Omni-Query Active Learning for Source-Free Domain Adaptive Cross-Modality 3D Semantic Segmentation

Jianxiang Xie[1], Yao Wu[1], Yachao Zhang[1], Zhongchao Shi[4], Jianping Fan[4], Yuan Xie[3,*], Yanyun Qu[1,2,*]

[1]School of Informatics, Xiamen University
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
[3]School of Computer Science and Technology, East China Normal University
[4]Lenovo Research
jianxiangxie@stu.xmu.edu.cn, yyqu@xmu.edu.cn

Abstract

Source-Free Domain Adaptation (SFDA) aims to transfer a pre-trained source model to the unlabeled target domain without accessing the source data, thereby effectively solving labeled data dependency and domain shift problems. However, the SFDA setting faces a bottleneck due to the absence of supervisory information. To mitigate this problem, Active Learning (AL) is introduced to combine with SFDA, endeavoring to actively label a small set of the most high-quality target points so that models with satisfactory performance can be obtained at an acceptable cost. Nevertheless, several issues remain unresolved, namely when to query new labels during training, what kind of samples deserve labeling to ensure rich information, and where the labels should be distributed to guarantee diversity. Thus we elaborate OmniQuery to omni-bearing address the "When, What, and Where" problems about active points querying in source-free domain adaptation for cross-modal 3D semantic segmentation. The method consists of three main components: Query Decider, Point Ranker, and Budget Slicer. The Query Decider determines the optimal timing to query new points by fitting the validation curves during training. The Point Ranker nominates points for annotation by calculating the ambiguity of neighboring points in the feature space. The Budget Slicer allocates the annotation quota, i.e., labeling percentage of the point cloud, to different semantic regions by utilizing the advanced 2D semantic segmentation capabilities of the Segment Anything Model (SAM). Extensive experiments demonstrate the effectiveness of our proposed method, achieving up to 99.64% of fully supervised performance with only 3% of labels, and consistently outperforming comparison methods across various scenarios.

Code — https://github.com/Kylin-XJX/ActiveSFDA

Introduction

The development of robotics and autonomous vehicles has brought increasing attention to 3D cross-modality perception (He et al. 2021). Meanwhile, the high cost of data annotation and the unrealistic assumption of independent and identically distributed data in deep learning (Hu et al. 2024; He et al. 2023; Hu et al. 2021) have attracted the interest of researchers to Unsupervised Domain Adaptation (UDA) methods (Wu et al. 2023, 2024a,b; Jaritz et al. 2023). These methods use both labeled source and unlabeled target data to tackle annotation and domain shift issues. However, due to privacy concerns, access to source data is often restricted, leading to Source-Free Domain Adaptation (SFDA), where pre-trained models are transferred to the target domain via self-training without source data access (Liang, Hu, and Feng 2020; Liu, Zhang, and Wang 2021; Wu et al. 2024c), thus their performance visibly lags behind fully supervised methods. Can we exchange a little bit of annotation cost on the target domain for more performance improvement?

Active Learning (AL) is a label-efficient paradigm that allows a model to query the knowledge source for annotation interactively. Deep learning methods typically require large amounts of annotated data, but traditional AL can only provide a small portion of labeled data by the initial model at the beginning, which conflicts with the deep learning process and faces a cold start problem (Ren et al. 2021). In the context of SFDA, the model is pre-trained, holds some inference capability, and has access to a large amount of unlabeled target data, so AL and SFDA are naturally compatible. Since the model receives only limited supervision from the target domain, AL methods must ensure the quality of labels, specifically their high information richness and diversity.

Numerous methods have been proposed to maximize the quality of labels in active source-free domain adaptation. For instance, in 3D segmentation, Annotator (Xie et al. 2023) quantifies the uncertainty of a voxel by computing the entropy of the class distribution among the points it contains. While such voxel-based (Xie et al. 2023) or similar region-based (Wu et al. 2021) selection methods can inevitably lead to inefficient use of label budget as low-information points may also be selected within these base units. Moreover, with a limited receptive field, they focus only on the neighbors of the point. It hinders their ability to effectively interact with points that are semantically related but spatially distant in 3D space, failing to model the uncertainty of points adequately. Another natural dilemma impacting label quality is timing. Querying all samples at the beginning of training (Ning et al. 2021) ensures full supervision throughout the

*Corresponding authors.

whole process, but selected points may be suboptimal due to unrelieved domain differences. An alternate approach is distributing the annotation quota over time, initially annotating a small amount and selecting superior samples later using an improved model (Samet et al. 2023). However, determining other exact times to query new labels during training is intractable. Incorrect query timing can directly affect the quality of annotated points and, ultimately, the model performance. Besides, diversity is a crucial aspect of high-quality (Ren et al. 2021). Existing methods usually query samples in specific regions, preventing the model from receiving plural supervision across various categories and leading to lower performance. Recently the Segment Anything Model (SAM) (Kirillov et al. 2023) exhibits advanced capability in segmentation, utilizing SAM to ensure diversity and then assist AL is promising, but such approaches remain unexplored.

In summary, the problem of how to select informative and diverse points for labeling can be broken down into three essential challenges according to the above analysis: 1) When to query: When should we annotate new points? i.e., allocating the budget across a temporal scale. 2) What to query: What criteria should be followed to select the most informative points and save the budget in the meantime? 3) Where to query: Where to distribute the budget? i.e., allocating the budget across a spatial scale to guarantee diverse supervision. Regarding the first challenge, inspired by previous work (Liu et al. 2022) addressing the "Early Learning" issue (Arpit et al. 2017; Liu et al. 2020), we decide to query new labels when the model starts to overfit noise among pseudo-labels. It is concretely detected by fitting mean Intersection over Union (mIoU) curves calculated with training set validation output and pseudo-labels. As for the second challenge, we consider the point, the finest granularity in a point cloud as the selection unit and calculate metrics in the feature space, in which semantic information is highlighted. Concerning the last challenge, with abundant 2D images, we resort to the powerful 2D semantic segmentation capabilities of the recently proposed vision foundation model, SAM. With box prompts, regions with semantic information are obtained by SAM, then diverse labels can be selected from these regions to provide multifarious supervision.

In this paper, we elaborate a method called OmniQuery to address the problems about active points querying in source-free domain adaptation for cross-modality 3D semantic segmentation. Our method contains three significant components: Query Decider, Point Ranker, and Budget Slicer, these components solve the high-quality label querying problem AL encountered in SFDA from all angles (When, What, and Where). Specifically, Query Decider is proposed to solve the "When to query" problem, by fitting the mIoU curves of the current training phase, Query Decider queries new labels when the change rate in the derivative reaches a certain threshold. Furthermore, to choose the most valuable points while saving budget, we design the Point Ranker, which nominates points for annotation by calculating the ambiguity of neighboring points in the feature space, aiming to solve the "What to query" problem. To ensure the model receives diverse guidance and to address the "Where to query" problem, we present a direct module called Budget Slicer to allocate the current annotation quota across different semantic regions generated by SAM. With these regions, a certain portion of the corresponding points is then selected according to the area of each region.

To sum up, our main contributions are as follows:

- We propose the method called OmniQuery for cross-modality 3D semantic segmentation. To the best of our knowledge, our work is the first to introduce AL methods into this field in the cross-modality scenario.
- We develop the Query Decider to address the less-explored query timing issue, design the Point Ranker to select the most valuable points and build the SAM-based Budget Slicer for label diversity, omni-bearing solving the high-quality label querying problems when incorporating AL to SFDA.
- Extensive experiments demonstrate the effectiveness of our method, with the model consistently outperforming comparison methods across all scenarios.

Related Works

Source-Free Domain Adaptation

With source data often unavailable in UDA settings, SFDA has garnered attention. SFDA methods can be divided into model-based and data-based approaches (Li et al. 2024): model-based methods refine pseudo-labels for better self-training, in contrast, data-based methods reconstruct virtual domains to ease source-free constraints. Hegde et al. introduced an uncertainty-aware mean teacher framework to filter out incorrect pseudo-labels implicitly. Saltori et al. suggested a pseudo-label refinement method based on reversible scale transformation and motion consistency. SUMMIT (Simons et al. 2023) proposed two specific pseudo-label refinement strategies tailored to the extent of domain differences. Kurmi, Subramanian, and Namboodiri presented a framework consisting of a generation module and an adaptation module. Despite these advancements, the performance of these methods generally lags behind fully supervised approaches due to the extremity of the source-free setting.

Active Learning

A key issue in AL methods is the design of selection criteria (Settles 2009). FEAL (Chen et al. 2024) introduces Dirichlet prior distribution to assess data uncertainty in federated learning. In 3D semantic segmentation, Annotator (Xie et al. 2023) uses the class entropy within a large voxel as the selection standard, while ReDAL (Wu et al. 2021) integrates prediction entropy, color discontinuity, and structural complexity into its selection criteria. In the SFDA setting, MHPL (Wang et al. 2023) selects "minimum happy points" characterized by neighbor-chaotic, individual-different, and source dissimilar for annotation. (Wang, Han, and Yin 2024) also proposes a framework with a Reciprocal Active Selection module and Partial Proxy Mixup module, aligning the distributions of inactive and active samples. However, AL methods in cross-modal SFDA remain unexplored. Additionally, equally crucial selection timing and budget allocation strategies have received little attention. Our work fills this gap.

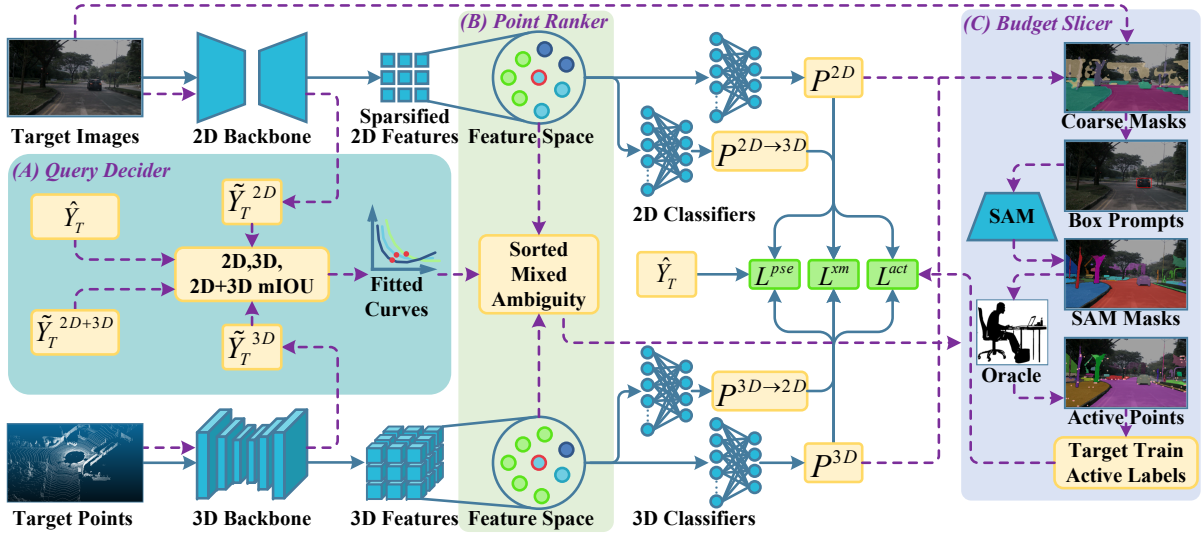

Figure 1: Overview of the workflow of OmniQuery, including Query Decider (Part A), Point Ranker (Part B), and Budget Slicer (Part C) for "When, What, Where" problems about high-quality label query in SFDA for cross-modality 3D semantic segmentation. The solid blue lines indicate the training and inference process and the purple dashed lines show the AL process only activated during training.

Method

Problem Setup

In the SFDA scenario, $\{X_S^{2D}, X_S^{3D}, Y_S\} \in S$ denotes the labeled source data that can not be obtained. We only have access to the source model $\{F_S^{2D}, F_S^{3D}\}$ pre-trained on the source data, and the unlabeled target data $\{X_T^{2D}, X_T^{3D}\} \in T$. $X^{2D} \in \mathbb{R}^{H\times W\times 3}$ stands for image and $X^{3D} \in \mathbb{R}^{N\times 3}$ for corresponding point cloud, $Y_S \in \{1, 2, ..., C\}$ are point-wise labels, where C denotes the number of categories. Note that X^{3D} includes only the points that are visible from the RGB camera and assumes the calibration between LiDAR and the camera is attainable and remains constant over time. OmniQuery aims to leverage the pre-trained source model, unlabeled target data, and a small portion of point annotation $Y_T \in \{1, 2, ..., C, \oslash\}$, $\oslash$ denotes unlabeled, acquired during training to obtain models $\{F_T^{2D}, F_T^{3D}\}$ that perform well in predicting target domain 3D labels.

Overview

The workflow of OmniQuery is illustrated in Figure 1. We begin by initializing the target model F_T using the source model F_S, including the backbone and classifiers for each modality. Subsequently, F_T generates point-wise pseudo-labels $\hat{Y}_T$ for the target domain. There are twice queries in OmniQuery with a total $p\%$ budget, i.e., label percent of the point cloud quantity. Before training, we allocate $b\%$ of the total budget to perform an initial query using the freshly initialized target model, obtaining Y_T. As training progresses, the Query Decider determines when to spend the remaining budget for another query. During training or the early stage of the query, 2D and 3D data from the target domain are fed into the corresponding backbone to obtain 2D and 3D features. For the 2D features, points in the 3D space are projected onto the 2D plane via the calibration matrix of the dataset, and 2D features in corresponding positions are sampled to achieve feature sparsification. Then the sparsified 2D features and the 3D features are passed through the classifiers to obtain point-wise segmentation probabilities for each modality. These segmentation results will be used to compute the loss or to query new 3D annotations.

Periodically, we evaluate the performance on the training set by computing mIoU between model label outputs $\tilde{Y}_T^{2D}, \tilde{Y}_T^{3D}, \tilde{Y}_T^{2D+3D}$ and pseudo-labels. The mIoU of each modality and their combinations are recorded and input to the Query Decider to detect another query timing. During the query process, the Point Ranker calculates mixed feature ambiguity incorporating multi-modal information (2D+3D) for active point selection. The Budget Slicer selects the top total $p\%$ points with the highest ambiguity in each SAM mask and turns to an Oracle for annotation. These few informative annotations are used to supervise the training of each modality. The specifics of the loss function and each module will be discussed in the following sections.

Query Decider

In previous AL methods, the query timing was rather arbitrary. Besides mentioned before, some methods (Ren et al. 2021) query new labels when the training converges. However, these approaches suffer from performance degradation due to the Early Learning problem in the self-training manner where noise is ubiquitous across pseudo-labels. Specifically, models tend to fit the training data with correct pseudo-labels during the early learning stage and eventually memorize instances with noise. By the time current training converges, the model has overfitted to these incorrect pseudo-labels. As illustrated in Figure 2, we analyze the learning process of 2 example classes. The solid curves

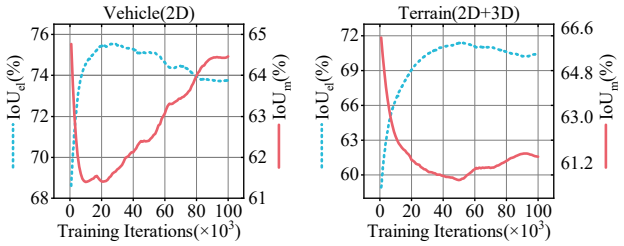

Figure 2: Examples of early learning phenomena for different categories in nuScenes-Lidarseg dataset.

represent the Intersection over Union (IoU) on the training set calculated between model outputs and pseudo-labels (IoU_m), while the dashed curves show the IoU with true labels (IoU_{el}). The two curves exhibit completely opposite trends. Initially, as learning progresses, the true IoU gradually increases while the pseudo-label-based IoU decreases. However, as the model starts to memorize the noise in the pseudo-labels, the pseudo-label-based IoU begins to rise, while the true IoU starts to decline. This observation also provides insight into determining when to query new labels. Queries should be made right before the memorization occurs, that is when the pseudo-label-based IoU curve reaches the bottom and the model desires new annotations.

To tackle this, we design the Query Decider. During the iteration-based training process, we validate the model using the training set for every m iteration and calculate the mIoU values for modality M with pseudo-labels. Here M is either 2D, 3D, or 2D+3D. After collecting n mIoU values for each modality, at current time $t \geq n$, we consider these t mIoU values as y and the indices in the list as x to fit the logarithmic function $f(x) = b\ln(ax+1) + c$ by the least squares method, and denote the derivative as y'. We then check if the change rate of y' relative to its initial value at t_1 for all previous valid points $i \in [n, t]$ exceeds a threshold τ in modality M, if so, the consensus counter C_M increases:

$$|y'_i - y'_{t1}|/|y'_{t1}| > \tau. \tag{1}$$

After checking all modalities, if the consensus is made between any two modalities or within a certain one, that is, $C_{2D} + C_{3D} + C_{2D+3D} \geq 2$, a query is triggered. A detailed algorithm can be found in supplementary material.

Point Ranker

In AL algorithms, the most crucial aspect is selecting samples for annotation. Traditional voxel-based (Xie et al. 2023) or region-based (Wu et al. 2021) query methods inevitably lead to budget wastage and inefficient uncertainty modeling. Therefore, we choose the point as the selection unit and focus on the feature space to address the issue of long-range dependencies in point clouds: insufficient interaction between semantically similar but spatially distant elements. Specifically, during querying, the current model is first employed to infer the ith paired image I_i and point cloud PC_i with N_i points in the target training set, yielding features $f_i^{2D} \in \mathbb{R}^{N_i \times D^{2D}}$, $f_i^{3D} \in \mathbb{R}^{N_i \times D^{3D}}$ and the segmentation probabilities $P_i^{2D}, P_i^{3D} \in \mathbb{R}^{N_i \times C}$. D^{2D} and D^{3D} represent feature dimensions. The segmentation labels are derived from averaged probabilities from the 2D and 3D networks: $\tilde{Y}_{T,i}^{2D+3D} = argmax((P_i^{2D} + P_i^{3D})/2)$. The K nearest neighbors of jth feature are then identified for both modalities i.e., $N_i^M(j)$ and divided into C groups based on the segmentation labels:

$$N_{i,c}^M(j) = \left\{ f_i^M(u) \mid f_i^M(u) \in N_i^M(j), \tilde{Y}_{T,i}^{2D+3D}(u) = c \right\} \tag{2}$$

here M is either $2D$ or $3D$. When the neighbors of a point belong to more categories in the feature space, it becomes more challenging for the classifier to distinguish it. Hence, labeling such points can provide stronger supervision. We calculate the Feature Ambiguity (FA) based on the entropy of the proportion of different neighbor categories:

$$FA_i^M(j) = -\sum_{c=1}^{C} \frac{\left|N_{i,c}^M(j)\right|}{K} \log \frac{\left|N_{i,c}^M(j)\right|}{K}, \tag{3}$$

where $|\cdot|$ denotes the size of the set. After obtaining FA for each modality, the final FA value is derived by summation: $FA_i(j) = FA_i^{2D}(j) + FA_i^{3D}(j)$. The Point Ranker sorts the points by FA value and nominates points with high FA for annotation. Furthermore, the labels $\tilde{Y}_T^{2D+3D}$ predicted by the trained model are of higher quality than the initially generated pseudo-labels $\hat{Y}_T$. Therefore, we replace the old pseudo-labels with the newly predicted labels.

Budget Slicer

After calculating the FA value for each point, directly selecting the top $p'\%$ (budget for current query) of unlabeled points with the highest FA values in a point cloud for annotation would result in annotations being concentrated in a few specific classes, hindering training due to lack of diverse supervision. Hence, we design the Budget Slicer, which utilizes segmentation masks of SAM to divide the image into several semantic regions. By considering the area of each region and the FA values of points within it, we balance diversity while selecting the most valuable points. Since the regions obtained by vanilla SAM or other area segmentation methods like SLIC (Achanta et al. 2012) lack semantic information, they can not fulfill our need for diverse supervision. So, we consider using the prediction results of the point cloud to build prompts for SAM, thereby imparting semantic information to the regions. Concretely, as shown in Figure 1 Part C, we first use the calibration matrix R_{PC2I} retrieved from the datasets to project the 3D points onto the 2D image: $PC_{i,img} = R_{PC2I} PC_i$. Based on the previously obtained segmentation labels $\tilde{Y}_{T,i}^{2D+3D}$, points on the 2D plane are then divided into C sparse masks:

$$mask_i^c = \left\{ PC_{i,img}(u) \mid \tilde{Y}_{T,i}^{2D+3D}(u) = c \right\}. \tag{4}$$

Subsequently, we apply morphological operations such as dilation and erosion to densify these sparse masks, Gaussian Blur (GB) is then used to eliminate the jagged edges of the masks, producing Coarse Masks (CM). We then calculate

the bounding boxes of each disconnected coarse mask and utilize them as box prompts for SAM.

$$\begin{aligned} CM_i &= GB(erosion(dilation(mask_i))), \\ BoxPrompts_i &= BoundingBox(Contours(CM_i)). \end{aligned} \tag{5}$$

By inputting these prompts along with the 2D image into SAM, we obtain delicate segmentation masks. We then traverse these m_i SAM Masks (SM) and select the top $p'\%$ Active Points with the highest FA values within each mask. We then turn to Oracle for annotation, getting updated Y_T.

$$\begin{aligned} ActivePoints_i = \bigcup_{j=1}^{m_i} \{ & PC_i(u) \mid Y_{T,i}(u) = \oslash, \\ & PC_{i,img}(u) \in SM_j, FA_i(u) \geq TOP_j \}, \end{aligned} \tag{6}$$

where TOP_j is the $p'\%$ top percentile of FA_i of unlabeled points whose corresponding $PC_{i,img}$ within SM_j. By slicing the labeling budget to each region, the diversity of selected points is guaranteed.

Overall Loss

The overall loss consists of three parts: active label loss, pseudo-label loss, and cross-modal loss. The cross-entropy-based active label and pseudo-label loss are as follows:

$$L^{seg}(Y, P) = -\frac{1}{N_i \times C} \sum_{i=1}^{N_i} \sum_{c=1}^{C} Y_{i,c} \log P_{i,c}, \tag{7}$$

$$L^{act} = L^{seg}(Y_T, P^{2D}) + L^{seg}(Y_T, P^{3D}), \tag{8}$$

$$L^{pse} = L^{seg}(\hat{Y}_T, P^{2D}) + L^{seg}(\hat{Y}_T, P^{3D}). \tag{9}$$

The double-head cross-modal alignment loss has been proven highly effective for domain adaptation in (Jaritz et al. 2023), so we adopt this loss. The formulation based on Kullback-Leibler (KL) divergence $D_{KL}(\cdot||\cdot)$ is as follows:

$$\begin{aligned} L^{xm} = & D_{KL}(P^{2D} \parallel P^{3D \to 2D}) \\ & + D_{KL}(P^{3D} \parallel P^{2D \to 3D}), \end{aligned} \tag{10}$$

where P^{2D} and P^{3D} are the target distribution from the main prediction during training which is to be estimated by the prediction $P^{3D \to 2D}$ and $P^{2D \to 3D}$ from the auxiliary head in another modality. The final loss is formulated as:

$$L = \alpha L^{act} + \beta L^{pse} + \gamma L^{xm}, \tag{11}$$

with α, β, γ being trade-off factors.

Experiments

Datasets

We utilize four public datasets: nuScenes-Lidarseg(Caesar et al. 2020), VirtualKITTI (Gaidon et al. 2016), SemanticKITTI(Behley et al. 2019), and A2D2(Geyer et al. 2020), to construct four domain adaptation experimental scenarios to validate our method. These scenarios include variations in urban landscapes, such as left-hand vs right-hand driving: USA→Singapore, and changes in lighting conditions: Day→Night, both derived from the nuScenes-Lidarseg dataset. Additionally, they include virtual to real-world adaptation: VirtualKITTI→SemanticKITTI, and differences in resolution and field of view caused by point cloud sensor setups: A2D2→SemanticKITTI. Detailed information is provided in the supplementary material.

Implementation Details

We use a U-Net (Ronneberger, Fischer, and Brox 2015) with ResNet-34 (He et al. 2016) as the basic block for our 2D backbone and the official SparseConvNet (Graham, Engelcke, and van der Maaten 2018) as our 3D backbone. We voxelize the point cloud to a size of 5cm to fit the input format of the 3D backbone. Notably, a size of 5cm ensures each voxel contains only one point as much as possible, and we retain redundant ones, so the point remains the basic selection unit during active querying. We initialize the target model using the source model trained only on the source domain, as published by (Jaritz et al. 2023), to meet the source-free setting. Except for the VirtualKITTI→SemanticKITTI scenario, which runs 30k iterations to prevent overfitting, all other scenarios are trained for 100k iterations. During the active querying phase, we directly use labels from the target training set to simulate Oracle annotations. We experiment under various label budgets p and set the initial percent b of the budget to 50. Except for our method, which uses the Query Decider to determine when to start another query, other methods perform an additional query at the manually set 30% milestone of the training process.

Quantitative and Qualitative Comparison

We compare our method with other AL approaches to verify its effectiveness, including traditional (Wang and Shang 2014) selection strategies such as random selection (Random), selection by point-wise minimum prediction confidence (Confidence), by the entropy of the predicted probabilities (Entropy), and by the margin between the highest and second-highest predicted probabilities (Margin). Additionally, we compare with two AL methods specifically designed for point clouds: Annotator (Xie et al. 2023) and ReDAL (Wu et al. 2021). We extend these methods to multimodal settings, using both 2D and 3D information to calculate the selection metrics to ensure a fair comparison. We report the 2D, 3D, and 2D+3D segmentation mIoU of each model on the test sets across different scenarios.

Results under the 3% budget are shown in Table 1, ST denotes Self-Training on the target domain, while Oracle here represents the upper bound of performance under full supervision of the target. Domain Gap represents the difference between Oracle and ST. As illustrated in the table, OmniQuery outperforms all comparison methods across scenarios. Compared to the baseline ST, only 3% of the labels result in significant improvements, with increases of 11.62%, 6.81%, 30.80%, and 15.29% in the 2D+3D mIoU, respectively. Notably, in the USA→Singapore scenario, OmniQuery achieves 99.64% of the fully supervised performance, nearly merging the domain gap. In the Day→Night scenario, with OmniQuery achieving the best performance, most methods outperform Oracle as well, this may be due to the particularity of the Day→Night scenario, where more supervision actually damages the knowledge in the source model. In the more challenging later two scenarios with larger domain gaps, OmniQuery reaches 92.78% and 91.06% of the fully supervised performance, respectively.

When compared with traditional AL methods, OmniQuery shows improvements of approximately 1.5% to 5%

Methods	USA→Sin.			Day→Night			Vir.K.→Sem.K.			A2D2→Sem.K.		
	2D	3D	2D+3D	2D	3D	2D+3D	2D	3D	2D+3D	2D	3D	2D+3D
ST	61.10	65.15	67.83	49.65	68.59	63.27	27.00	45.17	39.98	33.67	40.34	44.72
Random	68.53	72.72	77.00	58.43	69.15	68.64	47.28	68.00	67.73	44.54	58.61	58.36
Confidence	68.84	72.42	76.95	55.36	68.11	67.63	46.79	63.03	68.19	43.96	56.41	56.50
Entropy	66.77	70.98	75.56	54.45	68.51	67.53	44.27	60.32	65.69	41.41	53.86	54.27
Margin	68.52	72.67	76.96	56.55	68.93	68.63	48.63	64.59	68.73	44.35	57.89	57.36
ReDAL	68.80	72.93	76.87	56.24	68.65	68.48	45.04	66.59	66.45	43.73	59.59	54.96
Annotator	69.99	73.47	77.16	56.80	69.30	68.18	47.49	63.26	68.06	48.81	56.58	57.06
Ours	72.47	75.62	**79.45**	58.07	68.82	**70.08**	53.37	71.64	**70.78**	46.82	60.15	**60.01**
Oracle	75.38	76.24	79.74	60.28	68.64	68.16	59.47	78.11	76.29	55.24	66.51	65.90
Domain Gap	14.28	11.09	11.91	10.63	0.05	4.89	32.47	32.94	36.31	21.57	26.17	21.18

Table 1: Performance comparison of OmniQuery with other AL methods by mIoU(%) under 3% label budget. The abbreviations USA→Sin., Vir.K.→Sem.K., and A2D2→Sem.K. refer to USA→Singapore, VirtualKITTI→SemanticKITTI, and A2D2→SemanticKITTI, respectively. The best and the second best results of 2D+3D are marked in bold and underlined.

Methods	USA→Singapore / 2D + 3D						
	A	B	C	D	E	F	Overall
ST	82.03	94.72	39.29	52.59	60.32	78.01	67.83
Random	85.89	96.25	48.05	65.93	73.27	85.70	75.85
Confidence	86.70	96.06	48.92	65.70	68.90	83.18	74.91
Entropy	85.12	95.59	45.77	63.50	68.23	83.03	73.54
Margin	86.61	96.17	50.27	66.56	71.23	84.78	75.94
ReDAL	85.63	95.81	48.82	64.48	69.94	83.98	74.78
Annotator	84.58	96.52	51.73	65.19	70.27	84.85	75.53
Ours	**89.23**	**96.78**	**54.03**	**69.23**	**76.16**	**87.13**	**78.76**
Oracle	89.67	97.16	55.27	69.46	78.43	88.45	79.74

Table 2: Class-wise IoU(%) in USA→Sin. scenario under 1% label budget, A-F denote class names, i.e., Vehicle, Driveable Surface, Sidewalk, Terrain, Manmade, and Vegetation respectively. Overall represents the mIoU.

due to its advanced selection criteria, more suitable querying timing, and more reasonable allocation strategy. Against point-cloud-specific AL methods like Annotator, OmniQuery achieves performance gains of 2.29%, 1.90%, 2.72%, and 2.95% in the 2D+3D mIoU.

As shown in Table 2, even with a relatively extreme setting of 1% labels, OmniQuery demonstrates strong performance, reaching 98.77% of the upper bound and surpassing all comparison methods across all classes. Notably, for vehicles (Marked as A), that are crucial for autonomous driving, our method shows a significant improvement over the baseline (+7.2%) and comparison methods (+2.53% at least), essentially fulfilling our initial vision that "Models with satisfactory performance can be obtained at an acceptable cost".

We visualize some segmentation results in Figure 5, highlighting vehicle-related classes such as "car" and "bicyclist". Our method consistently achieves more accurate segmentation of these classes across various scenes, whether near or far, demonstrating exceptional performance.

Ablation Study and Analysis

To further verify the effectiveness of each module and explore the impact of different hyperparameters, we conduct in-depth ablation experiments and analysis.

Settings	Modules			USA→Singapore		
	PR	QD	BS	2D	3D	2D+3D
Random	-	-	-	68.53	72.72	77.00
1	✓	-	-	70.02	75.02	78.32
2	✓	✓	-	71.90	75.22	79.02
3	✓	-	✓	71.80	75.48	79.01
4	✓	✓	✓	72.47	75.62	**79.45**

Table 3: Ablation study of each module in OmniQuery.

Effectiveness of Each Module. As shown in Table 3, we incrementally introduce each proposed module on basic random selection. Replacing random selection with the Point Ranker (PR) results in a 1.32% improvement in 2D+3D mIoU. Further incorporating the Query Decider (QD) for querying decisions improves the quality of subsequent active labels and mitigates the early learning problem, leading to an additional 0.7% performance boost. Introducing the SAM-based Budget Slicer (BS) brings a more rational label allocation strategy resulting in a 0.69% increase. Combining all three components leads to significant improvements of 2.45% in 2D+3D mIoU over random selection.

Analysis in Query Decider. As shown in Figure 3 (a), the model achieves optimal performance when threshold τ is set to 0.85. Figure 4 illustrates that automatic queries (red hexagons) occur at about 10% of the training phase for most scenarios. Vir.K.→Sem.K. triggers query at about 30% due to fewer iterations. Although the query timing is similar to passive triggering, the superior selection and label allocation strategy still makes it stand out in comparison with other methods as shown in Table 1. The point plots in different markers show that our method, after fitting the validation curve and triggering queries, experiences a significant performance boost. Compared to the passive triggering in Annotator, Query Decider resolves the dilemma between long-term gains and label quality, resulting in better outcomes.

Impact of K in Point Ranker. Figure 3 (b) shows that the model achieves optimal performance when #neighbors K is set to 4. With too few neighbors, the entropy calculated around the feature space cannot effectively represent the un-

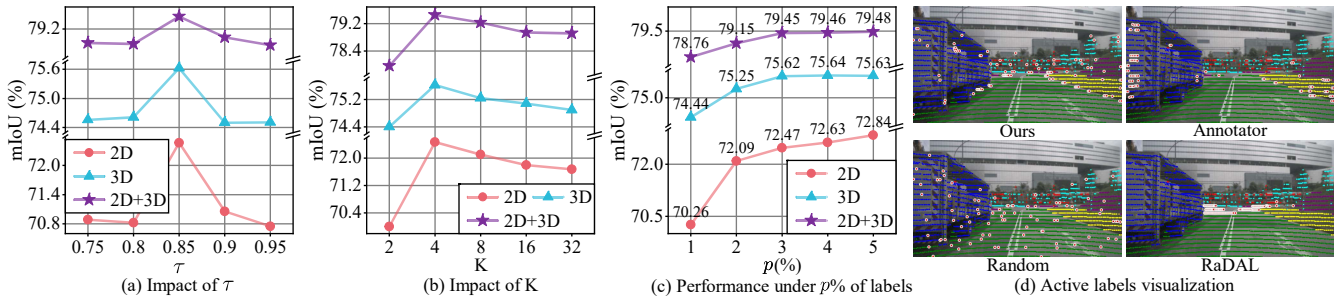

Figure 3: Experiments on hyperparameters and analysis. The red dots in subfigure (d) denote selected labels.

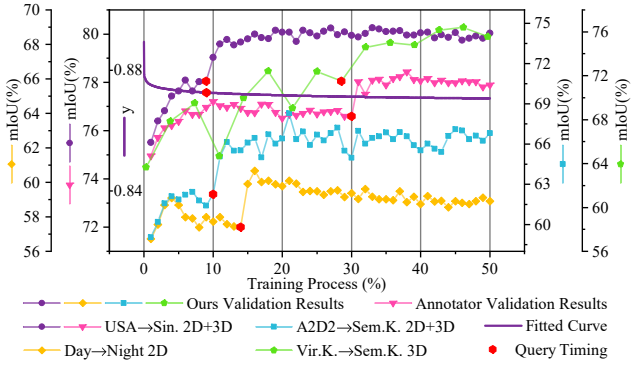

Figure 4: Schematic diagram of the fitted curve, results under the validation set, and the query timings of scenarios.

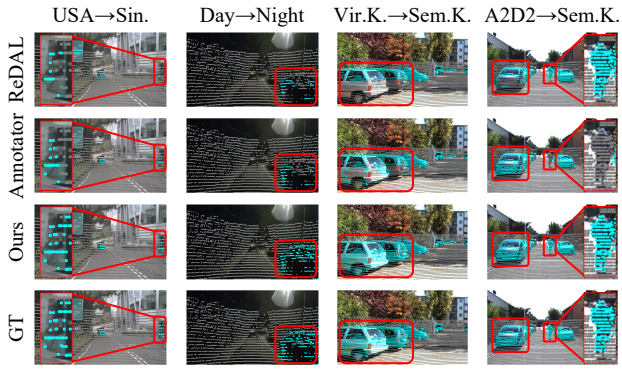

Figure 5: Qualitative Analysis. The abbreviations are the same as in Table 1, GT stands for Ground Truth.

Model	Arch.	USA→Singapore		
		2D	3D	2D+3D
SLIC	-	69.26	72.86	76.91
Vanilla SAM	ViT huge	72.21	75.59	79.11
Prompted SAM	ViT base	72.32	75.34	79.29
Prompted SAM	ViT large	72.31	75.39	79.29
Prompted SAM	ViT huge	72.47	75.62	**79.45**

Table 4: The impact of different 2D region segmentation models on the final performance.

certainty around the query point. As K gradually increases, the entropy of the neighboring classes approaches the entropy of a specific region or even the entire scene. Imagine if K is infinite, all points in the feature space would be included in the entropy calculation, resulting in a fixed value. This would harm the distinctness between different points, leading to a performance decline.

Impact of 2D Segmentation Models in Budget Slicer. Table 4 shows the results of experiments on the 2D segmentation models used in Budget Slicer. When employing the traditional region segmentation algorithm SLIC, or the vanilla SAM, suboptimal results are observed. However, when SAM is prompted with bounding boxes, the performance improves visibly. As the architecture of SAM becomes more complex, the performance gain becomes more pronounced. With the ViT (Dosovitskiy et al. 2021) huge-based SAM, our model achieves optimal performance.

Active Label Analysis. As shown in Figure 3 (c), the performance of the model increases generally with a larger label budget p, demonstrating its stability and ability to exploit the information from active labels. As illustrated in Figure 3 (d), we visualize the distribution of labels selected by different methods on 2D images. It is evident that voxel-based Annotator or region-based ReDAL selection strategies result in more concentrated labels, particularly in the latter one. On the other hand, the Random method yields more evenly distributed labels but is not that informative, easily classifiable road points are extensively selected. Our method balances information richness and diversity, providing the most effective supervision for training with premium active labels.

Conclusion

In this paper, we delve into the issues of source-free active domain adaptation for cross-modal 3D semantic segmentation. We design a method called OmniQuery to address the critical questions of "When, What, and Where" in high-quality active label querying. Specifically, a curve-fitting-based Query Decider for optimal timing, a feature-ambiguity-based Point Ranker for rich information, and a SAM-masks-based Budget Slicer for label diversity are proposed. Numerous experiments prove its effectiveness. We hope this work will provide some inspiration to the active learning and autonomous driving communities.

Acknowledgments

This work is supported by the NSFC (62176224, U2268217, 62176092, 62222602, 62106075, 62476090, 62431004, 62306165); Natural Science Foundation of Shanghai (23ZR1420400); Natural Science Foundation of Chongqing (CSTB2023NSCQ-JQX0007); China Computer Federation Lenovo Blue Ocean Research Fund; China Academy of Railway Sciences (2023YJ357).

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2274–2282.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 233–242. PMLR.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, J.; Ma, B.; Cui, H.; and Xia, Y. 2024. Think Twice Before Selection: Federated Evidential Active Learning for Medical Image Analysis with Domain Shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11439–11449.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Gaidon, A.; Wang, Q.; Cabon, Y.; and Vig, E. 2016. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A. S.; Hauswald, L.; Pham, V. H.; Mühlegg, M.; Dorn, S.; Fernandez, T.; Jänicke, M.; Mirashi, S.; Savani, C.; Sturm, M.; Vorobiov, O.; Oelker, M.; Garreis, S.; and Schuberth, P. 2020. A2D2: Audi Autonomous Driving Dataset. arXiv:2004.06320.

Graham, B.; Engelcke, M.; and van der Maaten, L. 2018. 3D Semantic Segmentation With Submanifold Sparse Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, C.; Li, K.; Zhang, Y.; Tang, L.; Zhang, Y.; Guo, Z.; and Li, X. 2023. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22046–22055.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, Y.; Yu, H.; Liu, X.; Yang, Z.; Sun, W.; Wang, Y.; Fu, Q.; Zou, Y.; and Mian, A. 2021. Deep learning based 3D segmentation: A survey. *arXiv preprint arXiv:2103.05423*.

Hegde, D.; Kilic, V.; Sindagi, V.; Cooper, A. B.; Foster, M.; and Patel, V. M. 2023. Source-free Unsupervised Domain Adaptation for 3D Object Detection in Adverse Weather. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 6973–6980.

Hu, R.; Liu, Y.; Gu, K.; Min, X.; and Zhai, G. 2021. Toward a no-reference quality metric for camera-captured images. *IEEE Transactions on Cybernetics*, 53(6): 3651–3664.

Hu, R.; Zhu, K.; Hou, Z.; Wang, R.; and Liu, F. 2024. Enhanced ADHD detection: Frequency information embedded in a visual-language framework. *Displays*, 83: 102712.

Jaritz, M.; Vu, T.-H.; de Charette, R.; Wirbel, E.; and Perez, P. 2023. Cross-Modal Learning for Domain Adaptation in 3D Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1533–1544.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Kurmi, V. K.; Subramanian, V. K.; and Namboodiri, V. P. 2021. Domain Impression: A Source Data Free Domain Adaptation Method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 615–625.

Li, J.; Yu, Z.; Du, Z.; Zhu, L.; and Shen, H. T. 2024. A Comprehensive Survey on Source-Free Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5743–5762.

Liang, J.; Hu, D.; and Feng, J. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6028–6039. PMLR.

Liu, S.; Liu, K.; Zhu, W.; Shen, Y.; and Fernandez-Granda, C. 2022. Adaptive Early-Learning Correction for Segmentation From Noisy Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2606–2616.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-Learning Regularization Prevents Memorization of Noisy Labels. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in*

Neural Information Processing Systems, volume 33, 20331–20342. Curran Associates, Inc.

Liu, Y.; Zhang, W.; and Wang, J. 2021. Source-Free Domain Adaptation for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1215–1224.

Ning, M.; Lu, D.; Wei, D.; Bian, C.; Yuan, C.; Yu, S.; Ma, K.; and Zheng, Y. 2021. Multi-Anchor Active Domain Adaptation for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9112–9122.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; and Wang, X. 2021. A Survey of Deep Active Learning. *ACM Comput. Surv.*, 54(9).

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N.; Hornegger, J.; Wells, W. M.; and Frangi, A. F., eds., *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. Cham: Springer International Publishing. ISBN 978-3-319-24574-4.

Saltori, C.; Lathuilière, S.; Sebe, N.; Ricci, E.; and Galasso, F. 2020. SF-UDA3D: Source-Free Unsupervised Domain Adaptation for LiDAR-Based 3D Object Detection. In *2020 International Conference on 3D Vision (3DV)*, 771–780.

Samet, N.; Siméoni, O.; Puy, G.; Ponimatkin, G.; Marlet, R.; and Lepetit, V. 2023. You Never Get a Second Chance To Make a Good First Impression: Seeding Active Learning for 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 18445–18457.

Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Simons, C.; Raychaudhuri, D. S.; Ahmed, S. M.; You, S.; Karydis, K.; and Roy-Chowdhury, A. K. 2023. SUMMIT: Source-Free Adaptation of Uni-Modal Models to Multi-Modal Targets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1239–1249.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, 112–119.

Wang, F.; Han, Z.; and Yin, Y. 2024. BIAS: Bridging Inactive and Active Samples for active source free domain adaptation. *Knowledge-Based Systems*, 284: 111151.

Wang, F.; Han, Z.; Zhang, Z.; He, R.; and Yin, Y. 2023. MHPL: Minimum Happy Points Learning for Active Source Free Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20008–20018.

Wu, T.-H.; Liu, Y.-C.; Huang, Y.-K.; Lee, H.-Y.; Su, H.-T.; Huang, P.-C.; and Hsu, W. H. 2021. ReDAL: Region-Based and Diversity-Aware Active Learning for Point Cloud Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15510–15519.

Wu, Y.; Xing, M.; Zhang, Y.; Luo, X.; Xie, Y.; and Qu, Y. 2024a. UniDSeg: Unified Cross-Domain 3D Semantic Segmentation via Visual Foundation Models Prior. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; Fan, J.; Shi, Z.; and Qu, Y. 2023. Cross-modal Unsupervised Domain Adaptation for 3D Semantic Segmentation via Bidirectional Fusion-then-Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 490–498.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; and Qu, Y. 2024b. CLIP2UDA: Making Frozen CLIP Reward Unsupervised Domain Adaptation in 3D Semantic Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 8662–8671.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; Wu, Z.; and Qu, Y. 2024c. Perturbed Progressive Learning for Semisupervised Defect Segmentation. *IEEE Trans. Neural Networks Learn. Syst.*, 35(5): 6118–6132.

Xie, B.; Li, S.; Guo, Q.; Liu, C.; and Cheng, X. 2023. Annotator: A Generic Active Learning Baseline for LiDAR Semantic Segmentation. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 48444–48458. Curran Associates, Inc.