Geometry-Guided Domain Generalization for Monocular 3D Object Detection

Fan Yang[1,2,3], Hui Chen[1,2*], Yuwei He[1,2], Sicheng Zhao[1,2], Chenghao Zhang[1,2], Kai Ni[4], Guiguang Ding[1,2*]

[1]Tsinghua University
[2]BNRist
[3]Hangzhou Zhuoxi Institute of Brain and Intelligence
[4]HoloMatic Technology
yfthu@outlook.com, {jichenhui2012, heyuwei403, schzhao}@gmail.com, ch-zhang20@mails.tsinghua.edu.cn, nikai97@gmail.com, dinggg@tsinghua.edu.cn

Abstract

Monocular 3D object detection (M3OD) is important for autonomous driving. However, existing deep learning-based methods easily suffer from performance degradation in real-world scenarios due to the substantial domain gap between training and testing. M3OD's domain gaps are complex, including camera intrinsic parameters, extrinsic parameters, image appearance, etc. Existing works primarily focus on the domain gaps of camera intrinsic parameters, ignoring other key factors. Moreover, at the feature level, conventional domain invariant learning methods generally cause the negative transfer issue, due to the ignorance of dependency between geometry tasks and domains. To tackle these issues, in this paper, we propose MonoGDG, a geometry-guided domain generalization framework for M3OD, which effectively addresses the domain gap at both camera and feature levels. Specifically, MonoGDG consists of two major components. One is geometry-based image reprojection, which mitigates the impact of camera discrepancy by unifying intrinsic parameters, randomizing camera orientations, and unifying the field of view range. The other is geometry-dependent feature disentanglement, which overcomes the negative transfer problems by incorporating domain-shared and domain-specific features. Additionally, we leverage a depth-disentangled domain discriminator and a domain-aware geometry regression attention mechanism to account for the geometry-domain dependency. Extensive experiments on multiple autonomous driving benchmarks demonstrate that our method achieves state-of-the-art performance in domain generalization for M3OD.

Introduction

Monocular 3D object detection (M3OD) enables inferring 3D bounding boxes from a single image, considerably reducing the perception cost for autonomous driving (Mousavian et al. 2017). Many research works have focused on deep learning-based methods for M3OD, but they often suffer from performance degradation in real-world scenarios due to the presence of domain gap (Hendrycks and Dietterich 2019; Recht et al. 2019). To cope with this challenge, recent studies have started exploring cross-domain techniques for M3OD. STMono3D (Li et al. 2022b) proposes the domain adaptation in M3OD. DGMono3D (Li et al. 2022a) investigates domain generalization (DG). Despite remarkable progress, existing methods still suffer from some limitations. In this context, we aim to highlight two vital issues in domain generalization for M3OD.

First, the domain gap in M3OD can be attributed to complex factors. We systematically analyze three crucial domain gaps in M3OD. (1) The intrinsic parameter gap, including focal length and field of view (FOV), influences the predicted depth by the detector (Fig. 1(a)). (2) Extrinsic parameter gap: (Fig. 1(b)) demonstrates that the camera's orientation can impact the M3OD results. (3) Image appearance gap: the variations in image style and environmental conditions, such as weather and lighting (Fig. 1(c)), can considerably affect the features extracted by the model. Existing approaches primarily focus on the first gap, i.e., camera intrinsic parameters, neglecting other domain gaps and thus lacking robustness in real-world scenarios.

Secondly, at the feature level, commonly used domain invariant learning methods often lead to the negative transfer issue in M3OD. Li et al. point out that DG aims to learn domain-invariant representations for varying domains (Li et al. 2018a). However, popular feature invariant learning techniques, such as domain adversarial training (Ganin et al. 2016; Chen et al. 2019) and statistical matching (Sun and Saenko 2016; Long et al. 2015), often result in a severe negative transfer problem, leading to a significant performance decay for M3OD (Fig. 2(a)). In light of this, we argue that the fundamental assumption of feature invariant learning, which assumes independence between domains and labels (Ghifary et al. 2017; Xie et al. 2017), does not hold true in M3OD. Specifically, M3OD exhibits a significant geometry-domain dependency, with substantial differences in geometry depth, dimensions, and object rotation across domains (Fig. 2). Through an entropy perspective analysis, we demonstrate that in M3OD, eliminating domain-specific features can disrupt geometry features, and using only domain-shared features is insufficient for geometry prediction. Moreover, depth prediction, widely recognized as

*Corresponding Authors.
https://MonoGDG.github.io/

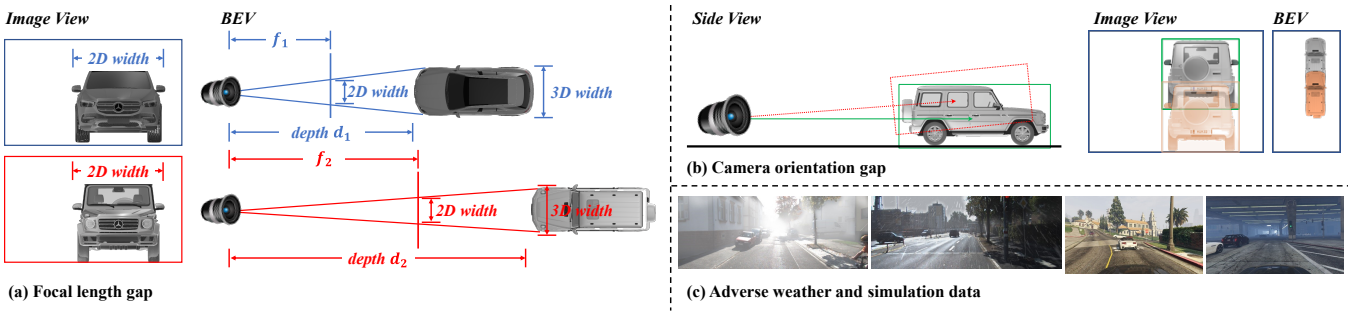

Figure 1: The domain gaps in monocular 3D object detection are very complex, including focal length gap, camera orientation gap, image appearance gap, etc. (a) Two vehicles of the same 2D and 3D size are taken at different focal lengths, and their depths vary dramatically. (b) A higher pitch angle of the camera causes objects to appear lower in the image, leading to the trained model predicting closer depths for the objects. (c) Variations in image appearance, such as adverse weather and simulation data can considerably affect the perceived contextual visual information for the M3OD model.

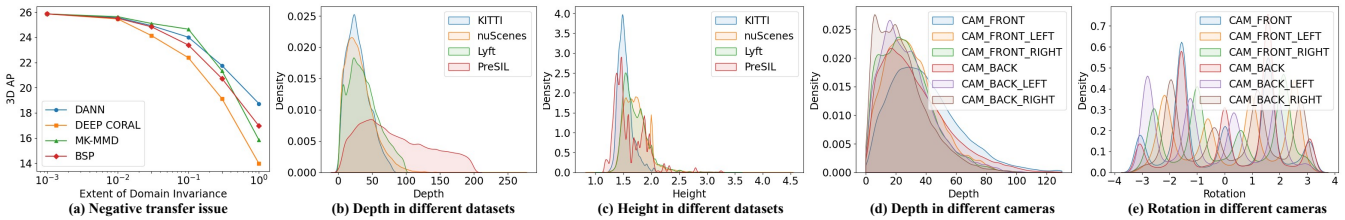

Figure 2: (a) Conventional domain invariant learning techniques often lead to the negative transfer issue in M3OD. The chart shows the DG performance of models trained on nuScenes and Lyft, tested on KITTI. As the extent of domain invariance increases, the accuracy of M3OD will significantly decrease. (b-e) M3OD demonstrates significant geometry-domain dependency, with notable disparity in the geometry distribution of various domains, such as objects' depth, dimension, and rotation.

the most critical task in M3OD, is severely affected by the misalignment caused by domain dependency, consequently impacting the generalization performance.

To tackle these challenges, we propose MonoGDG, a **G**eometry-Guided **D**omain **G**eneralization framework for **Mono**cular 3D Object Detection. We address the DG challenges in M3OD from the camera and the feature aspects. Specifically, at the camera level, we propose a geometry-based image reprojection strategy that unifies intrinsic parameters, randomizes camera orientations, and unifies the FOV range. These simple yet effective techniques significantly alleviate the camera domain gaps (Table 1). At the feature level, we propose a geometry-dependent feature disentanglement algorithm to mitigate the negative transfer issue. Rather than excluding domain-specific information, which can potentially disrupt geometry features, our algorithm disentangles and integrate both domain-shared and domain-specific features. Moreover, a depth-disentangled domain discriminator is utilized within the domain-shared branch to reduce misalignment among objects with varying depths. A domain-aware geometry regression attention is further used to integrate domain and geometry features. We conduct extensive experiments on multiple datasets for domain-generalizable M3OD in autonomous driving. The results demonstrate that our proposed method considerably outperforms existing methods, achieving state-of-the-art performance. Moreover, we utilize simulation data to enhance the model's DG performance in adverse weather conditions, improving its robustness in real-world scenarios.

In summary, our contributions are three-fold:

(1) We systematically analyze the complex domain gaps in M3OD, identify the negative transfer issue caused by geometry-domain dependency, and propose leveraging geometry strategy to guide the domain generalization.

(2) At the camera level, we introduce geometry-based image reprojection mechanism to address the camera parameter disparity. At the feature level, we propose geometry-dependent feature disentanglement to tackle the negative transfer issue.

(3) Through extensive experiments on various datasets, our proposed MonoGDG achieves state-of-the-art performance and significantly enhances the models' domain generalization capabilities.

Related Work

Domain Generalization

Domain generalization aims to generalize to unseen target domains (Erfani et al. 2016). Feature alignment is widely used for acquiring domain-invariant representations (Xiong et al. 2023). There are many approaches to achieving feature alignment (Zhou et al. 2023), such as minimizing moments (Muandet, Balduzzi, and Schölkopf 2013), minimiz-

Domain Gap	STMono3D	DGMono3D	MonoGDG (ours)
Focal Length	✓	✓	✓
FOV Distortion	✗	✓	✓
FOV Range	✗	✗	✓
Camera Orientation	✗	✗	✓
Image Appearance	✗	✗	✓
Target Data Free	✗	✓	✓

Table 1: Comparison of the proposed domain generalization method with other state-of-the-art baselines.

ing contrastive loss (Motiian et al. 2017), minimizing KL divergence (Wang, Loog, and van Gemert 2020), minimizing maximum mean discrepancy (Li et al. 2018a), and domain adversarial learning (Li et al. 2018b; Shao et al. 2019).

Domain Generalization in M3OD

FCOS3D (Wang et al. 2021) and CenterNet (Zhou, Wang, and Krähenbühl 2019) realize M3OD through a simple network architecture. Recently, researchers have attempted to incorporate geometry priors into 3D object detection and achieved encouraging results (Lu et al. 2021; Shi et al. 2021; Yang et al. 2022, 2023). Deep3DBox (Mousavian et al. 2017) employs a neural network to predict objects' rotation, dimension, and 2D bounding box, providing constraints for the 3D bounding box estimation. There is limited work on domain generalization in M3OD. STMono3D (Li et al. 2022b) explores domain adaptation in M3OD and utilizes the geometry-aligned multi-scale training and self-teacher methods. DGMono3D (Li et al. 2022a) explores single-source domain generalization in M3OD and employed object scaling and 2D-3D geometry-consistent strategies. OMNI3D (Brazil et al. 2023) combines existing M3OD datasets and proposes CubeRCNN, aimed at multi-dataset fully supervised training rather than domain generalization in unknown target domains. These methods primarily consider the domain gap within the camera, without consideration of the complex gap factors of M3OD, such as camera orientation, image appearance, etc.

Methodology

Our approach, as depicted in Fig. 3, comprises geometry-based image reprojection at the camera level and geometry-dependent feature disentanglement at the feature level.

Geometry-Based Image Reprojection

Variations in camera parameters across different domains can significantly impact M3OD (Fig. 1). To enhance the robustness of the detector, we propose an image reprojection mechanism, which aims to transform the image into a generalizable meta-camera, thereby improving the generalization capability at the camera level.

Intrinsic parameter unification To address the intrinsic parameters gap among different cameras, we utilize a reprojection approach that aligns all images from different domains to a common perspective meta-camera. This perspective meta-camera possesses uniform intrinsic parameters, effectively mitigating the domain gap in intrinsic parameters.

Given a source domain camera C^i, which has intrinsic parameter matrix $\mathbf{K^i}$, we could obtain its projection formula:

$$Z[x^i, y^i, 1]^\top = \mathbf{K^i}[X, Y, Z]^\top \tag{1}$$

$$\mathbf{K^i} = \begin{bmatrix} f_x^i & 0 & c_x^i \\ 0 & f_y^i & c_y^i \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where x^i, y^i are a pixel's image coordinate in camera C^i, X, Y, Z are its spatial position and $f_x^i, f_y^i, c_x^i, c_y^i$ are the intrinsic parameters of the camera C^i.

We define the intrinsic parameter matrix of the perspective meta-camera C^m as $\mathbf{K^m}$. $f_x^m, f_y^m, c_x^m, c_y^m$ are its intrinsic parameters. We project the above spatial points $[X, Y, Z]$ onto the meta-camera C^m's image pixel x^m, y^m:

$$Z[x^m, y^m, 1]^\top = \mathbf{K^m}[X, Y, Z]^\top \tag{3}$$

Using Eq. (1), Eq. (3) could be replaced:

$$Z[x^m, y^m, 1]^\top = \mathbf{K^m}\mathbf{K^i}^{-1} Z[x^i, y^i, 1]^\top \tag{4}$$

Since Z is a scalar, Z can be eliminated from both sides of the Eq. (4). By solving Eq. (4), we can obtain the reprojection conversion formula from source camera pixel coordinate x^i, y^i to perspective meta-camera pixel x^m, y^m:

$$\begin{aligned} x^m &= \frac{f_x^m}{f_x^i} x^i + c_x^m - \frac{f_x^m}{f_x^i} c_x^i \\ y^m &= \frac{f_y^m}{f_y^i} y^i + c_y^m - \frac{f_y^m}{f_y^i} c_y^i \end{aligned} \tag{5}$$

Eq. (5) is simply a linear transformation with scaling and translation. As a result, we can easily reproject images from different cameras to the same perspective meta-camera, achieving intrinsic parameters unification.

Camera orientation randomization Camera extrinsic parameters, including camera orientation and position, considerably impact 3D detection. However, perspective transformations on the camera position are not viable due to the lack of pixel-wise depth annotation information, dictated by geometry principles (Zhao, Kong, and Fowlkes 2021). Additionally, transforming the camera orientation is also a challenging process (Dubrofsky 2009).

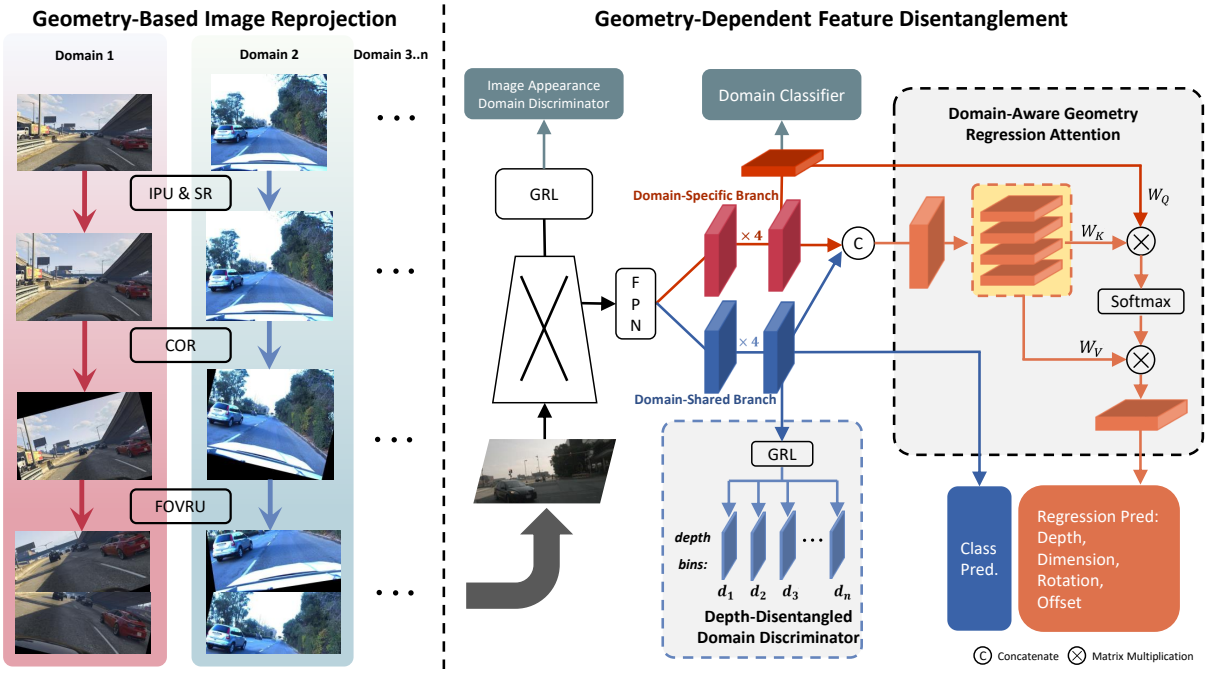

Figure 3: Overview of the proposed MonoGDG. At the camera level, the Geometry-Based Image Reprojection process is applied to images to address the domain gap of the camera, including Intrinsic Parameter Unification (IPU), Spherical Reprojection (SR), Camera Orientation Randomization (COR), and FOV Range Unification (FOVRU). The extracted features from images then undergo Geometry-Dependent Feature Disentanglement, which disentangles the feature into domain-shared and domain-specific branches. Depth-Disentangled Domain Discriminator disentangles the depth from domain alignment, and Domain-Aware Geometry Regression Attention is employed to integrate the domain and geometry features. GRL denotes the gradient reversal layer.

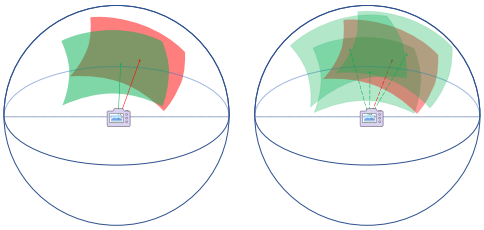

Figure 4: Left: The camera orientations and image field during training and testing are different. Right: Camera orientation randomization is performed in the spherical camera during training to make the model agnostic to camera orientation.

The spherical camera eliminates the variance in different view angles (Gu et al. 2021). By reprojecting images onto a spherical camera, we can achieve camera orientation randomization through simple image translation and rotation (Fig. 4), ensuring that the detector becomes agnostic to the camera orientation. Our approach does not introduce any perspective stretch or distortion, making it not only simple to implement but also mathematically rigorous.

First of all, we reproject the coordinates x^m, y^m from the perspective meta-camera onto the spherical meta-camera u^m, v^m:

$$u^m = f_x^m \arctan(\frac{x^m - c_x^m}{f_x^m}) + c_x^s$$
$$v^m = f_y^m \arctan(\frac{y^m - c_y^m}{f_y^m}) + c_y^s \quad (6)$$

where c_x^s, c_y^s are the new principal point of the spherical meta-camera, the spherical meta-camera has the same focal length as the perspective meta-camera.

The camera orientation can be described by Euler angles, including pitch, roll, and yaw. In the spherical meta-camera, We can randomize the camera's roll angle by rotating the image field, the pitch angle by translating the image field in the vertical direction, and the yaw angle by translating the image field in the horizontal direction.

Assuming the roll transformation is θ_r, the matrix of image rotation around the principal point c_x^s, c_y^s is as follows:

$$\mathbf{M_R} = \begin{bmatrix} \cos\theta_r & -\sin\theta_r & (1-\cos\theta_r) * c_x^s + \sin\theta_r * c_y^s \\ \sin\theta_r & \cos\theta_r & (1-\cos\theta_r) * c_y^s + \sin\theta_r * c_x^s \\ 0 & 0 & 1 \end{bmatrix} \quad (7)$$

	Cams	Images (train+val)	Resolution	Horizon FOV	Vertical FOV	Locations
KITTI	1	7481	1242x375	81	29	Germany
nuScenes	5	204894	1600x900	65	39	BO, SG
	1			90	59	
Lyft	6	136080	1224x1024	70	60	Palo Alto
			1920x1080	82	52	
PreSIL	1	51075	1920x1080	90	59	GTA V

Table 2: Different M3OD datasets have different cameras and fields of view. "BO" and "SG" are short for "Boston" and "Singapore", respectively.

Assuming the pitch transformation is θ_p, the yaw transformation is θ_y, then the image translation matrix is:

$$\mathbf{M_T} = \begin{bmatrix} 1 & 0 & f_x^m \theta_y \\ 0 & 1 & f_y^m \theta_p \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

Therefore, in the spherical meta-camera, the image transformation formula deriving from camera orientation randomization is as follows:

$$[\hat{u}^m, \hat{v}^m, 1]^\top = \mathbf{M_T} \mathbf{M_R} [u^m, v^m, 1]^\top \tag{9}$$

where $\hat{u}^m, \hat{v}^m$ are the image pixel after camera orientation randomization.

FOV range unification and image appearance domain discriminator Different cameras have different field of view (FOV) ranges (Table 2). To ensure that images from different domains possess a consistent FOV range, we just randomly crop the images to the same FOV. Aligning the FOV ranges is essential for domain adversarial learning at feature-level generalization. The discrepancies among different FOV ranges are highly noticeable, which can impact the ability of the domain discriminator to focus on discrepancies in image appearances and geometry clues. Additionally, to confuse the discriminator, the feature extractors would have to equalize the global information amount of images from different FOV ranges, which would undermine the global information.

To address image appearance domain gaps, like fog, rain, and simulation images, we propose using an image appearance domain discriminator and gradient reversal layer (GRL). This reduces the impact of texture changes on the neural network, enabling the detector to focus on object shapes instead of texture. It enhances the detector's generalization ability when facing unknown appearance variations.

Geometry-Dependent Feature Disentanglement

Theoretical analysis The distribution of objects' 3D geometry, including depth, dimension, and rotation, is dependent on the domain (Fig. 2). This correlation is influenced by multiple factors:

(1) Camera hardware and annotation standards vary in different datasets, affecting the ability to capture distant objects and the distribution of depth annotations. Newer datasets employ advanced cameras with high resolution, enabling better visibility and clearer annotations for distant objects. In contrast, older datasets primarily focus on objects nearby. Synthetic datasets, however, include annotations for objects at much further distances.

(2) Different datasets collected in diverse regions result in variations in vehicle sizes. And the depth distribution significantly varies between urban roads, rural roads, and highways.

(3) Cameras at different viewpoints capture different scenes. Object vehicles exhibit varied rotation angles from different viewpoints. Moreover, front-facing and rear-facing cameras tend to capture objects at greater depths, while side-facing cameras capture objects at closer depths.

Due to the existence of the geometry-domain dependency, using conventional domain invariant learning would undermine geometry features. Taking domain adversarial learning (Goodfellow et al. 2020; Li et al. 2018b) as an example, we define E, M, D as the parameters of the encoder, 3D detection head, and domain discriminator, respectively. Then the objective is as follows:

$$\min_{E,M} \max_{D} L(E, M, D) = \mathbb{E}_{p(x,d,y)}[-\lambda L_d + L_{3D}] \tag{10}$$

where x, d, y are the input data, domain, and geometry label. λ is a hyperparameter. L_{3D} is the 3D detection loss. L_d is the cross-entropy loss for the domain discriminator. According to Akuzawa et al (Akuzawa, Iwasawa, and Matsuo 2019), the optimization goal of the encoder is:

$$\min_{E} L(E) = -\lambda H(d|h) + \mathbb{E}_{p_E(h,d,y)} L_{3D} \tag{11}$$

where H, h denotes entropy, and latent feature, respectively.

The encoder aims to maximize $H(d|h)$, which has $H(d)$ as its upper bound in light of the entropy properties. Therefore, when achieving domain invariance, $H(d|h)$ equals $H(d)$. Additionally, the geometry-domain dependency implies that their mutual information entropy, denoted as $I(y, d)$, is greater than 0. Based on these two prerequisites, we propose the following theory:

Theorem 1 *if* $I(y, d) = H(d) - H(d|y) > 0$*, when* $H(d|h) = H(d)$*,* $H(y|h) > 0$*.*

Proof 1 *According to the properties of entropy:*

$$H(d|h) \leq H(d, y|h) = H(d|h, y) + H(y|h) \tag{12}$$

Since $H(d|h) = H(d)$*:*

$$H(d|h, y) = H(d|y) \tag{13}$$

Replace Eq. (12) with Eq. (13):

$$H(d|h) \leq H(d|y) + H(y|h) \tag{14}$$

Because $H(d|y) < H(d)$*, Eq. (14) can be expand as:*

$$H(d|h) \leq H(d|y) + H(y|h) < H(d) + H(y|h) \tag{15}$$

When $H(d|h) = H(d)$*, Eq. (15) can be substituted:*

$$H(d) < H(d) + H(y|h) \tag{16}$$

Therefore:

$$H(y|h) > 0 \tag{17}$$

Consequently, in the context of monocular 3D object detection, if the geometry label and domain are interdependent, then $I(y, d) > 0$. When the feature h removes domain information, $H(d|h) = H(d)$. According to Theorem 1, we have $H(y|h) > 0$, indicating that the geometry features suffer damage. As a result, a negative transfer issue will appear in traditional methods for M3OD.

Algorithm To tackle the negative transfer issue, we propose a geometry-dependent feature disentanglement approach. Rather than using conventional domain invariant learning to eliminate domain information, we disentangle the features into domain-specific and domain-shared branches, and leverage both features to enhance the geometry tasks effectively.

In the domain-specific branch, we use a domain classifier (without the gradient reversal layer(GRL)) to extract domain-specific features. For domain-shared features, we employ the GRL and object-level depth-disentangled domain discriminators. Depth is widely recognized as a crucial and challenging factor in M3OD, and misalignment of depth poses a significant issue. To tackle this, we propose the object-level depth-disentangled domain discriminator, decoupling domain adversarial learning from depth.

In detail, we partition the continuous depth into K bins: $[d_1, d_2, \ldots, d_K]$ and utilize K domain discriminators, assigning one discriminator to each bin. Each discriminator performs domain adversarial alignment to objects with depth in the dedicated bin. This ensures accurate alignment without misaligning objects with significantly different depths.

Semantic classification tasks in autonomous driving scenes are relatively simple, making aligning decision boundaries between domains easy. We use the feature from the domain-shared branch for semantic classification. In contrast, geometry regression tasks are continuous and challenging to align decision boundaries across domains due to strong domain dependency. We utilize features from both domain-specific and domain-shared branches.

Furthermore, we propose a domain-aware geometry regression attention mechanism to enhance the integration of domain-specific and domain-shared information for 3D geometry regression. In detail, as shown in Fig. 3, we first concatenate features from the domain-specific and domain-shared branches. These concatenated features are then divided into N groups, denoted as F_1, F_2, ..., F_N. Each group of features is responsible for the geometry regression task within a specific region of the feature space. We compute the keys and values (Vaswani et al. 2017) using these N groups of features:

$$K_i = F_i \times W_K, V_i = F_i \times W_V \tag{18}$$

We utilize the domain information h extracted by the domain classifier as the query:

$$Q = h \times W_Q \tag{19}$$

Finally, the domain-aware geometry regression attention feature Z is obtained through the following computation and used for geometry regression tasks:

$$Z = Softmax(Q\ K^T / \sqrt{d_k})\ V \tag{20}$$

Experiments

Setup and Implementation Details

Following (Li et al. 2022b,a), we subsample 1/4 data for nuScenes (Caesar et al. 2020), Lyft (Kesten et al. 2019), and PreSIL (Hurl, Czarnecki, and Waslander 2019) datasets, and use the FCOS3D as the detector for experiments. Following the evaluation metrics in STMono3D (Li et al. 2022b), we use the official AP_{11} and AP_{40} and the IoU 0.5 for KITTI (Geiger, Lenz, and Urtasun 2012). Both nuScenes and Lyft provide images from six different camera views. The images within a camera view constitute a source domain. NuScenes and Lyft datasets are each divided into 6 source domains. We employ cross-entropy loss for classification and SmoothL1Loss for regression task, with SGD optimizer and learning rate 0.001 (Ruder 2016).

Comparison with State-of-the-art Methods

DG performance in common autonomous driving benchmarks Table 3 illustrates the DG performance of different methods. Training Source Only on the source domain yields poor generalization performance. Oracle denotes full supervision training on the target domain. STMono3D follows the domain adaptation setting by incorporating target domain data during training. The KITTI, nuScenes, and Lyft datasets are collected from distinct cities with diverse road environments (Table 2), exhibiting considerable camera orientation and image appearance shifts. Unfortunately, the existing methods inadequately handle these gaps, resulting in mediocre performance. In contrast, our approach comprehensively tackles domain gaps and mitigates the geometry-domain dependency, achieving state-of-the-art performance. Notably, our method surpasses Oracle in most scenarios.

DG performance with simulation and adverse weather data Both Table 4 and Table 5 utilize the nuScenes and simulation dataset PreSIL as source domains. Table 5 shows the DG performance in heavy fog and rain. The simulation, heavy fog (Mai et al. 2021), and rain (Halder, Lalonde, and de Charette 2019) considerably affect the appearance of images, posing a challenge to the robustness of neural networks

nuScenes→ KITTI	AP_{11}						AP_{40}					
Method	BEV			3D			BEV			3D		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
Source Only	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Oracle	33.46	23.62	22.18	29.01	19.88	17.17	33.70	23.22	20.68	28.33	18.97	16.57
STMono3D	35.63	27.37	23.95	28.65	21.89	19.55	31.85	22.82	19.30	24.00	16.85	13.66
DGMono3D	34.22	28.99	27.82	28.77	24.82	23.67	31.90	26.33	25.60	23.20	18.55	17.72
MonoGDG (ours)	**39.50**	**32.39**	**32.07**	**33.48**	**27.14**	**26.37**	**35.73**	**28.16**	**27.71**	**29.30**	**22.01**	**21.40**

Lyft→KITTI	AP_{11}						Lyft→nuScenes	Metrics			
Method	BEV			3D			Method	AP	ATE	ASE	AOE
	Easy	Mod	Hard	Easy	Mod	Hard					
Source Only	0.00	0.00	0.00	0.00	0.00	0.00	Source Only	2.40	1.302	0.190	0.802
Oracle	33.46	23.62	22.18	29.01	19.88	17.17	Oracle	**28.20**	**0.798**	0.160	0.209
STMono3D	26.46	20.71	17.66	18.14	13.32	11.83	STMono3D	21.30	0.911	0.170	0.355
DGMono3D	36.18	28.30	27.16	30.03	23.38	22.23	DGMono3D	25.50	0.842	0.169	0.208
MonoGDG (ours)	**38.47**	**30.89**	**29.58**	**32.48**	**26.02**	**24.96**	MonoGDG (Ours)	25.97	0.828	**0.158**	**0.194**

Table 3: DG performance of various methods when generalizing from nuScenes to KITTI, Lyft to KITTI, and Lyft to nuScenes.

P+n→KIT	AP_{40} BEV			AP_{40} 3D		
Method	Easy	Mod	Hard	Easy	Mod	Hard
Oracle	33.70	23.22	20.68	28.33	18.97	16.57
STMono3D	32.47	23.35	19.81	24.43	17.37	14.29
DGMono3D	33.81	27.27	26.84	24.75	20.08	19.58
Ours	**38.35**	**29.48**	**28.71**	**31.55**	**23.31**	**22.25**

Table 4: DG performance of PreSIL+nuScenes→KITTI.

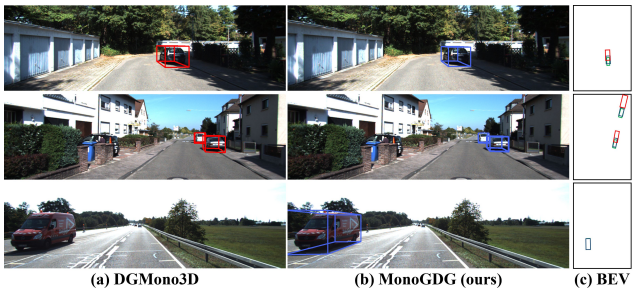

Figure 5: The 3D BBox and BEV prediction from DGMono3D (in red), MonoGDG (in blue), and ground truth (green in BEV) in PreSIL+nuScenes→KITTI setting. Zoom in for a clear comparison.

in handling texture variations. By utilizing our method's image appearance domain discriminator and feature decoupling, the neural network can recognize objects based on their shape rather than texture, thereby improving its generalization when facing changes in image appearance.

Ablation Study and Analysis

The effectiveness of geometry-based image reprojection. In Table 6, the baseline, Exp. (a), incorporates geometry-dependent feature disentanglement at the feature level but does not utilize any image reprojection techniques. The significant improvement from Exp. (a) to Exp. (b) illustrates that a unified camera intrinsic parameter can solve the intrinsic parameter gap, which is crucial for M3OD. Exp. (c) further improves the accuracy compared to Exp. (b) by employing FOV range unification. It allows the domain discriminator to focus on more fundamental domain discrepancy rather than simply distinguishing FOV ranges, thus enabling it to perform the intended function. Exp. (d) incorporates camera orientation randomization, which effectively prevents the detector from being biased towards specific camera orientation settings, thus achieving camera orientation agnosticism. By integrating the three aforementioned image reprojection techniques, Exp. (e) successfully eliminates domain gaps at the camera level, leading to the best experimental performance.

The effectiveness of geometry-dependent feature disentanglement. In Table 7, Exp. (a) does not use the feature disentanglement. Exp. (b) add domain invariant learning, decreasing AP due to the identified negative transfer issue. Exp. (c) disentangles the features and utilizes domain-specific and domain-shared features, resulting in much improvement. Comparing Exp. (d) and Exp. (c), the domain-aware geometry regression attention integrates domain and geometry features more effectively. Exp. (e) replace the domain discriminator with a depth-disentangled version, resulting in improvements over Exp. (c). The depth-disentangled discriminator effectively mitigates the misalignment issue among objects with varying depths. Exp. (f) is the complete version, tackling the geometry-domain dependency, and achieving the best performance.

The effectiveness of image appearance domain discriminator. As shown in Table 8, the image appearance domain discriminator encourages the detector to focus more on object shape rather than texture features, thus improving its generalization when facing appearance variations.

Training: Pre+nus	Test: fog KITTI AP_{40} 3D			Test: rain KITTI AP_{40} 3D		
Method	Easy	Mod	Hard	Easy	Mod	Hard
Oracle	22.25	15.88	14.10	25.58	15.71	14.21
STMono3D	19.59	14.63	13.35	23.18	16.46	13.59
DGMono3D	17.64	12.52	11.84	22.30	15.92	15.26
MonoGDG (ours)	**23.58**	**15.97**	**15.28**	**27.69**	**18.33**	**17.79**

Table 5: DG performance of PreSIL+nuScenes→ fog KITTI and PreSIL+nuScenes→ rain KITTI dataset.

n→K	IPU	COR	FOVRU	AP_{40} BEV			AP_{40} 3D		
Exp				Easy	Mod.	Hard	Easy	Mod.	Hard
(a)	-	-	-	13.76	8.92	7.53	9.48	5.27	4.84
(b)	✓	-	-	28.72	22.17	21.51	22.10	16.94	16.58
(c)	✓	-	✓	32.69	25.47	23.94	25.85	19.48	18.71
(d)	✓	✓	-	31.74	25.61	24.73	25.21	21.62	20.11
(e)	✓	✓	✓	**35.73**	**28.16**	**27.71**	**29.30**	**22.01**	**21.40**

Table 6: Ablation study on Intrisic Parameter Unification (IPU), Camera Orientation Randomization (COR), and Field of View Range Unification (FOVRU).

n→K	DIL	DR	DAGRA	DDDD	AP_{40} BEV			AP_{40} 3D		
Exp					Easy	Mod.	Hard	Easy	Mod.	Hard
(a)	-	-	-	-	26.43	19.42	19.06	20.37	15.82	15.25
(b)	✓	-	-	-	23.57	16.97	16.31	16.26	10.41	9.93
(c)	-	✓	-	-	28.43	20.95	20.14	22.15	16.36	15.87
(d)	-	✓	✓	-	31.54	27.02	26.25	26.82	20.73	20.02
(e)	-	✓	-	✓	31.88	26.45	25.89	26.92	20.59	19.30
(f)	-	✓	✓	✓	**35.73**	**28.16**	**27.71**	**29.30**	**22.01**	**21.40**

Table 7: Ablation study on domain invariant learning (DIL), disentangled representation (DR), domain-aware geometry regression attention (DAGRA), and depth-disentangled domain discriminator (DDDD).

P+n→KIT	AP_{40} BEV			AP_{40} 3D		
IADD	Easy	Mod	Hard	Easy	Mod	Hard
-	35.48	28.14	27.56	28.17	21.94	21.37
✓	**38.35**	**29.48**	**28.71**	**31.55**	**23.31**	**22.25**

Table 8: Effectiveness of image appearance domain discriminator (IADD).

n→K	AP_{40} BEV			AP_{40} 3D		
Method	Easy	Mod	Hard	Easy	Mod	Hard
GAMS	32.52	24.96	24.49	27.15	19.84	19.03
IPU	**35.73**	**28.16**	**27.71**	**29.30**	**22.01**	**21.40**

Table 9: For the camera focal length gap, the comparison between our IPU and STMono3D's GAMS.

Comparison with other focal length processing methods. Table 9 demonstrates the superiority of our proposed intrinsic parameter unification over the GAMS in STMono3D (Li et al. 2022b). GAMS preset multiple fixed focal lengths during training, which may lead to a performance decline when the focal length deviates from the preset values in unknown target domains.

Visualization Results

In Fig. 5, we compare the 3D BBox predictions from DGMono3D and MonoGDG. When the camera is tilted relative to the ground, DGMono3D exhibits inaccurate depth predictions due to its lack of consideration for camera orientation gaps. Moreover, DGMono3D fails to handle the image appearance gap, leading to a failure in detecting the red van in the third image. Additionally, our approach addresses the geometry-domain dependency, improving the generalization performance of rotation, dimensions, and depth of objects.

Conclusion

We propose MonoGDG to address the domain generalization challenges for M3OD. Firstly, we introduce geometry-based image reprojection to bridge domain gaps at the camera level. Furthermore, we propose geometry-dependent feature disentanglement to mitigate the negative transfer issue at the feature level. Extensive experimental results demonstrate the remarkable effectiveness of the proposed method.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 61925107, 62271281, U1936202, 62021002), Zhejiang Provincial Natural Science Foundation of China under Grant (No. LDT23F01013F01), and CCF-DiDi GAIA Collaborative Research Funds for Young Scholars.

References

Akuzawa, K.; Iwasawa, Y.; and Matsuo, Y. 2019. Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization. In *Machine Learning and Knowledge Discovery in Databases*, volume 11907, 315–331.

Brazil, G.; Kumar, A.; Straub, J.; Ravi, N.; Johnson, J.; and Gkioxari, G. 2023. Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13154–13164.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11618–11628.

Chen, X.; Wang, S.; Long, M.; and Wang, J. 2019. Transferability vs. Discriminability: Batch Spectral Penalization for Adversarial Domain Adaptation. In *International Conference on Machine Learning*, volume 97, 1081–1090.

Dubrofsky, E. 2009. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5.

Erfani, S. M.; Baktashmotlagh, M.; Moshtaghi, M.; Nguyen, V.; Leckie, C.; Bailey, J.; and Ramamohanarao, K. 2016. Robust Domain Generalisation by Enforcing Distribution Invariance. In *International Joint Conference on Artificial Intelligence*, 1455–1461.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1): 2096–2030.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3354–3361.

Ghifary, M.; Balduzzi, D.; Kleijn, W. B.; and Zhang, M. 2017. Scatter Component Analysis: A Unified Framework for Domain Adaptation and Domain Generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7): 1414–1430.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gu, Q.; Zhou, Q.; Xu, M.; Feng, Z.; Cheng, G.; Lu, X.; Shi, J.; and Ma, L. 2021. PIT: Position-Invariant Transform for Cross-FoV Domain Adaptation. In *IEEE/CVF International Conference on Computer Vision*, 8741–8750.

Halder, S. S.; Lalonde, J.; and de Charette, R. 2019. Physics-Based Rendering for Improving Robustness to Rain. In *IEEE/CVF International Conference on Computer Vision*, 10202–10211.

Hendrycks, D.; and Dietterich, T. G. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Hurl, B.; Czarnecki, K.; and Waslander, S. L. 2019. Precise Synthetic Image and LiDAR (PreSIL) Dataset for Autonomous Vehicle Perception. In *IEEE Intelligent Vehicles Symposium*, 2522–2529.

Kesten, R.; Usman, M.; Houston, J.; Pandya, T.; Nadhamuni, K.; Ferreira, A.; Yuan, M.; Low, B.; Jain, A.; Ondruska, P.; Omari, S.; Shah, S.; Kulkarni, A.; Kazakova, A.; Tao, C.; Platinsky, L.; Jiang, W.; and Shet, V. 2019. Lyft Level 5 AV Dataset 2019. https://level5.lyft.com/dataset/. Accessed: 2023-02-07.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018a. Domain Generalization With Adversarial Feature Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5400–5409.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018b. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In *European Conference on Computer Vision*, volume 11219, 647–663.

Li, Z.; Chen, Z.; Li, A.; Fang, L.; Jiang, Q.; Liu, X.; and Jiang, J. 2022a. Towards model generalization for monocular 3d object detection. arXiv:2205.11664.

Li, Z.; Chen, Z.; Li, A.; Fang, L.; Jiang, Q.; Liu, X.; and Jiang, J. 2022b. Unsupervised Domain Adaptation for Monocular 3D Object Detection via Self-training. In *European Conference on Computer Vision*, volume 13669, 245–262.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning Transferable Features with Deep Adaptation Networks. In *International Conference on Machine Learning*, volume 37, 97–105.

Lu, Y.; Ma, X.; Yang, L.; Zhang, T.; Liu, Y.; Chu, Q.; Yan, J.; and Ouyang, W. 2021. Geometry Uncertainty Projection Network for Monocular 3D Object Detection. In *IEEE/CVF International Conference on Computer Vision*, 3091–3101.

Mai, N. A. M.; Duthon, P.; Khoudour, L.; Crouzil, A.; and Velastin, S. A. 2021. 3D Object Detection with SLS-Fusion Network in Foggy Weather Conditions. *Sensors*, 21(20): 6711.

Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified Deep Supervised Domain Adaptation and Generalization. In *IEEE International Conference on Computer Vision*, 5716–5726.

Mousavian, A.; Anguelov, D.; Flynn, J.; and Kosecka, J. 2017. 3D Bounding Box Estimation Using Deep Learning and Geometry. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5632–5640.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain Generalization via Invariant Feature Representation. In *International Conference on Machine Learning*, volume 28, 10–18.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do ImageNet Classifiers Generalize to ImageNet? In *International Conference on Machine Learning*, volume 97, 5389–5400.

Ruder, S. 2016. An overview of gradient descent optimization algorithms. arXiv:1609.04747.

Shao, R.; Lan, X.; Li, J.; and Yuen, P. C. 2019. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10023–10031.

Shi, X.; Ye, Q.; Chen, X.; Chen, C.; Chen, Z.; and Kim, T. 2021. Geometry-based Distance Decomposition for Monocular 3D Object Detection. In *IEEE/CVF International Conference on Computer Vision*, 15152–15161.

Sun, B.; and Saenko, K. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *European Conference on Computer Vision*, volume 9915, 443–450.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021. FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection. In *IEEE/CVF International Conference on Computer Vision Workshops*, 913–922.

Wang, Z.; Loog, M.; and van Gemert, J. 2020. Respecting Domain Relations: Hypothesis Invariance for Domain Generalization. In *International Conference on Pattern Recognition*, 9756–9763.

Xie, Q.; Dai, Z.; Du, Y.; Hovy, E. H.; and Neubig, G. 2017. Controllable Invariance through Adversarial Feature Learning. In *Advances in Neural Information Processing Systems*, 585–596.

Xiong, Y.; Chen, H.; Lin, Z.; Zhao, S.; and Ding, G. 2023. Confidence-based Visual Dispersal for Few-shot Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11621–11631.

Yang, F.; Xu, X.; Chen, H.; Guo, Y.; Han, J.; Ni, K.; and Ding, G. 2022. Ground Plane Matters: Picking Up Ground Plane Prior in Monocular 3D Object Detection. arXiv:2211.01556.

Yang, F.; Xu, X.; Chen, H.; Guo, Y.; He, Y.; Ni, K.; and Ding, G. 2023. GPro3D: Deriving 3D BBox from ground plane in monocular 3D object detection. *Neurocomputing*, 562: 126894.

Zhao, Y.; Kong, S.; and Fowlkes, C. C. 2021. Camera Pose Matters: Improving Depth Prediction by Mitigating Pose Distribution Bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15759–15768.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2023. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4396–4415.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as points. arXiv:1904.07850.