Towards Universal LiDAR-Based 3D Object Detection by Multi-Domain Knowledge Transfer

Guile Wu, Tongtong Cao, Bingbing Liu*, Xingxin Chen, and Yuan Ren
Huawei Noah's Ark Lab
{guile.wu, caotongtong, liu.bingbing, xingxin.chen1, yuan.ren3}@huawei.com

Abstract

Contemporary LiDAR-based 3D object detection methods mostly focus on single-domain learning or cross-domain adaptive learning. However, for autonomous driving systems, optimizing a specific LiDAR-based 3D object detector for each domain is costly and lacks of scalability in real-world deployment. It is desirable to train a universal LiDAR-based 3D object detector from multiple domains. In this work, we propose the first attempt to explore multi-domain learning and generalization for LiDAR-based 3D object detection. We show that jointly optimizing a 3D object detector from multiple domains achieves better generalization capability compared to the conventional single-domain learning model. To explore informative knowledge across domains towards a universal 3D object detector, we propose a multi-domain knowledge transfer framework with universal feature transformation. This approach leverages spatial-wise and channel-wise knowledge across domains to learn universal feature representations, so it facilitates to optimize a universal 3D object detector for deployment at different domains. Extensive experiments on four benchmark datasets (Waymo, KITTI, NuScenes and ONCE) show the superiority of our approach over the state-of-the-art approaches for multi-domain learning and generalization in LiDAR-based 3D object detection.

1. Introduction

LiDAR-based 3D object detection aims to localize and recognize objects of interests from LiDAR point clouds. It has been used in a wide range of applications, such as autonomous driving and robotics. Although incredible success has been achieved in the past few years, most LiDAR-based 3D object detection methods focus on optimizing a specific model for a single domain [38, 42, 21, 34, 7, 23, 11], which usually shows poor generalization to other domains. To address this problem, some researchers resort to cross-domain adaptation [31, 16, 37, 36] to transfer knowledge from a source domain to a target domain, which improves generalization of a target model. However, domain adaptive 3D object detection methods usually ignore the importance of preserving generalization to the source domain, resulting in poor performance on the source domain.

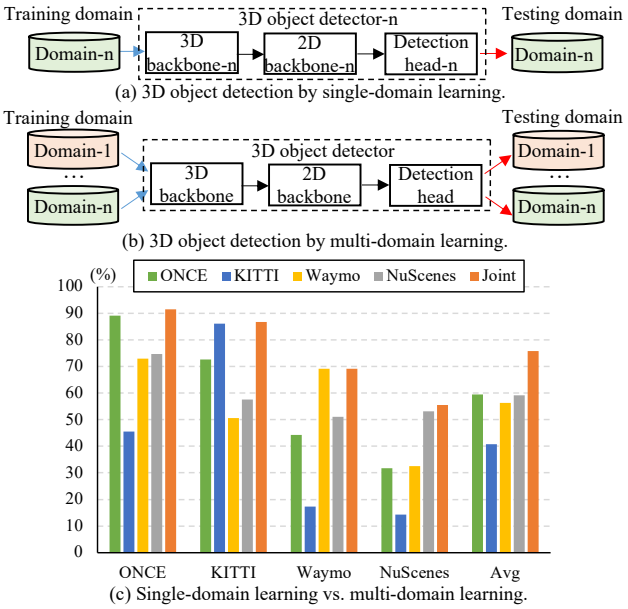


Figure 1. Comparisons of LiDAR-based 3D object detection by single-domain learning and multi-domain learning. (a) Conventional methods mainly train a specific model for each domain. (b) We propose to jointly optimize a universal model for all domains. (c) Comparison between oracle models and a joint-training model. Results are evaluated using CenterPoint [38] with KITTI metric [2] of AP_{BEV} and $IoU_{0.7}$ of the car category.

For automated driving systems, optimizing a specific LiDAR-based 3D object detection model for each domain (*e.g.*, locations, LiDAR sensors, etc.) is costly and lacks of scalability for real-world deployment. More importantly, optimizing each domain independently cannot make full use of knowledge of each domain to facilitate learning a universal detector. We summarize some domain gap infor-

*Bingbing Liu is the corresponding author.

Dataset	Location	Night/rain	LiDAR beam	Points per scene	Car average size (L,W,H)	VFOV
Waymo	USA	Yes	64-beam	160k	(4.80, 2.11, 1.79)	[-17.6°, 2.4°]
KITTI	Germany	No	64-beam	118k	(3.89, 1.62, 1.53)	[-23.6°, 3.2°]
NuScenes	USA/Singapore	Yes	32-beam	25k	(4.64, 1.96, 1.73)	[-30.0°, 10.0°]
ONCE	China	Yes	40-beam	70k	(4.36, 1.81, 1.56)	[-25.0°, 15.0°]

Table 1. Domain gap information of four autonomous driving datasets, namely Waymo [24], KITTI [2], NuScenes [1] and ONCE [17]. We resolve domain gaps by transferring knowledge across multiple domains to learn universal feature representations in a universal detector.

mation of four widely used autonomous driving datasets (Waymo [24], KITTI [2], NuScenes [1] and ONCE [17]) in Table 1 and show the performance of each oracle model (single-domain learning) on all domains in Fig. 1(c). Due to the significant domain gaps (*e.g*., LiDAR beams, locations, weather, object sizes, *etc*.), these single-domain learning models usually perform poorly on other domains. Thus, for real-world application, it is desirable to optimize a universal model with good generalization on all domains. However, there is no existing work proposed to address this problem.

In this work, we propose multi-domain learning and generalization for learning a universal LiDAR-based 3D object detector. First, we construct a joint-training baseline model using data from multiple domains. We validate the advantage of learning a variety of point clouds from different domains and show the superior performance of the joint-training model over the conventional single-domain learning model (see Fig. 1(c)). Then, to further explore informative knowledge across domains, we propose a multi-domain knowledge transfer (MDKT) framework with universal feature transformation for LiDAR-based 3D object detection. An overview of the proposed approach is depicted in Fig. 2. Specifically, we aim to learn universal feature representations in a universal detector by transferring informative knowledge across multiple domains. To this end, we employ multiple specific detectors trained on each domain and transfer knowledge from these specific detectors to a universal detector so that the universal detector can aggregate informative spatial-wise knowledge across domains. Meanwhile, to aggregate informative channel-wise knowledge across domains, we modulate channel information of feature representations in the universal detector. Together with spatial-wise and channel-wise knowledge transfer across domains, the universal feature transformation module facilitates to optimize a universal detector. Besides, we observe that learning to normalize intermediate BEV (Bird's Eye View) features can also facilitate model optimization when training data are from multiple domains. To some extent, this strategy is compatible with our MDKT approach by removing some noises in statistics across multiple domains. In summary, our **contributions** are:

- We propose multi-domain learning and generalization for LiDAR-based 3D object detection. To the best of our knowledge, this is the first work exploring multi-domain data for optimizing a universal LiDAR-based 3D object detector.
- We propose a multi-domain knowledge transfer framework with universal feature transformation for LiDAR-based 3D object detection. Our approach learns universal feature representations for LiDAR-based 3D object detection by aggregating informative spatial-wise and channel-wise knowledge across domains.
- We provide thorough experimental analyses on four autonomous driving benchmark datasets and demonstrate the superiority of our approach over the state-of-the-art approaches for multi-domain learning and generalization in LiDAR-based 3D object detection.

2. Related Work

LiDAR-Based 3D Object Detection. There have been many promising LiDAR-based 3D object detection methods [38, 42, 22, 35, 36] proposed in recent years. Conventional methods [38, 34, 23, 7, 11] mostly follow a single-domain learning paradigm to optimize a 3D object detector for each specific domain. Although Li *et al*. [11] use the term 'universal' to describe their 3D detector, their method still follows single-domain learning not universal to different domains. These single-domain learning methods usually show poor generalization to other domains. To resolve this problem, some researchers resort to domain adaptive 3D object detection by pre-training a detector on a labeled source domain and adapting the detector to an unlabeled target domain [36, 16, 31, 30]. Although these adaptation methods have shown good performance on a target domain compared with the source-only methods, they are elaborately designed for adaptation to the target domain so cannot guarantee the performance on the source domain. Recently, Lehner *et al*. [8] introduce single-domain generalization for 3D object detection and propose an adversarial augmentation method to deform objects by vector fields in point clouds. Different from these methods, in this work, we propose to optimize a universal LiDAR-based 3D object detector using training data from multiple domains. We present multi-domain knowledge transfer with universal feature transformation.

Multi-Domain Learning and Generalization. Optimizing a model using training data from multiple domains is a popular research field in computer vision. In multi-domain generalization [41, 9, 32, 40], there are different ways to improve the model generalization capability, *e.g*., mix style

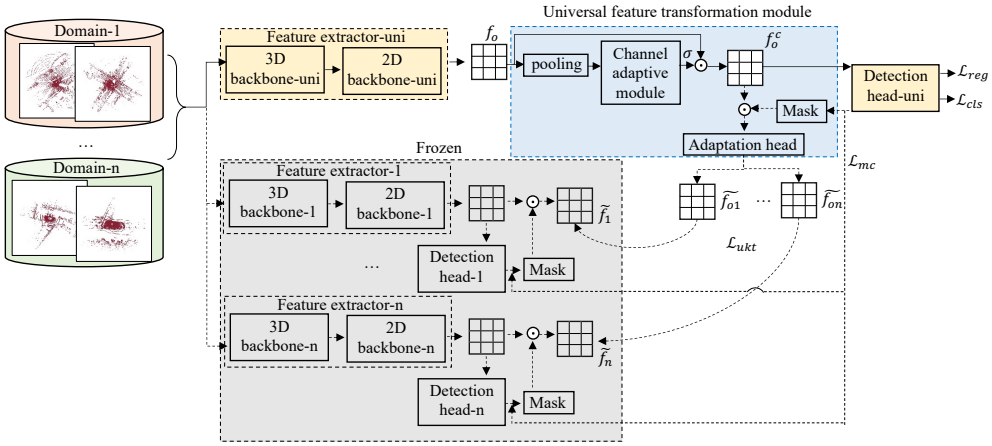


Figure 2. An overview of the proposed multi-domain knowledge transfer approach for LiDAR-based 3D object detection.

normalization [41], pseudo novel data synthesis [40], feature augmentation [9], *etc*. These methods mainly focus on image classification, while our work aims to learn a universal LiDAR-based 3D object detector. On the other hand, some attempts [18, 19, 10] have been proposed to optimize a universal model for multiple training domains. Rebuffi *et al*. [18, 19] propose series and parallel residual adapters to optimize a model from multiple domains for image classification, while Li *et al*. [10] align features from multiple pre-trained models to a single model with centered kernel alignment and adapt the model to a target task with a linear transformation layer for few-shot classification. Our work differs from [18, 19, 10] in that we explore informative spatial-wise and channel-wise knowledge across domains to learn universal feature representations in a universal 3D object detector, instead of learning domain-specific adapters or aligning image-level features for few-shot classification.

Domain Generalized Object Detection. There have been some domain generalized 2D object detection methods [12, 29, 39] proposed in recent years. Lin *et al*. [12] use a domain-invariant network with representation reconstruction to learn disentangled representations, while Wang *et al*. [29] present a domain attention adapter bank for image-based generalized 2D object detection. Different from these works, we present a multi-domain knowledge transfer framework to deal with point clouds from multiple domains for learning a universal 3D detector, instead of using representation reconstruction or an adapter bank for image-based 2D object detection. Besides, a concurrent work [28] presents camera-based single-DG for multi-view 3D object detection, while our work focuses on LiDAR-based 3D object detection with point clouds from multiple domains.

3. Methodology

Problem Statement. In this work, we focus on optimizing a universal model with training data from multiple domains for LiDAR-based 3D object detection. Suppose we have access to training point clouds data from n domains/datasets $\{X_i, Y_i\}_{i=1}^{n}$, where X_i and Y_i are the set of training point clouds and ground truth labels (category labels and 3D bounding box labels) of the ith domain, respectively. We aim to exploit all these training data for optimizing a 3D object detector capable of extracting universal feature representations of point clouds from different domains. This 3D object detector should generalize well to all seen training domains in the multi-domain learning setting and to any unseen new domain in the multi-domain generalization setting. In this work, we employ CenterPoint [38], a state-of-the-art 3D object detector, as the backbone due to its efficiency and outstanding performance, but our approach is generic and can be applied with other backbones.

3.1. Approach Overview

The framework of the proposed multi-domain knowledge transfer approach is depicted in Fig. 2. In our approach, there is a universal 3D object detector (consists of a feature extractor, a detection head and a universal feature transformation module) and n frozen specific 3D object detectors. Here, a feature extractor is usually composed of a 3D backbone to extract point clouds features and a 2D backbone to extract BEV features. The n frozen detectors are pre-trained on each specific domain using the conventional classification loss $\mathcal{L}_{cls}$ (*e.g*., a focal loss [13]) and regression loss $\mathcal{L}_{reg}$ (*e.g*., L1 regression loss [38]). With training data from multiple domains, we use a universal feature extractor to extract feature representations f_o and employ the frozen specific feature extractors to extract specific feature representations $\tilde{f}_i$. Then, we use a universal feature transformation module to transfer knowledge across domains and employ a universal knowledge transfer loss $\mathcal{L}_{ukt}$ and a mask consistency loss $\mathcal{L}_{mc}$ to facilitate the optimization of the universal 3D object detector. Together with $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$, the training objective for optimizing the universal 3D

object detector is formulated as:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha_1 \mathcal{L}_{reg} + \alpha_2 \mathcal{L}_{ukt} + \alpha_3 \mathcal{L}_{mc}, \quad (1)$$

where α_i is a weighting coefficient to balance losses.

In inference, the frozen specific models and the adaptation head are removed, so the universal detector is used to directly predict the categories and the 3D bounding boxes of objects of interests in point clouds from different domains.

3.2. Multi-Domain Knowledge Transfer

A Joint-Training Baseline Model. There is no existing work exploring multi-domain learning and generalization for LiDAR-based 3D object detection, so first we construct a baseline model and show the superiority of multi-domain learning over single-domain learning. As summarized in Table 1, different LiDAR point clouds datasets recorded with different LiDAR sensors contain different inherent characteristics, *e.g.*, ranges of point clouds, coordinate information, *etc*. Therefore, to construct a joint-training baseline model, it is required to align these inherent characteristics into a uniform range for multi-domain learning and generalization. To this end, we select one domain as the base and align inherent characteristics of other datasets to the base domain. In this work, we consider Waymo [24] is a relatively complete dataset, so we align inherent characteristics of all domains to Waymo, including the same range of point cloud, LiDAR coordinate system and voxel size. With this uniform range of input point clouds, we can randomly sample training data from different domains in each mini-batch and train a joint-training baseline model.

Interestingly, as shown in Fig. 1(c), this joint-training model outperforms most oracle models on each specific domain and achieves significantly better performance on average. This can be attributed to the collaboration of multiple domains which increases the diversity of training data for model optimization. This also shows the benefit of optimizing a LiDAR-based 3D object detector from multiple domains for deployment compared to single-domain learning.

Universal Feature Transformation. Although the vanilla joint-training model shows promising performance, it learns multi-domain knowledge only by assembling data across domains. This may be suboptimal for learning a universal LiDAR-based 3D object detector, because it fails to modulate specific knowledge contained in each domain and does not explore richer knowledge to facilitate model optimization. To resolve this problem, we propose to transfer informative spatial-wise and channel-wise knowledge across domains so as to learn universal feature representations for LiDAR-based 3D object detection.

Specifically, to learn informative spatial-wise knowledge across domains, inspired by few-shot classification from multiple domains [10], we pre-train n specific detectors for each specific domain and freeze them to extract specific feature representations. Since each specific detector is only trained with data from a specific domain, they have encoded informative knowledge within each domain. We therefore use the specific features as informative knowledge to guide the optimization of the universal features in the universal detector. Instead of directly aligning specific and universal features, first we generate a mask M of foregrounds and perform Hadamard product ($\odot$) between f_o and M to highlight objects of interests. Here, a heatmap obtained from the detection head is applied with a sigmoid function to generate M. Alternatively, when no mask is generated, M can be an identity function which includes more regions in f_o for knowledge transfer. Next, with $f_o \odot M$, we employ an adaptation head (consists of convolutional layers) to transform the extracted feature map into n transformed feature maps $\tilde{f}_{oi}$. Meanwhile, for each specific detector, object masks are also generated and applied to get $\tilde{f}_i$. We then compute $\mathcal{L}_{ukt}$ to transfer knowledge from the specific detectors to the universal detector as:

$$\mathcal{L}_{ukt} = \sum_{i=1}^{n} \left\| \phi(\tilde{f}_{oi}) - \phi(\tilde{f}_i) \right\|^2, \quad (2)$$

where $\phi(\cdot)$ performs L2 normalization of features. When we use heatmaps to generate masks, a mask consistency loss $\mathcal{L}_{mc}$ can be computed to constrain masks generated from the universal detector and the specific detectors, as:

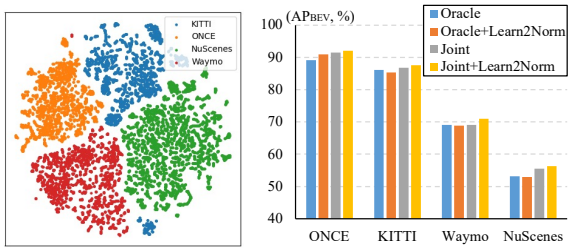
$$\mathcal{L}_{mc} = \sum_{i=1}^{n} \left(\frac{M_{gt} \left\| h_o - h_i \right\|^2}{sum(M_{gt})} \right), \quad (3)$$

where h_i are heatmaps obtained from detection heads, M_{gt} is a ground truth foreground mask (targets assigned as [38]), $sum(M_{gt})$ means the sum of the number of foregrounds.

Although spatial-wise knowledge aggregation transfers spatial information across domains for learning universal features, it does not modulate channel information which also encodes useful knowledge across domains to improve generalization. Therefore, to learn informative channel-wise knowledge across domains, in the universal model, we aggregate channel information of f_o to generate a channel modulated feature map f_o^c as:

$$f_o^c = f_o \odot \sigma(\mathcal{F}(\psi(f_o))), \quad (4)$$

where $\psi(\cdot)$ is global pooling, $\mathcal{F}$ is a channel-wise adaptive transformation module, and $\sigma(\cdot)$ is a sigmoid function. And we replace f_o with f_o^c to compute $f_o^c \odot M$. Then, $f_o^c \odot M$ is used as the input to the adaptation head to generate $\tilde{f}_{oi}$ for computing $\mathcal{L}_{ukt}$ in Eq. (2). Together with spatial-wise and channel-wise knowledge transfer, our MDKT approach learns universal feature representations in a universal 3D object detector for deployment in different domains.



(a) Statistics of BEV maps across domains.

(b) Learn2Norm in single-domain learning vs. multi-domain learning.

Figure 3. Statistics (concatenation of mean and variance) of BEV maps across domains obtained in the baseline (visualized by t-SNE [27]) and comparison between oracles and the baseline with learning to normalization (see § 4 for details of experiments).

Learning to Normalize BEV Features. In addition to spatial-wise and channel-wise knowledge transfer across domains, we observe that learning to normalize intermediate BEV features also benefits model optimization when training data are from multiple domains. And to some extent, this strategy is compatible with our MDKT approach.

Specifically, in traditional LiDAR-based 3D object detection, BatchNorm [6] is used in the 2D backbone to compute mean and variance of samples in a mini-batch to regularize model optimization. However, when training data are from multiple domains, each mini-batch usually contains samples from different domains, which may affect the regularization process. As shown in Fig. 3(a), statistics of BEV maps (*i.e.*, the input to the 2D backbone) obtained in the baseline are clearly separated across domains. This indicates that there are some gaps in statistics of intermediate BEV features across domains. Thus, more advanced normalization techniques [41, 20, 32] may be used to deal with this problem. In this work, we replace BatchNorm with SwitchNorm [15] in the 2D backbone so as to learn to normalize (Learn2Norm) intermediate BEV features across domains. Technically, SwitchNorm computes mean and variance of intermediate features in three-levels, *i.e.*, instance-wise, layer-wise and batch-wise, and employs weights to combine these statistics before applying affine parameters. The combination of multiple statistics helps to regularize model optimization. We refer readers to [15] for more details of SwitchNorm.

It is worth noting that SwitchNorm is not a generic technique to improve a 3D object detector. As shown in Fig. 3(b), oracle models with and without Learn2Norm achieve similar results. In this work, we observe that SwitchNorm can be applied on the 2D backbone to normalize BEV features across domains, not on the whole model or for single-domain learning. As shown in Fig. 3(b), when training data are from multiple domains, joint-training models with Learn2Norm outperforms those without Learn2Norm. See § 4 for more experimental analyses.

4. Experiments

Datasets. We conduct extensive experiments on four autonomous driving datasets, namely KITTI [2], Waymo [24], NuScenes [1] and ONCE [17]. KITTI [2] is an autonomous driving dataset with 3712 training point clouds samples and 3769 validation point clouds samples. Waymo [24] is a large-scale autonomous driving datasets consists of 798 training segments of 158081 point clouds frames and 202 validation segments of 39987 point clouds frames. We set the sample interval to 20 for training point clouds samples. NuScenes [1] is another large-scale autonomous driving datasets containing 700 training sequences of 28130 point clouds frames and 150 validation sequences of 6019 point clouds frames. Similar to Waymo, we set a uniform sampling interval to 5 for training point clouds samples. ONCE [17] is a recently introduced autonomous driving datasets containing 4961 training point clouds samples and 3321 validation point clouds samples. These four datasets contain significant domain gaps as summarized in Table 1, so we use them to evaluate multi-domain learning and generalization for LiDAR-based 3D object detection. Besides, we set some inherent characteristics of different datasets into a uniform range, including the same range of point cloud of $[-75.2m, -75.2m, -2m, 75.2m, 75.2m, 4m]$, the same LiDAR coordinate system (offset of z-axis is $1.8m$ and rotation is $\frac{\pi}{2}$ for NuScenes, and offset of z-axis is $1.75m$ and rotation is $\frac{\pi}{2}$ for ONCE, offset of z-axis is $1.6m$ for KITTI), and the same voxel size of [0.1, 0.1, 0.15]. These are done when loading data from multiple domains.

Evaluation Metrics. To comprehensively compare performance of different approaches across four domains, we use the KITTI evaluation metric. We mainly report results of the average precision (AP) over 40 recall positions of the commonly used car category (the vehicle category on Waymo) for both BEV IoUs and 3D IoUs with IoU thresholds of 0.7. On KITTI, results of the moderate mode are reported, while on other datasets, overall performance are reported. Metrics are computed in the aligned Waymo coordinates because training and testing coordinates need to be consistent.

Implementation Details. We implement our approach with Python and PyTorch using OpenPCDet [25]. We employ CenterPoint [38] as the backbone because of its superior performance. Data augmentation includes random world flipping, random world rotation, random world scaling, and GT sampling [34]. We set the batch size to 8, training epochs to 40 and a fixed random seed to 2022. We report the model performance of the last epoch for comparisons in all experiments. We use Adam as the optimizer with an initial learning rate of 3e-3, weight decay of 0.01, and set a one cycle learning rate scheduler. In channel-wise transformation, $\mathcal{F}$ consists of two 1×1 convolutional layers to modulate channel dimension and a ReLU function between

Method	ONCE		KITTI		Waymo		NuScenes		Avg	
	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}
Oracle-ONCE	89.16	78.40	72.63	52.03	44.18	20.67	31.77	10.69	59.44	40.45
Oracle-KITTI	45.45	25.44	86.07	76.09	17.25	3.88	14.36	4.76	40.78	27.54
Oracle-Waymo	72.95	45.95	50.52	14.31	69.07	53.97	32.60	14.35	56.29	32.15
Oracle-NuScenes	74.62	35.88	57.62	23.02	51.10	25.20	53.17	32.81	59.13	29.23
MixStyle [41]	76.71	48.10	67.68	37.92	49.57	25.09	33.15	14.84	56.78	31.49
FeatAug [9]	75.13	36.04	72.41	43.56	46.17	20.05	34.47	14.11	57.05	28.44
URL [10]	76.90	48.45	69.94	38.89	50.28	25.32	34.50	14.96	57.83	31.91
AdvAlign [3, 26]	75.58	45.35	66.96	35.16	52.38	26.27	34.37	**15.15**	57.32	30.48
USE [29]	81.05	51.03	77.53	53.59	53.20	28.10	**35.20**	14.58	61.75	36.83
Baseline (Joint-training)	77.04	47.73	69.79	36.63	50.18	25.82	34.25	14.89	57.82	31.27
MDKT (ours)	**81.40**	53.16	80.28	58.63	56.09	32.75	35.02	14.49	63.20	39.76
MDKT (w/ Learn2Norm, ours)	81.18	**53.32**	**80.65**	**61.77**	**56.77**	**33.10**	34.81	14.55	**63.35**	**40.69**
OT* [30]	76.91	51.17	67.42	31.46	53.85	29.93	31.38	10.95	57.39	30.88
SN* [30]	77.83	52.24	78.60	62.51	54.39	**36.68**	34.90	**17.30**	61.43	42.18
MDKT (w/ Learn2Norm, ours) + SN*	**81.12**	**56.79**	**83.68**	**70.71**	**56.24**	34.88	**35.95**	16.86	**64.25**	**44.81**

Table 2. Comparison with the state-of-the-art methods in multi-domain generalization to unseen new domains on ONCE, KITTI, Waymo and NuScenes. The leave-one-domain-out protocol is adopted. Models trained on seen domains are directly tested on an unseen testing domain without fine-tuning (Oracle is trained on one domain and tested on all domains). 'Avg' is the average performance of AP_{BEV}/AP_{3D} across all domains. *OT/SN is weakly supervised because object size statistics from the unseen domain are used for adjusting/fine-tuning.

them, inspired by [4]. In spatial-wise transformation, the adaptation head consists of n parallel convolutional layers with kernel size 1×1. When using Learn2Norm, both the universal detector and the specific detectors are trained with SwitchNorm. Besides, our approach also supports implementation with MindSpore [5], a new deep learning computing framework, in which more comprehensive analyses are left for future work.

4.1. Multi-Domain Generalization Evaluation

To evaluate the generalization capability of a 3D object detector to any unseen new domain, we adopt the leave-one-domain-out protocol [41, 9, 32] by selecting one dataset as an unseen new testing domain and using the remaining datasets as seen training domains and repeating this process for all datasets. In inference, a model trained on seen domains are directly tested on an unseen testing domain without using data from the unseen domain for fine-tuning. In addition to the joint-training baseline, we re-implement several state-of-the-art methods based on the baseline, including: *MixStyle* [41] which applies feature statistic mixing to the BEV map of the baseline; *FeatAug* [9] which applies random noises to the BEV map of the baseline; *URL* [10] which co-aligns feature representations from multiple pre-trained models to a single model with centered kernel alignment based on the baseline; *AdvAlign* [3, 26] which uses adversarial learning of features to perform global and center-aware alignments across domains in the baseline; *USE* [29] which inserts residual domain attention adapters after every block in the 2D backbone of the baseline; *OT* [30] and *SN* [30] which employ object size statistics of unseen testing domains as prior information for adjusting/fine-tuning.

As shown in Table 2, our MDKT and MDKT w/ Learn2Norm significantly improve the joint-training baseline by a large margin (more than 5% AP_{BEV} and 8% AP_{3D} on average). This can be attributed to the multi-domain knowledge transfer framework with universal feature transformation and the learning to normalization strategy, which explores informative knowledge across domains to improve model generalization capability. Moreover, compared with the state-of-the-arts, our approach still achieves superior performance on average. Although OT [30] and SN [30] use statistic information of the unseen testing domain as prior knowledge, our MDKT and MDKT w/ Learn2Norm still perform better than OT and are on par with SN. When combined with SN, our approach w/ SN achieves the best results of 64.25% AP_{BEV} and 44.81% AP_{3D}. This indicates that our approach has encoded knowledge to recognize objects in BEV on unseen new domains, while SN can further improve our approach for 3D box size estimation.

4.2. Multi-Domain Learning Evaluation

In multi-domain learning for LiDAR-based 3D object detection, we optimize a model with training data from multiple domains and evaluate the model on all these domains. In this experiment, *MixStyle* [41], *FeatAug* [9], *URL* [10], *AdvAlign* [3, 26] and *USE* [29] are still compared, but *SN* [30] and *OT* [30] are not compared because object size statistics of all testing domains have been trained. Besides, we also compare with *SRA* [18] and *PRA* [19] which insert series and parallel residual adapters in the 2D backbone of the baseline, respectively.

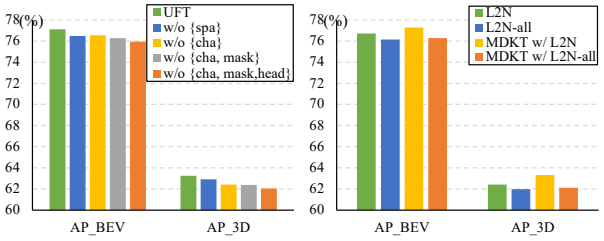
From Table 3, we can see that: (1) The joint-training baseline model consistently outperforms the oracle models. This shows the effectiveness of learning from multiple domains for LiDAR-based 3D object detection. (2) Our proposed approach significantly improves the performance of the baseline in both AP_{BEV} and AP_{3D}. Also,

Method	ONCE		KITTI		Waymo		NuScenes		Avg	
	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}
Reference oracle (Single best)	89.16	78.40	86.07	76.09	69.07	53.97	53.17	32.81	74.37	60.32
MixStyle [41]	91.14	78.73	87.30	79.12	68.83	53.97	54.68	34.96	75.49	61.70
FeatAug [9]	91.09	78.19	86.59	78.16	68.80	51.93	53.69	33.11	75.04	60.35
URL [10]	91.68	79.25	88.61	79.34	69.15	54.18	55.96	35.60	76.35	62.09
SRA [18]	91.91	80.95	86.59	78.30	69.23	54.24	56.26	36.32	76.00	62.45
PRA [19]	91.86	80.83	88.43	78.52	69.29	54.37	56.09	36.31	76.42	62.51
AdvAlign [3, 26]	91.62	79.13	86.92	79.37	69.21	54.36	55.79	35.17	75.89	62.01
USE [29]	91.81	81.27	88.57	79.03	69.25	54.38	56.03	36.40	76.42	62.77
Baseline (Joint-training)	91.46	78.97	86.74	79.11	69.11	54.28	55.57	35.51	75.72	61.97
MDKT (ours)	92.02	81.48	88.64	**79.45**	70.67	**54.71**	**57.16**	37.31	77.12	63.24
MDKT (w/ Learn2Norm, ours)	**92.26**	**82.07**	**88.87**	79.07	**70.96**	54.56	57.09	**37.54**	**77.30**	**63.31**

Table 3. Comparison with the state-of-the-art methods in multi-domain learning on ONCE, KITTI, Waymo and NuScenes. 'Reference oracle' is trained and tested on each specific domain (reported best results), while other methods are trained and tested on all domains.

Method	MDL		MDG	
	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}
Base	75.72	61.97	57.82	31.27
Base + UFT (MDKT)	77.12	63.24	63.20	39.76
Base + Learn2Norm	76.72	62.42	62.64	39.16
Base + UFT + Learn2Norm	**77.30**	**63.31**	**63.35**	**40.69**

Table 4. Component effectiveness analysis (average performance) in multi-domain learning (MDL) and generalization (MDG).



(a) UFT analysis. (b) Learn2Norm analysis.

Figure 4. Evaluation of (a) universal feature transformation variants and (b) learning to normalize BEV features variants in multi-domain learning. The average performance is reported.

our approach consistently surpasses the oracle models on all datasets. These can be attributed to the collaboration of multiple domains, universal feature transformation as well as learning to normalization, which results in a universal 3D object detector. (3) Compared with the state-of-the-art methods, our approach still achieves better performance.

4.3. Further Analysis and Discussion

Component Effectiveness Analysis. In Table 4, we evaluate the effectiveness of main components of our approach. It can be seen that multi-domain knowledge transfer with universal feature transformation improves the baseline by approximately 1.4% AP_{BEV} and 1.3% AP_{3D} on multi-domain learning and by around 5.4% AP_{BEV} and 8.5% AP_{3D} on multi-domain generalization. Besides, learning to normalize BEV features can also improve the baseline. Furthermore, MDKT w/ Learn2Norm achieves the best performance of 77.30% AP_{BEV} and 63.31% AP_{3D} on MDL and 63.35% AP_{BEV} and 40.69% AP_{3D} on MDG.

Universal Feature Transformation Analysis. In Fig. 4(a), we evaluate variants of universal feature transformation in MDKT. It can be observed that only learning channel-wise information (*i.e.*, UFT w/o {spatial}) or only learning spatial-wise information (*i.e.*, UFT w/o {channel}) leads to model performance degradation. Besides, in spatial-wise information learning, without using the object mask and its consistency (*i.e.*, UFT w/o {channel, mask}) reduces model performance while further removing the adaptation head from the model (UFT w/o {channel, mask, head}) results in the worst performance.

Learning to Normalize BEV Features Analysis. In Fig. 4(b), we further evaluate Learn2Norm applied to the whole backbone model. We can observe that applying Learn2Norm to the whole model results in performance degradation compared with only applying Learn2Norm to the 2D backbone. From Fig. 3(b) and Fig. 4(b), we know that Learn2Norm with SwitchNorm is not a generic strategy to improve 3D object detectors in single-domain learning, but it can be applied to normalize the intermediate BEV features across domains to facilitate model learning. Note that, Learn2Norm with SwitchNorm does not always bring improvement to MDKT (*e.g.*, as shown in Table 3), so there may be more suitable normalization strategies to use, which we leave for future work.

Qualitative Comparison. We present qualitative comparison of oracle, baseline and ours in Fig. 5. In the 1st row, oracle misses some objects while baseline and ours successfully recognize these objects; In the 2nd row, oracle and baseline generate some false positives while ours is not disturbed by these background noises; In the 3rd row, all approaches are confused by some distant objects but ours still predicts more true positives; In the 4th row, oracle and baseline predict incorrect 3D bounding boxes for some objects while ours provides more accurate predictions. The superior performance of our approach can be attributed to the

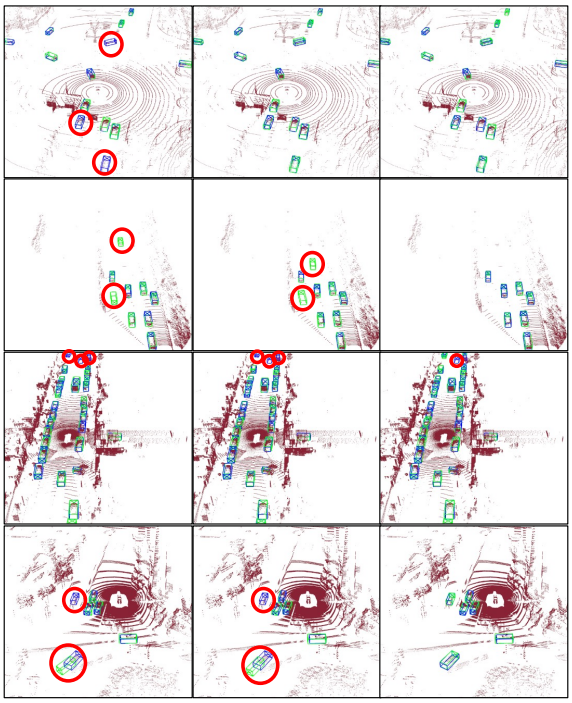


Figure 5. Qualitative comparison of oracle (1st column), baseline (2nd column) and ours (3rd column) in multi-domain learning. Results on the 1st/2nd/3rd/4th rows are ONCE, KITTI, Waymo and NuScenes. We mark predictions in green boxes, ground truths in blue boxes and some noticeable regions in red ellipses. Note that green and blue boxes are overlapped for true positives.

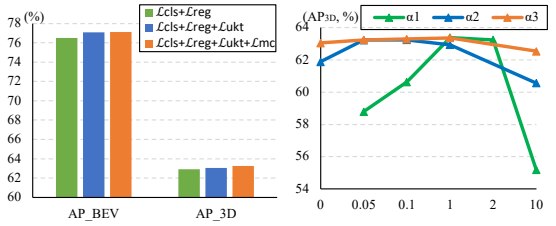
collaboration of multiple domains for knowledge transfer.

Effect of Training Objective. In Fig. 6(a), we study different training losses for multi-domain knowledge transfer with universal feature transformation. It can be seen that the baseline ($\mathcal{L}_{cls}$+$\mathcal{L}_{reg}$) without $\mathcal{L}_{ukt}$ and $\mathcal{L}_{mc}$ yields the worst performance, while our approach with all losses ($\mathcal{L}_{cls}$+$\mathcal{L}_{reg}$+$\mathcal{L}_{ukt}$+$\mathcal{L}_{mc}$) performs the best.

Effect of Weighting Coefficients. In Fig. 6(b), we test MDKT with different values of weighting coefficients (α_1, α_2 and α_3). It can be seen that large α_1 and α_2 and small α_1 cause significant model performance degradation, while the model performance is less sensitive to α_3.

Evaluation with Various Backbones. In Table 5, we evaluate our approach with various backbones, including CenterPoint [38], PointPillars [7] and PillarNet [22] (w/ center head [38]). With different backbones, our approach consistently outperforms the baseline, which shows the compatibility of our approach with different backbones.

Single-DG vs. Multi-DG. In Table 6, we compare with some single-DG methods, including 3D-VField[8], IGL2[33], ChaA[14] and AdvG[33]. Following [8], PointPillars [7] is used as the backbone and single-DG methods are trained on KITTI and tested on KITTI/Waymo, while



(a) Effect of losses. (b) Effect of weighting coefficients.

Figure 6. Evaluating (a) training losses and (b) weighting coefficients in multi-domain learning (average performance).

Method	CenterPoint		PointPillars		PillarNet	
	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}	AP_{3D}
Baseline	75.72	61.97	74.16	59.21	75.69	60.78
Ours	**77.30**	**63.31**	**75.98**	**61.35**	**76.86**	**62.23**

Table 5. Evaluation with various backbones in multi-domain learning. The average performance is reported.

Datasets	IGL2[33]	ChaA[14]	AdvG[33]	3D-VField[8]	Ours
KITTI	76.92	77.05	76.39	77.13	**78.11**
→Waymo	39.86	40.54	40.55	44.61	**56.12**

Table 6. Single-DG methods vs. our multi-DG approach from KITTI (AP of $3DIoU_{0.7}$, moderate) to Waymo (AP of $3DIoU_{0.5}$).

our approach is trained on KITTI+ONCE+NuScenes and tested on KITTI/Waymo. Results of single-DG are cited from [8]. From Table 6, we can observe that on both seen domain (KITTI) and unseen domain (Waymo), our approach surpasses single-DG methods, which shows the advantage of generalizing from multiple domains.

Discussion. Despite the promising results, there are some limitations in this work. First, following the common practice [8, 36], we evaluate on the car/vehicle category, but multi-domain learning and generalization for multiple classes are also important, which we leave for future work. Second, we choose Waymo as the base to set the uniform range since it is a relatively complete dataset, but more designs of the base is worthy of further study. For example, we also explored using NuScenes or other ranges as the base, and overall a large range yields better results.

5. Conclusion

In this work, we propose to explore multi-domain learning and generalization for LiDAR-based 3D object detection. To learn a universal LiDAR-based 3D object detector, we present multi-domain knowledge transfer with universal feature transformation to aggregate spatial-wise and channel-wise informative knowledge across domains. Extensive experiments on four autonomous driving datasets show the superiority of our approach.

Acknowledgement. We thank the reviewers for their valuable suggestion. Also, we thank MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor for the support to further evaluate this research.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[3] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020.

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[5] Huawei. Mindspore. `https://www.mindspore.cn/`, 2020.

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

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

[8] Alexander Lehner, Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Mohammad-Ali Nikouei Mahani, Nassir Navab, Benjamin Busam, and Federico Tombari. 3d-vfield: Adversarial augmentation of point clouds for domain generalization in 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17295–17304, 2022.

[9] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021.

[10] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representation learning from multiple domains for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9526–9535, 2021.

[11] Zhichao Li, Feng Wang, and Naiyan Wang. Lidar r-cnn: An efficient and universal 3d object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7546–7555, 2021.

[12] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8771–8780, 2021.

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

[14] Daniel Liu, Ronald Yu, and Hao Su. Adversarial shape perturbations on 3d point clouds. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020.

[15] Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, and Jingyu Li. Differentiable learning-to-normalize via switchable normalization. In *International Conference on Learning Representations*, 2019.

[16] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. Unsupervised domain adaptive 3d detection with multi-level consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8866–8875, 2021.

[17] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, et al. One million scenes for autonomous driving: Once dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[18] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017.

[19] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018.

[20] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *European Conference on Computer Vision*, pages 68–83. Springer, 2020.

[21] Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, and Min-Jian Zhao. Improving 3d object detection with channel-wise transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2743–2752, 2021.

[22] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *European Conference on Computer Vision*, pages 35–52. Springer, 2022.

[23] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[24] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[25] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. `https://github.com/open-mmlab/OpenPCDet`, 2020.

[26] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[27] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[28] Shuo Wang, Xinhai Zhao, Hai-Ming Xu, Zehui Chen, Dameng Yu, Jiahao Chang, Zhen Yang, and Feng Zhao. Towards domain generalization for multi-view 3d object detection in bird-eye-view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13333–13342, 2023.

[29] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7289–7298, 2019.

[30] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020.

[31] Yi Wei, Zibu Wei, Yongming Rao, Jiaxin Li, Jie Zhou, and Jiwen Lu. Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. In *European Conference on Computer Vision*, 2022.

[32] Guile Wu and Shaogang Gong. Collaborative optimization and aggregation for decentralized domain generalization and adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6484–6493, 2021.

[33] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9136–9144, 2019.

[34] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[35] Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards efficient 3d object detection with knowledge distillation. *Advances in Neural Information Processing Systems*, 2022.

[36] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2021.

[37] Zeng Yihan, Chunwei Wang, Yunbo Wang, Hang Xu, Chaoqiang Ye, Zhen Yang, and Chao Ma. Learning transferable features for point cloud detection via 3d contrastive cotraining. *Advances in Neural Information Processing Systems*, 34:21493–21504, 2021.

[38] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[39] Xingxuan Zhang, Zekai Xu, Renzhe Xu, Jiashuo Liu, Peng Cui, Weitao Wan, Chong Sun, and Chen Li. Towards domain generalization in object detection. *arXiv preprint arXiv:2203.14387*, 2022.

[40] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020.

[41] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021.

[42] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *European Conference on Computer Vision*, pages 496–513. Springer, 2022.