2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

BEVUDA: Multi-geometric Space Alignments for Domain Adaptive BEV 3D Object Detection

Jiaming Liu[1,3*], Rongyu Zhang[1,2*], Xiaoqi Li[1*], Xiaowei Chi[1], Zehui Chen[1], Ming Lu[1], Yandong Guo[3], Shanghang Zhang[1] ✉

***Abstract*— Vision-centric bird-eye-view (BEV) perception has shown promising potential in autonomous driving. Recent works mainly focus on improving efficiency or accuracy but neglect the challenges when facing environment changing, resulting in severe degradation of transfer performance. For BEV perception, we figure out the significant domain gaps existing in typical real-world cross-domain scenarios and comprehensively solve the Domain Adaption (DA) problem for multi-view 3D object detection. Since BEV perception approaches are complicated and contain several components, the domain shift accumulation on multiple geometric spaces (i.e., 2D, 3D Voxel, BEV) makes BEV DA even challenging. In this paper, we propose a Multi-space Alignment Teacher-Student (MATS) framework to ease the domain shift accumulation, which consists of a Depth-Aware Teacher (DAT) and a Geometric-space Aligned Student (GAS) model. DAT tactfully combines target lidar and reliable depth prediction to construct depth-aware information, extracting target domain-specific knowledge in Voxel and BEV feature spaces. It then transfers the sufficient domain knowledge of multiple spaces to the student model. In order to jointly alleviate the domain shift, GAS projects multi-geometric space features to a shared geometric embedding space and decreases data distribution distance between two domains. To verify the effectiveness of our method, we conduct BEV 3D object detection experiments on three cross-domain scenarios and achieve state-of-the-art performance. Code: https://github.com/liujiaming1996/BEVUDA.**

I. Introduction

Camera-based 3D object detection, particularly in the context of autonomous driving [1], [3], [4], [5], has garnered increasing attention. Notably, advancements have been evident, primarily driven by Bird-Eye-View (BEV) perception methods [6], [7], when the test data distribution aligns with the training data. However, real-world machine perception systems typically operate in dynamic and ever-changing environments [8], [9], [10], [11], as illustrated in Fig.1. In these scenarios, we observe a substantial domain gap in typical real-world cross-domain situations [12], [13], resulting in significant performance degradation of LSS-based BEV method [2], [14]. Recently, though mono-view 3D detection methods [15], [13] study the domain shift problems of different camera parameters or annotation methods variation, domain adaptation on many real-world scenarios are still unexplored in both Mono-view [16], [17] and Multi-view[18], [19], [20]. Therefore, our endeavor is to delve into unsupervised domain adaptation (UDA) challenges within the realm of BEV perception.

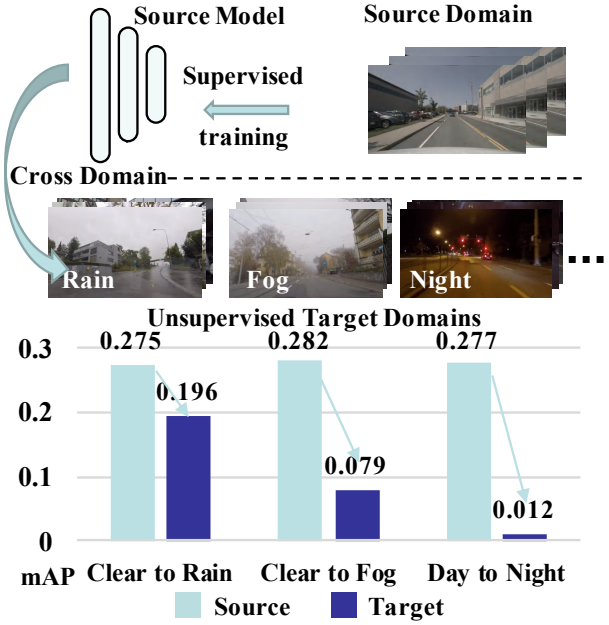

Fig. 1. BEV 3D detectors exhibit impressive performance when tested under data distributions closely resembling their training data. However, real-world machine perception systems (i.e., autonomous driving [1]) operate in non-stationary and constantly changing environments, which leads to tremendous performance degradation. All methods are built on BevDepth [2] with a ResNet-50 backbone and evaluated on the unlabeled target domain.

Multi-view LSS-based methods [6], [2] tend to be intricate, comprising numerous components. This complexity, coupled with domain shift accumulation across various geometric spaces, poses significant challenges for BEV-oriented UDA: (1) *2D images geometric space.* Since multi-view images contain abundant semantic information, it will result in a manifold domain shift when the environment changes. (2) *3D Voxel geometric space.* Voxel features, formed from image features and potentially unreliable depth predictions from a source pre-trained model, contribute to even more pronounced domain shift in the target domain. (3) *BEV geometric space.* Due to the shift in the above spaces, the further constructed BEV feature results in an accumulation of domain shift and leads to noises for final prediction.

To this end, we propose a BEV-oriented Multi-space Alignment Teacher-Student (MATS) framework to disentangle accumulated domain shift problems, which consists of a Depth-Aware Teacher (DAT) and a Geometric-space Aligned Student (GAS) model. Since BEV feature construction heavily

[1] Jiaming Liu, Xiaoqi Li, Xiaowei Chi, Zehui Chen, Ming Lu, and Shanghang Zhang are with National Key Laboratory for Multimedia Information Processing, School of CS, Peking University. [2] Rongyu Zhang is with Nanjing University. [3] Jiaming Liu, Yandong Guo are with *AI²Robotics*.
*: Equal Contribution: jiamingliu@stu.pku.edu.cn
✉ Corresponding Author: shanghang@pku.edu.cn

relies on the accuracy of depth [2], [21], DAT tactfully combines target lidar data and reliable depth prediction to compose depth-aware information. Specifically, the reliable depth prediction is screened out by the uncertainty scheme, selecting the depth pixels with lower uncertainty values and performing stable estimation during the cross-domain phase. In this way, we are able to construct reliable corresponding voxel and BEV features by depth-aware information and extract target domain-specific knowledge in the DAT. Then, the target domain knowledge is transferred from DAT to the student model, aiming to further address the distribution shift between the two domains. Since multi-geometric features (i.e., 2D image, 3D voxel, and BEV) are of geometric consistency, we propose GAS to project multi-latent space features to a shared geometric embedding space, facilitating joint alignment of source and target domain feature representations.

To evaluate the effectiveness of our method, we design three UDA scenarios, which are **Scene** (from Boston to Singapore), **Weather** (from clear to rainy), and **Day-night** in [22]. The main contributions are summarized as follows:

1) We explore the UDA problem for BEV perception of Multi-view 3D object detection. We propose a Multi-space Alignment Teacher-Student (MATS) framework to address the domain shift accumulation on multi-geometric spaces.

2) In MATS, we propose a Depth-Aware Teacher (DAT) model to fully extract target domain-specific knowledge by leveraging depth-aware information. To take full advantage of the domain knowledge, we propose a Geometric-space Aligned Student (GAS) model that projects multi-latent space features to a shared geometric embedding space and jointly decreases data distribution distance between two domains.

3) We conduct extensive experiments on the three UDA scenarios, achieving SOTA performance compared with previous Mono-view 3D and 2D detection UDA methods.

II. Related work

A. Camera-based 3D object detection.

Nowadays, 3D Object Detection plays an important role in autonomous driving and machine scene understanding. Two paradigms are prominent in this aspect: Single-view [16], [17], [23], [24], [25], [26], [27], [28], [29], [30] and Multi-view[6], [31], [18], [32], [33], [34], [7], [35], [19], [20], [2], [36], [21]. In Single-view detection, previous works can be categorized into several streams, i.e. leveraging CAD models [25], [26], [27], setting prediction targets as key points [28], [29], and disentangling transformation for 2D and 3D detection [30]. Specifically, FCOS3D [17] can predict 2D and 3D attributes synchronously. M3D-RPN [23] considers single-view 3D object detection task as a standalone 3D region proposal network. [16] calculates the depth of the objects by integrating the actual height of the objects. To better utilize the depth information in the process, [37] proposes an end-to-end depth-aware transformer network. However, taking into account the precision and practicality of detection, more and more multi-view 3D object detectors are proposed.

The Multi-view paradigm can be categorized into two branches, namely transformer-based [38] and LSS-based [6]. First of all, to extend DETR [38] into 3D detection, DETR3D [31] first predicts 3D bounding boxes with a transformer network. Inspired by DETR3D, some works adopt object queries [18], [32], [33], [34] or BEV grid queries [7] to extract features from images and utilize attention method, resulting in better 2D-to-3D transformation. However, transformer-based methods don't project image features to BEV representation. Following LSS [6], some methods [35], [19], [20] predict a distribution over lidar depth and generate a point cloud with multi-view image features for 3D detection. Specifically, BevDepth [2] introduces depth supervision and speeds up the operation of voxel pooling. Bevdet4d [36] and BevStereo [21] thoroughly explore temporal information in the task and concatenate volumes from multiple time steps. In this paper, we adopt BevDepth [2] as the baseline 3D object detector for its simple and powerful workflow, along with its great potential in cross-domain feature extraction.

B. UDA in 3D object detection.

Domain Adaptive Faster R-CNN [39] first probes the cross-domain problem in object detection. Based on [40], most previous works [41], [42], [12], [43], [44], [45], [46] follow the cross-domain alignment strategy and explore the influence of domain shift in multi-level features. As for 3D object detection, [47], [13], [48] investigate Unsupervised Domain Adaptation (UDA) strategies for point cloud 3D detectors. In particular, [47], [48] adopt alignment methods to align the feature and instance level information between two domains. STM3D [13] develop self-training strategies to realize UDA by consistent and high-quality pseudo labels. Recently, some works [49], [50], [51], [52] investigate the cross-domain strategies in BEV perception, which aim to reduce the simulation-to-real domain shift. In terms of camera-based monocular 3D object detection, [15], [13] first attempt to disentangle the camera parameters and guarantee the geometry consistency in the cross-domain phase. In contrast, we dedicate to solving the domain shift accumulation problem in multi-view 3D object detection tasks, which infer 3D scenes from the BEV perspective.

III. Methods

A. Problem Formulation

For the UDA setting [53], we are access to labeled source domain $D_s = \{\{I_s^i\}_{j=1}^M, L_s^i, G_s^i\}_{i=1}^{N_s}$ and unlabeled target domain $D_t = \{\{I_t^i\}_{j=1}^M, L_t^i\}_{i=1}^{N_t}$ of N samples and M camera views, in which I^i, L^i, and G^i denote images, lidar, and detection ground truth respectively. Following camera-based works [2], [21], we only utilize lidar supervision during training.

B. Overall Framework and Motivation

The overall framework is depicted in Fig .2. Initially, we utilize encoders to extract features from multi-view image observations in both the teacher and student models. In the Depth-Aware Teacher model, these features are transformed into 3D Voxel and BEV representations, incorporating depth-aware information. Our goal is to extract sufficient target domain knowledge, which is subsequently transferred to the

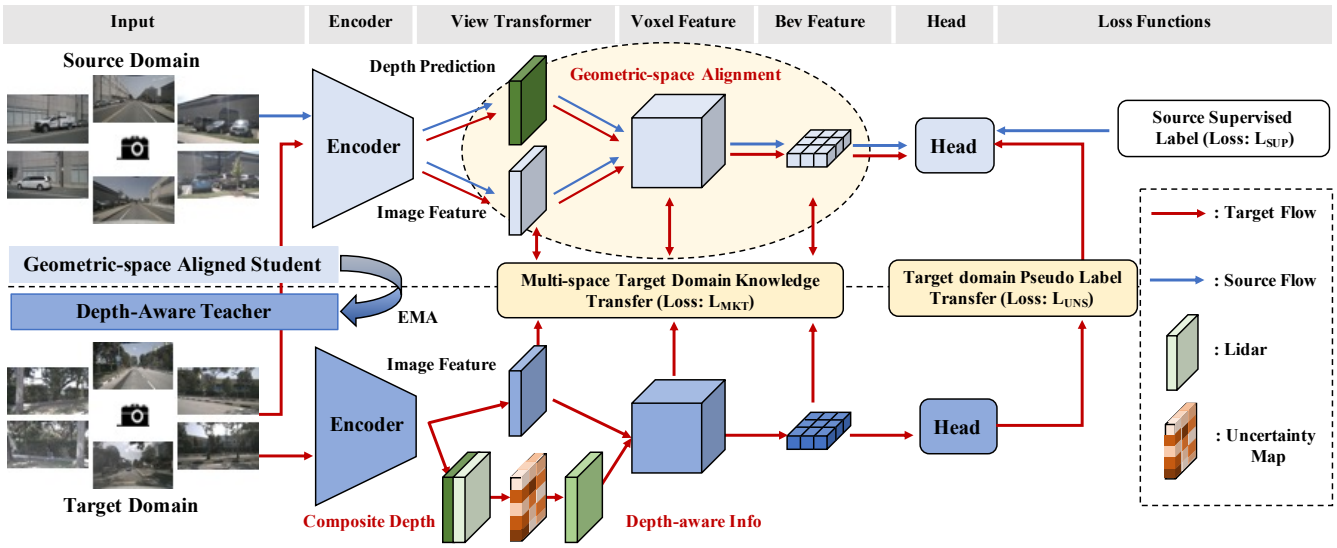

Fig. 2. The framework of Multi-space Alignment Teacher-Student (MATS) is composed of the Depth-Aware Teacher (DAT) and Geometric-space Aligned Student (GAS) model. In **the bottom part**, the DAT model takes target domain input and adopts depth-aware information to construct Voxel and BEV features with sufficient target domain knowledge, which is further transferred to the student model in the multi-latent space (i.e., 2D image, 3D voxel, and BEV). In **the upper part**, the GAS model takes two domains input and decreases data distribution distance in a shared geometric embedding space. MATS framework aims to comprehensively address the multi-geometric space domain shift accumulation problem.

student model within the multi-latent space. The Geometric-space Aligned Student model then minimizes the distribution gap between the two domains within a shared geometric embedding space. Finally, we incorporate a task-specific head to facilitate 3D object detection.

Teacher-Student framework Inspired by the observation that mean teacher predictions often exhibit higher quality than standard models [54], we employ a teacher model to provide more accurate pseudo labels during the domain adaptation process. Additionally, considering the robustness of teacher-student framework in dynamic environments [55], we leverage it to maintain stability in target domains.

Depth-Aware Teacher. In the context of LSS-based BEV perception, the accuracy of depth information is pivotal, as it forms the cornerstone of BEV feature construction alongside 2D features [2], [21]. To enhance the reliability of depth information in the target domain, we introduce DAT. This model cleverly combines target lidar data with reliable depth predictions, creating depth-aware information to extract valuable knowledge specific to the target domain.

Geometric-space Aligned Student. Given that the BEV features are obtained through 3D Voxel pooling, and Voxel features are constructed from 2D features and depth, these latent spaces exhibit geometric consistency [6]. This insight prompts us to address all space domain shifts simultaneously, as opposed to tackling them individually in each geometric space. To achieve this, we project multi-latent space features into a shared geometric embedding space, effectively reducing the distribution gap between the two domains.

C. Depth-aware teacher

As shown in Fig .3, we construct composite depth-aware information in the teacher model by combining sparse lidar data with reliable depth prediction. Since lidar data can reflect the most accurate depth information, we adopt lidar data if

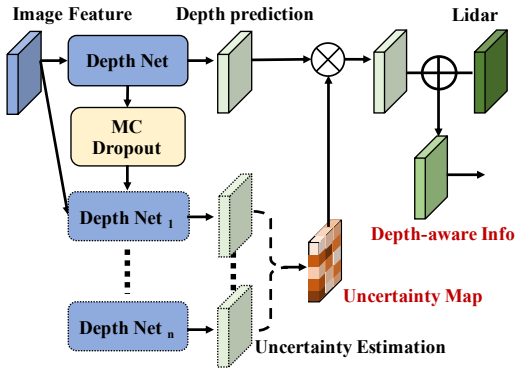

Fig. 3. The detailed process of constructing depth-aware information. The uncertainty map is estimated by MC Dropout [56].

the pixel holds. For pixels without lidar data, we reserve the depth prediction if it is of reliable accuracy to avoid noises. Though confidence is a straightforward measurement to reflect reliability, it is trustless in pixel-wise cross-domain scenarios. Therefore, we explore a new solution in dense prediction domain adaptation tasks to reduce the noise in training, which is adopting an uncertainty mechanism to select reliable depth estimations. Specifically, we adopt the Dropout method [56] to realize m times (e.g., m = 10) forward propagation and obtain m group probabilities for each pixel. We calculate the uncertainty map of the depth prediction and figure out how it is influenced by domain shift:

$$\mathscr{U}(x_j) = \left(\frac{1}{m} \sum_{i=1}^{m} \| p_i(y_j|x_j) - \mu \|^2 \right)^{\frac{1}{2}} \quad (1)$$

, where $p_i(y_j|x_j)$ is the input pixel x_j probability of i^{th} forward propagation, μ is the mean probability (m rounds) of x_j. $\mathscr{U}(x_j)$ represents the uncertainty of the depth sub-network for pixel-wise target input x_j. Predictions with low uncertainty will be reserved since they are more adaptive in the target

domain. Therefore, DAT can better extract target domain knowledge with the help of depth-aware information, which is then transferred to the student model for domain alignment.

The rest of the teacher model is built with exponential moving average (EMA) [54]. The initial weights of teacher $\mathscr{T}_{DAT}$ and student models $\mathscr{S}_{TGT.}$ are loaded from the source pre-trained model. The EMA equation is shown below, where α is a smoothing coefficient and t is the training iteration.

$$\mathscr{T}^{t}_{DAT.} = \alpha \mathscr{T}^{t-1}_{DAT.} + (1-\alpha)\mathscr{S}^{t}_{TGT.} \tag{2}$$

D. Geometric-space Aligned Student

After receiving the target domain knowledge transferred from DAT, we further introduce Geometric-space Aligned Student (GAS) to fully exploit the transferred knowledge and jointly address the domain shift accumulation. Due to the geometric consistency, we project multi-geometric features (i.e., 2D images, Voxel, and BEV) to a shared geometric embedding space and decrease the data distribution distance between two domains. Specifically, in the source domain, we utilize respective MLPs to project the three features to a shared embedding space, in which the dimension is $3 \times C \times n$, channel dimension C is set to 256, and n is equal to the number of categories. We then rearrange the feature dimension to $3C \times n$ and use a shared MLP to aggregate the category feature to the source domain prototype ($256 \times n$). Meanwhile, the target domain prototype is also projected in the same process. We adopt alignment loss ($\mathscr{L}_{GAS}$) [57] to pull close the two domain prototypes. $\mathscr{L}_{GAS}$ as shown in Eq.3, where F_s and F_t demonstrate the source and target domain prototype respectively and D denotes the domain discriminator.

$$\mathscr{L}_{GAS}(F_s, F_t) = \log D(F_s) + \log(1 - D(F_t)) \tag{3}$$

We employ the traditional detection loss ($\mathscr{L}_{UNS}$) [2] to update the student model, with this loss being applied in conjunction with the target domain pseudo-label as a penalty.

E. Training objectives and inference

When adopting the DAT model to transfer multi-space features to the student, the knowledge transfer loss $\mathscr{L}_{MKT}$ is:

$$\mathscr{L}_{MKT} = \sum_{l \in L} \frac{1}{W'_l \times H'_l} \sum_{i \in \mathscr{P}} ||F^i_{Te,l} - F^i_{St,l}||^2 \tag{4}$$

, where $F^i_{Te,l}$ and $F^i_{St,l}$ stand for the i^{th} pixel value from DAT model and student model at l geometric space, $L \in \{2Dimages, 3DVoxel, BEV\}$. W'_l and H'_l stand for width and height of the transferred features, $\mathscr{P} = \{1, 2, .., W'_l \times H'_l\}$. Meanwhile, the integrated domain adaptation loss $\mathscr{L}_{DA}$ is shown in Eq.5.

$$\mathscr{L}_{DA} = \lambda_1 * \mathscr{L}_{UNS} + \lambda_2 * \mathscr{L}_{SUP} + \lambda_3 * \mathscr{L}_{MKT} + \lambda_4 * \mathscr{L}_{GAS} \tag{5}$$

, where $\mathscr{L}_{SUP}$ is the detection loss [2] penalized by source domain detection label. In order to maintain the balance of loss penalties, λ_1 and λ_2 are set to 1, λ_3, and λ_4 are set to 0.1. During inference, same with other camera-based methods [2], [7], [21], we only adopt multi-view camera data.

TABLE I

RESULTS OF DIFFERENT METHODS FOR SCENE ADAPTATION SCENARIO ON THE VALIDATION SET [22], FROM BOSTON TO SINGAPORE. DA MEANS UTILIZING THE DOMAIN ADAPTION METHOD, AND R50 AND R101 MEAN ADOPTING RESNET 50 AND 101 AS THE BACKBONE.

	Method	Backbone	NDS ↑	mAP ↑
Baseline	BEVDet[19]	R50	0.126	0.117
	BEVDepth[2]	R50	0.174	0.115
	BEVDepth[2]	R101	0.187	0.115
DA	SFA[12](BEVDepth)	R50	0.181	0.124
	STM3D[13](BEVDepth)	R50	0.183	0.129
DA	Ours(BEVDepth)	R50	**0.208**	**0.148**
	Ours(BEVDepth)	R101	**0.211**	**0.166**

IV. EVALUATION

In Sec IV-A, the details of the setup of UDA scenarios and implementation details are given. In Sec IV-B, we evaluate the effectiveness of MATS in three UDA scenarios. The comprehensive ablation studies are conducted in Sec IV-C, which investigate the impact of each component. Finally, we provide qualitative analysis to further evaluate our proposed framework in Sec IV-D.

A. Experimental setup

1) Datasets and adaptation scenarios: We evaluate our proposed framework on nuscenes [22], which is a large-scale autonomous-driving dataset. In order to pave the way for Unsupervised Domain Adaptation (UDA) in multi-view 3D object detection, we split the nuscenes into different paired source-target domain data. We introduce three classical cross-domain scenarios: **Scene**, **Weathers**, and **Day-Night**.

Scene Adaptation We set Boston as the source scene data and realize UDA on the Singapore target domain. Since scene layouts are frequently changing in autonomous driving, the domain gap occurs in multiple scenes [45], [43].

Weathers Adaptation The sunny weather is considered as source domain data, while rainy weather is considered as target domain data. Various weather conditions are common phenomena in the real world, and BEV detection should be reliable under such conditions [41], [12].

Day-Night Adaptation We design daytime as the source domain and realize UDA on the target domain (night data). Since the camera-based method has a tremendous domain gap from day to night, it is essential to explore the domain adaptation method in the day-night scenario [58], [59].

2) Implementation details: MATS framework is built based on BEVDepth [2]. According to previous work [2], [35], [19], [20], ResNet-50 and ResNet-101 [60] serve as backbone to extract image features respectively. We adopt 256×704 as image input size and the same data augmentation methods as [2]. We apply AdamW [61] optimizer with the 2e-4 learning rate and without any decay. For training, the source domain pre-training and UDA transfer training are set to 24 and 12 epochs, respectively. During inference, our method infers without any test time augmentation or model ensemble. We report the evaluation metrics following previous 3D detection works[2], [19], including NuScenes Detection Score (NDS) and mean Average Precision (mAP). All experiments are conducted on NVIDIA Tesla V100 GPUs.

TABLE II

RESULTS OF DIFFERENT METHODS FOR WEATHER ADAPTATION SCENARIOS ON THE VALIDATION SET [22], FROM CLEAR TO RAINY

	Method	Backbone	NDS ↑	mAP ↑
Baseline	BEVDet[19]	R50	0.232	0.207
	BEVDepth[2]	R50	0.268	0.196
	BEVDepth[2]	R101	0.272	0.212
DA	SFA[12](BEVDepth)	R50	0.281	0.200
	STM3D[13](BEVDepth)	R50	0.276	0.212
DA	Ours(BEVDepth)	R50	**0.305**	**0.243**
	Ours(BEVDepth)	R101	**0.308**	**0.247**

TABLE III

RESULTS OF DIFFERENT METHODS FOR DAY-NIGHT ADAPTATION SCENARIO ON THE VALIDATION SET [22].

	Method	Backbone	NDS ↑	mAP ↑
Baseline	BEVDet[19]	R50	0.010	0.009
	BEVDepth[2]	R50	0.050	0.012
	BEVDepth[2]	R101	0.062	0.036
DA	SFA[12](BEVDepth)	R50	0.092	0.032
	STM3D[13](BEVDepth)	R50	0.070	0.035
DA	Ours(BEVDepth)	R50	**0.132**	**0.054**
	Ours(BEVDepth)	R101	**0.188**	**0.127**

B. Main results

We compare our proposed method with other BEV perception methods [19], [2] to verify the superior performance. To further demonstrate our special design in addressing domain shift of multi-view 3D object detection, we reproduce other promising 2D and mono-view 3D detection UDA methods on BEVDepth [2], i.e., SFA [12] and STM3D [13].

Scene Adaptation As shown in Tab .I, MATS outperforms all the baseline methods, which obviously exceeds BEVDepth [2] of R50 and R101 backbone by 3.4% and 2.4% NDS. It thus demonstrates that our proposed method can effectively address the multi-geometric spaces domain shift caused by scene and environmental change. Compared with other SOTA DA methods, MATS outperforms SFA and STM3D by 2.7% and 2.5% NDS respectively. The comparison further demonstrates that our proposed method is tailored for LSS-based 3D object detection, addressing the domain shift accumulation problem in multi-geometric space. **Weathers Adaptation** As shown in Tab. II, in the Clear to Rainy adaptation scenario, MATS outperforms other methods by a significant margin. Compared with SFA and STM3D, MATS improves NDS by 2.4% and 2.9% since it can extract multi-space target domain knowledge to realize a better alignment between source and target data distribution. Moreover, with the MATS framework attaining mAP scores of 0.243% and 0.247% using R50 and R101 backbones respectively, this further underscores the robustness of our method in tackling domain shift across diverse target domain data distributions. **Day-Night Adaptation** The Day-Night adaptation is the most challenging scenario for camera-based methods, MATS significantly improves the detection performance and solves the domain shift in the Night domain. In Tab. III, the tremendous domain gap makes baseline methods perform extremely poorly with only 6.2% NDS and 3.6% mAP under R101. While MATS, especially with R101 as its backbone, can achieve 18.8% NDS and 12.7% mAP. Even compared with other DA methods, it also achieves a superior improvement of more than 4.0% and 6.2% NDA. Since previous DA methods like STM3D and STA ignore the inaccuracy depth estimation in Night data, they can not effectively extract target domain knowledge.

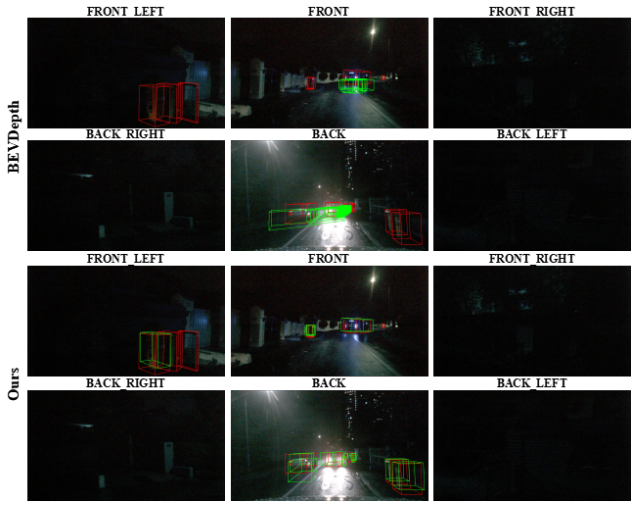

Fig. 4. Qualitative results: The upper and bottom parts are visualization of BevDepth [2] and our proposed method respectively. The results are visualized on the weather adaptation scenario.

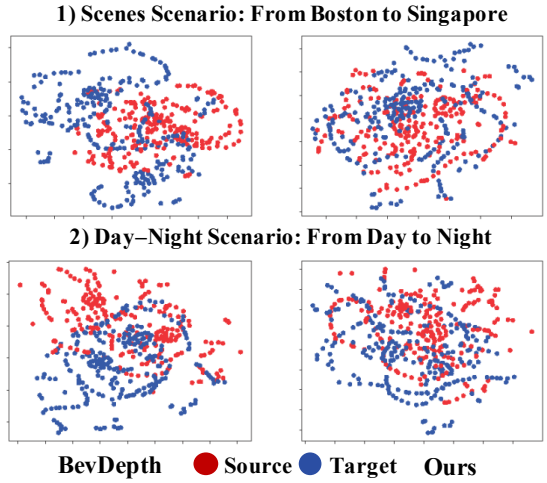

Fig. 5. Visualization of feature distributions using T-SNE [62]. The blue spots denote the source features, while red spots represent target features.

C. Ablation study

To better reflect the role of each component in MATS, we conduct ablation experiments on **Clear-Rainy** adaptation scenario to analyze how each component can deal with domain shift for LSS-based BEV perception.

The effectiveness of DAT and GAS. In Tab. IV, vanilla BEVDepth (Ex_0) can only achieve 26.8% NDS and 19.6% mAP when the scenario is transformed from clear to the rainy domain. For DAT, it transfers multi-latent space target domain knowledge to the student model, which are constructed by depth-aware information. As shown in Ex_1, the student model can absorb feature and pseudo-label level knowledge from DAT, thus improving NDS, and mAP by 1.5% and 3.5% respectively. As shown in Ex_2, the EMA updating only brings a trivial improvement around 0.4% mAP, which shows that

TABLE IV

ABLATION STUDIES ON THE CLEAR TO RAINY SCENARIO. DAT CONSISTS THREE COMPONENTS, INCLUDING DEPTH-AWARE INFORMATION(DA), TEACHER MODEL EMA, AND MULTI-SPACE KNOWLEDGE TRANSFER(KT). FOR GAS, THE GEOMETRIC EMBEDDING SPACE CAN BE CONSTRUCTED BY BEV(BA), IMAGE(IA), AND VOXEL(VA) FEATURE.

Name	DA	EMA	KT	BA	IA	VA	NDS ↑	mAP ↑
Ex_0	-	-	-	-	-	-	0.268	0.196
Ex_1	✓	-	✓	-	-	-	0.283	0.231
Ex_2	✓	✓	✓	-	-	-	0.286	0.235
Ex_3	-	-	-	✓	-	-	0.276	0.200
Ex_4	-	-	-	✓	✓	-	0.282	0.204
Ex_5	-	-	-	✓	✓	✓	0.288	0.207
Ex_6	✓	✓	✓	✓	✓	✓	0.305	0.243

TABLE V

THE ABLATION STUDY ON THE EFFECTIVENESS OF EACH COMPONENT IN DEPTH-AWARE INFORMATION (DAT). PRED MEANS DIRECTLY UTILIZING DEPTH PREDICTION, AND CON AND UG MEAN ADAPTIVE CONFIDENCE- AND UNCERTAINTY-GUIDED DEPTH SELECTION RESPECTIVELY.

Depth-aware:	Lidar	Pred	Con	UG	NDS ↑	mAP ↑
Ex_{2-1}	✓	-	-	-	0.275	0.223
Ex_{2-2}	✓	✓	-	-	0.278	0.228
Ex_{2-3}	✓	✓	✓	-	0.280	0.229
Ex_2	✓	✓	-	✓	0.286	0.235

the main performance improvement does not come from the usages of the EMA scheme. By gradually incorporating more multi-space features ($Ex_3 - Ex_5$) in the shared geometric space, the student model will get a 2% improvement in NDS, which demonstrates that it is crucial to jointly address all space domain shift. When we combine DAT and GAS in MATS (Ex_6), NDS reaches 30.5% while mAP achieves 24.3%. The results prove that all components in DAT and GAS can jointly address the domain shift accumulation problem.

Detailed ablation study of DAT We study the effectiveness of depth-aware information composition and multi-geometric space knowledge transferring in DAT. As shown in Tab. V, only taking lidar ground truth to replace depth prediction (Ex_{2-1}) can improve 0.7% NDS and 2.7% mAP compared with Ex_0. The obviously increased mAP demonstrates that lidar data plays an important role in target domain-specific Voxel feature construction. However, due to the sparse property of lidar data, we utilize all dense depth predictions to composite sparse lidar. In (Ex_{2-2}), NDS and mAP can achieve 27.9% and 22.8%, which only have limited improvement since the original predictions contain noises. We further adopt traditional confidence scores to select reliable depth prediction. In (Ex_{2-3}), NDS and mAP can achieve 0.2% and 0.1% improvement compared with Ex_{2-2}, since confidence is trusting less in the pixel-wise cross-domain scenario. Therefore, we introduce uncertainty guidance to adaptively select more reliable and task-relevant depth predictions. Ex_2 has obvious performance progress compared with Ex_{2-1} - Ex_{2-3}, demonstrating the uncertainty mechanism can reflect the reliability of depth prediction. And it reduces the noises of depth information and can further ease the domain shift influence. As shown in Tab. VI, transferring target domain knowledge in different geometric spaces can be beneficial to DAT. With pseudo label, BEV, voxel, and image feature transferred between DAT and student model, mAP is gradually improved from 19.6% to 23.5%. The improved performance proves that the transferred multi-space target domain knowledge is essential for the student model to align the distribution between two domains.

TABLE VI

THE ABLATION STUDY ON THE EFFECTIVENESS OF EACH COMPONENT IN MULTI-LATENT SPACE KNOWLEDGE TRANSFER. PL MEANS TRANSFERRING INSTANCE-LEVEL PSEUDO LABELS. BEV, VOXEL, AND IMAGE STAND FOR TRANSFERRING ON CORRESPONDING LATENT SPACES.

Latent Space:	PL	BEV	Voxel	Image	NDS ↑	mAP ↑
Ex_{2-4}	✓	-	-	-	0.280	0.213
Ex_{2-5}	✓	✓	-	-	0.283	0.222
Ex_{2-6}	✓	✓	✓	-	0.285	0.230
Ex_2	✓	✓	✓	✓	0.286	0.235

D. Qualitative analysis

As shown in Fig. 4, it is quite clear that the BEVDepth fails to locate the objects well, while MATS yields more accurate localization results as its predicted green box overlap better with the ground truth red box. We can also observe that MATS can detect objects that baseline ignores, demonstrating the superiority of MATS in object detection and presenting great potential in deploying to real-world autonomous driving applications. The visualization in Fig. 5, as a clear separation can be seen in the clusters of the source and target distributions produced by BEVDepth, the features generated by MATS get closer distribution between two domains, which further demonstrates the ability of our proposed method in addressing domain shift.

V. CONCLUSION AND DISCUSSION OF LIMITATIONS

Our Multi-space Alignment Teacher-Student (MATS) framework is designed to effectively mitigate domain shift accumulation in LSS-based BEV perception. The Depth-Aware Teacher (DAT) extracts reliable target domain-specific knowledge across multiple latent spaces using depth-aware information, which is then transferred to the student model. Meanwhile, the Geometric-space Aligned Student (GAS) model leverages knowledge from both the source and target domains to reduce the data distribution gap between them. The combined efforts of DAT and GAS help tackle the domain shift accumulation challenge, resulting in MATS achieving SOTA performance in three challenging Unsupervised Domain Adaptation (UDA) scenarios. For limitations, the teacher-student framework brings more computational costs during training. However, the student model keeps the same memory and computational cost as the baseline in inference.

VI. ACKNOWLEDGEMENT

Shanghang Zhang is supported by the National Key Research and Development Project of China (No.2022ZD0117801).

REFERENCES

[1] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[2] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," *arXiv preprint arXiv:2206.10092*, 2022.

[3] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[4] J. Li, M. Lu, J. Liu, Y. Guo, Y. Du, L. Du, and S. Zhang, "Bev-lgkd: A unified lidar-guided knowledge distillation framework for multi-view bev 3d object detection," *IEEE Transactions on Intelligent Vehicles*, 2023.

[5] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, Y. Guo, and S. Zhang, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," *arXiv preprint arXiv:2312.14074*, 2023.

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

[7] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *arXiv preprint arXiv:2203.17270*, 2022.

[8] R. Zhang, Y. Luo, J. Liu, H. Yang, Z. Dong, D. Gudovskiy, T. Okuno, Y. Nakata, K. Keutzer, Y. Du *et al.*, "Efficient deweather mixture-of-experts with uncertainty-aware feature-wise linear modulation," *arXiv preprint arXiv:2312.16610*, 2023.

[9] S. Yang, J. Wu, J. Liu, X. Li, Q. Zhang, M. Pan, and S. Zhang, "Exploring sparse visual prompt for cross-domain semantic segmentation," *arXiv preprint arXiv:2303.09792*, 2023.

[10] J. Liu, S. Yang, P. Jia, M. Lu, Y. Guo, W. Xue, and S. Zhang, "Vida: Homeostatic visual domain adapter for continual test time adaptation," *arXiv preprint arXiv:2306.04344*, 2023.

[11] J. Liu, R. Xu, S. Yang, R. Zhang, Q. Zhang, Z. Chen, Y. Guo, and S. Zhang, "Adaptive distribution masked autoencoders for continual test-time adaptation," *arXiv preprint arXiv:2312.12480*, 2023.

[12] W. Wang, Y. Cao, J. Zhang, F. He, Z.-J. Zha, Y. Wen, and D. Tao, "Exploring sequence feature alignment for domain adaptive detection transformers," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1730–1738.

[13] Z. Li, Z. Chen, A. Li, L. Fang, Q. Jiang, X. Liu, and J. Jiang, "Unsupervised domain adaptation for monocular 3d object detection via self-training," *arXiv preprint arXiv:2204.11590*, 2022.

[14] X. Chi, J. Liu, M. Lu, R. Zhang, Z. Wang, Y. Guo, and S. Zhang, "Bev-san: Accurate bev 3d object detection via slice attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 461–17 470.

[15] Z. Li, Z. Chen, A. Li, L. Fang, Q. Jiang, X. Liu, and J. Jiang, "Towards model generalization for monocular 3d object detection," *arXiv preprint arXiv:2205.11664*, 2022.

[16] Y. Cai, B. Li, Z. Jiao, H. Li, X. Zeng, and X. Wang, "Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 478–10 485.

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

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

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

[20] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *arXiv preprint arXiv:2206.00630*, 2022.

[21] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo," *arXiv preprint arXiv:2209.10248*, 2022.

[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[23] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9287–9296.

[24] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, and P. Luo, "Learning depth-guided convolutions for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1000–1001.

[25] Z. Liu, D. Zhou, F. Lu, J. Fang, and L. Zhang, "Autoshape: Real-time shape-aware monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 641–15 650.

[26] F. Manhardt, W. Kehl, and A. Gaidon, "Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2069–2078.

[27] I. Barabanau, A. Artemov, E. Burnaev, and V. Murashkin, "Monocular 3d object detection via geometric reasoning on keypoints," *arXiv preprint arXiv:1905.05618*, 2019.

[28] Z. Li, Z. Qu, Y. Zhou, J. Liu, H. Wang, and L. Jiang, "Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2791–2800.

[29] Y. Zhang, J. Lu, and J. Zhou, "Objects are different: Flexible monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3289–3298.

[30] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.

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

[32] Y. Liu, J. Yan, F. Jia, S. Li, Q. Gao, T. Wang, X. Zhang, and J. Sun, "Petrv2: A unified framework for 3d perception from multi-camera images," *arXiv preprint arXiv:2206.01256*, 2022.

[33] S. Chen, X. Wang, T. Cheng, Q. Zhang, C. Huang, and W. Liu, "Polar parametrization for vision-based surround-view 3d detection," *arXiv preprint arXiv:2206.10965*, 2022.

[34] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y.-G. Jiang, "Polarformer: Multi-camera 3d object detection with polar transformers," *arXiv preprint arXiv:2206.15398*, 2022.

[35] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564.

[36] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[37] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "Monodtr: Monocular 3d object detection with depth-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4012–4021.

[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

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

[40] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[41] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 457–11 466.

[42] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6956–6965.

[43] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 724–11 733.

[44] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 355–12 364.

[45] J. Yu, J. Liu, X. Wei, H. Zhou, Y. Nakata, D. Gudovskiy, T. Okuno, J. Li, K. Keutzer, and S. Zhang, "Cross-domain object detection with mean-teacher transformer," *arXiv preprint arXiv:2205.01643*, 2022.

[46] ——, "Mttrans: Cross-domain object detection with mean teacher transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 629–645.

[47] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8866–8875.

[48] W. Zhang, W. Li, and D. Xu, "Srdan: Scale-aware and range-aware domain adaptation network for cross-dataset 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6769–6779.

[49] A. Barrera, J. Beltrán, C. Guindel, J. A. Iglesias, and F. García, "Cycle and semantic consistent adversarial domain adaptation for reducing simulation-to-real domain shift in lidar bird's eye view," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 3081–3086.

[50] D. Acuna, J. Philion, and S. Fidler, "Towards optimal strategies for training self-driving perception models in simulation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1686–1699, 2021.

[51] M. H. Ng, K. Radia, J. Chen, D. Wang, I. Gog, and J. E. Gonzalez, "Bev-seg: Bird's eye view semantic segmentation using geometry and semantic point cloud," *arXiv preprint arXiv:2006.11436*, 2020.

[52] K. Saleh, A. Abobakr, M. Attia, J. Iskander, D. Nahavandi, M. Hossny, and S. Nahvandi, "Domain adaptation for vehicle detection from bird's eye view lidar point cloud data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[53] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia *et al.*, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[54] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[55] M. Döbler, R. A. Marsden, and B. Yang, "Robust mean teacher for continual and gradual test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7704–7714.

[56] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[57] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[58] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 765–10 775.

[59] Q. Wang, O. Fink, L. V. Gool, and D. Dai, "Continual test-time domain adaptation," *ArXiv*, vol. abs/2203.13591, 2022.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[61] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[62] L. Van Der Maaten, "Barnes-hut-sne," *arXiv preprint arXiv:1301.3342*, 2013.