DualCross: Cross-Modality Cross-Domain Adaptation for Monocular BEV Perception

Yunze Man[1], Liangyan Gui[1] and Yu-Xiong Wang[1]

***Abstract*— Closing the domain gap between training and deployment and incorporating multiple sensor modalities are two challenging yet critical topics for self-driving. Existing work only focuses on single one of the above topics, overlooking the simultaneous domain and modality shift which pervasively exists in real-world scenarios. A model trained with multi-sensor data collected in Europe may need to run in Asia with a subset of input sensors available. In this work, we propose DualCross, a cross-modality cross-domain adaptation framework to facilitate the learning of a more robust monocular bird's-eye-view (BEV) perception model, which transfers the point cloud knowledge from a LiDAR sensor in one domain during the training phase to the camera-only testing scenario in a different domain. This work results in the first open analysis of cross-domain cross-sensor perception and adaptation for monocular 3D tasks in the wild. We benchmark our approach on large-scale datasets under a wide range of domain shifts and show state-of-the-art results against various baselines. Our project webpage is at https://yunzeman.github.io/DualCross.**

I. Introduction

In recent years, multimodal 3D perception has shown outstanding performance and robustness over its single-modality counterpart, achieving leading results for various 3D perception tasks [12], [23], [27], [30], [35], [38] on large-scale multi-sensor 3D datasets [1], [13], [33]. Despite the superiority in information coverage, the introduction of more sensor modalities also poses additional challenges to the perception system. On the one hand, generalizing the model between datasets becomes hard, because each sensor has its unique properties, such as field-of-view (FoV) for cameras, density for LiDAR, *etc*. On the other hand, the operation of the model is conditioned on the presence and function of more sensors, making it hard to work on autonomous agents with less sensor types or under sensor failure scenarios.

More specifically, transferring knowledge among different data domains is still an open problem for autonomous agents in the wild. In the self-driving scenario, training the perception models offline in a source domain with annotation while deploying the model in a different target domain without annotation is very common in practice. As a result, a model has to consider the domain gap between source and target environments or datasets, which usually involves different running locations, different sensor specifications, different illumination and weather conditions, *etc*.

Meanwhile, in addition to domain shift, modality shift is another factor which challenges the successful deployment of models. The widely adopted assumption that all sensors are available during training, validation, and deployment time is not always true in reality. Due to the cost and efficiency trade-off, or sensor failure scenarios, in many cases we can have fewer sensors available in the target domain during testing than what we have in the source domain during training. A typical scenario is having camera and LiDAR sensors in the large-scale training phase while only having cameras for testing, as shown in Figure 1. It is not clear how to facilitate the camera-only 3D inference with the help of a LiDAR sensor only in the source domain during training.

The challenges above raise an important question: *Can we achieve robust 3D perception under both domain shift and sensor modality shift?* Existing methods either study cross-domain scenarios assuming consistent modality [8], [12], [15], [19], [21], [42], [44], or study cross-modality scenarios assuming the same domain during training and validation [4], [6], [9], [11], [17], [18], [41]. However, simultaneous domain and modality shift poses additional challenges of large domain discrepancy and exacerbates the ill-posed nature of 3D inference from monocular information due to the misaligned sensory data. As we will discuss in Sec. III-B, our new setting requires a novel methodology in using LiDAR without increasing the domain discrepancy.

To tackle the above challenges, we propose DualCross, a cross-modality cross-domain adaptation framework for bird's-eye-view (BEV) perception. Our model addresses the monocular 3D perception task between different domains, and utilizes additional modalities in the source domain to facilitate the evaluation performance. Motivated by the fact that image and BEV frames are bridged with 3D representation, we first design an efficient backbone to perform 3D depth estimation followed by a BEV projection. Then, to learn from point clouds without explicitly taking them as model inputs, we propose an implicit learning strategy, which distills 3D knowledge from a LiDAR-Teacher to help the Camera-Student learn better 3D representation. Finally, in order to address the visual domain shift, we introduce adversarial learning on the student to align the features learned from source and target domains. Supervision from the teacher and feature discriminators are designed at multiple layers to ensure an effective knowledge transfer.

By considering the domain gap and effectively leveraging LiDAR point clouds in the source domain, our proposed method is able to work reliably in more complicated, uncommon, and even unseen environments. Our model achieves state-of-the-art performance in four very different domain shift settings. Extensive ablation studies are conducted to investigate the contribution of our proposed components, the robustness under different changes, and other design choices.

[1]University of Illinois Urbana-Champaign

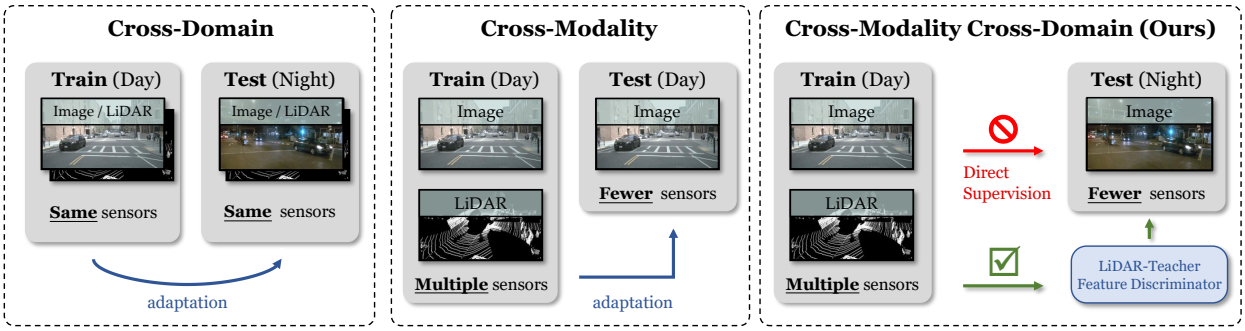

Fig. 1: **Left & Middle**: Existing models assume a fixed sensor or modality during training and testing phases. **Right**: We introduce a more realistic setting which considers cross-modality cross-domain shift. **Surprisingly**, using source-only LiDAR as depth supervision leads to worse performance (**6.3** in IoU) than the image-only model (**6.7**). Thus, we propose DualCross to reduce the domain discrepancy, achieving state-of-the-art performance (**17.0**).

The main contributions of this paper are as follows. (1) We introduce mixed domain and modality mismatch, an overlooked but realistic problem setting in 3D domain adaptation in the wild, leading to a robust camera-only 3D model that works in complicated and dynamic scenarios with minimum sensors available. (2) We propose a novel LiDAR-Teacher and Camera-Student knowledge distillation model, which considerably outperforms state-of-the-art LiDAR supervision methods. (3) Extensive experiments in challenging domain shift settings demonstrate the capability of our method in leveraging source domain point cloud information for accurate monocular 3D perception.

II. Related Work

Multi- and Cross-Modality 3D Perception. Considerable research has examined leveraging signals from multiple modalities, especially images and point clouds, for 3D perception tasks. Early work [20] projects point clouds to the BEV frame and fuses them with 2D RGB features to generate proposals and regress bounding boxes. Later work [43], [46], [22] explores deep fusion between points and images. Under the umbrella of the cross-modality setting, 2DPASS [41] transfers features learned from images to the LiDAR. BEVDepth [18] obtains reliable depth estimation by exploiting camera parameters with image features during training. More recently, a line of work explores knowledge distillation from one sensor to another for 3D object detection [4], [6], [18], [41]. On the contrary, our method explores a more realistic yet challenging setting, where we use LiDAR data in one domain (*e.g.,* Boston/Sunny/Daylight) during training to help the camera-only model during inference in another domain (*e.g.,* Singapore/Rainy/Night). As a result, we analyze and improve the actual usefulness of additional sensors under domain shift settings.

Cross-Domain 3D Perception. While extensive research has been conducted on domain adaptation for 2D tasks, the field of domain adaptation for 3D perception in the real world has received relatively less attention. Some prior work adapts depth estimation from synthetic to real image domains [15], [45]. Working on point clouds, PointDAN [31] designs a multi-scale adaptation model for 3D classification. For 3D semantic segmentation, SqueezeSeg [39] projects point clouds to the 2D view, while other work [8], [12], [16] leverages point clouds and images data together. Recent work [21], [44] explores cross-domain 3D object detection from point clouds. SRDAN [44] employs adversarial learning to align the features between different domains. Although prior work [12], [19] explores various domain adaptation techniques for different sensor modalities, these methods only adopt the same modalities to learn the domain shift between source and target data. In contrast, our approach achieves robust 3D perception in a more general scenario, where the model can perform accurate 3D inference in the target domain by adapting information encoded in source-exclusive modalities.

3D Inference in Bird's-Eye-View Frame. Inferring 3D scenes from the BEV perspective has recently received a large amount of interest due to its effectiveness. MonoLayout [24] estimates the layout of urban driving scenes from images in the BEV frame and uses an adversarial loss to enhance the learning of hidden objects. Another work [2] proposes to employ graphical representation and temporal aggregation for better inference in the driving scenarios using on-board cameras. Recently, using BEV representation to merge images from multiple camera sensors has become a popular approach [10], [26]. Following the monocular feature projection proposed by Orthographic Feature Transform (OFT) [32], Lift-Splat-Shoot [29] disentangles feature learning and depth inference by learning a depth distribution over pixels to convert camera image features into BEV. Unlike the above work performing BEV analysis in settings with more controlled premises, we are the first to explore cross-domain and cross-sensor settings, leading to a more robust and more realistic 3D inference methodology.

III. Approach

In this work, we consider the task of learning BEV representation of scenes with domain shift and modality mismatch. Specifically, the model will be given annotated LiDAR point clouds and camera images in the source domain, but only unannotated camera images in the target domain. And the model seeks to achieve highest performance on the unsupervised target domain. This problem setting is common and worthwhile, especially considering the existence of many existing public multi-modality datasets and the rise of many camera-only vehicle scenarios.

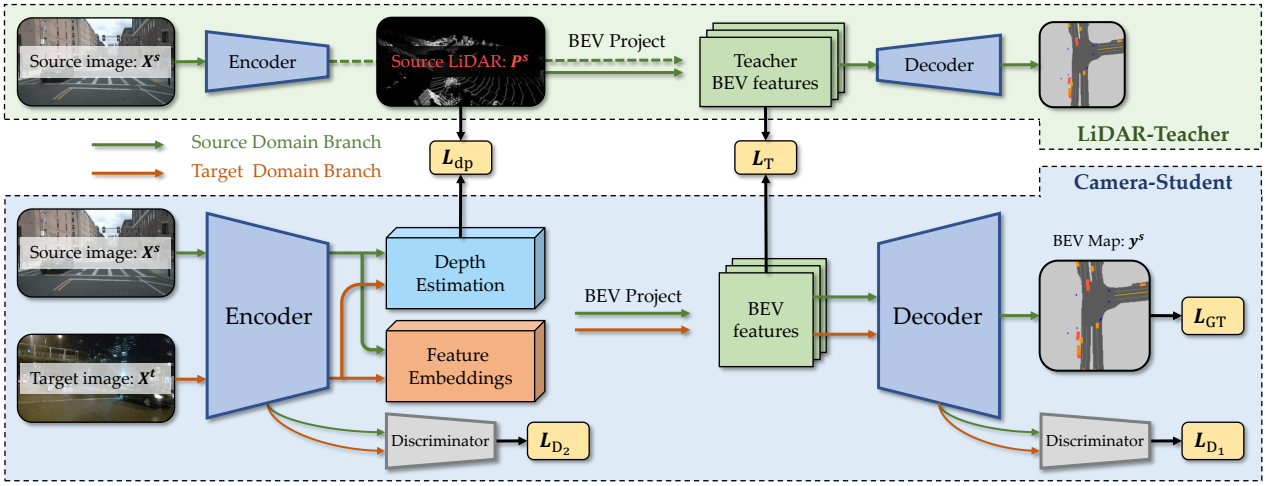

Fig. 2: **Overview of our DualCross framework**. DualCross includes three components. (1) **LiDAR-Teacher** uses voxelized LiDAR point clouds to transform the image features to BEV frame. It provides essential knowledge on how to guide image learning given LiDAR information. (2) **Camera-Student** is supervised by the teacher model as well as the LiDAR ground truth. (3) **Discriminators** are used to align features from source and target domains.

Formally, **for the source domain**, we are given *labeled* data with N^s multi-modality samples, $\mathcal{D}^s = \{(\boldsymbol{X}_i^s, \boldsymbol{P}_i^s, \boldsymbol{y}_i^s)\}_{i=1}^{N^s}$, where s represents the source domain. Here $\boldsymbol{X}_i^s = \{\boldsymbol{x}_{ik}^s\}_{k=1}^n$ consists of n camera images $\boldsymbol{x}_{ik}^s \in \mathbb{R}^{3\times H\times W}$. The number of cameras n can take any integer as small as one, depending on the dataset or cameras deployed on the vehicle. In addition, each camera image has an intrinsic matrix and an extrinsic matrix. $\boldsymbol{P}_i^s$ is a point cloud containing multiple unordered points $\boldsymbol{p} \in \mathbb{R}^3$ represented by 3D coordinate values. And label $\boldsymbol{y}_i^s$ represents rasterized representation of the scenes in the BEV coordinate. **For the target domain**, we are given *unlabeled* data with N^t image samples, $\mathcal{D}^t = \{\boldsymbol{X}_i^t\}_{i=1}^{N^t}$, where t represents the target domain, and we want to estimate $\{\boldsymbol{y}_i^t\}_{i=1}^{N^t}$, the BEV representation of the scenes in the target domain.

An overview of our method DualCross is illustrated in Figure 2. DualCross is designed to extract features from monocular images and project the features into the BEV frame (Section III-A), using estimated or ground truth 3D depth information. The model is composed of a LiDAR-Teacher and a Camera-Student (Section III-B), where the teacher encodes how to learn better representation given point clouds, and transfers that knowledge to the camera-only student using multi-level teacher-student supervision. Finally, to bridge the domain gap between source and target domains, we leverage adversarial discriminators at different feature layers to align the distributions across two domains in the camera-student model (Section III-C). In addition, we describe learning objectives and loss designs (Section III-D).

A. Learning BEV from Images

In order to achieve 3D perception under the cross-modality setting, our first challenge is to unify the image coordinates, point cloud coordinates, and BEV coordinates into a joint space. We follow LSS [29] to transform the image features from the perspective view into the BEV view. Specifically, we tackle this problem by constructing a 3D voxel representation of the scene for each input image. We discretize the depth axis into N_d bins and lift each pixel of the images into multiple voxels (frustums), where each voxel is represented by the 3D coordinate of its center location. For a given pixel $\text{px} = (h, w)$ on one of the camera image, it corresponds to a set of N_d voxels at different depth bins:

$$V_{\text{px}} = \{v_i = M^{-1}[d_i h, d_i w, d_i]^T | i \in \{1, 2, \cdots, N_d\}\}, \quad (1)$$

where M is the camera matrix and d_i is the depth of the i-th depth bin. The feature vector of each voxel v_i in V_{px} is the base feature $\boldsymbol{f}_{\text{px}}$ of pixel px scaled by the depth value α_i. More specifically, $\boldsymbol{f}_{v_i \in V_{\text{px}}} = \alpha_i \cdot \boldsymbol{f}_{\text{px}}$, where the pixel feature $\boldsymbol{f}_{\text{px}}$ is extracted by an image encoder. And the depth value α_i is obtained either from LiDAR point clouds or by estimation, in the teacher and the student model, respectively. The acquirement of α_i is introduced in Section III-B.

After getting the feature for each of the voxels, we project the voxels onto the BEV and aggregate the features to get the BEV feature map. The BEV frame is rasterized into (X, Y) 2D grids, and for each grid, its feature is constructed from the features of all the 3D voxels projected into it using mean pooling. This projection allows us to transform an arbitrary number of camera images into a unified BEV frame. Finally, we obtain an image-like BEV feature embedding, which is used to estimate the final representation using a convolutional neural network (CNN) decoder.

This architecture design bridges the image and LiDAR modalities through an intermediate 3D voxelized representation. Hence, we can take LiDAR point clouds as input into the model to directly guide the BEV projection without having to change the overall pipeline. This further enables the distillation of knowledge from the point clouds to images using a teacher-student model.

B. Cross-Modality Transfer via Teacher-Student Distillation

The co-existence of domain and modality gaps poses additional challenges to the adaptation task. Although the LiDAR sensor in the source domain provides 3D knowledge to the model, it also increases the discrepancy between the two domains, which hurts the model adaptation (as we will see in Section IV-D and Table VI). Hence, the unique difficulty of our work lies in exploiting the *LiDAR* during training to guide the camera model for better 3D estimation.

Depth Supervision by Point Clouds. The main advantage of point clouds over the image modality is the accurate 3D positional information coming from the depth measurement. Due to the lack of LiDAR during evaluation, we cannot use point clouds as direct input of the model. Hence, one alternative approach to using point clouds is to supervise the depth estimation in the model. As in Eq. 1, for each pixel, we calculate the features of its corresponding voxels by multiplying the pixel feature with a depth value α_i. We use another head to predict a depth distribution $\boldsymbol{\alpha}_{\text{px}} = \{\alpha_1, \alpha_2, \dots, \alpha_{N_d}\}$ over N_d depth bins for each pixel px.

The ground truth depth supervision for this estimation task is generated by LiDAR point clouds as follows: When projected to the image frame, the points corresponding to one pixel can have three conditions. If the pixel has, **(1) no point inside**: the ground truth depth distribution of it is omitted; **(2) only one point inside**: the ground truth depth distribution is a one-hot vector, with value one being in the voxel that the point lies in; **(3) multiple points inside**: the ground truth depth distribution α_i of this point is calculated by counting the number of points in each depth bin, and dividing them by V_{px}, the total number of points in $\boldsymbol{\alpha}_{\text{px}}$: $\alpha_i = \frac{\text{Number of points in depth bin } v_i}{\text{Total number of points in } V_{\text{px}}}$.

Using a distribution-based depth representation effectively accounts for the ambiguity when objects of different depth occur in one pixel. This happens at the boundary of objects, and becomes more severe during feature encoding processing when images get down-sampled and each pixel represents larger space. Moreover, a probabilistic depth representation considers uncertainty during depth estimation, and degenerates to pseudo-LiDAR methods [37] if the one-hot constraint is added.

Learning from LiDAR-Teacher. Despite being intuitive and straightforward, direct depth supervision is not optimal for two reasons. First, LiDAR supervision is only on the intermediate feature layer, providing no supervision on the second half of the model. Also, while LiDAR provides accurate depth measurement, "depth estimation" is still different from our overall objective on BEV representation. Motivated by this, as shown in Figure 2, we propose to use a pretrained LiDAR oracle model to supervise the image model at the final BEV feature embedding, such that the supervision of LiDAR is provided to the whole model and aligns better with the final objective. We call the model using ground truth point cloud information "LiDAR-Teacher," and the model to be supervised "Camera-Student." This boils down to a knowledge distillation problem where the 3D inference knowledge of the LiDAR-teacher is distilled to the camera-only student. Note that the classic problem of "better teacher, worse student" [5], [25], [47] in knowledge distillation due to capacity mismatch does not exist in this model, because the LiDAR-Teacher and Camera-Student models in DualCross are almost identical.

Overall, this teacher-student mechanism allows the camera model to learn better 3D representation from the point clouds, leading to better LiDAR supervision at different stages, while still keeping the model image-centric for image-only inference.

C. Cross-Domain Adaptation with Adversarial Alignment

Since the BEV annotations and the LiDAR ground truth are only available in the source data, the model will be heavily biased to the source distribution during teacher-student supervision. Hence, we bridge the target and source domains using adversarial training. Specifically, we place one discriminator D_1 at the BEV decoder CNN blocks, and another D_2 at the image encoder CNN blocks, to align the features of two domains by optimizing over discriminator losses. While the final-layer discriminator D_1 is constantly useful to align features learned from the LiDAR-Teacher and final ground truth, we find that the middle-layer discriminator D_2 is very effective under certain domain gaps where images have great changes but LiDAR remains robust.

To achieve adversarial learning, given a feature encoder E and an input sample X, a domain discriminator D is used to discriminate whether the feature $E(X)$ comes from the source domain or the target domain. The target and source domain samples are given the label $d = 1$ and $d = 0$, respectively. And $D(E(X))$ outputs the probability of the sample X belonging to the target domain. Hence, the discriminator loss is formulated by a cross-entropy loss:

$$\mathcal{L}_{\text{dis}} = d \log D(E(X)) + (1 - d) \log(1 - D(E(X))). \quad (2)$$

Moreover, in order to learn domain-invariant features, our feature encoder E should try to extract features that fool the discriminator D, while the discriminator D tries to distinguish the right domain label of the samples. This adversarial strategy can be formulated as a "min-max" optimization problem: $\mathcal{L}_{\text{D}} = \min_E \max_D \mathcal{L}_{\text{dis}}$. The "min-max" problem is achieved by a Gradient Reverse Layer (GRL) [7], which produces reverse gradient from the discriminator D to learn the domain-invariant encoder E. The loss form is the same for both D_1 and D_2.

D. Full Objective and Inference

The overall objective of our model is composed of the supervision from the BEV ground truth, the LiDAR-Teacher, and the domain alignment discriminators. Given the output rasterized BEV representation map $\boldsymbol{y} \in \mathbb{R}^{X \times Y \times C}$, the ground truth (GT) loss term $\mathcal{L}_{\text{GT}}$ can be formulated as a cross-entropy loss between the estimated source domain BEV map $\tilde{\boldsymbol{y}}^s$ and the GT label $\boldsymbol{y}^s$:

$$\mathcal{L}_{\text{GT}}(\tilde{\boldsymbol{y}}^s, \boldsymbol{y}^s) = -\sum_{i=1}^{X} \sum_{j=i}^{Y} \sum_{k=1}^{C} y^s_{(i,j,k)} \log \tilde{y}^s_{(i,j,k)}. \quad (3)$$

TABLE I: DualCross leads to significant improvements under ***day-to-night*** domain shift, and also achieves the best results under ***dry-to-rain*** domain shift in IoU. ***DA*** and ***CM*** denote whether a model considers domain adaptation and cross-modality in design, respectively.

Day → Night	*DA*	*CM*	Vehicle	Road	Lane
MonoLayout [24]	✗	✗	5.9	37.7	5.9
OFT [32]	✗	✗	6.6	40.5	6.0
LSS [29]	✗	✗	6.7	41.2	7.1
Wide-range Aug.	✓	✗	10.3	46.0	10.4
Vanilla DA	✓	✗	11.2	48.8	11.1
Depth-Supv DA	✓	✓	15.7	50.5	14.2
Input-fusion Teacher	✓	✓	14.9	48.8	13.1
DualCross (Ours)	✓	✓	**17.0**	**51.8**	**16.9**

Dry → Rain	*DA*	*CM*	Vehicle	Road	Lane
MonoLayout [24]	✗	✗	20.6	68.7	13.1
OFT [32]	✗	✗	24.1	79.8	16.2
LSS [29]	✗	✗	27.8	71.0	16.8
Wide-range Aug.	✓	✗	28.2	71.2	17.2
Vanilla DA	✓	✗	29.1	70.8	18.3
Depth-Supv DA	✓	✓	**29.6**	71.8	19.1
Input-fusion Teacher	✓	✓	29.5	71.0	18.8
DualCross (Ours)	✓	✓	**29.6**	**71.9**	**19.5**

TABLE II: DualCross achieves the best performance under ***city-to-city*** shift in IoU.

Boston → Singapore	*DA*	*CM*	Vehicle	Road	Lane
MonoLayout [24]	✗	✗	14.2	35.9	7.5
OFT [32]	✗	✗	16.8	37.9	9.6
LSS [29]	✗	✗	17.6	38.2	10.6
Wide-range Aug.	✓	✗	17.9	40.5	12.4
Vanilla DA	✓	✗	13.0	31.4	9.1
Depth-Supv DA	✓	✓	19.0	42.8	14.9
Input-fusion Teacher	✓	✓	18.6	42.7	14.1
DualCross (Ours)	✓	✓	**20.5**	**43.1**	**15.6**

TABLE III: DualCross achieves the best performance under ***dataset-to-dataset*** domain gaps in IoU.

nuScenes → Lyft	*DA*	*CM*	Vehicle
MonoLayout [24]	✗	✗	11.8
OFT [32]	✗	✗	16.5
LSS [29]	✗	✗	19.9
Wide-range Aug.	✓	✗	21.9
Vanilla DA	✓	✗	22.5
Depth-Supv DA	✓	✓	23.4
Input-fusion Teacher	✓	✓	22.8
DualCross (Ours)	✓	✓	**24.4**

The supervision from the LiDAR-Teacher is composed of a direct depth estimation loss $\mathcal{L}_{\text{dp}}$ and a teacher feature supervision $\mathcal{L}_{\text{T}}$. As described in Sec. III-A, given the 3D depth volume $\boldsymbol{\alpha} \in \mathbb{R}^{H \times W \times N_d}$, the direct depth supervision term $\mathcal{L}_{\text{dp}}$ is formulated as a cross-entropy loss between the estimated 3D depth distribution volume $\tilde{\boldsymbol{\alpha}}^s$ in the source domain, and the GT depth volume $\boldsymbol{\alpha}^s$ calculated from LiDAR point clouds as described in Sec. III-B:

$$\mathcal{L}_{\text{dp}}(\tilde{\boldsymbol{\alpha}}^s, \boldsymbol{\alpha}^s) = -\sum_{i=1}^{H}\sum_{j=i}^{W}\sum_{k=1}^{N_d} \alpha^s_{(i,j,k)} \log \tilde{\alpha}^s_{(i,j,k)}. \tag{4}$$

And for the LiDAR-Teacher feature supervision: $\mathcal{L}_{\text{T}}(\boldsymbol{F}^{\text{te}}, \boldsymbol{F}^{\text{st}}) = \mathcal{L}_2(\boldsymbol{F}^{\text{te}}, \boldsymbol{F}^{\text{st}})$ is an $\mathcal{L}_2$ loss, where $\boldsymbol{F}^{\text{te}}$ and $\boldsymbol{F}^{\text{st}}$ are the feature maps of teacher and student models, respectively. Finally, the domain adaptation loss contains $\mathcal{L}_{\text{D}_1}$ and $\mathcal{L}_{\text{D}_2}$ with the form described in Eq. 2.

The final objective is a multi-task optimization problem:

$$\mathcal{L}_{\text{DualCross}} = \mathcal{L}_{\text{F}} + \lambda_{\text{T}}\mathcal{L}_{\text{T}} + \lambda_{\text{dp}}\mathcal{L}_{\text{dp}} + \lambda_{\text{D}_1}\mathcal{L}_{\text{D}_1} + \lambda_{\text{D}_2}\mathcal{L}_{\text{D}_2}, \tag{5}$$

where $\lambda_{\text{T}}, \lambda_{\text{dp}}, \lambda_{\text{D}_1}$, and λ_{D_2} are weights for the corresponding loss terms. Our model is trained end-to-end using the loss term in Eq. 5. During inference, target samples are fed into the Camera-Student model to output the final BEV representation. More training details are provided in Sec. IV.

IV. Experiments

A. Datasets, Domain Settings, and Implementation

We evaluate DualCross with four unique domain shift settings constructed from two large-scale datasets, nuScenes [1] and Lyft [13], following existing LiDAR-based domain adaptation work, including SRDAN [44], ST3D [42], UDA3D [21], and xMUDA [12]. Specifically, for the *day-to-night*, *city-to-city*, and *dry-to-rain* settings, we use the sentence in the nuScenes dataset and filter the keywords to split the dataset into corresponding subsets to create the intra-class adaptation scenarios. For the *dataset-to-dataset* setting, we use the official split of the nuScenes dataset, and the split provided in ST3D [42] for the Lyft dataset. All adaptation settings follow the assumption that the source has access to camera and LiDAR sensors, while the target only has cameras. We use all six cameras provided by the nuScenes dataset. We also analyze surprising observations on cross-modality performance in the ablation study.

Following [29], we use EfficientNet [34] pretrained on ImageNet as our image encoder backbone. Two heads are applied to estimate pixel features and pixel-wise depth distribution from the $8\times$ down-sampled feature map. The 3D feature maps are projected to the BEV frame using mean pooling. For the BEV decoder we use ResNet-18 as the backbone, and upsample the features learned from the first three meta-layers of ResNet to the final BEV output. The D_1 and D_2 domain discriminators are applied to the output feature layers of EfficientNet and ResNet backbones, respectively. We use a light-weight discriminator architecture, which is composed of a global averaging pooling layer, followed by two fully-connected layers, and outputs the domain label. For input, we resize and crop input images to size 128×352. For output, we consider a 100 meters$\times100$ meters range centered at the ego-vehicle, with the grid size set to be 0.5 meters $\times\, 0.5$ meters. The depth bin is set to be 1.0 meter between 4.0 meters and 45.0 meters range. The whole model is trained end-to-end, with $\lambda_{\text{T}} = 1.0, \lambda_{\text{dp}} = 0.05, \lambda_{\text{D}_1} = 0.1,$ and $\lambda_{\text{D}_2} = 0.01$. We train DualCross using the Adam [14] optimizer with learning rate 0.001 and weight decay $1e$-7 for 50K steps for the teacher model, and 200K for the student model. We use horizontal flipping, random cropping, rotation, and color jittering augmentation during training. The whole model is implemented using the PyTorch framework [28].

B. BEV Segmentation Results and Comparisons

Baselines. We compare our method with state-of-the-art BEV 3D layout perception work MonoLayout [24], OFT [32], LSS [29], as well as other baseline methods in domain adaptation and cross-modality learning. *Wide-range Aug.* means using a wide range of random scaling

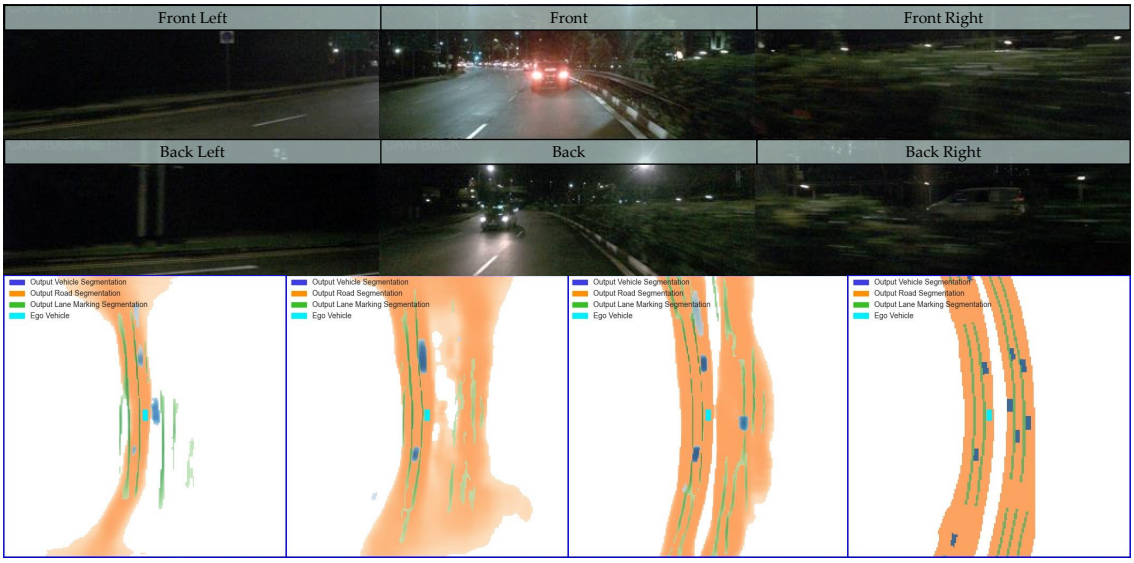

Fig. 3: **Qualitative results in the Day → Night setting** (model is trained with daytime data, and validated with night data). We notice that DualCross performs significantly better than other baselines for vehicles, drivable roads, and lane marking classes. From **left** to **right**: (1) Vanilla adversarial learning; (2) LiDAR as depth supervision with adversarial learning; (3) our DualCross model; and (4) ground truth. **Best viewed in color**.

augmentation which potentially includes the target domain scale. For *Vanilla DA*, we adapt camera-only DA-Faster [3] to our BEV perception setting. *Depth-Supv DA* stands for depth-supervised domain adaptation. We use source domain LiDAR as ground truth to supervise the depth estimation during training, without LiDAR-Teacher supervision (only $\mathcal{L}_{\text{dp}}$ without $\mathcal{L}_{\text{T}}$). *Input-fusion Teacher* is an alternative way of designing the LiDAR-Teacher, where we directly fuse point (x, y, z) coordinates into their corresponding image pixels as additional channels in the teacher model, similar to Pointpainting [35]. We use ***DA*** and ***CM*** to denote whether a model considers domain adaptation and cross-modality (making use of source-exclusive Lidar data) in design, respectively. Results are reported on vehicle, drivable roads, and lane classes using intersection-over-union (IoU).

Day-to-Night Adaptation. As shown in Table I on the left, we observe that our DualCross model achieves the best performance on all classes. We notice that the improvement under the Day → Night setting is exceptionally high. This is because the initial domain gap between day and night scenarios is very large in the camera modality space. Moreover, the LiDAR sensor is robust under illumination changes, due to its active imaging mechanism as opposed to camera's passive one. Thus, incorporating LiDAR point cloud information helps the model to learn a more robust, illumination-invariant representation from the image inputs.

Dry-to-Rain Adaptation. As shown in Table I on the right, under this setting we also observe that our DualCross model achieves the best performance on all classes. We notice that the improvement under the Dry → Rain setting is not as big as the previous setting. This is because the domain gap between dry and rain scenarios is not big in the image modality. Hence, baseline methods OFT and LSS are already able to obtain decent results even without domain adaptation. Furthermore, rainy weather is known to cause great domain shift in the LiDAR modality [40]. As a result, the knowledge learned from source LiDAR suffers from an unknown domain shift which hinders its usefulness. This can potentially cancel out the benefit of 3D information learned from point clouds and explains for the smaller improvement.

Dataset-to-Dataset Adaptation. As shown in Table III, we also observe that our DualCross model achieves the best performance in the nuScenes → Lyft setting. Following [29], because Lyft does not provide road segment and lane marking information in the high-definition (HD) map, we report results on the vehicle class. Compared with baselines with and without domain adaptation or cross-modality learning, our DualCross demonstrates superior performance in leveraging and adapting LiDAR information.

City-to-City Adaptation. As shown in Table II, we observe that our DualCross model achieves the best performance on all classes for two inter-city transfer settings. Without domain adaptation, baseline approaches MonoLayout, OFT, and LSS all suffer from performance degradation. Direct depth supervision and alternative input-fusion teacher models do not bring as much improvement as DualCross. The results clearly demonstrate the effectiveness of our method by distilling and aligning the LiDAR information for cross-modality 3D BEV perception.

Qualitative Results. As shown in Figure 3, under the Day → Night domain shift setting, our model achieves significantly better monocular 3D perception than other baselines. We observe that DualCross provides more clearly defined road boundaries and lane markings. The depth and size of the vehicles and the road on the right side are also predicted more accurately. DualCross only misses some vehicles that are hardly visible in the camera due to occlusion and distance. Overall, the qualitative results validate the effectiveness of

TABLE IV: DualCross achieves the best performance under simultaneous modality and domain shift for the 3D object detection task. Domain shift is **Singapore** → **Boston**.

Modality-shift + Domain Shift	mAP ↑	NDS ↑
LSS [29]	16.0	20.3
MonoDistill [6]	16.5	21.9
Depth-Supv DA	19.1	23.5
DualCross (Ours)	**22.5**	**26.1**

TABLE V: DualCross achieves great performance under only modality shift for the 3D object detection task. EfficientNet and ResNet50 are backbones for image feature extraction.

Modality-shift Only	mAP ↑	NDS ↑
Set2Set [36]	33.1	41.0
MonoDistill [6]	34.3	41.2
DualCross-EfficientNet	34.5	41.5
DualCross-ResNet50	**35.2**	**42.4**

TABLE VI: Our proposed components all contribute to the final performance. We report results on vehicle class under ***day-to-night*** domain gap in IoU. *WA, AD, LS, LT* stand for Wide Augmentation, Adversarial Discriminators, LiDAR Supervision, and LiDAR-Teacher, respectively.

Baseline	WA	AD	LS	LT	Results	*diff*
✓					6.7	0
✓			✓		6.4	−0.3
✓				✓	8.9	+2.2
✓	✓				10.3	+3.6
✓	✓	✓			11.2	+4.5
✓	✓	✓	✓		15.7	+9.0
✓	✓	✓	✓	✓	**17.0**	+10.3

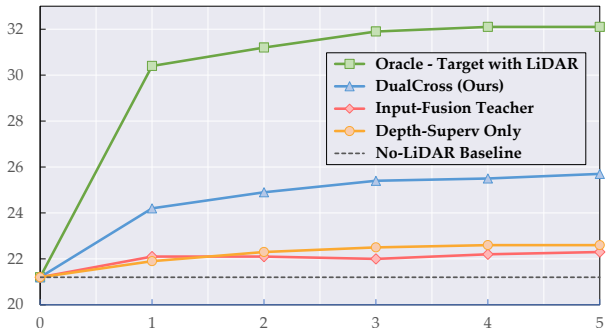

Fig. 4: Results of DualCross improve as the number of LiDAR scans increases.

DualCross in closing the gap between data domains and leveraging point cloud information for better 3D inference.

C. 3D Detection Results and Comparisons

Baselines. In addition to the BEV segmentation task, we compare DualCross with existing cross-modality 3D detection models, MonoDistill [6] and Set2Set [36]. MonoDistill is designed for the single-camera setting on the KITTI dataset. We extend it into a multi-camera setting for a fair comparison. We evaluate with mean Average Precision (mAP) and Nuscenes Detection Score (NDS) metrics [1].

Modality-Shift Only. As shown in Table V, our model achieves the best performance. By using ResNet-50, we can achieve more than 1% improvement in mAP and NDS metrics. Moreover, our model still runs 42ms per frame at test time, faster than MonoDistill which runs 80ms per frame.

Modality-Shift + Domain-Shift. As shown in Table IV, under concurrent modality and domain gaps, DualCross outperforms previous baselines by a large margin, demonstrating a more robust and trustworthy model in real-world scenarios.

D. Analysis and Ablation Study

Direct Lidar Supervision Leads to Worse Performance. It is naturally believed that introducing multiple sensors in the perception model is bound to increase the model performance. Surprisingly, experiments shown in Table VI negate this naive intuition. When we introduce the LiDAR sensor in the source domain as depth supervision, the result decreases by 0.3, and we observe constant decreases among different domain shift settings. As we described in Sec. III-B, the domain distribution divergence increases after introducing the sensor-modality shift. As a result, we propose multiple components in DualCross to account for the visual and sensor domain shifts. Experiments show that while the wide augmentation strategy and adversarial discriminator both achieve better results than the baseline (11.2 vs. 6.7 in IoU), our LiDAR-Teacher further boosts the result to 17.0 by leveraging effective LiDAR knowledge distillation.

LiDAR Density & Comparison with Oracle Model. As shown in Figure 4, we validate that our model achieves higher performance when denser LiDAR is available. This can be accomplished by grouping continuous scans of LiDAR point clouds (from 1 to 5) into a single unit, to have a denser 3D representation of the scene. We observe that other cross-modality baselines including *Input-Fusion Teacher* and *Depth-Supv* models cannot effectively leverage the LiDAR knowledge, even with dense point clouds available. We also compare our model with the LiDAR oracle model (target domain also has the LiDAR modality) and find that the gap between the upper bound result and the No-LiDAR baseline is significantly reduced. The remaining performance gap is caused by the unknown LiDAR domain gap which we hope to further reduce in future work.

Dealing with Mixed Domain Shift. Another common but under-explored question we observe in the 3D domain adaption setting is the mixed domain shift problem, where multiple types of gaps between source and target domains occur concurrently. For example, in the nuScenes dataset, the Boston data are collected exclusively during daytime, whereas the Singapore data encompass both day and night captures. This leads to a mixture of city-wise and lighting-wise domain shifts. As shown in Table VII, we find that directly leveraging domain adaptation in this scenario leads to worse performance than direct inference, because mixed domains in the target confuse the discriminator. Hence, we propose a progressive learning mechanism, where we first perform adaptation with city-wise data for $100K$ steps, and then train the model on the full target domain dataset for another $150K$ steps. This effectively alleviates the mixed domain shift problem, and helps DualCross achieve leading results than other baselines.

TABLE VII: The proposed progressive learning strategy effectively addresses the challenge caused by mixed domain gap scenario (***Boston-to-Singapore*** mixed with the ***day-to-night***) on nuScenes.

Mixed Domain Gap	Vehicle	Road	Lane
Direct Inference	17.6	38.2	10.6
Vanilla DA	13.0	31.4	9.1
Progressive DA	18.8	41.5	13.2
DualCross (Ours)	**20.5**	**43.1**	**15.6**

TABLE VIII: DualCross achieves great perception results with efficient inference time compared with the baseline methods.

	#Params (M)	Frame-per-Second (FPS)
OFT [32]	22	25
LSS [29]	14	35
DualCross (Ours)	15	33

Computational Complexity Table VIII summarizes the number of parameters and inference speed for prior baselines and our model. Our Lidar-Teacher distillation and multi-level adversarial learning modules do not affect the inference efficiency of DualCross, compared with the baselines. Our total number of parameters is 15M, and our inference time is 33 Frame-per-Second (FPS) on a V100 GPU, which is on par with the baseline LSS [29]. The training time for our model is around 20 hours on 4×V100 GPUs.

V. Conclusion

In this paper, we propose DualCross to estimate 3D scene representation in BEV under domain shift and modality change. To achieve this, we construct a LiDAR-Teacher and distill knowledge from it into a Camera-Student by feature supervision. We further propose to align feature space between the domains using multi-stage adversarial learning. Results on large-scale datasets with diverse domain gaps demonstrate the effectiveness of our approach, which marks a significant step towards robust 3D perception in the wild.

Acknowledgement. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, the NCSA Fellows program, the IBM-Illinois Discovery Accelerator Institute, the Illinois-Insper Partnership, and the Amazon Research Award.

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.
[2] Y. B. Can, A. Liniger, *et al.*, "Structured bird's-eye-view traffic scene understanding from onboard images," in *ICCV*, 2021.
[3] Y. Chen, W. Li, C. Sakaridis, D. Dai, *et al.*, "Domain adaptive faster R-CNN for object detection in the wild," in *CVPR*, 2018.
[4] Z. Chen, Z. Li, S. Zhang, L. Fang, *et al.*, "BEVDistill: Cross-modal BEV distillation for multi-view 3D object detection," in *ICLR*, 2023.
[5] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *ICCV*, 2019.
[6] Z. Chong, X. Ma, H. Zhang, *et al.*, "MonoDistill: Learning dpatial features for monocular 3D object detection," in *ICLR*, 2022.
[7] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.
[8] R. Gong, D. Dai, *et al.*, "mDALU: Multi-source domain adaptation and label unification with partial datasets," in *ICCV*, 2021.
[9] X. Guo, S. Shi, *et al.*, "Liga-stereo: Learning LiDAR geometry aware representations for stereo-based 3D detector," in *ICCV*, 2021.
[10] N. Hendy, C. Sloan, F. Tian, P. Duan, N. Charchut, *et al.*, "Fishing net: Future inference of semantic heatmaps in grids," in *CVPR*, 2020.
[11] Y. Hong, H. Dai, and Y. Ding, "Cross-modality knowledge distillation network for monocular 3D object detection," in *ECCV*, 2022.
[12] M. Jaritz, T.-H. Vu, *et al.*, "xmuda: Cross-modal unsupervised domain adaptation for 3D semantic segmentation," in *CVPR*, 2020.
[13] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, *et al.*, "Level 5 perception dataset 2020," https://level-5.global/level5/data/, 2019.
[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[15] J. N. Kundu, P. K. Uppala, *et al.*, "Adadepth: Unsupervised content congruent adaptation for depth estimation," in *CVPR*, 2018.
[16] M. Li *et al.*, "Cross-domain and cross-modal knowledge distillation in domain adaptation for 3D semantic segmentation," in *ACMMM*, 2022.
[17] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3D object detection," 2022.
[18] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, *et al.*, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," *AAAI*, 2023.
[19] Y.-J. Li, X. Dai, C.-Y. Ma, Y.-C. Liu, K. Chen, B. Wu, *et al.*, "Cross-domain adaptive teacher for object detection," in *CVPR*, 2022.
[20] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3D object detection," in *CVPR*, 2019.
[21] Z. Luo, Z. Cai, C. Zhou, G. Zhang, H. Zhao, S. Yi, S. Lu, H. Li, S. Zhang, and Z. Liu, "Unsupervised domain adaptive 3D detection with multi-level consistency," in *ICCV*, 2021.
[22] Y. Man, L.-Y. Gui, and Y.-X. Wang, "BEV-guided multi-modality fusion for driving perception," in *CVPR*, 2023.
[23] Y. Man, X. Weng, P. K. Sivakumar, M. O'Toole, and K. M. Kitani, "Multi-echo LiDAR for 3D object detection," in *ICCV*, 2021.
[24] K. Mani, S. Daga, S. Garg, S. S. Narasimhan, *et al.*, "Monolayout: Amodal scene layout from a single image," in *WACV*, 2020.
[25] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, *et al.*, "Improved knowledge distillation via teacher assistant," in *AAAI*, 2020.
[26] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *RA-L*, 2020.
[27] J. Park, X. Weng, Y. Man, and K. Kitani, "Multi-modality task cascade for 3D object detection," in *BMVC*, 2021.
[28] A. Paszke, S. Gross, F. Massa, A. Lerer, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.
[29] J. Philion *et al.*, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *ECCV*, 2020.
[30] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas, "Imvotenet: Boosting 3D object detection in point clouds with image votes," in *CVPR*, 2020.
[31] C. Qin, H. You, L. Wang, *et al.*, "Pointdan: A multi-scale 3D domain adaption network for point cloud representation," in *NeurIPS*, 2019.
[32] T. Roddick *et al.*, "Orthographic feature transform for monocular 3D object detection," *arXiv preprint arXiv:1811.08188*, 2018.
[33] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.
[34] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.
[35] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3D object detection," in *CVPR*, 2020.
[36] Y. Wang and J. M. Solomon, "Object DGCNN: 3D object detection using dynamic graphs," in *NeurIPS*, 2021.
[37] X. Weng and K. Kitani, "Monocular 3D object detection with pseudo-LiDAR point cloud," in *ICCVW*, 2019.
[38] X. Weng, Y. Wang, Y. Man, and K. M. Kitani, "GNN3DMOT: Graph neural network for 3D multi-object tracking with 2d-3D multi-feature learning," in *CVPR*, 2020.
[39] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud," in *ICRA*, 2019.
[40] Q. Xu, Y. Zhou, *et al.*, "Spg: Unsupervised domain adaptation for 3D object detection via semantic point generation," in *ICCV*, 2021.
[41] X. Yan, J. Gao, C. Zheng, *et al.*, "2DPASS: 2D priors assisted semantic segmentation on LiDAR point clouds," in *ECCV*, 2022.
[42] J. Yang, S. Shi, Z. Wang, *et al.*, "ST3D: Self-training for unsupervised domain adaptation on 3D object detection," in *CVPR*, 2021.
[43] J. H. Yoo, Y. Kim, J. S. Kim, and J. W. Choi, "3D-CVF: Generating joint camera and LiDAR features using cross-View spatial feature fusion for 3D object detection," in *ECCV*, 2020.
[44] W. Zhang *et al.*, "Srdan: Scale-aware and range-aware domain adaptation network for cross-dataset 3D object detection," in *CVPR*, 2021.
[45] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *CVPR*, 2019.
[46] M. Zhu *et al.*, "Cross-modality 3D object detection," in *WACV*, 2021.
[47] Y. Zhu and Y. Wang, "Student customized knowledge distillation: Bridging the gap between student and teacher," in *ICCV*, 2021.