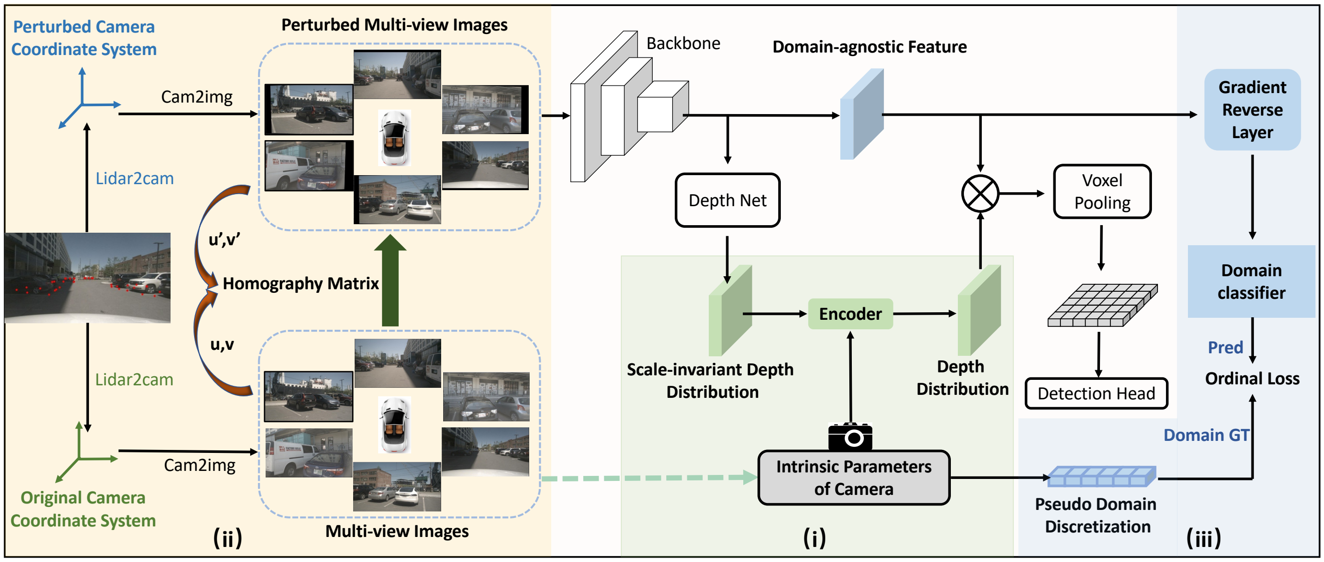


Figure 3. The overall framework of our approach DG-BEV. Building on top of BEVDepth, we propose three efficient strategies to improve the domain generalization ability: (i) intrisics-decoupled depth estimation in Sec. 3.3.1 (ii) dynamic perspective augmentation in Sec. 3.3.2, and (iii) domain-invariant feature learning in Sec. 3.3.3.

Random scaling of an image is one of the widely used augmentation methods. When an image is randomly scaled, the intrinsic parameter can be denoted as

$$K = \begin{bmatrix} r_x & r_y & 1 \end{bmatrix} \begin{bmatrix} f_x & 0 & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

where r_x and r_y are resize rates, f and p are the focal length and optical center, x and y indicates image coordinate axes, respectively.

According to Eq. (2), we can get the images of different intrinsic parameters by adjusting the resize rates. Motivated by DD3D [28], we decouple depth from different focal lengths and acquire the scale-invariant depth as

$$d = \frac{s}{c} \cdot d_m, \tag{3}$$

$$s = \sqrt{\frac{1}{f_x^2} + \frac{1}{f_y^2}}, \tag{4}$$

where d_m is the metric depth, s is the original pixel size, and c is a constant representing the pixel size at a given reference focal length. Through Eq. (3), the pixel scale of the reference focal length is regarded as the basis for depth estimation, which makes the predicted depth consistent with the size of the object in the image. During the training, we modify the intrinsic parameters and image resolutions simultaneously. Once obtaining the scale-invariant depth, we utilize the actual focal length to encode the estimated depth to the metric depth, which greatly alleviates the problem caused by different intrinsic parameters among domains.

3.3.2 Dynamic Perspective Augmentation

Camera poses relative to the ego car are usually divergent among different domains. As noted in [43], monocular depth predictors are naturally biased $w.r.t$ the distribution of camera poses, which inevitably impairs the accuracy of depth estimates when inferring on the unseen target domain. Transforming the image perspective in the source domain can be a feasible solution to obtain more robust depth predictions. However, direct perturbation of image perspective (*i.e.* camera pose) like PDA [43] is not feasible due to the unavailability of pixel-wise depth, instead, we propose dynamic perspective augmentation by leveraging homography [9] to heuristically generate various perspective images for model learning.

A homography is a mapping between two planar surfaces which is widely used for perspective conversion. The homography matrix $H \in \mathbb{R}^{3\times3}$ projects p_1 on one plane to p_2 on another plane,

$$sp_2 = Hp_1, \tag{5}$$

where $p = [x, y, 1]^T$ is the homogeneous coordinate of a 2D point in a plane and s is the scale factor. Since the homography matrix has natural attributes with 8 degrees of freedom, at least 4 corresponding point pairs are needed for recovering the matrix.

Suppose camera pose relative to the ego car as $P_i = (y_i, p_i, r_i)$, where i is the index of camera, y_i, p_i and r_i denote yaw, pitch and roll respectively. Then we perturb the camera pose as

$$\hat{P}_i = (y_i + \Delta y_i, p_i + \Delta p_i, r_i + \Delta r_i), \tag{6}$$

where Δy_i, Δp_i and Δr_i are the random perturbation. We opt to use the homography matrix to describe the projection relationship between the original camera pose and the scrambled camera pose. Let $B_{gt} = \{b_1, \cdots, b_n\}$ represent the 3D ground truth boxes, where n is the number of boxes. We pick up five bottom points $Q_{gt} = [x_{gt}, y_{gt}, z_{gt}]^T$ of 3D ground truth box b_i as representatives, including one bottom center point and four bottom corner points. The selected points Q_{gt} will be transformed into the original image plane by spatial mapping relationship, which is defined by

$$d \cdot q = K(\Phi(P) \cdot Q_{gt} + T), \tag{7}$$

where d is the actual depth, Φ is the transformation of Euler angles into a rotation matrix, T is the transformation matrix from the ego car to the camera and K denotes the intrinsic matrix. At the same time, the identical points Q_{gt} will be transformed into the scrambled image plane by

$$\hat{d} \cdot \hat{q} = K(\Phi(\hat{P}) \cdot Q_{gt} + T). \tag{8}$$

If q and $\hat{q}$ are both in the range of image size, this pair of points will be kept. The formulation of the transformation of the two perspectives is defined as

$$\hat{q} = Hq, \tag{9}$$

When more than four pairs of points are reversed for a camera, we can acquire the estimated homography matrix H by applying the least square method. Then we can utilize the homography matrix to roughly convert the original image into the one after the camera pose perturbation. More details about the homography principles and implementation can be found in the *supplementary materials*.

3.3.3 Domain-Invariant Feature Learning

Domain-related annotations [4, 12], which are used to extract domain-agnostic representations containing intrinsic characteristics, is helpful for improving the generalization capability of models. However, when the target domain is inaccessible, how to well extract the domain-agnostic representations remains under-explored.

Intrinsic characteristics in the domain include image style, illumination, object scale, *etc*. And in MV3D-Det, one of the most significant differences among domains is the object scale caused by the diverse intrinsic parameters, which is also an important reason for the shift of feature distribution. From Sec. 3.3.1, we conclude that random scaling of images indicates a corresponding change in the intrinsic parameters. Hence, to acquire domain-invariant feature representation, we enforce the network to classify the domain itself by explicitly constructing pseudo-domain categories based on the focal length values.

Considering the wide range of the focal length values of camera intrinsic parameters, we quantize the focal length interval [α, β] into K sub-intervals by uniform discretization (UD), where α and β denote the minimum and maximum values of the interval, respectively. The discretization thresholds $t_i \in \{t_0, t_1, \cdots, t_K\}$ can be formulated as:

$$t_i = \alpha + \frac{(\beta - \alpha) * i}{K}, \tag{10}$$

Assuming that different focal length sub-intervals represent different pseudo-domains, it is obvious that the pseudo-domains form a well-ordered set with a strong ordinal correlation. However, typical classification losses (*e.g.* CrossEntropy Loss, Focal Loss [22]) ignore the ordered information among the discrete labels. Motivated by [11], we treat the pseudo-domain classification as a sequential process and adopt an ordinal loss to make the most of the ignored information. Due to the intrinsic properties of ordinal classification, we opt to take ranges on both sides of the interval into consideration instead only closed intervals. As shown in Fig. 4, if the interval originally has K sub-intervals, it means that there are $K + 1$ discretization thresholds and $K + 2$ categories.

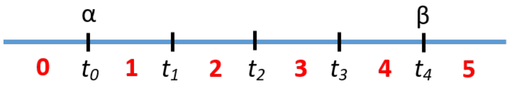


Figure 4. Illustration of the relationships among sub-intervals, discrete focal distance values and categories. t_i represents the discrete value, and $\{0, 1, 2, 3, 4, 5\}$ denote corresponding the domain categories. In this figure, there are 4 sub-intervals, 5 discretization thresholds, and 6 categories.

Let x denote the feature map given an image, then we can acquire the ordinal outputs by

$$y = \phi(x, \theta), \tag{11}$$

where y is a $2(K + 1)$-dimensional vector and θ is the parameters of the domain classifier. Let $l \in \{0, 1, \cdots, K+1\}$

denote the discrete label.Our ordinal loss can be defined as

$$\mathcal{L}(y,l) = \sum_{k=0}^{K+1} \gamma(k,l)log(P^k) + (1-\gamma(k,l))log(1-P^k),$$

$$\gamma(k,l) = \begin{cases} 1, & l \leq k \\ 0, & l > k \end{cases} \tag{12}$$

$$P^k = \frac{e^{y(2k)}}{e^{y(2k)} + e^{y(2k+1)}},$$

where $\gamma(k,l)$ indicates whether the actual focal length value is less than the k-th discrete focal length threshold and P^k denotes the probability that the domain classifier discriminates the focal length less than the k-th discrete values.

To align the domain distribution, we simultaneously optimize the domain classifier to minimize the ordinal loss and the base network to maximize this loss. For the implementation we use the gradient reverse layer (GRL) [12] to invert the gradient back from the domain classifier.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments on three widely used autonomous driving datasets: nuScenes [3], Waymo [34], and Lyft [15]. Each dataset has a diverse set of cameras with different intrinsic parameters and extrinsic parameters. We summarize the dataset information in detail in the *supplementary material*.

Comparision Methods. In our experiments, we compare our DG-BEV with three counterparts: (*i*) **Source Only** indicates directly using the model trained by the source domain to evaluate on the target domain. (*ii*) **Oracle** indicates the fully supervised model trained on the target domain. (*iii*) **CAM-Convs** [10] is a new type of convolution that improves the generalization capabilities of depth prediction networks considerably.

Evaluation Metrics. The mAP defined by nuScenes is based on the matching of 2D center distance on the ground plane instead of the Intersection over Union (IoU), which measures the error of ranging. Hence, we adopt the same validation metrics predefined officially by nuScenes for all datasets for simplicity. Since the attribute labels and the velocity labels are different from each other, we discard the Average Attribute Error (mAAE) and the Average Velocity Error (mAVE) in the case that nuScenes is not the source domain, and report the Average Precision (mAP), the Average Translation Error (mATE), the Average Scale Error (mASE) and the Average Orientation Error (mAOE) for other cases. Due to the lack of necessary metrics (*i.e.* mAAE and mAVE), we develop NDS* as an alternative, which is defined as:

$$\text{NDS}^* = \frac{1}{6}[3\,\text{mAP} + \sum_{\text{mTP}\in\mathbb{TP}} (1 - \min(1, \text{mTP}))], \tag{13}$$

We focus on the commonly used vehicle category, and more specifically, the 'car', 'truck', 'construction vehicle', 'bus', and 'trailer' of nuScenes, the 'vehicle' of Waymo and the 'car' of Lyft. What is more, to maintain consistency during training and validation, we only validate results in the range [-50m, 50m] like nuScenes.

Implementation Details. To validate the effectiveness of our DG-BEV, we adopt BEVDepth as our base model. Following [14], models are trained with AdamW [24] optimizer, in which gradient clip is exploited with learning rate 2e-4, a total batch size of 64 on 8 Tesla V100s. We use $W_{in} \times H_{in}$ to denote the width and height of the input image and $W \times H$ represents the origin resolution. Then the original image will be processed by random flipping, random scaling with a range of $s \in [W_{in}/W - 0.04, W_{in}/W + 0.18]$ for nuScenes and Lyft and $s \in [W_{in}/W - 0.08, W_{in}/W + 0.08]$ for Waymo, random rotating with a range of $r \in [-5.4°, 5.4°]$ and finally cropping to a size of $W_{in} \times H_{in}$. Due to the different aspect ratios of different datasets, We use 704×256 as the input size for nuScenes, 704×320 for Waymo and 704×384 for Lyft. More implementation details are shown in the *supplementary material*.

4.2. Main Results

As shown in Tab. 1, we compare the detection performance with Source Only, Oracle, and CAM-Convs. Our method outperforms the Source Only baseline and CAM-Convs baseline under four different settings. We can observe that CAM-Convs hardly improve the performance of the model on the target domain. On nuScenes→Waymo and Waymo→nuScenes tasks, the Source Only model cannot detect 3D objects where the mAP almost drops to 0 caused by the huge domain gap. Our approach can greatly enhance the generalization ability of the model and achieves 64% and 80% of Oracle performance (NDS*) in nuScenes→Waymo and Waymo→nuScenes, respectively. As for the nuScenes→Lyft and Lyft→nuScenes tasks, the Source Only model still maintains a certain level of detection capability, which indicates that there is no extremely large domain gap between the two. The reason is that the camera intrinsic parameters of the two datasets are close to each other and the six multi-view camera poses are similar. In this situation, our method can greatly improve the performance of the model in the unknown domain, *e.g.* 0.296 NDS* →0.437 NDS* in nuScenes→Lyft and 0.213 NDS* →0.374 NDS* in Lyft→nuScenes.

4.3. Ablation Studies and Analysis

In this section, we explore the role of each module in DG-BEV through more detailed ablation studies. If not

Nus → Waymo	Source Domain (nuScenes)					Target Domain (Waymo)				
Method	mAP↑	mATE↓	mASE↓	mAOE↓	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS$\hat{*}$↑
Oracle	-	-	-	-	-	0.552	0.528	0.148	0.085	0.649
Source Only	0.328	0.666	0.274	**0.560**	0.407	0.040	1.303	0.265	0.790	0.178
CAM-Convs [10]	0.328	0.681	0.273	0.571	0.397	0.045	1.301	0.253	0.773	0.185
DG-BEV (Ours)	**0.337**	**0.647**	**0.272**	0.567	**0.407**	**0.297**	**0.822**	**0.216**	**0.372**	**0.415**
Waymo → Nus	Source Domain (Waymo)					Target Domain (nuScenes)				
Method	mAP↑	mATE↓	mASE↓	mAOE↓	NDS$\hat{*}$↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS$\hat{*}$↑
Oracle	-	-	-	-	-	0.475	0.577	0.177	0.147	0.587
Source Only	0.552	0.528	**0.148**	0.085	0.649	0.032	1.305	0.768	0.532	0.133
CAM-Convs [10]	0.549	0.532	0.148	0.080	0.648	0.038	1.308	0.316	0.506	0.215
DG-BEV (Ours)	**0.568**	**0.519**	0.149	**0.078**	**0.660**	**0.303**	**0.689**	**0.218**	**0.171**	**0.472**
Nus → Lyft	Source Domain (nuScenes)					Target Domain (Lyft)				
Method	mAP↑	mATE↓	mASE↓	mAOE↓	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS$\hat{*}$↑
Oracle	-	-	-	-	-	0.602	0.471	0.152	0.078	0.684
Source Only	0.328	0.666	0.274	0.560	0.407	0.112	0.997	0.176	0.389	0.296
CAM-Convs [10]	0.328	0.681	0.273	0.571	0.397	0.145	0.999	0.173	0.368	0.316
DG-BEV (Ours)	**0.341**	**0.655**	**0.273**	**0.538**	**0.409**	**0.287**	**0.771**	**0.170**	**0.302**	**0.437**
Lyft → Nus	Source Domain (Lyft)					Target Domain (nuScenes)				
Method	mAP↑	mATE↓	mASE↓	mAOE↓	NDS$\hat{*}$↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS$\hat{*}$↑
Oracle	-	-	-	-	-	0.401	0.651	0.179	0.484	0.482
Source Only	0.602	0.471	0.152	0.078	0.684	0.102	1.143	0.239	0.789	0.213
CAM-Convs [10]	**0.611**	**0.465**	**0.149**	**0.075**	**0.691**	0.098	1.198	0.209	1.064	0.181
DG-BEV (Ours)	0.590	0.488	0.153	0.079	0.675	**0.268**	**0.764**	**0.205**	**0.591**	**0.374**

Table 1. Performance of DG-BEV on four source-target pairs. For the source domain, we report the average metric for the 10 categories of nuScenes. For the target domain, we report the average metric for the car category when Lyft→nuScenes and the metric of five categories when Waymo→nuScenes, including car, truck, construction vehicle, bus and trailer. As for Lyft and Waymo, we report the results of the vehicle category and car category, respectively. All results are on the validation subset of the corresponding dataset.

specified, all experiments are conducted with BEVDepth-R50 on the task of training on nuScenes and validating on 1/2 subset of Waymo.

4.3.1 Main Ablations

In order to understand of how each component contributes to the final performance, we subsequently add the proposed module and report the performance in Tab. 2. The vanilla baseline starts from 0.178 NDS*, which incurs a drastic performance drop compared to the source domain. When intrinsic-decoupled depth estimation module is added, the detection accuracy improves from 0.178 to 0.393, indicating the necessity of the disentanglement between depth estimation and camera intrinsic. Then, we apply the dynamic perspective augmentation strategy, which further gains 1.4% NDS*. Finally, when domain-invariant feature learning is introduced, the performance achieves 0.415 NDS*, yielding an enhancement of 24% NDS*.

IDD	DPA	DIFL	mAP ↑	NDS$\hat{*}$ ↑
			0.040	0.178
✓			0.272	0.393
✓	✓		0.297	0.408
✓	✓	✓	0.297	0.415

Table 2. Ablation studies on the effectiveness of each component in DG-BEV. "IDD" denotes intrinsic-decoupled depth estimation, "DPA" denotes dynamic perspective augmentation, and "DIFL" denotes domain-invariant feature learning.

Perturbation	nuScenes → Waymo mAP↑	nuScenes → Lyft mAP↑
0	0.272	0.242
Δ p=0.01	0.284	0.260
Δ p=0.02	0.259	**0.293**
Δ p=0.03	0.258	0.270
Δ p=0.04	0.247	0.275
Δ y=0.02	0.286	0.252
Δ y=0.04	0.281	0.248
Δ y=0.06	0.277	0.254
Δ y=0.08	0.283	0.252
Δ r=0.02	**0.290**	0.263
Δ r=0.04	0.286	0.263
Δ r=0.06	0.283	0.270
Δ r=0.08	0.286	0.263

Table 3. Ablation study of Dynamic Perspective Augmentation. Δ p, Δ y and Δ r denote the perturbed range of pitch, yaw and roll, respectively.

4.3.2 Dynamic Perpsective Augmentation

In order to investigate the effect of camera pose perturbation on the generalization ability of the model, we conducted experiments under two settings, nuScenes→Waymo (6 cameras→5 cameras) and nuScenes→Lyft (6 camera→6 cameras), respectively. The results are shown in Tab. 3. Overall, perspective augmentation can improve the generalization capability of the model in both settings. Specifically, for nuScenes→Waymo, perturbation of pitch within a certain range can lead to improved results (0.284 mAP), and perturbation of yaw always brings a definite gain. The reason is that the difference in extrinsic parameters between nuScenes and Waymo is primarily due to different yaws of the camera relative to the ego car. For nuScenes→Lyft, a perturbation of pitch can promote the detection ability greatly, where the possible reason is that the heights of the cameras are different when collecting the two datasets, so adjusting the pitch can mitigate the difference to some extent. Also since nuScenes and Lyft are both multi-view 6 cameras and each camera has a similar orientation, perturbing the yaw only obtains a little improvement. Moreover, since roads are not always flat, adjusting the roll range allows the model to adapt to different slopes, and enhance the generalization capability of the model in both settings.

4.3.3 Domain-Invariant Feature Learning

The domain shift among different domains contains many factors, and we think that one of the most important factors in 3D detection is scale variation caused by the focal length. To verify the effectiveness of our method, we adopt the classical method of DA [12] including both source and target domain and conduct detailed experiments in Tab. 4. No matter whether the target domain is included or not, using domain classifiers to align feature distributions among different domains does not bring any gains (0.171 NDS* & 0.002 NDS*). This indicates that the domain shift across domains can not be resolved only from the feature dimension. Moreover, under the setting of decoupling depth, our proposed method can achieve the improvement of 1.4 NDS* (0.393 → 0.407), while the result has a serious degradation (0.198 NDS*) after the target domain is available. The reasons can be two-fold: (i) directly aligning the feature distributions between the source and target domain would make the network focus on all inter-domain differences without distinction, while most of the inter-domain differences may not be conducive to accurately estimating the spatial information of the objects. (ii) the images of different domains have different aspect ratios. When both are input to the network with the same size, it is necessary to pad the image of one of the domains, which makes it easy for the domain classifier to identify domains based on whether images are padding. As a result, the image features are filled with noise after passing the GRL, which causes the degradation of the results.

DC	IDD	Target Domain	mAP↑	NDS$\hat{*}$↑
			0.040	0.178
✓			0.032	0.171
✓		✓	0.004	0.002
	✓		0.272	0.393
✓	✓		**0.292**	**0.407**
✓	✓	✓	0.054	0.198

Table 4. Ablation study of Domain-Invariant Feature Learning. "DC" denotes the domain classifier and "IDD" denotes intrisic-decoupling depth prediction.

Domain classification loss. Domain classifier with GRL usually performs a classification task to align the feature distributions between the source and target domain. In this paper, we divide the input images into different domains according to the ranges of the focal lengths. Since the focal length has the property of order, we explore the effect of different classification losses on performance. The experiment is based on the premise of intristic-decoupled depth, and the results are shown in Tab. 5. We can find that the CrossEntropy loss and the focal loss can not work well (0.397 NDS* & 0.401 NDS*) in such a ordinal classification. In contrast, with the help of the ordinal loss, our model reaches 0.407 NDS*.

Loss	mAP↑	$\text{NDS}^{\hat{*}}$↑
Cross Entropy Loss	0.282	0.397
Focal Loss	0.281	0.401
Oridinal Loss	**0.292**	**0.407**

Table 5. Ablation study on three different domain classification loss.

5. Conclusion

In this paper, we have proposed a novel domain-general BEV perception method named DG-BEV which can alleviate the performance drop on the unseen target domain. We observe that current BEV perception methods are all for specific domain, which will greatly limit the application in the industry. We decouple the BEV feature distribution with specific domain by the proposed instrinsics-decoupled depth prediction and domain-invariant feature learning. Extensive experiments on various public datasets, including Waymo, nuScenes, and Lyft, demonstrate the generalization and effectiveness of our approach. We hope that the proposed method DG-BEV could improve the implementation of BEV perception in the industry.

References

[1] David Acuna, Jonah Philion, and Sanja Fidler. Towards optimal strategies for training self-driving perception models in simulation. *Advances in Neural Information Processing Systems*, 34:1686–1699, 2021. 3

[2] Morris Antonello, Marco Carraro, Marco Pierobon, and Emanuele Menegatti. Fast and robust detection of fallen people from a mobile robot. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4159–4166. IEEE, 2017. 1

[3] 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. 2, 6

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 2, 3, 5

[5] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Graph-detr3d: Rethinking overlapping regions for multi-view 3d object detection. *arXiv preprint arXiv:2204.11582*, 2022. 3

[6] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinghong Jiang, Feng Zhao, Bolei Zhou, and Hang Zhao. Autoalign: Pixel-instance feature aggregation for multimodal 3d object detection. *arXiv preprint arXiv:2201.06493*, 2022. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[9] Elan Dubrofsky. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5, 2009. 4

[10] 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, 6, 7

[11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 5

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3, 5, 6, 8

[13] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *European conference on computer vision*, pages 309–324. Springer, 2020. 3

[14] 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. 1, 2, 6

[15] 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. 6

[16] Daniel Keysers, Thomas Deselaers, Christian Gollan, and Hermann Ney. Deformation models for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1422–1435, 2007. 2

[17] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1

[18] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 2, 3

[19] 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. 1, 2

[20] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, and Junjun Jiang. Unsupervised domain adaptation for monocular 3d object detection via self-training. *arXiv preprint arXiv:2204.11590*, 2022. 3

[21] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 1

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[23] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022. 3

[24] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6

[25] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*, 2022. 2

[26] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 2, 3

[27] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 2

[28] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021. 4

[29] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 2

[30] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[32] Martijn J Schuemie, Peter Van Der Straaten, Merel Krijn, and Charles APG Van Der Mast. Research on presence in virtual reality: A survey. *CyberPsychology & Behavior*, 4(2):183–201, 2001. 1

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. 1

[34] 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. 2, 6

[35] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[36] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021. 2

[37] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 3

[38] Yingjie Wang, Qiuyu Mao, Hanqi Zhu, Yu Zhang, Jianmin Ji, and Yanyong Zhang. Multi-modal 3d object detection in autonomous driving: a survey. *arXiv preprint arXiv:2106.12735*, 2021. 1

[39] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 847–856, 2022. 2, 3

[40] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 3

[41] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 2

[42] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, pages 86–102. Springer, 2020. 3

[43] Yunhan Zhao, Shu Kong, and Charless Fowlkes. Camera pose matters: Improving depth prediction by mitigating pose distribution bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15759–15768, 2021. 4

[44] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2