ADAS: A Simple Active-and-Adaptive Baseline for Cross-Domain 3D Semantic Segmentation

Ben Fei[1], Siyuan Huang[3], Jiakang Yuan[1], Botian Shi[2], Bo Zhang[†,2], Tao Chen[1], Min Dou[2], Yu Qiao[2]
[1]Fudan University [2]Shanghai AI Laboratory, [3]Shanghai Jiaotong University
bfei21@m.fudan.edu.cn, siyuan_sjtu@sjtu.edu.cn, jkyuan22@m.fudan.edu.cn

Abstract

State-of-the-art 3D semantic segmentation models are trained on the off-the-shelf public benchmarks, but they often face the major challenge when these well-trained models are deployed to a new domain. In this paper, we propose an Active-and-Adaptive Segmentation (ADAS) baseline to enhance the weak cross-domain generalization ability of a well-trained 3D segmentation model, and bridge the point distribution gap between domains. Specifically, before the cross-domain adaptation stage begins, ADAS performs an active sampling operation to select a maximally-informative subset from both source and target domains for effective adaptation, reducing the adaptation difficulty under 3D scenarios. Benefiting from the rise of multi-modal 2D-3D datasets, ADAS utilizes a cross-modal attention-based feature fusion module that can extract a representative pair of image features and point features to achieve a bi-directional image-point feature interaction for better safe adaptation. Experimentally, ADAS is verified to be effective in many cross-domain settings including: 1) Unsupervised Domain Adaptation (UDA), which means that all samples from target domain are unlabeled; 2) Unsupervised Few-shot Domain Adaptation (UFDA) which means that only a few unlabeled samples are available in the unlabeled target domain; 3) Active Domain Adaptation (ADA) which means that the selected target samples by ADAS are manually annotated. Their results demonstrate that ADAS achieves a significant accuracy gain by easily coupling ADAS with self-training methods or off-the-shelf UDA works. Our code is available at https://github.com/Fayeben/ADAS

1. Introduction

In recent years, 3D semantic segmentation models [11, 14, 36] have achieved remarkable performance gains, owing to the large-scale annotated benchmarks, such as SemanticKITTI [1] and nuScenes [3], *etc.* But current 3D semantic segmentation models still suffer from a severe performance degradation issue when they are directly deployed to a novel domain. Actually, one straightforward method to alleviate such a performance drop issue is to build a specific dataset for the novel domain, by extensively collecting target-domain data and performing labour-intensive human annotation, and refine the segmentation model on the newly-constructed domain [8, 30]. But this is impractical in many real-world applications such as autonomous driving.

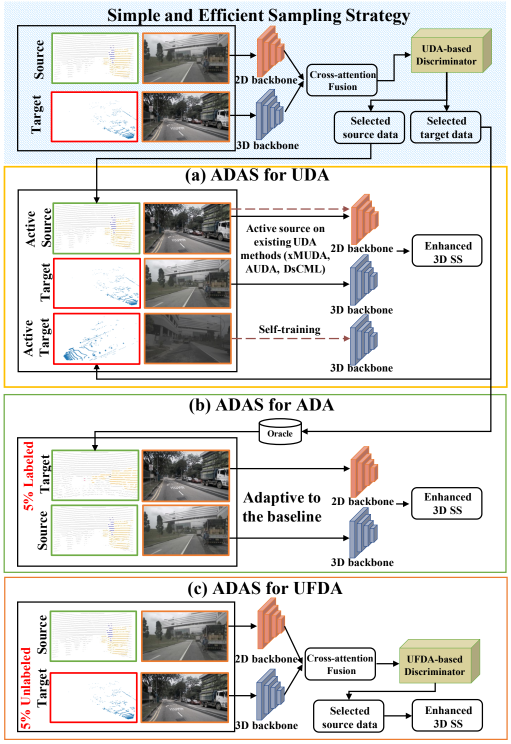

Figure 1. Our ADAS employs a **unified** multi-modal sampling strategy that can be effectively applied to various DA settings: (a) UDA, (b) UFDA, and (c) ADA.

Domain Adaptation (DA), as one of the typical techniques in the transfer learning community, aims to tackle the above-mentioned performance drop issue by learning

†Corresponding author. Email: bo.zhangzx@gmail.com

domain-invariant representations [23, 33, 40], which includes Unsupervised Domain Adaptation (UDA), Unsupervised Few-shot Domain Adaptation(UFDA), and Active Domain Adaptation (ADA) settings according to the condition of target domain. Recently, inspired by UDA study in 2D image recognition [2, 39, 46], some researchers [16, 20, 25] have tried to address the 3D point cloud-induced domain discrepancies under the UDA setting. The existing methods [20, 25], *e.g.* xMUDA [16], attempt to extract 2D and 3D features using two different network branches, exploiting the inter-modal feature complementarity by 2D-3D modal feature matching. Orthogonal to these UDA 3D segmentation works focusing on leveraging cross-modal data only under **one DA setting** (*i.e.* UDA), we aim to investigate how to exploit multi-modal features to achieve a high-efficient and safe 3D segmentation model adaptation from **multiple DA settings**.

To design a unified method that can be effective under multiple DA settings, we start from the perspective of sub-domain data sampling, which means that we are expected to pick up a maximally-informative subset from both source and target domains to reduce the difficulty of model transfer. Specifically, on the one hand, due to the intra-domain feature variations, some samples from the source domain may present a large data distribution difference with the target domain data, causing a severe model adaptation interference caused by these irrelevant source-domain samples. On the other hand, in autonomous driving scenarios, samples within the same target sequence have a similar data distribution, which may result in redundant target-domain adaptation by self-training UDA methods [22, 45, 47].

Motivated by this observation, we propose a simple Active-anD-Adaptive Segmentation (ADAS) baseline to revisit the 3D cross-domain segmentation problem. For fully leveraging the multi-modal data from different domains, we design a cross-modal attention-based feature fusion module that can encode features from a single image or point cloud modality, and then perform an image-to-point and point-to-image feature-level information interaction by a symmetrical cross-branch attention structure. Furthermore, the learned cross-modal features are utilized as an important proxy for subsequent multi-modal subset sampling process, which is versatile to many DA settings. As illustrated in Fig. 1, such a subset sampling way can be easily combined with the off-the-shelf DA or UDA variants, achieving a better source-domain pre-trained performance only using the sampled source data and a promising target-domain performance using both the sampled source-and-target data.

We conduct extensive experiments on several public benchmarks, including nuScenes [3], A2D2 [10], and SemanticKITTI [1] under three DA settings: 1) UDA setting where all data from the target domain are unlabeled; 2) UFDA setting where we can only access few-shot unlabeled samples from the target domain; 3) ADA setting where a portion of unlabeled target data is selected to be annotated by an oracle. The experimental results demonstrate that our ADAS can be easily applied to the UDA, UFDA, and ADA settings, to enhance the model transferability, outperforming the existing UDA-based 3D segmentation works xMUDA [16], AUDA [20], and DsCML [25] by 7.75%, 7.92%, and 4.61% on USA-to-Singapore setting.

2. Related Works

2.1. Active Domain Adaptation

Active learning aims to develop annotation-efficient algorithms via sampling the most representative samples to be labeled by an oracle [28]. Most recently, active learning coupled with domain adaptation, termed as Active Domain Adaptation (ADA), has great practical significance. Nevertheless, only a few previous researchers focus on addressing the problem, pioneered by active adaptation in the area of sentiment classification for text data [27]. Rita et al. [4] choose target samples to learn importance weights for source instances by solving a convex optimization problem of minimizing Maximum Mean Discrepancy (MMD). Recently, Su et al. [31] try to study ADA problem in the context of Convolution Neural Networks (CNN), and instances are selected based on their designed uncertainty and "targetness". However, these sampling strategies are designed based on the 2D image domain, and it is intractable to directly apply these 2D image-based sampling strategies to the 3D image-point multi-modal task. In our work, for the first time, we design a novel active-and-adaptive segmentation baseline to sample the most informative 2D-3D pairs to enhance the weak cross-domain generalization ability of a well-trained 3D segmentation model.

2.2. 3D Semantic Segmentation

In 3D semantic segmentation, 3D point clouds are often represented as voxels. For instance, SSCNs [11] and the following works [7, 15, 41] leverage hash tables to convolve only on sampled voxels, allowing for very high resolution with typically only one point per voxel. Point-based methods conduct computation in continuous 3D space and can directly take point clouds as input. PointNet++ [26] utilizes point-wise convolution and max-pooling to extract global features and local neighborhood aggregation for hierarchical learning with CNN. Following that, continuous convolutions [37] and deformable kernels [32] have been proposed. DGCNN [38] and LDGCNN [44] perform convolution on the edges of a point cloud. In this work, SparseConvNet [11] is chosen as a 3D backbone, which is widely utilized in the UDA methods [16, 20, 25].

3. The Proposed Method

Our method is proposed to enhance the adaptability of a 3D semantic segmentation model by assuming the presence of 2D images and 3D point clouds. In this section, we will first define the problem and illustrate the network architecture in Sec. 3.1. Following that, the ADAS framework is described in Sec. 3.2.

3.1. Preliminary

Problem Definition. Suppose that $\mathcal{S}$ and $\mathcal{T}$ define a source domain and a target domain, both of which contain different modalities including 2D images $\boldsymbol{x}_{\mathrm{2D}}^s$ and $\boldsymbol{x}_{\mathrm{2D}}^t$, and 3D point clouds $\boldsymbol{x}_{\mathrm{3D}}^s$ and $\boldsymbol{x}_{\mathrm{3D}}^t$. The purpose of cross-domain 3D semantic segmentation is to adapt a well-trained segmentation baseline from a labeled source domain $\mathcal{S}$ to a new target domain $\mathcal{T}$ with the data distribution shift.

To design a unified cross-domain 3D semantic segmentation pipeline under different target-domain conditions, we consider the following adaptation settings: 1) Unsupervised Domain Adaptation (UDA), where we can access all unlabeled 2D-3D sample pairs from the target domain; 2) Unsupervised Few-shot Domain Adaptation (UFDA), where only a few data (*e.g.,* 5% target data) from the unlabeled target domain are available; 3) Active Domain Adaptation (ADA), where one can sample a subset from the full set of the unlabeled target domain and perform the manual annotation process.

3D Segmentation Architecture. Given the label in the source domain, the segmentation network is trained in a supervised manner with cross-entropy loss function for the 2D image $\boldsymbol{x}_{2D}^s$ and 3D point cloud $\boldsymbol{x}_{3D}^s$, which can be formulated as:

$$\begin{aligned}\mathcal{L}_{2D}^{seg}\left(x_{2D}^s, \phi(y_{3D}^s)\right) &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\phi(y_{(n,c)}^s)\log P_{2D}^{(n,c)}, \\ \mathcal{L}_{3D}^{\mathrm{seg}}\left(x_{3D}^s, y_{3D}^s\right) &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{(n,c)}^s \log P_{3D}^{(n,c)},\end{aligned} \tag{1}$$

where $y_{(n,c)}^s$ and $P^{(n,c)}$ stand for the ground-truth label and prediction of the point n for the class c, respectively. ϕ is the projection of the point cloud to the front view. Herein, the overall objective of the source domain can be formulated as:

$$\min_{\theta_{2D},\theta_{3D}} \frac{1}{N^s}\sum_{x_s\in S}\mathcal{L}_{seg}^{2D}\left(x_{2D}^s, \phi(y_{3D}^s)\right) + \mathcal{L}_{seg}^{3D}\left(x_{3D}^s, y_{3D}^s\right), \tag{2}$$

where θ_{2D} and θ_{3D} are the parameters of the 2D sub-network and the 3D sub-network, respectively.

3.2. ADAS: Active-and-Adaptive 3D Segmentation Baseline

Previous works mainly focus on cross-model learning within a single dataset, facing unforeseen cross-model performance hurt when they are directly used in cross-domain applications. Thus, ADAS is proposed to reduce the domain discrepancies of multi-modality data from two aspects: 1) **Active 3D Segmentation**, meaning that we actively mine a subset of source-and-target data that are representative and transferable for dynamically changing target-domain distribution, and 2) **Adaptive 3D Segmentation**, meaning that we adapt a well-trained cross-modal baseline to the target domain according to the above-sampled subset of data.

Active 3D Segmentation: A Cross-modal Attention-based Feature Fusion Module. Although inter-domain differences exist, we found that many frames in the source domain have a similar data distribution to those in the target domain. This phenomenon motivates us to mine target-domain-like frames from the source domain to enhance the model adaptability, regarded as **source-domain sampling**. On the other hand, for 3D segmentation scenarios, there are many semantically-duplicated samples between adjacent frames. To this end, to reduce the cost of target data acquisition and annotation, we propose to select a maximally-informative subset of a given unlabeled target domain to perform the pseudo-label or manual annotation process regarded as **target-domain sampling**.

In this part, we study how to design a unified sampling strategy to pick up samples from both domains, to reduce the feature gaps between domains and enhance the model transferability. An effective method to alleviate the inter-domain data distribution differences of 3D point clouds is leveraging the multi-modality residing in the 3D segmentation dataset. Motivated by this, a cross-modal cross-attention-based feature fusion module is exploited to train a domain discriminator that can dynamically evaluate the representativeness of each sample from image-point cloud modalities. Given a pair of images and point clouds (x_{2D}, x_{3D}), the 2D and 3D backbones are respectively used to extract a pair of high-level features (f_{2D}, f_{3D}), where $f_{2D} \in \mathbb{R}^{N\times F_{2D}}$ and $f_{3D} \in \mathbb{R}^{N\times F_{3D}}$. Although the feature pairs (f_{2D}, f_{3D}) extracted by the backbone network contain rich semantic information of a single modal, the cross-modality semantic relations between the images and point clouds are not taken into consideration.

Different from previous works [16, 20, 25] that align cross-modal features using a well-designed KL divergence loss, ADAS aims to exploit the relations between modalities by Transformer [43]. This approach achieves better cross-modal feature fusion performance and thus is beneficial to pick up data representative for both image-and-point modalities. The feature maps from each modality $f \in \mathbb{R}^{N\times F_{2D}}$ can be described as $f = \left[f^1, f^2, \ldots, f^N\right]$, where $f_{2D}^i \in \mathbb{R}^{F_{2D}}$ and $f_{3D}^i \in \mathbb{R}^{F_{3D}}$. In our implementation, the F_{2D} is equal to F_{3D}. Specifically, given feature maps with weight parameters $W_{2D}^q, W_{2D}^k, W_{2D}^v$ for the image branch $2D$, and parameters $W_{3D}^q, W_{3D}^k, W_{3D}^v$ for point

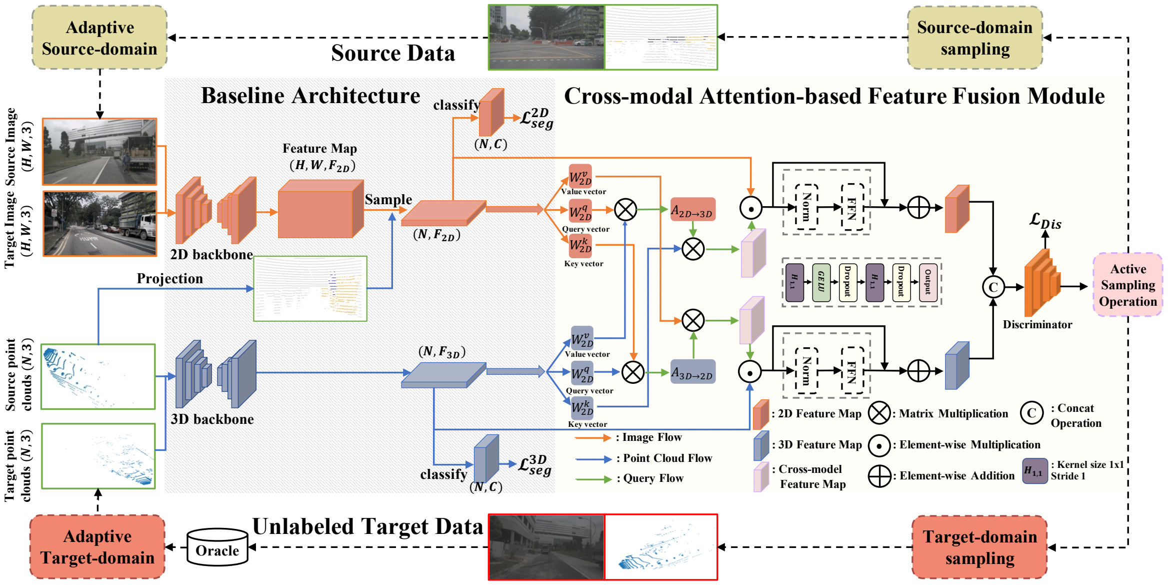

Figure 2. The network architecture of ADAS, which consists of a 3D semantic segmentation baseline, the Cross-modal Attention-based Feature Fusion Module, and the Active Sampling Operation. The 3D segmentation baseline comprises a 2D U-Net-Style ConvNet [12] backbone, which takes an image as input, and a 3D U-Net-Style SparseConvNet [11] backbone, which receives a point cloud as input. The Cross-modal Attention-based Feature Fusion Module can leverage the features (F_{2D} and F_{3D}) to exploit a representative pair of image features and point cloud features to achieve a bi-directional image-point feature interaction. The last active sampling module utilizes the cross-attention features to perform both the source-domain sampling and target-domain sampling via the unified module.

cloud branch $3D$, the query vector q^i, key vector k^i, and value vector v^i can be calculated as follows:

$$q_{2D}^i := W_{2D}^q f_{2D}^i, \quad k_{2D}^i := W_{2D}^i f_{2D}^i, \quad v_{2D}^i := W_{2D}^v f_{2D}^i, \quad (3)$$

$$q_{3D}^i := W_{3D}^q f_{3D}^i, \quad k_{3D}^i := W_{3D}^i f_{3D}^i, \quad v_{3D}^i := W_{3D}^v f_{3D}^i. \quad (4)$$

After that, the symmetrical cross-attention is leveraged in a bi-direction manner: 1) The 2D branch-related features are obtained using the value vector v_{3D}^i from $3D$ backbone branch, formulated as **3D** $\rightarrow$ **2D**; 2) Similarly, the 3D branch-related features are acquired using the value vector v_{2D}^i of the 2D backbone branch, formulated as **2D** $\rightarrow$ **3D**.

Specifically, for **3D** $\rightarrow$ **2D**, $\mathbf{A}_{3D\rightarrow 2D} \in \mathbb{R}^{N\times N}$ denotes the attention score matrix obtained via the matrix multiplication as follows:

$$\mathbf{A}_{3D\rightarrow 2D} = K_{2D} V_{3D}^{\mathrm{T}}, \quad (5)$$

where $K_{2D} = \left[k_{2D}^1, \ldots, k_{2D}^N\right] \in \mathbb{R}^{N\times F_{2D}}$ and $V_{3D} = \left[v_{3D}^1, \ldots, v_{3D}^N\right] \in \mathbb{R}^{N\times F_{3D}}$, which can be obtained through Eq. 3, Eq. 4. Moreover, a softmax layer combined with a scaling operation is leveraged to carry out the normalization for attention scores and find the semantically-related regions according to information from another modality $3D$ branch, which can be calculated as follows:

$$R_{3D\rightarrow 2D} = \mathrm{Softmax}\left(\frac{\mathbf{A}_{3D\rightarrow 2D}}{\sqrt{F_{2D}}}\right) V_{2D}, \quad (6)$$

where $V_{2D} = \left[v_{2D}^1, v_{2D}^2, \ldots, v_{2D}^N\right] \in \mathbb{R}^{N\times F_{2D}}$, and $R_{3D\rightarrow 2D} \in \mathbb{R}^{N\times F_{2D}}$ represents the encoded semantic relations from 3D to 2D features. After that, the $R_{3D\rightarrow 2D}$ is reshaped to the same dimension as the backbone features $f_{2D} \in \mathbb{R}^{F_{2D}\times N}$ and then fused with f_{2D} by an element-wise addition, in order to enhance the semantically similar backbone features.

For **2D** $\rightarrow$ **3D**, the $R_{2D\rightarrow 3D} \in \mathbb{R}^{N\times F_{3D}}$ can be easily obtained by performing a symmetrical process described above, which can be written as follows:

$$\begin{aligned} \mathbf{A}_{2D\rightarrow 3D} &= K_{3D} V_{2D}^{\mathrm{T}}, \\ R_{2D\rightarrow 3D} &= \mathrm{Softmax}\left(\frac{\mathbf{A}_{2D\rightarrow 3D}}{\sqrt{F_{3D}}}\right) V_{3D}. \end{aligned} \quad (7)$$

Given single-modal features (f_{2D}, f_{3D}) and cross-modal features $(R_{3D\rightarrow 2D}, R_{2D\rightarrow 3D})$, the enhanced representations can be calculated as follows:

$$\begin{aligned} \hat{f}_{2D} &= \mathrm{FFN}\left(\mathrm{Norm}\left(f_{2D} \odot R_{3D,2D}\right)\right), \\ \hat{f}_{3D} &= \mathrm{FFN}\left(\mathrm{Norm}\left(f_{3D} \odot R_{2D,3D}\right)\right), \end{aligned} \quad (8)$$

where $\odot$ denotes an element-wise multiplication operation. And the FFN module allows the backbone to focus on the cross-branch modality discrepancies over the predicted semantically-similar regions. Overall, $\hat{f}_{2D}$ and $\hat{f}_{3D}$ are defined as the output features and will be concatenated and then fed into a well-designed domain discriminator, as illustrated in Fig. 2.

Furthermore, a 3-layer fully-convolutional domain discriminator $\mathcal{D}$ with parameters θ_{Dis} is constructed, which takes the cross-attention features $[\hat{f}_{2D}, \hat{f}_{3D}]$ as input and is trained to distinguish the source data from the target ones. We label the source domain and the target domain as '0' and '1', respectively. Let $\mathcal{L}_{\mathcal{D}}$ represent the domain classification loss of the discriminator. The training objective of the discriminator can be written as follows:

$$\begin{aligned} \mathcal{L}_{Dis} = \min_{\theta_{Dis}} \frac{1}{N^s} \sum_{x^s_{2D}, x^s_{3D}} \mathcal{L}_{\mathcal{D}} \left([\hat{f}_{2D}, \hat{f}_{3D}], 1 \right) \\ + \frac{1}{N^t} \sum_{x^s_{2D}, x^s_{3D}} \mathcal{L}_{\mathcal{D}} \left([\hat{f}_{2D}, \hat{f}_{3D}], 0 \right). \end{aligned} \tag{9}$$

Once the trained domain discriminator is obtained, we can leverage the output of the domain discriminator to score each frame, which represents the domainness of a sample belonging to the source or target domain. Since the active sampling strategy is one general approach, we can use it for both the source and target domains sampling. For the source domain, the higher score means that the frame complies with the distribution of the target domain. For the target domain, when the frame possesses the more informative characteristic, it will get a higher score. After the scoring stage is finished, all frames are sorted by the calculated scores, and the budget $\mathcal{B}$ frames are chosen as the sampled source data and target data.

Adaptive 3D Segmentation: Effective Model Adaptation Strategy. When an informative subset jointly sampled from both source and target domains is determined, an effective adaptation strategy is designed to fully leverage these samples to enhance 3D segmentation model's adaptability. In this part, we investigate two representative target-oriented adaptation strategies to explore how to perform an effective model transfer based on these informative samples.

1) Self-training using pseudo-labels: Cross-modal learning is complementary to the pseudo-labeling technique, which is originally employed in semi-supervised and unsupervised domain adaptation tasks [18, 42]. In detail, given a pre-trained source model, the pseudo-labels can be obtained by selecting a portion of highly confident predicted results according to the pre-trained model. It should be emphasized that the **Pseudo-Labeling methods (PL)** generate labels for the whole target domain data yet bring many label noises. By comparison, we can leverage the ADAS to pick up samples with high domainness scores. Then, we only pseudo-label those selected samples to perform the subsequent self-training process. Such a sampling-based pseudo-labeling method is termed as **Active Pseudo-Labeling (APL)**. Based on those pseudo-labeled target frames, the model is further adapted to the target domain using the following segmentation loss:

$$\min_{\theta} \left[\frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}^s} \left(\mathcal{L}_{\text{seg}} \left(\boldsymbol{x}^s, \boldsymbol{y}^s_{3\text{D}} \right) \right) + \frac{1}{|\mathcal{T}|} \sum_{\boldsymbol{x}^t} \left(\mathcal{L}_{\text{seg}} \left(\boldsymbol{x}^t, \hat{\boldsymbol{y}}_{3\text{D}} \right) \right) \right] \tag{10}$$

where $\hat{\boldsymbol{y}}^{3\text{D}}$ represents the pseudo-labels

2) Off-the-shelf UDA techniques: Furthermore, our cross-modal attention-based source-and-target sampling strategy can be easily integrated into the existing methods, such as xMUDA [16], AUDA [20], and the state-of-the-art DsCML [25], to further boost their cross-domain segmentation accuracy. Specifically, we first utilize the proposed sampling strategy to pick up the maximally-informative samples from both source and target domains, and construct a subset domain. Then, these UDA-based techniques are employed to adapt the baseline model using the constructed subset to achieve a more effective and safe adaptation result.

4. Experiments

4.1. Dataset Description

Our ADAS is evaluated under 3 real-to-real adaptation scenarios. The first setting is the day-to-night adaptation, which presents a domain gap caused by the illumination changes, where the laser beams are almost invariant to illumination conditions while the camera suffers from a low-light environment. The second scenario is the country-to-country adaptation, representing a large domain difference where the 3D shapes and images might change frequently. The last setting is the dataset-to-dataset adaptation, containing the variations in the sensor setup. The widely-used autonomous driving datasets nuScenes [3], A2D2 [10], and SemanticKITTI [1] are leveraged, where the LiDAR and camera are synchronized and calibrated to obtain the projection between a 3D point and its corresponding 2D image pixel. All these datasets consist of 3D annotations. And only the front camera images and the LiDAR points are utilized for simplicity and consistency across datasets. For nuScenes, the point-wise labels for 3D semantic segmentation are obtained by assigning the corresponding object label if that point lies inside an annotated 3D bounding box. If not, that point will be labeled as background. Following the works [16, 25], two scenarios of Day/Night and USA/Singapore are conducted. For the adaptation from A2D2 to SemanticKITTI, the point-wise labels on A2D2 are directly provided, and the 10 categories shared between the two datasets are selected.

4.2. The Designed Baselines

In order to validate the effectiveness of our ADAS, 5 baselines are designed: **Baseline 1** only leverages 2D features extracted by the 2D backbone. The 2D scores ob-

	Methods	USA/Singapore			Day/Night			A2D2/SemanticKITTI		
		2D	3D	Softmax avg.	2D	3D	Softmax avg.	2D	3D	Softmax avg.
Domain adaption methods	Oracle	66.4	63.8	71.6	48.6	47.1	55.2	58.3	71.0	73.7
	MinEnt (CVPR-19) [35]	53.4	47.0	59.7	44.9	43.5	51.3	38.8	38.0	42.7
	PL (CVPR-19) [19]	55.5	51.8	61.5	43.7	45.1	48.6	37.4	44.8	47.7
	CyCADA (ICML-18) [13]	54.9	48.7	61.4	45.7	45.2	49.7	38.2	44.3	43.9
	AdaptSegNet (CVPR-18) [34]	56.3	47.7	61.8	45.3	44.6	49.6	38.8	44.3	44.2
	CLAN (CVPR-19) [21]	57.8	51.2	62.5	45.6	43.7	49.2	39.2	44.7	44.5
	xMUDA (CVPR-20) [16]	59.3	52.0	62.7	46.2	44.2	50.0	36.8	43.3	42.9
	xMUDA + PL (CVPR-20) [16]	61.1	54.1	63.2	47.1	46.7	50.8	43.7	48.5	49.1
	AUDA [20]	59.7	51.7	63.0	48.7	46.2	55.7	43.3	43.3	47.3
	AUDA + PL [20]	59.8	52.0	63.1	49.0	47.6	54.2	43.0	43.6	46.8
	DsCML (CVPR-21) [25]	61.3	53.3	63.6	48.0	45.7	51.0	39.6	45.1	44.5
	DsCML + CMAL (CVPR-21) [25]	63.4	55.6	64.8	49.5	48.2	52.7	46.3	50.7	51.0
	DsCML + CMAL + PL (CVPR-21) [25]	63.9	56.3	65.1	50.1	48.7	53.0	46.8	51.8	52.4
Sampling source with ADAS	Source only	53.4	46.5	61.3	42.2	41.2	47.8	34.2	35.9	40.4
	Source only + Source-domain sampling	55.2	49.7	63.1	45.4	40.8	51.7	35.1	38.0	44.7
	Improvement	***+1.8***	***+3.2***	***+1.8***	***+3.2***	***+0.4***	***+3.9***	***+0.9***	***+2.1***	***+4.3***
	Source only + Source-domain sampling + PL	59.4	53.6	63.3	46.2	42.6	51.4	42.0	35.9	45.8
	Improvement	***+5.0***	***+7.1***	***+2.0***	***+4.0***	***+1.4***	***+3.6***	***+7.8***	***+5.4***	***+5.4***
	Source only + Source-domain sampling + APL	60.9	53.1	66.8	47.0	42.2	51.9	41.9	41.4	46.7
	Improvement	***+7.5***	***+6.6***	***+5.5***	***+4.8***	***+1.0***	***+3.1***	***+7.7***	***+5.5***	***+6.3***
UDA with ADAS	xMUDA + Source-domain sampling	55.9	50.1	63.4	45.6	42.2	51.2	43.6	47.6	50.8
	Improvement	-	-	***+0.7***	-	-	***+1.2***	-	-	***+7.9***
	xMUDA + Source-domain sampling + PL	60.8	51.4	63.6	46.2	42.6	51.4	42.7	43.3	50.6
	Improvement	-	-	***+0.6***	-	-	***+1.4***	-	-	***+1.5***
	xMUDA + Source-domain sampling + APL	57.6	54.6	63.8	48.4	42.8	51.5	45.1	46.3	50.6
	Improvement	-	-	***+0.6***	-	-	***+0.7***	-	-	***+1.5***
	AUDA + Source-domain sampling	54.7	50.3	63.2	49.1	48.4	54.4	43.7	42.3	47.0
	Improvement	-	-	***+0.0***	-	-	***+0.2***	-	-	***+0.2***
	AUDA + Source-domain sampling + PL	59.4	53.59	63.3	49.1	41.3	54.0	42.6	37.9	47.1
	Improvement	-	-	***+0.1***	-	-	***-0.2***	-	-	***+0.3***
	AUDA + Source-domain sampling + APL	58.3	52.4	63.6	48.5	41.6	54.7	42.0	39.7	47.9
	Improvement	-	-	***+0.5***	-	-	***-0.5***	-	-	***+1.1***
	DsCML +CMAL + Source-domain sampling	55.6	52.0	65.0	49.3	41.6	53.5	43.5	46.3	51.1
	Improvement	-	-	***+0.2***	-	-	***+0.8***	-	-	***+0.1***
	DsCML +CMAL + Source-domain sampling + PL	59.9	55.2	65.7	49.1	41.5	53.3	44.7	47.0	52.7
	Improvement	-	-	***+0.6***	-	-	***+0.3***	-	-	***+0.3***
	DsCML +CMAL + Source-domain sampling + APL	61.8	55.7	66.2	49.2	42.6	53.4	47.3	46.1	54.3
	Improvement	-	-	***+1.1***	-	-	***+0.4***	-	-	***+1.9***
Ours + UFDA	Ours + UFDA-10%	55.0	49.3	63.0	45.1	41.0	51.3	35.0	37.9	44.2
	Ours + UFDA-5%	54.6	48.9	62.2	43.7	40.7	50.5	34.5	37.2	43.6

Table 1. Segmentation results (mIoU) for different DA settings including UDA, UDA methods coupled with ADAS, and UFDA under three cross-modal domain adaptation scenarios. “ Avg ” represents the performance which is obtained by computing the mean of the predicted 2D and 3D probabilities after softmax operation. “ PL ” denotes the pseudo-labeling operation for all target data, while “ APL ” is proposed in this work to pseudo-label the actively sampled target-domain data.

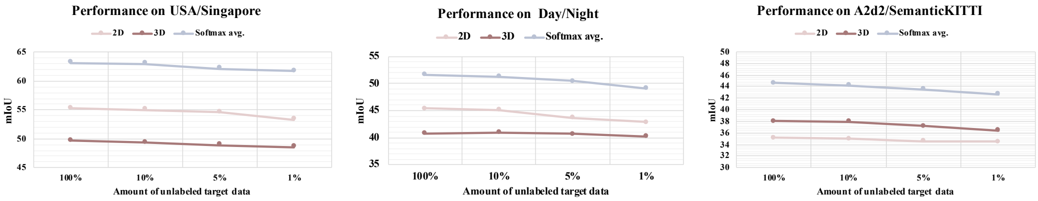

Figure 3. Segmentation results (mIoU) under the UFDA setting, where we study the discriminator performance under different amounts of samples from the target domain: 100%, 10%, 5%, and 1% target domain samples.

tained by the domain discriminator $\mathcal{D}$ are used to sample the source and target data. **Baseline 2** utilizes the 3D features for sampling, where the sampling process will depend on the 3D scores from the $\mathcal{D}$. **Baseline 3** naively averages the 2D and 3D scores for the sampling strategy. Since the scores are calculated from a single modality, the sampled data might be distributed in a single modality. For instance, the data sampled by the 2D scores will concentrate more on the image information and vice versa. **Baseline 4** integrates the 2D active domain adaptation method CLUE [6], to be compared with our ADAS in 3D semantic segmentation. **Baseline 5** further applies the LabOR [29] (a representative method for active segmentation model) to pick up data from the target domain. The module-wise ablation study of all these baselines can be found in Table 3, and please refer to the Appendix for the network visualization

	Methods	USA/Singapore			Day/Night			A2D2/SemanticKITTI		
		2D	3D	Softmax avg.	2D	3D	Softmax avg.	2D	3D	Softmax avg.
	source only	53.4	46.5	61.3	42.2	41.2	47.8	34.2	35.9	40.4
UDA	xMUDA + PL [16]	61.1	54.1	63.2	47.1	46.7	50.8	43.7	48.5	49.1
	AUDA + PL [20]	59.8	52.0	63.1	49.0	47.6	54.2	43.0	43.6	46.8
	DsCML + CMAL +PL [25]	63.9	56.3	65.1	50.1	48.7	53.0	46.8	51.8	52.4
ADA	Ours + Source-domain sampling	55.2	49.7	63.1	45.4	40.8	51.7	35.1	38.0	44.7
	Ours + Random sampling (Target domain)	55.5	58.7	64.9	47.5	41.0	53.0	42.8	43.1	48.2
	Ours + Source-domain sampling + Target-domain sampling	63.6	52.8	67.4	47.1	41.5	53.6	43.9	47.8	51.1
	Ours + Source-domain sampling + Target-domain sampling + APL	62.7	56.3	68.1	49.8	41.2	54.6	46.2	45.7	52.8
	Oracle	66.4	63.8	71.6	48.6	47.1	55.2	58.3	71.0	73.7

Table 2. Segmentation results (mIoU) for the ADA setting under 5% target-domain annotation budget, where Random sampling denotes that we randomly sample 5% target domain to perform the manual annotation.

of all designed baselines.

4.3. Implementation Details

Network Baseline. For a fair comparison with the state-of-the-art multi-modal 3D domain adaptation frameworks, we leverage the ResNet34 [12] pre-trained on ImageNet [9] as the encoder for the 2D network branch and the SparseConvNet [11] with U-Net architecture for 3D network branch. Moreover, the voxel size is set to 5 cm in the 3D network to ensure that only one 3D point lies in a single voxel. The models are trained and evaluated with PyTorch toolbox [24]. We train ADAS with one Tesla A100 GPU.

Parameter Settings. In the training stage, the batch size is set to 8, and the Adaptive Moment Estimation (Adam) [17] is used as the optimizer with β_1 = 0.9 and β_2 = 0.999. We set the learning rate of $1 \times e^{-3}$ initially and follow the poly learning rate policy [5] with a poly power of 0.9. We set the max training iteration as 100k.

4.4. Experimental Results

In this part, we conduct experiments on UDA, UFDA, and ADA tasks, and comprehensively show the generality and effectiveness in addressing 3D segmentation model's domain discrepancies.

Results on 3D UDA task. *1) The effectiveness of source-domain sampling*: As shown in Table 1, when the segmentation baseline model is deployed to a new domain (*e.g.*, from USA to Singapore), its segmentation accuracy is seriously degraded (only 61.3% compared with 71.6% achieved by fully-supervised model). One cost-free solution provided by our ADAS is to use the designed source-domain sampling strategy, which can actually sample some target-domain-like source data to bridge the large domain shift from different modal data, achieving 1.8%, 3.9%, and 4.3% accuracy gains for different cross-domain scenarios. The accuracy gain achieved by only source sampling strategy even surpasses some representative uni-modal UDA methods such as MinEnt [35], PL [19], CyCADA [13], AdaptSegNet [34], and CLAN [21].

Furthermore, we would like to emphasize that another advantage of our ADAS is that it can be plugged and played into the existing UDA models (*e.g.*, xMUDA [16], AUDA [20], and DsCML [25]), to further strengthen these UDA models' adaptability. For example, xMUDA coupled with our ADAS achieves 0.7%, 1.2%, and 7.9% segmentation accuracy gain compared with xMUDA itself, for the USA/Singapore, Day/Night, and A2D2/SemanticKITTI scenarios. Also, we conduct extensive validations by inserting our ADAS into many UDA models, such as AUDA [20] and DsCML [25], and observe consistent segmentation accuracy gains.

2) The effectiveness of target-domain sampling: APL means that we only perform the pseudo-labeling process for the selected target samples by ADAS. Also, it can be seen from Table 1 that, compared with the widely-used Pseudo-Labeling (PL) strategy, APL can obtain a relatively high mIoU in A2D2/SemanticKITTI scenario. This is mainly because the selected target data (SemanticKITTI) by ADAS present less noise and can be pseudo-labeled more accurately. By merging the merits of the APL and self-training strategy, ADAS can further enhance the model adaptability.

Results on 3D UFDA task. Motivated by the outstanding performance of ADAS in UDA task, we try to reduce the number of unlabeled target data to check the domain-related representation ability of the designed discriminator with the cross-model attention-based feature fusion module. Such a task setting can be regarded as UFDA task. Specifically, we randomly select a small number of target data (*e.g.* 10%, 5%, and even 1%) to train the discriminator, and the results are shown in Fig. 3. It can be seen that owing to the robust source-domain sampling strategy, only a few-shot unlabeled target data also can effectively train the designed discriminator and achieve a comparable target-domain segmentation accuracy, providing an option to eliminate the need for unlabeled target data.

Results on 3D ADA task. In the above DA tasks, all samples from the target domain are unlabeled. Although performing the APL or PL based on self-training methods on the unlabeled target domain can improve the target-domain

Methods	USA/Singapore			Day/Night			A2D2/SemanticKITTI		
	2D	3D	Softmax avg.	2D	3D	Softmax avg.	2D	3D	Softmax avg.
Baseline1: ADAS-2D	55.2	47.1	62.8	45.2	40.9	50.6	34.8	36.6	43.0
Baseline2: ADAS-3D	53.8	48.8	61.8	44.9	41.2	51.0	35.5	38.4	42.7
Baseline3: ADAS-2D-3D-Avg	53.4	48.9	61.8	45.0	41.2	51.1	35.7	36.9	43.2
Baseline4: ADAS-CLUE [6]	53.2	46.3	60.1	42.8	41.0	49.4	35.2	36.0	40.6
Baseline5: ADAS-LabOR [29]	52.2	50.0	58.9	44.0	40.1	49.2	31.3	32.0	38.6
Ours: ADAS-2D-3D-Attention	55.2	49.7	63.1	45.4	40.8	51.7	35.1	38.0	44.7

Table 3. The module-level ablation studies of ADAS under three cross-modal domain adaptation scenarios.

segmentation accuracy, there is still a large accuracy gap between the unsupervised target-domain model and the fully-supervised one. To this end, in this work, we also study the ADA task, which assumes that a portion of unlabeled target data selected by an active learning algorithm can be annotated by an oracle or human expert. Instead of using pseudo-labeled target samples, we utilize an oracle to annotate all selected target samples to investigate if our ADAS can be applied to ADA task.

As shown in Table 2, the baseline model fine-tuned on the sampled source data and annotated target data can significantly enhance the model transferability, obtaining 67.4%, 53.6%, and 51.1% segmentation accuracies under different scenarios. Furthermore, we conduct experiments on changing the budget of the target-domain annotation. The results can be found in Fig. 4, and we observe that the model performance in the target domain is improved with the increase of the annotation budget. We also observe that the 5% annotation budget can be regarded as a good trade-off between the annotation cost and the model adaptability.

4.5. Further Analyses

Module-level Ablation Studies. To validate the effectiveness of the cross-attention module, ablation studies are conducted to gain insight into the cross-attention module. As shown in Table 3, the sampled source data by only the 2D scores tend to boost the accuracy of baseline on 2D metric, while the samples selected only by 3D scores are more beneficial to improve the 3D segmentation results. When the mean sampling strategy (Baseline 3) is utilized, the average accuracy is able to be improved. By employing the designed cross-attention module, satisfactory overall results can be obtained, showing the effectiveness and strength of our cross-attention based sampling strategy.

Segmentation Accuracy for Both Domains. To validate the multi-domain generalization ability of our ADAS, further experiments are conducted to observe the segmentation accuracy of the model on both source and target domains. As shown in Table 4, the source-only baseline, which means that the model is trained only on the source domain and directly tested on the target domain, has the best segmentation accuracy for the source domain, but it is hard to be generalized to the target domain. On the other hand, the xMUDA baseline (UDA model) achieves better results on the target domain but not performs well on the source domain. Our ADAS perform a fine-tuning process on the selected source data (labeled) and the target data (pseudo-labeled or annotated), achieving better generalization on both the source and target domains.

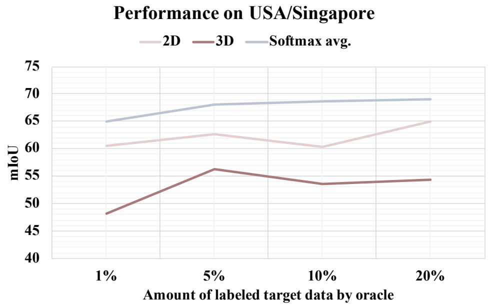

Figure 4. The influence of different target-domain annotation budget $\mathcal{B}$ (20%, 10%, 5%, and 1%) on the ADA task. Here, we take the USA-to-Singapore setting for example.

	Source 2D/3D/Avg.	Target 2D/3D/Avg.	Average 2D/3D/Avg.
Source only	**53.4/46.5/61.3**	31.4/43.4/49.1	42.4/45.0/55.2
xMUDA	36.6/43.8/48.6	55.9/50.1/63.4	46.3/47.0/56.0
ADAS	47.8/43.1/54.9	63.6/52.8/67.4	**55.7/48.0/61.2**

Table 4. Generalization ability of ADAS. The results tested on both source and target domains are reported. Avg. denotes the softmax average performance.

5. Conclusion

In this paper, we have presented an Active-and-Adaptive Segmentation (ADAS) baseline to tackle the domain discrepancies of 3D semantic segmentation model under many DA settings including UDA, UFDA, and ADA tasks. By the merits of the designed cross-modal attention-based feature fusion module, ADAS can fully leverage multi-modal feature interaction to achieve an effective multi-modal sample selection. Experiments are conducted on widely-used cross-domain 3D semantic segmentation scenarios, and show the superiority of ADAS in many DA settings.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 1, 2, 5

[2] Jan-Aike Bolte, Markus Kamp, Antonia Breuer, Silviu Homoceanu, Peter Schlicht, Fabian Huger, Daniel Lipinski, and Tim Fingscheidt. Unsupervised domain adaptation to improve image segmentation quality both in the source and target domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[3] 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. 1, 2, 5

[4] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *International conference on machine learning*, pages 253–261. PMLR, 2013. 2

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7

[6] Yixin Chen, James Ze Wang, and Robert Krovetz. Clue: Cluster-based retrieval of images by unsupervised learning. *IEEE transactions on Image Processing*, 14(8):1187–1201, 2005. 6, 8

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2

[8] Yuhang Ding, Xin Yu, and Yi Yang. Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1246–1254, 2021. 1

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 7

[10] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 2, 5

[11] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 1, 2, 4, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 7

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 6, 7

[14] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14373–14382, 2021. 1

[15] Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15488–15498, 2021. 2

[16] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12605–12614, 2020. 2, 3, 5, 6, 7

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[18] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 5

[19] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 6, 7

[20] Wei Liu, Zhiming Luo, Yuanzheng Cai, Ying Yu, Yang Ke, José Marcato Junior, Wesley Nunes Gonçalves, and Jonathan Li. Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:211–221, 2021. 2, 3, 5, 6, 7

[21] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 6, 7

[22] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pages 415–430. Springer, 2020. 2

[23] A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with do-

main density transformations. *Advances in Neural Information Processing Systems*, 34:5264–5275, 2021. 2

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *Openreview*, 2017. 7

[25] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7108–7117, 2021. 2, 3, 5, 6, 7

[26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[27] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010. 2

[28] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. 2

[29] Inkyu Shin, Dong-Jin Kim, Jae Won Cho, Sanghyun Woo, Kwanyong Park, and In So Kweon. Labor: Labeling only if required for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8588–8598, 2021. 6, 8

[30] Vishwanath A Sindagi and Vishal M Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29:323–335, 2019. 1

[31] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020. 2

[32] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[33] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: a review. *Technologies*, 8(2):35, 2020. 2

[34] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 6, 7

[35] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 6, 7

[36] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10296–10305, 2019. 1

[37] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. Deep parametric continuous convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2589–2597, 2018. 2

[38] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 2

[39] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020. 2

[40] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020. 2

[41] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. Learning with noisy labels for robust point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6443–6452, 2021. 2

[42] Yu-Ting Yen, Chia-Ni Lu, Wei-Chen Chiu, and Yi-Hsuan Tsai. 3d-pl: Domain adaptive depth estimation with 3d-aware pseudo-labeling. In *European Conference on Computer Vision*, pages 710–728. Springer, 2022. 5

[43] Bo Zhang, Jiakang Yuan, Baopu Li, Tao Chen, Jiayuan Fan, and Botian Shi. Learning cross-image object semantic relation in transformer for few-shot fine-grained image classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2135–2144, 2022. 3

[44] Kuangen Zhang, Ming Hao, Jing Wang, Clarence W de Silva, and Chenglong Fu. Linked dynamic graph cnn: Learning on point cloud via linking hierarchical features. *arXiv preprint arXiv:1904.10014*, 2019. 2

[45] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 2

[46] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[47] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2