Towards Generalizable Multi-Camera 3D Object Detection via Perspective Rendering

Hao Lu[1,2], **Yunpeng Zhang**[3], **Guoqing Wang**[4], **Qing Lian**[2], **Dalong Du**[3], **Yingcong Chen**[1,2,*]

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] The Hong Kong University of Science and Technology
[3] PhiGent Robotics
[4]Shanghai Jiao Tong University

Abstract

Detecting and localizing objects in 3D space using multiple cameras, known as Multi-Camera 3D Object Detection (MC3D-Det), has gained prominence with the advent of bird's-eye view (BEV) approaches. However, these methods often struggle with the serious domain gaps caused by various viewpoints and environments between the training and testing domains. To address this challenge, we propose a novel framework that aligns 3D detection with 2D camera plane results by perspective rendering, thus achieving consistent and accurate results when facing serious domain shifts. Our approach consists of two main steps in both source and target domains: 1) rendering diverse view maps from BEV features by leveraging implicit foreground volumes and 2) rectifying the perspective bias of these maps. This design promotes the learning of perspective- and context-independent features, crucial for accurate object detection across varying viewpoints, camera parameters, and environmental conditions. Notably, our model-agnostic approach preserves the original network structure without incurring additional inference costs, facilitating seamless integration across various models and simplifying deployment. Worth noting is that our approach achieves satisfactory results in real data when trained only with virtual datasets, eliminating the need for real scene annotations. Experimental results on both Domain Generalization (DG) and Unsupervised Domain Adaptation (UDA) demonstrate its effectiveness.

Code —
https://github.com/EnVision-Research/Generalizable-BEV

Introduction

Multi-Camera 3D Object Detection (MC3D-Det) refers to the task of detecting and localizing objects in 3D space using multiple cameras (Ma et al. 2022; Li et al. 2022a). Through the complementary information of multi-view images from various perspectives, MC3D-Det emerges as a potent approach, yielding heightened precision and robust object detection results, especially in scenarios where objects may be occluded or partially visible from certain viewpoints. In recent years, bird's-eye view (BEV) approaches have gained tremendous attention for the MC3D-Det task (Ma et al. 2022; Li et al. 2022a; Liu et al. 2022; Wang et al. 2022; Jiang et al. 2024). Despite the efficacy of multi-camera information fusion strategies, their performance drops severely when the testing environment is significantly different from the training ones.

Two promising directions to alleviate the aforementioned challenge are domain generalization (DG) and unsupervised domain adaptation (UDA). DG methods typically disentangle and eliminate the domain-specific features to improve the generalization capability of unseen domains (Wang et al. 2023a). As for UDA, recent advancements address the domain shifts by generating pseudo labels (Li et al. 2022c; Yuan et al. 2023; Yang et al. 2021; Sun et al. 2024; Cheng and Sun 2024; Jiang et al. 2022) or aligning latent features distribution (Xu et al. 2023b; Wang et al. 2023d; Sun 2024). However, the limited availability of training data across various viewpoints, camera settings, and environmental conditions poses a formidable challenge for visual perception striving to learn robust perspective- and environment-independent features.

Our observations indicate that 2D detection in a single-view (camera plane) often demonstrates better generalization capabilities than multi-camera 3D object detection on the BEV plane, primarily due to inaccurate depth estimation, as illustrated in Fig. 1 (a). However, directly establishing 2D-to-3D consistency in cross-domain scenarios faces the following challenges: (1) Relying solely on existing viewpoints in the source domain for supervision hampers the network to learn perspective-invariant features, especially for different camera parameters (Lian et al. 2022; Yang et al. 2022; Wang et al. 2023e). Hence, leveraging the results of 3D projection is necessary to constrain rendering from a broader range of viewpoints. (2) Due to the absence of 2D supervision in the target domain, it is essential to employ a robust 2D detector.

To address the aforementioned issues, we rethink the two key steps in MC3D-Det task: multi-view image features extraction and BEV space transformation. These steps are prone to overfitting due to limited viewpoints, camera parameters, and similar environments, leading to a sharp decrease in generalization performance on cross-domain protocols. We turn these above spatial biases into the bias of a single perspective and define it as the perspective bias.

*Corresponding author.

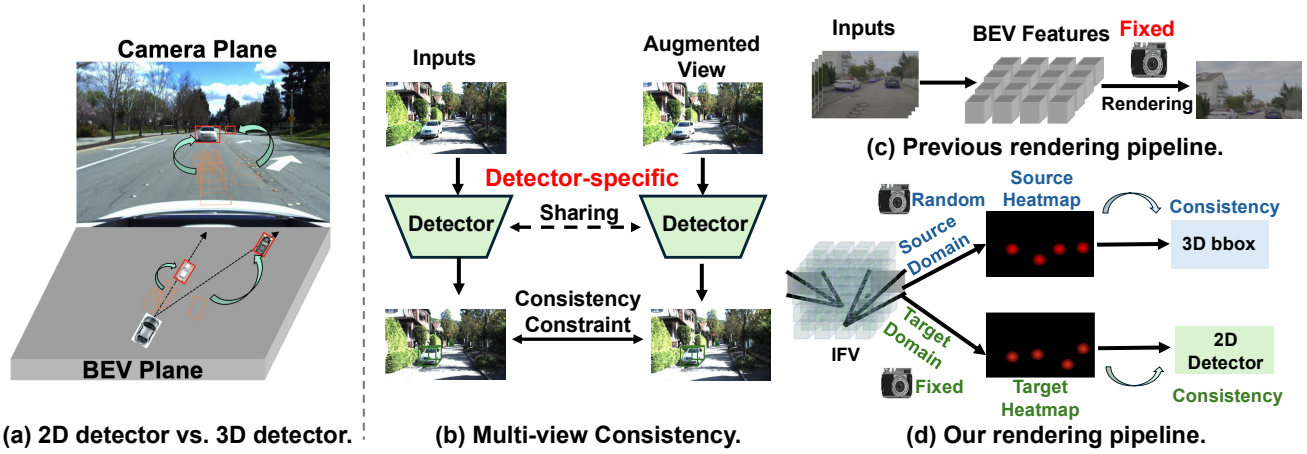

Figure 1: (a) The diagram of our motivation that 2D detectors exhibit remarkable cross-domain performance compared with 3D detectors, which can be effectively utilized to constrain the newly rendered perspective and rectify the spurious BEV features in the target domain. (b) The multi-view consistency methods (Yang et al. 2022; Lian et al. 2022) leverage detector-specific loss to enforce consistency constraints on detection results. (c) The previous rendering pipeline (Wang et al. 2023e) re-renders the BEV features to a new perspective based on the fixed camera parameters. (d) Our proposed rendering pipeline utilizes various camera parameters to re-render the implicit foreground volume features to different heatmaps.

The detailed definition, proof, and discussion are in and supplementary material. Our core insight is that optimizing perspective bias can help alleviate domain shift.

To this end, we introduce a perspective rendering framework to modify the geometry position of BEV space directly, enabling the learning of perspective- and context-invariant features against domain shifts. Our approach involves two main steps: 1) rendering diverse view maps from BEV features and 2) rectifying the perspective bias of these maps. The first step leverages implicit foreground volumes (IFV) to relate the camera and BEV planes, allowing for rendering view maps with varied camera parameters. The second step, in the source domain, uses random camera positions and angles to supervise the camera plane map rendered from IFV, promoting the learning of perspective- and context-independent features. Similarly, in the target domain, a pre-trained 2D detector aids in rectifying BEV features. Notably, our model-agnostic approach preserves the original network structure without incurring additional inference costs, facilitating seamless integration across various models and simplifying deployment. This reduces development and maintenance complexity and ensures efficiency and resource conservation, which are crucial for real-time applications and long-term, large-scale deployments.

Moreover, our framework achieves desirable properties beyond previous 2D-to-3D consistency methods: (1) **The unified method for any detection head.** MC3D-Det supports both anchor-based (Centerpoint) and end-to-end (DETR) detection heads, unlike prior methods (Yang et al. 2022, 2023; Lian et al. 2022; Wang et al. 2023e), which are limited to a single head due to their diverse structures (Fig. 1(b)). (2) **Rendering new perspectives by various camera parameters.** Unlike (Wang et al. 2023e; Pan et al. 2024; Xu et al. 2023a), which rely on fixed camera parameters, our method re-renders BEV features into different heatmaps, supervised by a reliable 2D detector (Fig. 1(c) and (d)). (3) **Significant improvement on source and target domains.** By supervising new perspectives through rendering, our approach surpasses previous 2D-3D consistency methods (Lian et al. 2022; Wang et al. 2023e) on both domains. (4) **Geometric feature correction with less semantic destruction.** Our rendering accurately aligns BEV features with 2D images, refining geometric positions while preserving semantics, unlike prior constraints that disrupt the network and degrade information.

We summarize our core contributions as follows.

- We propose a generalizable MC3D-Det framework based on perspective rendering, which can not only help the model learn the perspective- and context-invariant features in the source domain but also utilize the 2D detector further to correct the spurious geometric features in the target domain.
- We make the first attempt to study unsupervised domain adaptation on MC3D-Det and establish a benchmark. Our approach achieved the state-of-the-art results on both UDA and DG protocols.
- We explore the training on a virtual engine without the real scene annotations to achieve real-world MC3D-Det tasks for the first time.

Related Works

Vision-based 3D Object Detection

In recent years, there has been a growing focus on directly predicting 3D object detection from a single-view image (Wang et al. 2021). However, single-view-based prediction does not effectively combine information from

multiple cameras and lack a holistic perception of the surrounding environment. Therefore, multi-view 3D object detection has gradually become the mainstream research direction. Multi-camera 3D object detection (MC3D-Det) targets to identify and localize objects in 3D space, received widespread attention (Ma et al. 2022; Li et al. 2022a). Recently, most of MC3D-Det methods extract image features and project them onto the bird 's-eye view (BEV) plane in integrating the spatial-temporal feature (Roddick, Kendall, and Cipolla 2019; Philion and Fidler 2020; Li et al. 2022d; Huang et al. 2021; Li et al. 2023b,a; Liu et al. 2022, 2023b). These methods have achieved satisfactory results on the in-distribution dataset but may show very poor results under cross-domain protocols.

Cross-domain Object Detection

Many cross-domain approaches have been designed for 2D detection, such as feature distribution alignment or pseudo-label methods (Muandet, Balduzzi, and Schölkopf 2013; Li et al. 2018; Dou et al. 2019; Facil et al. 2019; Chen et al. 2018; Xu et al. 2020; He and Zhang 2020; Zhao et al. 2020; Yuan et al. 2024b,a). These methods can only solve the domain shift problem caused by environmental changes like rain or low light. For the MC3D-Det task, there is only one study for alleviating the domain shift, which demonstrates that an important factor for MC3D-Det is the overfitting of camera parameters (Wang et al. 2023a). Essentially, the fixed observation perspective and similar road structures in the source domain lead to spurious and deteriorated geometric features. However, without additional supervision, it is very difficult to further extract perspective- and context-independent features on the target domain.

Virtual Engine for Autonomous Driving

The virtual engine has better controllability and can generate various scenarios and samples: domain shift (Sun et al. 2022), vehicle-to-everything (Xu et al. 2022; Li et al. 2022b), corner case (Kim et al. 2022; Wang et al. 2023c). So, breaking the domain gap between virtual and real datasets can further facilitate the closed-loop form of visually-oriented planning (Jia et al. 2023). To our best knowledge, there are no studies that only use a virtual engine without real scenes labels for MC3D-Det.

Preliminaries

Problem Setup

We explore two widely used and practical protocols, namely, domain generalization (DG) and unsupervised domain adaptation (UDA).

• For DG on MC3D-Det task, our primary objective is to leverage solely the labeled data from the source domain $D_S = \{X_s^i, Y_s^i, K_s^i, E_s^i\}$ to improve the generalization of the model. Here, the i-th sample contains N multi-view images $X^i = \{I_1, I_2, ..., I_N\}$ (superscript is omitted for clarity) and the corresponding intrinsic K^i and extrinsic E^i parameters. The source domain label Y_s^i is a set of 3D bounding boxes, each including location, size in each dimension, and orientation.

• For UDA on MC3D-Det task, additional unlabeled target domain data $D_T = \{X_t^i, K_t^i, E_t^i\}$ can be utilized to further improve the generalization of the model. The only difference between DG and UDA is whether the unlabeled data of the target domain can be utilized.

Perspective Bias

To detect the object's location $L = [x, y, z]$ on the BEV space, corresponding to the image plane $[u, v]$, most MC3D-Det methods involve two essential steps: (1) get the image features from multi-view cameras by the image encoder F_{img}. (2) map these features into the BEV space and fuse them to get the final location of objects by the BEV encoder F_{bev}:

$$\begin{aligned} L &= F_{bev}(F_{img}(I_1), ..., F_{img}(I_N), K, E) \\ &= L_{gt} + \Delta L_{img} + \Delta L_{bev}, \end{aligned} \tag{1}$$

where L_{gt}, ΔL_{img}, and ΔL_{bev} are the ground-truth location, the bias from F_{img}, and the bias from F_{bev}. Both ΔL_{img} and ΔL_{bev} are caused by overfitting the limited viewpoints, camera parameters, and similar environments. Without additional supervision in the target domain, ΔL_{img} and ΔL_{img} are difficult to be mitigated. So we turn these spatial biases into the bias of a single perspective. We compute the perspective bias $[\Delta u, \Delta v]$ on the uv image plane as:

$$[\Delta u, \Delta v] = [\frac{k_u(u - c_u) + b_u}{d(u, v)}, \frac{k_v(v - c_v) + b_v}{d(u, v)}]. \tag{2}$$

where $\{k_u, b_u, k_v, b_v\}$ are related to the domain bias of BEV encoder ΔL_{BEV}, and $d(u, v)$ represents the predicted depth information of the model. c_u and c_v represent the coordinates of the camera's optical center in the uv image plane. The detailed proof and discussion are in supplementary material. Eq. 2 provides us with the important inferences: the presence of the final position shift can be reflected in perspective bias, indicating that optimizing perspective bias can help alleviate domain shift. This perspective bias can simultaneously model an incorrect understanding of different camera settings and scenes.

Intuitively, the domain shift generates misplaced BEV features and thereby inaccurate detection results, which arise due to overfitting with limited viewpoint and camera parameters. To mitigate this issue, it is crucial to re-render novel views from BEV features, thereby enabling the network to learn perspective- and environment-independent features. In light of this, we aim to address the perspective bias associated with different rendered viewpoints to enhance the generalization ability of the model.

Method

To reduce the bias stated in Eq. 2, we tailored a generalizable framework (PR-BEV) based on perspective rendering as shown in Fig. 2. Our framework is model-agnostic. For the convenience of demonstration, we select BEVDepth (Li et al. 2023a) as the main pipeline to instantiate our entire process.

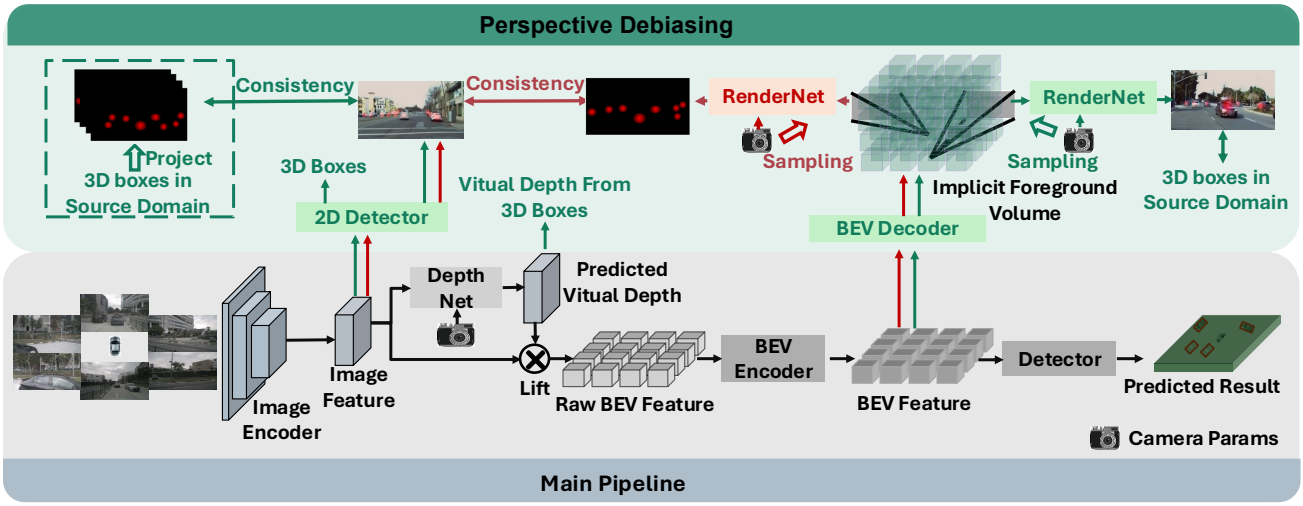

Figure 2: The overview of the perspective rendering framework (PR-BEV) for MC3D-Det. The main pipeline of BEV perception (BEVDepth) is shown in the bottom part of the figure. With the supervision of heatmaps and virtual depth, the semantic and geometric knowledge is incorporated into initial image features. Then, implicit foreground volume (IFV) is tailored as a carrier for the camera plane and the BEV plane. The rendered heatmaps from IFV are supervised by the projected 3D boxes in the source domain and by the pre-trained 2D detector in the target domain. The green and red flows indicate the supervision from the source and target domain, respectively. The RenderNet shares the same parameters.

Semantic Rendering

We first introduce how to establish the connection between the 2D image plane and BEV space. Most MC3D-Det methods utilize the BEV plane without height information, which leads to a loss of geometric information in the height dimension (Huang et al. 2021; Li et al. 2023a,b). We propose the implicit foreground volume to enable new viewpoints rendering from the BEV plane. Specifically, we use a geometry-aware decoder D_{geo} to transform the BEV features $F_{bev} \in \mathbb{R}^{C\times X\times Y}$ into the intermediate features $F_{bev}^{'} \in \mathbb{R}^{C\times 1\times X\times Y}$ and $F_{height} \in \mathbb{R}^{1\times Z\times X\times Y}$, and these features is lifted from BEV plane to an implicit foreground volume $V_{ifv} \in \mathbb{R}^{C\times Z\times X\times Y}$:

$$V_{ifv} = \text{sigmoid}(F_{height}) \cdot F_{bev}. \tag{3}$$

Eq. 3 lifts the object on the BEV plane into 3D volume with the estimated height distribution sigmoid(F_{height}), which represents whether there is an object at the corresponding height. X, Y, and Z represent the three-dimensional size of the established feature volume. Ideally, the volume V_{ifv} contains all the foreground object information in the corresponding position.

To render semantic features of different viewpoints, we propose the Multi-View Semantic Rendering (MVSR). Specifically, we first randomly perturb the camera's position $(x + \triangle x, y + \triangle y, z + \triangle z)$ and orientation $(\theta_{yaw} + \triangle\theta_{yaw}, \theta_{pitch} + \triangle\theta_{pitch}, \theta_{roll} + \triangle\theta_{roll})$. Based on the camera's position and observation orientation, we generate the coordinates of multiple rays $r_i^{w,h} = [x^{w,h}, y^{w,h}, z^{w,h}]$ of image plane w, h to sample from the implicit foreground volumes V_{ifv} and aggregate them into the camera plane features F_{render}:

$$F_{render}(w,h) = \sum_{i=1}^{n} V_{ifv}(x^{w,h}, y^{w,h}, z^{w,h}), \tag{4}$$

where $r_i^{w,h} = [x^{w,h}, y^{w,h}, z^{w,h}]$ represents the ray coordinates of w-th row and h-th column camera plane in the implicit foreground volumes V_{ifv}. The rendered camera plane features F_{render} is then fed into the RenderNet, which is the combination of several 2D convolutional layers, to generate the heatmaps $h_{render} \in \mathbb{R}^{N_{cls}\times W\times H}$ and attributes $a_{render} \in \mathbb{R}^{N_{atri}\times W\times H}$. N_{cls} means the number of categories. The detailed structure of RenderNet is introduced in the supplementary material. The semantic heatmaps and attributes can be constrained on the source and target domains to eliminate perspective bias $[\Delta u, \Delta v]$.

Perspective Rendering on Source Domain

To reduce perspective bias as stated in Eq. 2, the 3D boxes of the source domain can be utilized to supervise the predicted heatmaps and attributes on the rendered novel views. In addition, we also utilize the normalized depth information to help the image encoder learn better geometry information.

Novel View Semantic Supervision Based on , the heatmaps and attributes from different perspectives (the output of RenderNet) can be rendered. Here we will explain how to regularize them to eliminate the perspective bias in Eq. 2. Specifically, we project each object's box from ego coordinate to the j-th 2D image plane using the intrinsic K_j' and extrinsic parameters E_j' from the rendering process: $\hat{P}_j = (ud, vd, d) = K_j' E_j' P$, where $\hat{P}_j$ and P stand for

the object on 2.5D image plane and 3D space, d represents the depth between the object and the view's optical center. Based on the position of the object on the image plane, the category heatmaps $h_{gt} \in \mathbb{R}^{N_{cls} \times W \times H}$ can be generated (Yin, Zhou, and Krahenbuhl 2021). The object's dimensions (length, width and height) $a_{gt} \in \mathbb{R}^{N_{atri} \times W \times H}$ are also projected to the UV plane. Following (Yin, Zhou, and Krahenbuhl 2021), focal loss (Lin et al. 2017) $\mathcal{L}_{focal}$ and L1 loss $\mathcal{L}_1$ are used to supervise the class information and object dimensions on source domain:

$$\mathcal{L}_{render} = \mathcal{L}_{focal}(h_{render}, h_{gt}) + \mathcal{L}_1(a_{render}, a_{gt}). \quad (5)$$

Additionally, we also train a 2D detector for the image features using 3D boxes by $\mathcal{L}_{ps}$, which uses the same mapping and supervision methods as above. The only difference is that the 3D boxes are projected using the original intrinsics K and extrinsic E of the camera. 2D detectors can be further applied to correct the spurious geometry in the target domain.

Perspective Geometry Supervision Providing explicit depth information can be effective in improving the performance of multi-camera 3D object detection (Li et al. 2023a). However, the depth prediction tends to overfit the intrinsic parameters. Therefore, following (Park et al. 2021; Wang et al. 2023a), we force the DepthNet to learn normalized virtual depth $D_{virtual}$:

$$\begin{aligned} \mathcal{L}_{pg} &= \mathcal{L}_{BCE}(D_{pre}, D_{virtural}), \\ D_{virtual} &= \frac{\sqrt{\frac{1}{f_u^2} + \frac{1}{f_v^2}}}{U} D, \end{aligned} \quad (6)$$

where $\mathcal{L}_{BCE}$ means the binary cross entropy loss, and D_{pre} represents the predicted depth of DepthNet. f_u and f_v are focal lengths of the image plane, and U is a constant. It is worth noting that the depth D here is the foreground depth information provided using 3D boxes rather than the point cloud. By doing so, The DepthNet is more likely to focus on the depth of foreground objects. Finally, when using the actual depth information to lift semantic features into BEV plane, we use Eq. 6 to convert the virtual depth back to the actual depth.

Perspective Rendering on Target Domain

Unlike the source domain, there are no 3D labels in the target domain, so the perspective semantic supervision cannot be directly applied. Fortunately, we subtly utilize the pre-trained 2D detector to modify spurious geometric BEV features on the target domain. To achieve this, we render the heatmaps h_{render} from the implicit foreground volume with the original camera parameters. Focal loss is used to constrain the consistency between the pseudo labels from the 2D detector and the rendered maps:

$$\begin{aligned} \mathcal{L}_{con} &= \mathcal{L}_{focal}(h_{render}, h_{pseudo}), \\ h_{pseudo} &= \begin{cases} 1, & h > c \\ h, & else \end{cases}, \end{aligned} \quad (7)$$

where $\mathcal{L}_{focal}$ is the vanilla focal loss (Lin et al. 2017). $\mathcal{L}_{con}$ can effectively use accurate 2D detection to correct the position of foreground targets in the BEV space, which is an unsupervised regularization on the target domain. Following (Yang et al. 2021), we enhanced the confidence of the predicted heatmaps in a pseudo way.

Overall Framework

Although we have added some networks to aid in training, these networks are not needed in inference. In other words, our method is suitable for most MC3D-Det algorithms to learn perspective-invariant features. To verify the effectiveness of our framework, BEVDepth (Li et al. 2023a) is instantiated as our main pipeline. The original detection loss $\mathcal{L}_{det}$ of BEVDepth is used as the main 3D detection supervision on the source domain, and the depth supervision of BEVDepth has been replaced by the proposed $\mathcal{L}_{pg}$. In summary, the final loss function with our framework is:

$$\mathcal{L} = \lambda_s \mathcal{L}_{det} + \lambda_s \mathcal{L}_{render} + \lambda_s \mathcal{L}_{pg} + \lambda_s \mathcal{L}_{ps} + \lambda_t \mathcal{L}_{con}, \quad (8)$$

where λ_s sets to 1 for the source domain and sets to 0 for the target domain, and it is the opposite for λ_t. In other words, $\mathcal{L}_{con}$ is not used under the DG protocol. where β is a hyper-parameter, $iter_{num}$ denotes the current iteration count, and max_{iter} represents the maximum number of iterations. $\bar{c}$ is the average across all feature channels, while $\bar{c_i}$ refers to the average of the ith feature channel. By dynamically weighting feature channels, we can exert control over the real-time updates of specific feature channels.

Experiment

To verify the effectiveness, we elaborately use both DG and UDA protocols for MC3D-Det. The details of datasets, evaluation metrics, and implementation details are elaborated in the supplementary materials.

Domain Generalization Benchmark

For DG protocol, we compare our framework with the reproduced DG-BEV (Wang et al. 2023a) and BEVDepth (Li et al. 2023a). As shown in Table 1, our method has significantly improved the performance on the target domain. It demonstrates that IFV as a bridge can help learn perspective-invariant features against domain shifts. In addition, our approach does not sacrifice performance in the source domain and even has some improvement in most cases. It is worth mentioning that DeepAccident (Wang et al. 2023b) was collected from the CARLA virtual engine, and our algorithm also achieved satisfactory generalization ability by training on DeepAccident. In addition, we have tested other MC3D-Det methods, and their generalization performance is very poor without special designs as shown in Table 2.

Unsupervised Domain Adaptation Benchmark

To further validate the effect of debiasing on the target domain, we also established a UDA benchmark and applied widely-used UDA methods (including Pseudo Label, Coral (Sun and Saenko 2016), and AD (Ganin

nuScenes → Lyft		Source Domain (nuScenes)					Target Domain (Lyft)				
Method	Target-Free	mAP↑	mATE↓	mASE↓	mAOE↓	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS* ↑
Oracle		-	-	-	-	-	0.598	0.474	0.152	0.092	0.679
BEVDepth	✓	0.326	0.689	0.274	0.581	0.395	0.114	0.981	0.174	0.413	0.296
DG-BEV	✓	0.330	0.692	**0.272**	0.584	0.397	0.284	0.768	0.171	0.302	0.435
PR-BEV	✓	**0.334**	**0.688**	0.276	**0.579**	**0.399**	**0.304**	**0.709**	**0.169**	**0.289**	**0.458**
Pseudo Label		0.320	0.694	0.276	0.598	0.388	0.294	0.743	0.172	0.304	0.443
Coral		0.318	0.696	0.283	0.592	0.387	0.281	0.768	0.174	0.291	0.435
AD		0.312	0.703	0.288	0.596	0.381	0.277	0.771	0.174	0.288	0.381
PR-BEV$^+$		**0.331**	**0.686**	**0.275**	**0.591**	**0.396**	**0.316**	**0.684**	**0.165**	**0.241**	**0.476**
Lyft → nuScenes		Source Domain (Lyft)					Target Domain (nuScenes)				
Method	Target-Free	mAP↑	mATE↓	mASE↓	mAOE↓	NDS*↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS*↑
Oracle		-	-	-	-	-	0.516	0.551	0.163	0.169	0.611
BEVDepth	✓	**0.598**	**0.474**	0.152	0.092	**0.679**	0.098	1.134	0.234	1.189	0.176
DG-BEV	✓	0.591	0.491	0.154	0.092	0.672	0.251	0.751	0.202	0.813	0.331
PR-BEV	✓	0.593	0.478	**0.150**	**0.084**	0.677	**0.263**	**0.746**	**0.186**	**0.790**	**0.344**
Pseudo Label		0.580	0.538	0.153	**0.079**	0.657	0.261	0.744	0.201	0.819	0.306
Coral		0.574	0.511	0.164	0.105	0.649	0.244	0.767	0.212	0.919	0.302
AD		0.568	0.521	0.161	0.126	0.649	0.247	0.761	0.223	0.902	0.309
PR-BEV$^+$		**0.589**	**0.489**	**0.150**	0.091	**0.672**	**0.280**	**0.733**	**0.182**	**0.776**	**0.358**
DeepAccident → nuScenes		Source Domain (DeepAccident)					Target Domain (nuScenes)				
Method	Target-Free	mAP↑	mATE↓	mASE↓	mAOE↓	NDS*↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS*↑
Oracle		-	-	-	-	-	0.516	0.551	0.163	0.169	0.611
BEVDepth	✓	0.334	0.517	0.741	0.274	0.412	0.087	1.100	0.246	1.364	0.169
DG-BEV	✓	0.331	0.519	0.757	0.264	0.408	0.159	1.075	0.232	1.153	0.207
PR-BEV	✓	**0.345**	**0.499**	**0.735**	**0.251**	**0.425**	**0.187**	**0.931**	**0.229**	**0.967**	**0.239**
Pseudo Label		0.312	0.522	0.785	0.271	0.393	0.151	1.112	0.238	1.134	0.202
Coral		0.314	0.544	0.796	0.274	0.388	0.164	1.045	0.242	1.104	0.208
AD		0.312	0.539	0.787	0.263	0.391	0.166	1.013	0.251	1.073	0.207
PR-BEV$^+$		**0.344**	**0.488**	**0.737**	**0.248**	**0.426**	**0.207**	**0.862**	**0.235**	**0.962**	**0.260**
DeepAccident → nuScenes		Source Domain (DeepAccident)					Target Domain (Lyft)				
Method	Target-Free	mAP↑	mATE↓	mASE↓	mAOE↓	NDS*↑	mAP↑	mATE↓	mASE↓	mAOE↓	NDS*↑
Oracle		-	-	-	-	-	0.598	0.474	0.152	0.092	0.679
BEVDepth	✓	0.334	0.517	0.741	0.274	0.412	0.045	1.219	0.251	1.406	0.147
DG-BEV	✓	0.331	0.519	0.757	0.264	0.408	0.135	1.033	0.269	1.259	0.189
PR-BEV	✓	**0.345**	**0.499**	**0.735**	**0.251**	**0.425**	**0.151**	**0.941**	**0.242**	**1.130**	**0.212**
Pseudo Label		0.323	0.531	0.768	0.271	0.399	0.132	1.113	0.281	1.241	0.185
Coral		0.308	0.573	0.797	0.284	0.378	0.145	1.004	0.254	1.129	0.196
AD		0.304	0.554	0.796	0.274	0.381	0.148	0.997	0.262	1.189	0.197
PR-BEV$^+$		**0.330**	**0.517**	**0.737**	**0.240**	**0.416**	**0.171**	**0.871**	**0.212**	**1.043**	**0.238**

Table 1: Comparison of different approaches on DG and UDA protocols. Target-Free means DG protocol. Pseudo Label, Coral, and AD are applied in DG-BEV on UDA protocol. Following (Wang et al. 2023a), nuScenes is evaluated according to the original NDS as source domain. Other results are evaluated only for the 'car' category by NDS*. $^+$ represents that fine-tuning uses unsupervised data from the target domain.

	w/o ours		w ours	
nuScenes → Lyft	mAP ↑	NDS*↑	mAP↑	NDS*↑
DETR3D* (Wang et al. 2022)	0.008	0.044	0.028	0.076
PETR* (Liu et al. 2022)	0.012	0.051	0.032	0.091
SparseBEV* (Liu et al. 2023a)	0.016	0.059	0.038	0.097
BEVDepth (Li et al. 2023a)	0.114	0.296	0.304	0.458
BEVDet (Huang et al. 2021)	0.104	0.275	0.296	0.446
BEVFormer (Li et al. 2022d)	0.084	0.246	0.208	0.355
FB-OCC (Li et al. 2023b)	0.113	0.294	0.301	0.454

Table 2: The plug-and-play capability testing of our method. We tested more MC3D-Det algorithms under the DG and tried to add our algorithm for further improvement. Methods with an asterisk (*) do not have BEV representation, while all others do.

and Lempitsky 2015)) on DG-BEV. As shown in 1, our algorithm achieved significant performance improvement. This is mainly attributed to perspective rendering, which fully utilizes the 2D detector with better generalization performance to correct the spurious geometric information of the 3D detector. Additionally, we found that most algorithms tend to degrade performance on the source domain, while our method is relatively gentle. It is worth mentioning that we found that AD and Coral show significant improvements when transferring from a virtual dataset to a real dataset, but exhibit a decline in performance when evaluated on real-to-real testings. This is because these two algorithms are designed to address style changes, but in scenarios with small style changes, they may disrupt semantic information. As for the pseudo-label algorithm, it can improve the model's generalization performance by increasing confidence in some easy samples on target domains, but blindly increasing confidence in target domains can actually make the model worse.

Ablation Study

To further demonstrate the effectiveness of our proposed algorithm, we conducted ablation experiments on three key components: 2D information injection $\mathcal{L}_{ps}$ (DII), source domain debiasing $\mathcal{L}_{render}$ (SDB), and target domain debiasing $\mathcal{L}_{con}$ (TDB). DII and SDB are designed for the source domain, while TDB is designed for the target domain. In other words, we report the results under the UDA protocol only when using TDB, while the results of other components are reported under the DG protocol. As presented in Tab 3, each component has yielded improvements, with SDB and TDB exhibiting relatively significant improvements. SDB can better capture perspective invariance and more generalizable features, while TDB leverages the strong generalization ability of 2D to facilitate the correction of spurious geometric features of the 3D detector in the target domain. DII makes the network learn more robust features by adding geometric supervision to the image features. These findings underscore the importance of each component in our algorithm and highlight the potential of our approach for addressing the challenges of the domain gap in MC3D-Det.

			nuScenes → Lyft		DeepAccident → Lyft	
DII	SDB	TDB	mAP ↑	NDS* ↑	mAP↑	NDS*↑
			0.279	0.433	0.132	0.188
✓			0.290	0.438	0.143	0.205
	✓		0.300	0.453	0.147	0.209
✓	✓		0.304	0.458	0.151	0.212
✓	✓	✓	**0.316**	**0.476**	**0.171**	**0.238**

Table 3: Ablation study on different modules of PR-BEV, including 2D information injection (DII), source domain debiasing (SDB), and target domain debiasing (TDB). The bottom line is the UDA result, and the rest is the DG result.

Further Discussion

Here we try to migrate our framework to more MC3D-Det methods to prove its universal capability. Any MC3D-Det algorithm with the image features and BEV features can be embedded with our algorithm. The results are shown in Tab 2. For these methods without BEV representation, we leverage the LSS mechanism in BEVDet (Huang et al. 2021) to build additional BEV representations so that our algorithm can be used to improve these methods. As the results show, we draw several conclusions: (1) Methods without BEV presentation perform very poorly across domains compared to other BEV-based methods. This can be attributed to their tendency to overfit camera extrinsic parameters in a learning way. Conversely, methods with BEV representation solely employ camera extrinsic parameters to project 2D image features into 3D space through a physical modeling form. This approach is highly resilient to variations in camera extrinsic parameters, thereby increasing its robustness and reliability. In conclusion, the BEV representation can effectively establish the connection of different perspectives through physical modeling, as opposed to learning. This way enables the model to have superior cross-domain generalization. (2) Our method has the potential to enhance the performance of DETR3D (Wang et al. 2022), PETR (Liu et al. 2022), and SparesBEV (Liu et al. 2023a) algorithms by rendering the new viewpoints from an auxiliary BEV presentation. This is because the process of rerendering new perspectives compels the network to acquire a generalizable BEV representation. The generalizable BEV representation, in turn, encourages the network to learn more robust visual features that mitigate the impact of overfitting camera parameters.

Conclusion

This paper proposes a framework for multi-camera 3D object detection (MC3D-Det) based on perspective rendering to address the issue of poor generalization for unseen domains. We render the semantic maps of different views from BEV features. We then use 3D boxes or pre-trained 2D detectors to correct the spurious BEV features. Our framework is model-agnostic, and we demonstrate its effectiveness by optimizing multiple MC3D-Det methods. Our algorithms have achieved significant improvements in both DG and UDA protocols.

References

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3339–3348.

Cheng, S.; and Sun, H. 2024. SPT: Sequence Prompt Transformer for Interactive Image Segmentation. arXiv:2412.10224.

Dou, Q.; Coelho de Castro, D.; Kamnitsas, K.; and Glocker, B. 2019. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32.

Facil, J. M.; Ummenhofer, B.; Zhou, H.; Montesano, L.; Brox, T.; and Civera, J. 2019. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11826–11835.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189.

He, Z.; and Zhang, L. 2020. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *European conference on computer vision*, 309–324. Springer.

Huang, J.; Huang, G.; Zhu, Z.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Jia, X.; Gao, Y.; Chen, L.; Yan, J.; Liu, P. L.; and Li, H. 2023. DriveAdapter: Breaking the Coupling Barrier of Perception and Planning in End-to-End Autonomous Driving. In *Proceedings of the IEEE international conference on computer vision*.

Jiang, C.; Du, D.; Liu, J.; Zhu, S.; Liu, Z.; Ma, Z.; Liang, Z.; and Zhou, J. 2024. NeuroGauss4D-PCI: 4D Neural Fields and Gaussian Deformation Fields for Point Cloud Interpolation. *arXiv preprint arXiv:2405.14241*.

Jiang, C.; Wang, G.; Miao, Y.; and Wang, H. 2022. 3-d scene flow estimation on pseudo-lidar: Bridging the gap on estimating point motion. *IEEE Transactions on Industrial Informatics*, 19(6): 7346–7354.

Kim, H.; Lee, K.; Hwang, G.; and Suh, C. 2022. Crash to Not Crash: Learn to Identify Dangerous Vehicles Using a Simulator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2583–2589.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5400–5409.

Li, H.; Sima, C.; Dai, J.; Wang, W.; Lu, L.; Wang, H.; Xie, E.; Li, Z.; Deng, H.; Tian, H.; Zhu, X.; Chen, L.; Gao, Y.; Geng, X.; Zeng, J.; Li, Y.; Yang, J.; Jia, X.; Yu, B.; Qiao, Y.; Lin, D.; Liu, S.; Yan, J.; Shi, J.; and Luo, P. 2022a. Delving into the Devils of Bird's-eye-view Perception: A Review, Evaluation and Recipe. *arXiv preprint arXiv:2209.05324*.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023a. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, Y.; Ma, D.; An, Z.; Wang, Z.; Zhong, Y.; Chen, S.; and Feng, C. 2022b. V2X-Sim: Multi-Agent Collaborative Perception Dataset and Benchmark for Autonomous Driving. *IEEE Robotics and Automation Letters*, 7(4): 10914–10921.

Li, Z.; Chen, Z.; Li, A.; Fang, L.; Jiang, Q.; Liu, X.; and Jiang, J. 2022c. Unsupervised Domain Adaptation for Monocular 3D Object Detection via Self-training. In *European conference on computer vision*, 245–262. Springer.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; and Dai, J. 2022d. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. *European conference on computer vision*.

Li, Z.; Yu, Z.; Wang, W.; Anandkumar, A.; Lu, T.; and Alvarez, J. M. 2023b. FB-OCC: Forward-Backward View Transformations for Occupancy Prediction. In *Proceedings of the IEEE international conference on computer vision*.

Lian, Q.; Xu, Y.; Yao, W.; Chen, Y.; and Zhang, T. 2022. Semi-supervised monocular 3d object detection by multi-view consistency. In *European Conference on Computer Vision*, 715–731. Springer.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. 2980–2988.

Liu, H.; Teng, Y.; Lu, T.; Wang, H.; and Wang, L. 2023a. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18580–18590.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, Q.; Wang, T.; Zhang, X.; and Sun, J. 2023b. PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images. *Proceedings of the IEEE international conference on computer vision*.

Ma, Y.; Wang, T.; Bai, X.; Yang, H.; Hou, Y.; Wang, Y.; Qiao, Y.; Yang, R.; Manocha, D.; and Zhu, X. 2022. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 10–18.

Pan, M.; Liu, J.; Zhang, R.; Huang, P.; Li, X.; Xie, H.; Wang, B.; Liu, L.; and Zhang, S. 2024. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 12404–12411. IEEE.

Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; and Gaidon, A. 2021. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3142–3152.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, 194–210. Springer.

Roddick, T.; Kendall, A.; and Cipolla, R. 2019. Orthographic feature transform for monocular 3d object detection. In *BMVC*.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 443–450. Springer.

Sun, H. 2024. Ultra-High Resolution Segmentation via Boundary-Enhanced Patch-Merging Transformer. arXiv:2412.10181.

Sun, H.; Xu, L.; Jin, S.; Luo, P.; Qian, C.; and Liu, W. 2024. PROGRAM: PROtotype GRAph Model based Pseudo-Label Learning for Test-Time Adaptation. In *The Twelfth International Conference on Learning Representations*, 15173–15183.

Sun, T.; Segu, M.; Postels, J.; Wang, Y.; Van Gool, L.; Schiele, B.; Tombari, F.; and Yu, F. 2022. SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21371–21382.

Wang, S.; Zhao, X.; Xu, H.-M.; Chen, Z.; Yu, D.; Chang, J.; Yang, Z.; and Zhao, F. 2023a. Towards Domain Generalization for Multi-View 3D Object Detection in Bird-Eye-View. 13333–13342.

Wang, T.; Kim, S.; Ji, W.; Xie, E.; Ge, C.; Chen, J.; Li, Z.; and Luo, P. 2023b. DeepAccident: A Motion and Accident Prediction Benchmark for V2X Autonomous Driving. *arXiv preprint arXiv:2304.01168*.

Wang, T.; Kim, S.; Ji, W.; Xie, E.; Ge, C.; Chen, J.; Li, Z.; and Ping, L. 2023c. DeepAccident: A Motion and Accident Prediction Benchmark for V2X Autonomous Driving. *arXiv preprint arXiv:2304.01168*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 913–922.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 180–191.

Wang, Y.; Yin, J.; Li, W.; Yang, R.; and Shen, J. 2023d. SSDA3D: Semi-supervised Domain Adaptation for 3D Object Detection from Point Cloud. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, Z.; Huang, Z.; Fu, J.; Wang, N.; and Liu, S. 2023e. Object as Query: Lifting Any 2D Object Detector to 3D Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3791–3800.

Xu, C.; Wu, B.; Hou, J.; Tsai, S.; Li, R.; Wang, J.; Zhan, W.; He, Z.; Vajda, P.; Keutzer, K.; et al. 2023a. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23320–23330.

Xu, C.-D.; Zhao, X.-R.; Jin, X.; and Wei, X.-S. 2020. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11724–11733.

Xu, R.; Xia, X.; Li, J.; Li, H.; Zhang, S.; Tu, Z.; Meng, Z.; Xiang, H.; Dong, X.; Song, R.; Yu, H.; Zhou, B.; and Ma, J. 2023b. V2V4Real: A Real-World Large-Scale Dataset for Vehicle-to-Vehicle Cooperative Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13712–13722.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In *2022 International Conference on Robotics and Automation*, 2583–2589.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17830–17839.

Yang, J.; Shi, S.; Wang, Z.; Li, H.; and Qi, X. 2021. ST3D: Self-Training for Unsupervised Domain Adaptation on 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10368–10378.

Yang, J.; Wang, T.; Ge, Z.; Mao, W.; Li, X.; and Zhang, X. 2022. Towards 3D Object Detection with 2D Supervision. *arXiv preprint arXiv:2211.08287*.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. 11784–11793.

Yuan, J.; Zhang, B.; Yan, X.; Chen, T.; Shi, B.; Li, Y.; and Qiao, Y. 2023. Bi3D: Bi-Domain Active Learning for Cross-Domain 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15599–15608.

Yuan, Z.; Cao, J.; Li, Z.; Jiang, H.; and Wang, Z. 2024a. SD-MVS: Segmentation-Driven Deformation Multi-View Stereo with Spherical Refinement and EM Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6871–6880.

Yuan, Z.; Cao, J.; Wang, Z.; and Li, Z. 2024b. Tsar-mvs: Textureless-aware segmentation and correlative refinement guided multi-view stereo. *Pattern Recognition*, 154: 110565.

Zhao, G.; Li, G.; Xu, R.; and Lin, L. 2020. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, 86–102. Springer.