4D Point Cloud Segmentation via Active Test-Time Adaptation

Mingrong Gong[1], Chaoqi Chen[1*], Luyao Tang[2], Yuxi Wang[3], Sergio Escalera[4]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Department of Electrical and Electronic Engineering, The University of Hong Kong
[3]Institute of Artificial Intelligence, Ocean University of China
[4]Department of Mathematics and Informatics, Universitat de Barcelona
{gmr52333,cqchen1994,lytang1999}@gmail.com, yuxi.wang@ouc.edu.cn, sescalera@ub.edu

Abstract

4D point cloud segmentation is crucial for autonomous driving with continuous LiDAR streams. While test-time adaptation (TTA) is the standard approach for handling dynamic environments, current methods suffer from catastrophic error accumulation due to over-reliance on pseudo-labels. Active learning could provide reliable annotations for critical samples, but combining it with TTA faces severe challenges: real-time processing requirements and expensive 3D labeling costs. In this paper, we propose ATTA-4DSeg, the first framework to achieve efficient active test-time adaptation for 4D point cloud segmentation under extreme budget constraints. Our key insight is a self-reinforcing loop: oracle annotations refine adaptation prototypes, which then guide the selection of subsequent high-value samples from regions with severe distribution shifts, maximizing each annotation's impact. Specifically, we propose three key innovations: (1) dual-prototype comparison that precisely localizes distribution shift boundaries to narrow annotation scope, (2) Class-Inverse Budget Allocation (CIBA) ensuring balanced adaptation across all categories, coupled with hybrid uncertainty scoring combining voxel-level geometry and point-wise variance for optimal sample selection, and (3) a refinement strategy leveraging sparse oracle annotations to improve predictions on unlabeled points, maximizing annotation utility. Extensive experiments show ATTA-4DSeg improves mIoU by 18.87%, 19.92%, and 3.6% on three domain adaptation benchmarks using only 1% annotation budget. Our method operates 2.28× faster than state-of-the-art methods. Remarkably, our approach reaches 90% of fully-supervised performance using only 5% annotation budget.

Introduction

Point cloud segmentation is crucial in real-world applications such as autonomous driving (Li et al. 2021), robotics (Chen et al. 2021b) and scene understanding (Wang et al. 2024; Hu et al. 2024). However, models trained on synthetic or curated datasets often fail when deployed due to distribution shifts from environmental changes (Chen et al. 2023). The real-time processing demands of 4D LiDAR streams exacerbate this problem, making offline adaptation methods impractical (Saltori et al. 2022; Zou et al. 2024). A natural direction to address this issue is Test-Time Adaptation (TTA) (Wang et al. 2021; Sivaprasad and Fleuret 2021; Saltori et al. 2022; Zou et al. 2024; Shin et al. 2022; Jiang et al. 2024; Chen, Tang, and Huang 2024), which enables online model adjustment using unlabeled target data. While existing TTA methods attempt to solve this problem, they face critical limitations: entropy-based approaches (Wang et al. 2021; Sivaprasad and Fleuret 2021) often fail due to the inherent sparsity of point clouds, and pseudo-labeling strategies (Saltori et al. 2022; Zou et al. 2024; Jiang et al. 2024) are prone to error accumulation from unreliable predictions.

Crucially, mispredicted samples in pseudo labels often indicate regions where distribution shifts occur and the model struggles most (Wang, Peng, and Zhang 2021; Wang, Liang, and Zhang 2024). Rather than treating all pseudo labels equally, Active Test-Time Adaptation (ATTA) leverages active learning (Ren et al. 2021; Prabhu et al. 2021; Ma, Gao, and Xu 2021; Xie et al. 2022; mathelin et al. 2022) to strategically identify these challenging samples for oracle annotation. By obtaining ground truth labels for the most difficult cases where the model fails, ATTA enables targeted adaptation to distribution shifts. While this approach is theoretically promising, existing ATTA methods like SimATTA (Gui, Li, and Ji 2024) and HILTTA (Li et al. 2024), despite success in 2D applications, fail in 3D segmentation. SimATTA (Gui, Li, and Ji 2024) cannot handle the unordered nature and geometric properties of 3D point clouds, and its incremental clustering algorithm struggles with dense point clouds. HILTTA (Li et al. 2024) employs model selection techniques that are computationally prohibitive for large-scale stream data in autonomous driving scenarios (Behley et al. 2019; Caesar et al. 2020). Even if these technical limitations were addressed, the challenge intensifies in 4D LiDAR streams where both real-time constraints and the inherently expensive nature of 3D annotation severely limit annotation budgets. Despite each frame containing tens of thousands of points, only a minimal fraction ($<1\%$) can be practically annotated. This scarcity makes every annotation precious, as selecting outliers or points with minimal distribution shifts can perform even worse than random selection (Mittal et al. 2019; Munjal et al. 2022). This motivates the critical research question:

How to maximize adaptation effectiveness with minimal annotation budget in large-scale and sparse point cloud segmentation?

In this paper, we propose ATTA-4DSeg, a novel active test-

*Corresponding author.

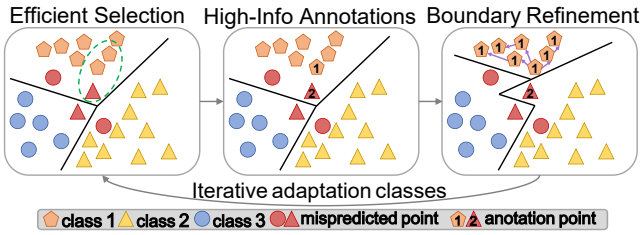

Figure 1: ATTA-4DSeg framework overview.

time adaptation framework specifically designed for 4D point cloud segmentation. Our key innovation is the elegant and efficient integration of active learning and test-time adaptation that creates a self-reinforcing cycle. Specially, ATTA-4DSeg operates through a closed-loop adaptation cycle integrating four key components: First, we leverage dual prototypes to first filter outliers, then precisely locate samples near distribution shift boundaries - regions where shifts are most severe and the model struggles most. Prioritizing annotation of these boundary samples enables rapid model adaptation to difficult cases, maximizing the impact of minimal annotation budgets by focusing resources on the most informative samples that represent unhandled distribution shifts, as shown in the green circle of Fig. 1 *left*. Second, from the previously identified target regions, we leverage point cloud geometric properties and hybrid uncertainty entropy - combining voxel-level semantic uncertainty with point-level prediction variance - to precisely select the most informative points for annotation, effectively handling point cloud irregularities (Fig. 1 *middle*). Finally, we maximize oracle-annotated samples' utility by refining unlabeled data with adaptive prototype updates and conservative pseudo-labeling in stable regions. This significantly expands the pool of reliable labels beyond limited oracle annotations (Fig. 1 *right*), creating a self-reinforcing cycle for continuous adaptation to evolving 4D streams.

Method

Preliminary

Test-time Adaptation (TTA) for point cloud segmentation. Given a segmentation source model F_s pre-trained on source domain $\mathcal{D}s$, TTA initializes a target model Ft and online adapts its parameters θ_t using unlabeled target data during inference to handle distribution shifts. For 4D point cloud segmentation, the target model processes streaming point clouds and must continuously adapt its parameters to maintain segmentation performance across changing environments.

Active Learning strategically selects the most informative samples from an unlabeled pool for oracle annotation to maximize model performance with minimal labeling cost. Given annotation budget $\mathcal{B}$, the goal is to identify samples that provide maximum information gain for model improvement.

Active Test-Time Adaptation (ATTA) combines both paradigms by actively selecting high-value samples from streaming test data for oracle annotation during the adaptation process. Within each point cloud frame, the model iteratively performs active selection for each pseudo-predicted class. Specifically, for the t-th pseudo-predicted class in the current frame, the model selects a subset $\mathcal{A}^{(t)} \subset X^{(t)}$ for annotation subject to budget constraint $|\mathcal{A}^{(t)}| \leq \mathcal{B}^{(t)}$, then updates parameters using both oracle labels and refined pseudo-labels:

$$\theta_t = \arg\min_{\theta_t} \mathcal{L}_{oracle}(\mathcal{A}^{(t)}) + \lambda \mathcal{L}_{pseudo}(\mathcal{R}^{(t)}),$$

where $\mathcal{R}^{(t)}$ represents reliable pseudo-labeled points. This active selection process is conducted once for each pseudo-predicted class, resulting in T total selections per frame, where T corresponds to the number of distinct pseudo-predicted classes present in the current point cloud frame.

Overview

Our ATTA-4DSeg framework operates through three synergistic phases that create a self-reinforcing adaptation cycle:

- **Shift Detection** (Step 1): Efficiently narrows the annotation search space by precisely identifying points at distribution shift boundaries, reducing computational overhead while focusing on the most informative regions.
- **Uncertainty Fusion** (Step 2): Combines voxel-level geometric structure with point-wise uncertainty to select high-quality annotation candidates that boost adaptation effectiveness.
- **Knowledge Propagation** (Step 3): Amplifies the impact of limited annotations by propagating oracle knowledge to refine pseudo-labels in stable regions, creating a multiplier effect for each annotated sample.

The pipeline as shown in Fig. 2. The complete ATTA-4DSeg pipeline integrating shift detection, active selection, and knowledge propagation. Detailed algorithmic procedures are provided in the Appendix.

Prototype-Driven Shift Detection (Step 1)

Let $X = \{x_i\}_{i=1}^N$ be points in the current frame with pseudo-labels $\hat{y}$ from F_t. Following the fixed prototypical boundary strategy (Chen et al. 2021a; Gong et al. 2025), our dual-prototype framework serves two purposes: (1) identifying high-value regions with distribution shift for active annotation, and (2) detecting stable regions suitable for pseudo-labeling. The **source prototype** $\mathbf{q}_k = \boldsymbol{W}_k \in \mathbb{R}^d$ for class k uses the target model's segmentation weights, remaining fixed during testing to preserve source domain boundaries (*i.e.,* the final segmentation head weights will be frozen). The **target prototype** $\mathbf{p}_k^{(t)} \in \mathbb{R}^d$ is computed as $\mathbf{p}_k^{(t)} = \frac{1}{|\mathcal{X}_t|} \sum_{x_i \in \mathcal{X}_t} \mathbf{f}_i$, where $\mathcal{X}_t = \{x_i | \hat{y}_i = t\}$ and $\mathbf{f}_i = F_t(x_i)$. We iterate through all classes sequentially over $T = K$ iterations, updating the target prototype $\mathbf{p}_k^{(t+1)}$ after each iteration using oracle-annotated samples from $\mathcal{D}_{shift}^{(t)}$ (detailed in next Section). For each point x_i with predicted class k, we measure its alignment with both domains through normalized cosine similarities: $s_{q,i} = \frac{\langle \mathbf{f}_i, \mathbf{q}_k \rangle}{\|\mathbf{f}_i\| \|\mathbf{q}_k\|}$ and $s_{p,i} = \frac{\langle \mathbf{f}_i, \mathbf{p}_k^{(t)} \rangle}{\|\mathbf{f}_i\| \|\mathbf{p}_k^{(t)}\|}$. These similarities directly measure distributional shifts, we partition points into two candidate pools:

$$\mathcal{D}_{shift}^{(t)} = \{x_i \mid s_{q,i} \geq \tau_q \wedge s_{p,i} < \tau_p\},$$
$$\mathcal{D}_{stable}^{(t)} = \{x_i \mid s_{q,i} \geq \tau_q \wedge s_{p,i} \geq \tau_p\}. \quad (1)$$

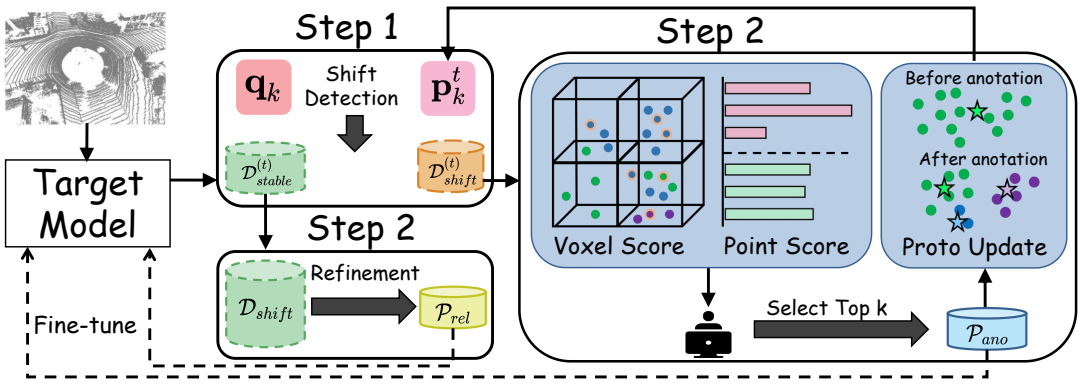

Figure 2: ATTA-4DSeg system architecture. The target model processes point clouds through three phases: (Step 1) Shift detection using dual prototypes, (Step 2) Active update with CIBA budget allocation and uncertainty fusion, (Step 3) Pseudo-label refinement. The framework creates a closed-loop adaptation cycle iteration t times.

The intuition is that points in $\mathcal{D}_{shift}^{(t)}$ maintain source domain characteristics ($s_{q,i} \geq \tau_q$) but deviate from current target distribution ($s_{p,i} < \tau_p$), indicating informative boundary regions where the model struggles most. Points in $\mathcal{D}_{stable}^{(t)}$ remain consistent between source and target domains, representing regions suitable for reliable pseudo-labels. We adaptively set thresholds $\tau_q = Q_\gamma(s_q)$ and $\tau_p = Q_\gamma(s_p)$ using the γ-percentile, where $\gamma = \frac{\langle \mathbf{q}_k, \mathbf{p}_k^{(t)} \rangle}{|\mathbf{q}_k||\mathbf{p}_k^{(t)}|}$ measures source-target prototype alignment. When domains are well-aligned, we use higher percentiles to focus on significant shifts; when poorly aligned, lower percentiles capture broader variations.

Guided Active Update (Step 2)

From the candidate pool $\mathcal{D}_{shift}^{(t)}$, we select a subset of high-value points for oracle annotation through three steps: (1) allocate annotation budget across classes using our Class-Inverse Budget Allocation (CIBA) method, (2) identify the most informative points within each class using a hybrid uncertainty measure, and (3) update the target prototype $\mathbf{p}_k^{(t)}$ with the selected oracle-annotated points.

Class-Inverse Budget Allocation (CIBA). Our CIBA method addresses a critical challenge: minority classes often suffer severe distribution shifts yet receive insufficient attention in uniform sampling. Given a total annotation budget $\mathcal{B} = \lfloor \alpha N \rfloor$ (typically $\alpha = 1\%$), we allocate budget inversely proportional to class frequency:

$$\mathcal{B}_k = \lfloor w_k \mathcal{B} \rfloor, \quad \text{where} \quad w_k = \frac{1/n_k}{\sum_{j=1}^{K} 1/n_j}. \tag{2}$$

Here, $n_k = |\{x_i | \hat{y}_i = k\}|$ denotes the number of points predicted as pseudo class k. This ensures underrepresented classes receive proportionally more annotations. For instance, if vehicle class contains 10,000 points while pedestrian has 1,000, pedestrian receives $10\times$ more budget per point, enabling balanced adaptation across all semantic categories.

Voxel Uncertainty Score. Point-wise uncertainty alone can be noisy due to the sparse and non-uniform nature of 3D point clouds. We therefore aggregate local geometric context through voxelization. We partition the space into voxels $\mathcal{V} = \{v_m\}_{m=1}^{M}$ with size κ, considering only voxels that intersect with $\mathcal{D}_{shift}^{(t)}$. For each voxel v_m, we compute the pseudo class distribution:

$$p_{m,k} = \frac{n_{m,k}}{\sum_{j=1}^{K} n_{m,j}}, \tag{3}$$

where $n_{m,k} = |\{x_i \in v_m : \hat{y}_i = k\}|$. The voxel's semantic uncertainty is measured by normalized entropy:

$$\mathcal{S}_{\text{voxel}}(m) = \frac{H(m)}{\log(C(m))}, \quad H(m) = -\sum_{k=1}^{K} p_{m,k} \log(p_{m,k}). \tag{4}$$

where $C(m) = \sum_{k=1}^{K} \mathbb{I}(n_{m,k} > 0)$ counts the number of unique classes. The normalization factor $\log(C(m))$ ensures fair comparison: a voxel with 2 classes uniformly distributed (uncertainty = 1) is considered more uncertain than a voxel with 3 classes where one dominates.

Point Uncertainty Score. While voxel uncertainty captures local semantic confusion, we also need fine-grained prediction variance. Using Monte Carlo Dropout (Gal and Ghahramani 2016; Saltori et al. 2022), we perform J forward passes with different dropout masks for each point x_i, obtaining prediction vectors $\{p_i^j \in \mathbb{R}^K\}_{j=1}^{J}$:

$$\mathcal{S}_{\text{point}}(i) = \frac{1}{J} \sum_{j=1}^{J} \|p_i^j - \bar{p}_i\|^2, \quad \text{where} \quad \bar{p}_i = \frac{1}{J} \sum_{j=1}^{J} p_i^j \tag{5}$$

Higher variance indicates greater epistemic uncertainty in the model's predictions.

Acquisition Function. For each point $x_i \in \mathcal{D}_{shift}^{(t)}$ with pseudo label $\hat{y} = k$, we combine both uncertainty measures:

$$\mathcal{S}_k(i) = \mathcal{S}_{\text{point}}(i) + \mathcal{S}_{\text{voxel}}(m), \quad \text{where } x_i \in v_m \tag{6}$$

This score leverages both geometric context and prediction confidence. Within pseudo class k, we select the top $\mathcal{B}_k$ points with highest $\mathcal{S}_k(i)$ scores to form the annotation set $\mathcal{P}_k$.

Target Prototype Update. With oracle-annotated points from $\mathcal{P}_k$, we update the target prototype $\mathbf{p}_k^{(t)}$ to incorporate new domain knowledge while maintaining stability. We compute the oracle-guided prototype:

$$\hat{\mathbf{p}}_k = \frac{1}{|\mathcal{P}_k|} \sum_{x_i \in \mathcal{P}_k} \mathbf{f}_i. \tag{7}$$

The updated target prototype balances new annotations with existing knowledge:

$$\mathbf{p}_k^{(t+1)} = \beta \hat{\mathbf{p}}_k + (1-\beta)\mathbf{p}_k^{(t)}, \quad \text{where} \quad \beta = \frac{|\mathcal{P}_k|}{n_k}. \tag{8}$$

Here, n_k represents the total number of points belonging to class k in the target domain. That is, the more annotated points available, the more reliable the updated prototype $\hat{\mathbf{p}}_k^{(t+1)}$ becomes, allowing the model to trust oracle annotations more. This oracle-guided update progressively refines the target prototype representation, improving the accuracy of $\mathcal{D}_{shift}^{(t+1)}$ and $\mathcal{D}_{stable}^{(t+1)}$ partitioning in subsequent iterations.

Oracle-Guided Refinement (Step 3)

After processing all K classes, we leverage oracle annotations to refine pseudo-labels in stable regions, amplifying the impact of our limited annotation budget. We first consolidate the point sets:

$$\mathcal{P}_{oracle} = \bigcup_{k=1}^{K} \mathcal{P}_k, \qquad \mathcal{D}_{stable} = \bigcup_{k=1}^{K} \mathcal{D}_{stable}^{(k)}. \tag{9}$$

Using the voxel grid, we compute local class distributions from oracle-annotated points. For each voxel v_m containing oracle annotations:

$$\mathcal{A}_m = \{x_i \in v_m \cap \mathcal{P}_{oracle}\}, \quad q_{m,k} = \frac{|x_i \in \mathcal{A}_m : y_i = k|}{|\mathcal{A}_m|} \tag{10}$$

For each point $x_j \in \mathcal{D}_{stable}$ in voxel v_m, we compute a hybrid refinement score combining local voxel-level class distribution $q_{m,k}$ with global feature alignment $s_{j,k} = \frac{\langle \mathbf{f}_j, \mathbf{p}_k^{(K)} \rangle}{\|\mathbf{f}_j\| \|\mathbf{p}_k^{(K)}\|}$:

$$\phi_j(k) = q_{m,k} + s_{j,k}, \qquad \hat{y}_j^{ref} = \arg\max_k \phi_j(k). \tag{11}$$

We apply conservative filtering to maintain high confidence:

$$\mathcal{P}_{reliable} = \{x_j \in \mathcal{D}_{stable} : \hat{y}_j^{ref} = \hat{y}_j\} \tag{12}$$

This consistency check ensures we only propagate oracle knowledge to points in stable regions where refinement agrees with original predictions, effectively multiplying the value of each annotation.

Training Objective

Following prior works (Saltori et al. 2022; Zou et al. 2024), we optimize a composite soft dice loss and time consistance loss function:

$$\mathcal{L} = \mathcal{L}_{dice}^{oracle} + \lambda \mathcal{L}_{dice}^{reliable} + \mathcal{L}_{reg} \tag{13}$$

$\mathcal{L}_{dice}^{oracle}$ applies the soft Dice loss to high-quality oracle-annotated points in $\mathcal{P}_{oracle}$, while $\mathcal{L}_{dice}^{reliable}$ applies the same loss to refined pseudo-labeled points in $\mathcal{P}_{reliable}$ with confidence weighting λ. We further incorporate $\mathcal{L}_{reg}$, a temporal consistency regularization term that ensures smooth predictions across consecutive frames. More details are provided in the Appendix.

Experiments

Setup

Dataset: We utilize synthetic **SynLiDAR** (Xiao et al. 2022) and **Synth4D** (Saltori et al. 2022) for training, and real-world **SemanticKITTI** (Behley et al. 2019) and **nuScenes** (Caesar et al. 2020) for validation. Our method addresses (i) synthetic-to-real adaptation (e.g., SynLiDAR $\rightarrow$ SemanticKITTI), and (ii) cross-condition generalization (e.g., weather and sensor variations). Experiments cover diverse environments, point densities, and class distributions. **Evaluation Metric:** Following the standard assessment framework established in GIPSO (Saltori et al. 2022) and HGL (Zou et al. 2024), we evaluate performance using semantic segmentation metrics. We assess the intersection-over-union (IoU) for individual semantic categories as well as the overall mean intersection-over-union (mIoU) across all classes. Results show performance gains relative to the baseline source model to demonstrate adaptation effectiveness.

Implementation Details: We adapt MinkowskiNet (Choy, Gwak, and Savarese 2019) as the backbone for point cloud segmentation and utilize the same training recipe for preparing the source model as GIPSO (Saltori et al. 2022). We use the Adam optimizer with learning rate of 0.0004. Unless specified, all experiments are conducted with voxel size $\kappa = 30$, $\lambda = 0.8$ in Eq. (14), and budget $\alpha = 1\%$ in Eq. (2).

Baseline Methods For test-time adaptation, we compare against classic baselines that have proven successful in 2D applications, including ProDA (Zhang et al. 2021), SHOT (Liang, Hu, and Feng 2020), TENT (Wang et al. 2021), and ConjugatePL (Goyal et al. 2022), as well as state-of-the-art 3D methods GIPSO (Saltori et al. 2022) and HGL (Zou et al. 2024). For active learning, we evaluate against classic methods including Random, Entropy (Wang and Shang 2014), and Margin sampling (Joshi, Porikli, and Papanikolopoulos 2009), along with VCD (Xie et al. 2023) which demonstrates high efficiency in 3D scenarios. Full experimental details are available in the Appendix.

Main Results

We evaluate our ATTA-3DSeg across three domain adaptation scenarios, comparing TTA and AL methods.

SynLiDAR $\rightarrow$ SemanticKITTI. In Table 1, Our method achieves 18.87% average mIoU improvement, substantially outperforming the best TTA method HGL at 6.72% and AL method Random at 12.59%. Traditional TTA methods suffer from catastrophic negative transfer by relying on entropy minimization without distinguishing reliable pseudo-labels from domain-induced noise. Pure AL methods show positive gains but lack domain-specific adaptation. Our +13.27% improvement on Pedestrian class is particularly noteworthy,

Model	Vehicle	Pedestrian	Road	Sidewalk	Terrain	Manmade	Vegetation	Avg.
Source	59.80	14.20	34.90	53.50	31.00	37.40	50.50	40.19
Test-time Adaptation Methods								
ProDA (Zhang et al. 2021)	-53.30	-13.79	-33.83	-52.78	-30.52	-36.68	-49.29	-38.60
SHOT (Liang, Hu, and Feng 2020)	-57.83	-12.64	-24.80	-46.02	-30.80	-36.83	-49.32	-36.89
TENT (Wang et al. 2021)	-0.27	-3.54	+1.63	+1.49	-0.33	+4.96	+4.15	+1.15
ConjugatePL (Goyal et al. 2022)	+4.16	-0.73	+1.82	+1.80	-1.36	+5.27	+4.95	+2.27
GIPSO (Saltori et al. 2022)	+13.95	-6.76	+3.26	+5.01	+3.00	+3.34	+4.08	+3.70
HGL (Zou et al. 2024)	+14.76	+5.66	+1.83	+5.43	+7.33	+5.64	+6.40	+6.72
Active Learning Methods								
Random	+15.58	-11.63	+7.13	+15.99	+23.71	+22.54	+14.82	+12.59
Entropy (Wang and Shang 2014)	+15.86	-11.95	+7.43	+16.27	+25.39	+21.48	+11.49	+12.28
Margin (Joshi, Porikli, and Papanikolopoulos 2009)	+15.40	-11.57	+7.21	+14.47	+22.18	+21.31	+13.24	+11.75
VCD (Xie et al. 2023)	+9.19	-6.84	+4.49	+6.60	+18.16	+15.47	+11.16	+8.32
Ours	**+15.77**	**+13.27**	**+8.30**	**+18.10**	**+28.72**	**+24.38**	**+23.58**	**+18.87**

Table 1: mIoU improvement (%) comparison in **SynLiDAR** $\rightarrow$ **SemanticKITTI** test-time adaptation.

Model	Vehicle	Pedestrian	Road	Sidewalk	Terrain	Manmade	Vegetation	Avg.
Source	63.90	12.60	38.10	47.30	20.20	26.10	43.30	35.93
Test-time Adaptation Methods								
ProDA (Zhang et al. 2021)	-57.77	-12.34	-37.36	-46.95	-19.97	-25.62	-42.48	-34.64
SHOT (Liang, Hu, and Feng 2020)	-62.44	-12.00	-28.27	-40.20	-20.00	-25.47	-42.55	-32.99
TENT (Wang et al. 2021)	+5.40	-0.30	-2.40	-3.95	-0.95	+5.73	+3.42	+0.99
ConjugatePL (Goyal et al. 2022)	+5.93	-0.03	-1.69	-1.86	+1.43	+1.62	+4.98	+1.48
GIPSO (Saltori et al. 2022)	+13.12	-0.54	+1.19	+2.45	+2.78	+5.64	+5.54	+4.31
HGL (Zou et al. 2024)	**+13.24**	+3.84	+0.79	+1.95	+5.27	+10.98	+8.73	+6.40
Active Learning Methods								
Random	+11.14	-7.75	+2.87	+18.67	+28.21	+22.77	+15.10	+13.00
Entropy (Wang and Shang 2014)	+11.06	-9.02	+2.98	+18.78	+28.71	+23.42	+14.69	+12.95
Margin (Joshi, Porikli, and Papanikolopoulos 2009)	+11.35	-8.28	+2.40	+16.52	+26.63	+22.41	+13.91	+12.14
VCD (Xie et al. 2023)	+6.99	+2.45	-0.01	+9.30	+25.04	+26.98	+20.31	+13.01
Ours	+11.50	**+9.23**	**+3.72**	**+20.24**	**+33.95**	**+32.24**	**+28.59**	**+19.92**

Table 2: mIoU improvement (%) comparison in **Synth4D** $\rightarrow$ **SemanticKITTI** test-time adaptation.

Model	Vehicle	Pedestrian	Road	Sidewalk	Terrain	Manmade	Vegetation	Avg.
Source	22.45	14.38	42.03	28.39	15.58	38.18	54.14	30.75
Test-time Adaptation Methods								
ProDA (Zhang et al. 2021)	+0.57	-1.40	+0.73	+0.09	+0.71	+0.40	+0.91	+0.29
SHOT (Liang, Hu, and Feng 2020)	+0.82	-1.77	+0.68	-0.05	-0.70	-0.54	+1.09	-0.07
TENT (Wang et al. 2021)	-0.16	**-0.20**	-1.25	-0.29	+0.02	-0.12	-0.34	-0.34
ConjugatePL (Goyal et al. 2022)	+1.14	-0.41	+0.15	+0.67	+0.41	+0.58	+1.40	+0.57
GIPSO (Saltori et al. 2022)	+0.55	-3.76	+1.64	+1.72	+2.28	+1.18	+2.36	+0.85
HGL (Zou et al. 2024)	+1.42	-2.58	**+5.57**	+2.80	+2.16	+1.02	+2.32	+1.87
Active Learning Methods								
Random	+2.69	-2.45	+2.09	+5.14	+6.45	+2.38	+3.57	+2.88
Entropy (Wang and Shang 2014)	**+2.96**	-3.01	+2.33	+4.93	+6.16	+2.23	+3.72	+2.82
Margin (Joshi, Porikli, and Papanikolopoulos 2009)	+2.76	-3.17	+2.09	+4.98	+6.06	+2.12	+3.37	+2.66
VCD (Xie et al. 2023)	+2.17	-2.23	+1.39	+3.15	+3.29	+0.79	+2.45	+1.62
Ours	+2.01	-1.78	**+3.98**	**+6.20**	**+6.92**	**+2.80**	**+4.93**	**+3.62**

Table 3: mIoU improvement (%) comparison in **Synth4D** $\rightarrow$ **nuScenes** test-time adaptation.

as pedestrians consistently represent the worst performing class across all baselines due to severe class imbalance. This demonstrates our CIBA strategy's effectiveness - by allocating annotation budget inversely proportional to class frequency, we ensure comprehensive coverage of all semantic categories, particularly those that confidence-based filtering approaches struggle with due to insufficient samples.

Synth4D $\rightarrow$ SemanticKITTI. As shown in the Table 2, We achieve 19.92% improvement, representing a 3.1× enhancement over HGL's 6.40%. Traditional AL methods show positive gains but have clear limitations under domain shift. Looking at AL baseline results, they achieve very similar performance, showing that standard uncertainty measures fail to work properly in cross-domain settings. This occurs because domain shift makes uncertainty estimation unreliable: entropy-based methods mistake domain differences for model uncertainty, selecting samples that seem "hard" but are actually just noisy (Karamcheti et al. 2021). Our dual-prototype approach models both domain features directly, allowing it to distinguish truly informative samples from domain-caused false uncertainty.

Synth4D $\rightarrow$ nuScenes. Despite cross-sensor challenges, our method achieves +3.62% improvement, outperforming HGL (+1.87%) and VCD (+13.01%) (See Table 3). The smaller gains reflect compounded challenges from sensor gaps: 64 vs 32-beam differences and varying point densities. Yet our approach maintains resilience on difficult classes, showing our dual-level uncertainty quantification (geometric struc-

ture + point-wise prediction) enables more robust sample selection than baseline single-metric approaches. Our hybrid method delivers 2-3× gains through superior AL score design. Unlike entropy-based methods relying solely on prediction confidence, our multi-scale uncertainty fusion captures geometric consistency and prediction reliability, enabling stable, informative selection at 1% budget.

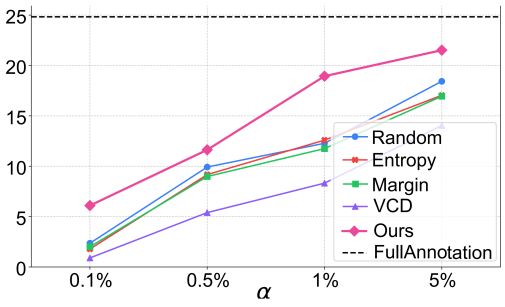

Figure 3: Budget analysis on **Synth4D → SemanticKITTI**

Budget

To evaluate our method's efficiency under varying annotation constraints, we conduct comprehensive budget analysis across different annotation percentages of 0.1%, 0.5%, 1%, and 5% on the Synth4D → SemanticKITTI adaptation task, as illustrated in Fig. 3. Our method consistently outperforms all active learning baselines across all budget levels. At the minimal 0.1% budget, approximately 80 points per frame, our approach achieves 6.2% mIoU. Notably, our method approaches the Full Annotation performance of 22.8% mIoU more rapidly than baselines. At 5% budget, we achieve 22.3% mIoU, reaching 90% of full supervision performance. This rapid convergence validates that our prototype-guided selection identifies the most critical samples early in the annotation process, demonstrating high practicality for real-world deployment where annotation resources are severely limited.

Ablation Study

Selection Strategy. To validate our Class-Inverse Budget Allocation (CIBA), we examine class coverage impact on Synth4D → SemanticKITTI (Fig. 4). Results show monotonic improvement with increasing coverage. Allocating budget solely to the frequent Vehicle class degrades performance (-12.1% avg), boosting Vehicle (+5.2%) but hurting rare classes (e.g., Sidewalk -29.3%). Performance reaches +18.9% mIoU with full 7-class coverage. Diminishing returns at higher levels (6→7: +1.2% vs. 3→4: +2.8%) confirm CIBA efficiently prioritizes critical classes. Thus, inverse-frequency coverage is essential for adaptation, preventing performance drops in underrepresented categories common in frequency-based methods.

Active Learning Score. To validate our hybrid uncertainty measurement, we conduct an ablation study examining individual components of our active learning score (Table 4). The full pipeline (#4) achieves 18.87% mIoU. Without prototype updates (#1), performance drops to 17.12%, highlighting the importance of adaptive refinement across iterations.

#ID	$\mathcal{S}_{\text{voxel}}$	$\mathcal{S}_{\text{point}}$	update $\mathbf{p}_k$	mIoU
1	✓	✓	✗	17.12
2	✗	✓	✓	13.90
3	✓	✗	✓	15.89
4	✓	✓	✓	18.87

Table 4: Ablation study of our pipeline.

Method	GIPSO	HGL	Entropy	VCD	Ours
Inference time (s)	0.96	2.76	0.70	1.79	1.21
Selection Complexity	-	-	$\mathcal{O}(n \log n)$	$\mathcal{O}(n)$	$\mathcal{O}(n + c \log c)$

Table 5: Computational efficiency comparison.

Among individual uncertainty measures, voxel uncertainty (#3) achieves 15.89% mIoU, outperforming point uncertainty (#2) at 13.90%. This 1.99% gap demonstrates that voxel-level scoring better captures spatial context in sparse 3D point clouds. Combining both uncertainties (#4) yields a 2.98% improvement over each individual method, as voxel scoring provides geometric consistency while point scoring captures prediction confidence. The prototype update mechanism contributes +1.75% improvement (comparing #1 vs #4), ensuring selection criteria evolve with the target domain. This ablation confirms that each component addresses different uncertainty aspects necessary for robust sample selection under extreme annotation constraints.

Analysis of Time Efficiency

Table 5 presents the computational efficiency comparison on Synth4D → SemanticKITTI. Our method achieves 1.21s inference time, 2.28× faster than HGL (2.76s) while maintaining superior performance. Compared to pure active learning methods, our approach exhibits moderate computational overhead (+0.58s vs VCD) because we perform sample selection only within designated region candidates, while VCD requires evaluating the entire point cloud of each frame. For selection complexity, traditional AL methods require $\mathcal{O}(nlogn)$ for global sorting. Our method achieves $\mathcal{O}(n+clogc)$, where c ≪ n represents shift region candidates. The $\mathcal{O}(n)$ handles prototype similarity for all points, while $\mathcal{O}(clogc)$ covers sorting within the reduced set. Since shift regions typically contain <10% of points, the overall complexity becomes around $\mathcal{O}(n)$. This efficiency makes our approach practical for real-time 4D LiDAR task with low latency constraints.

Visualization

Fig. 5 shows qualitative results on SemanticKITTI. Our method achieves more accurate segmentation, particularly for challenging minority classes. In the circled regions, GIPSO and HGL exhibit significant misclassifications and boundary confusion, while our approach produces cleaner class boundaries closer to ground truth. This visual improvement demonstrates how our dual-prototype mechanism and CIBA strategy effectively handle distribution shifts and class imbalance in real-world scenarios.

Related Work

Test-time Adaptation (TTA). TTA are designed to mitigate distribution discrepancies between source and target domain

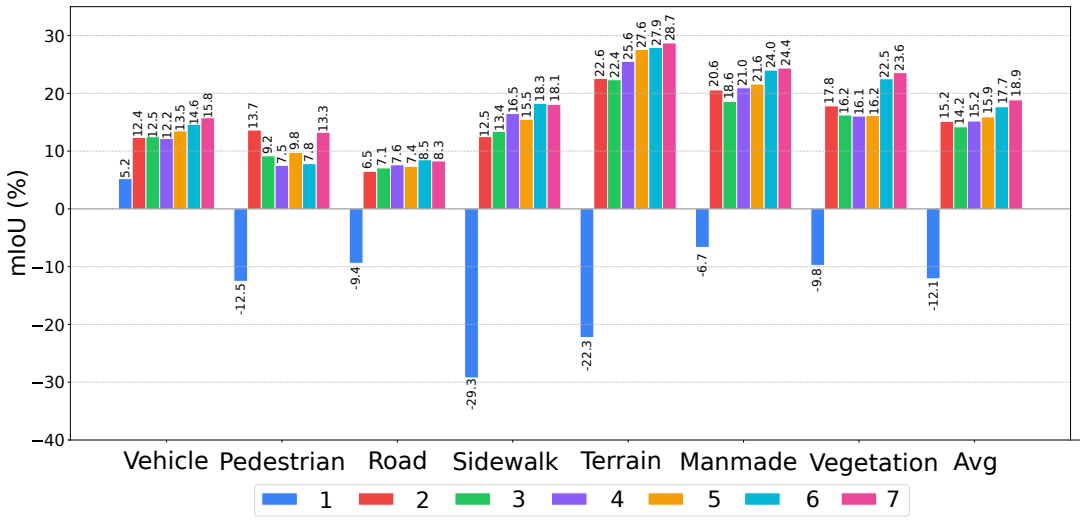

Figure 4: Class coverage analysis on **SynLiDAR** → **SemanticKITTI**.

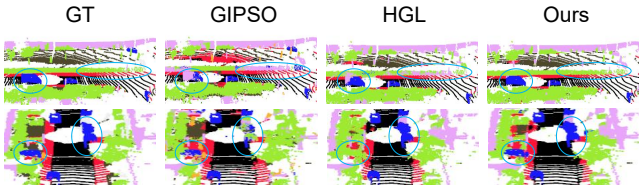

Figure 5: Qualitative results on SemanticKITTI.

datasets. For two-dimensional image processing, TTA frameworks employ entropy-based optimization (Wang et al. 2021), pseudo-label generation (Wang et al. 2022), or augmentation invariance principles (Sun et al. 2020) to refine network weights across diverse 2D applications. Despite achieving considerable success in 2D visual tasks, the direct application of these methodologies to 3D scenarios frequently results in subpar performance, primarily due to inadequate consideration of the distinctive geometric structural properties inherent in 3D point clouds (Shin et al. 2022). CloudFixer (Zhang et al. 2021) achieves model adaptation by optimizing geometric transformation parameters. GIPSO (Saltori et al. 2022) employs geometric and temporal information for pseudo-label propagation to mitigate domain shift. HGL (Zou et al. 2024) utilizes a hierarchical geometric structure in point-cloud streams and a temporal consistency regularization module to adapt the pre-trained model. CoMAC (Cao et al. 2023) leverages class-Wise momentum queues prevent catastrophic forgetting. MOS (Chen et al. 2025) introduced dynamic checkpoint selection for LiDAR-based detection. While these methods achieve adaptation through self-supervised learning with pseudo-labels, they suffer from error accumulation due to unreliable pseudo-labels in regions with severe distribution shifts. To address this limitation, our work integrates active learning with TTA to strategically incorporate oracle annotations for enhanced model adaptability.

Active Learning. The high cost of 3D annotation drove innovation in active learning strategies that achieve near-full performance with minimal labeled data (Luo et al. 2023). HPAL (Xu et al. 2023) introduced point-level selection with hierarchical context. VCD (Xie et al. 2023) introduces a voxel-centric active learning baseline that leverages voxel confusion degree to exploit local topology relations and point cloud structures. These active learning methods have made significant contributions to the 3D domain, but most of them cannot meet the requirements of TTA methods due to real-time constraints.

Active Test-time Adaptation (ATTA). ATTA is a powerful synthesis of active learning and test-time adaptation, providing theoretical guarantees while maintaining practical efficiency. SimATTA (Gui, Li, and Ji 2024) provided the first formal ATTA framework with learning theory guarantees. Their real-time sample selection uses entropy balancing to prevent catastrophic forgetting while incorporating limited labeled test instances. EATTA (Wang and Ding 2025) minimized annotation burden to one sample per batch. Their border sample identification uses feature perturbation to find samples feasible for single-step adaptation. In this work, we mainly focus on how to improve the utilization of oracle annotations under low budget constraints on 3D applications.

Conclusion

In this paper, we introduce ATTA-4DSeg, a novel active test-time adaptation framework for 4D point cloud segmentation under severe annotation constraints. By efficiently identifying informative samples at distribution boundaies while maintaining balanced adaptation across all classes, ATTA-4DSeg achieves effective adaptation with minimal annotation budget. Extensive experiments demonstrate significant performance improvements over baselines while maintaining computational efficiency, validating its effectiveness for real-world LiDAR deployment. This work opens promising directions for extending active adaptation to broader 3D perception tasks and autonomous driving scenarios.

Acknowledgements

This work was supported in parts by NSFC (U21B2023), ICFCRT (W2441020), Guangdong Basic and Applied Basic Research Foundation (2023B1515120026), and Scientific Development Funds from Shenzhen University; and in part by the Spanish project PID2022-136436NB-I00 and by ICREA under the ICREA Academia programme.

References

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 *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 *CVPR*.

Cao, H.; Xu, Y.; Yang, J.; Yin, P.; Yuan, S.; and Xie, L. 2023. Multi-Modal Continual Test-Time Adaptation for 3D Semantic Segmentation. In *ICCV*.

Chen, C.; Tang, L.; and Huang, H. 2024. Reconstruct and Match: Out-of-Distribution Robustness via Topological Homogeneity. In *NeurIPS*.

Chen, C.; Zheng, Z.; Huang, Y.; Ding, X.; and Yu, Y. 2021a. I3Net: Implicit Instance-Invariant Network for Adapting One-Stage Object Detectors. In *CVPR*.

Chen, S.; Xu, H.; Li, R.; Liu, G.; Fu, C.; and Liu, S. 2023. SIRA-PCR: Sim-to-Real Adaptation for 3D Point Cloud Registration. In *ICCV*, 14348–14359.

Chen, X.; Li, S.; Mersch, B.; Wiesmann, L.; Gall, J.; Behley, J.; and Stachniss, C. 2021b. Moving Object Segmentation in 3D LiDAR Data: A Learning-Based Approach Exploiting Sequential Data. *IEEE Robotics Autom. Lett.*, 6(4): 6529–6536.

Chen, Z.; Meng, J.; Baktashmotlagh, M.; Zhang, Y.; Huang, Z.; and Luo, Y. 2025. MOS: Model Synergy for Test-Time Adaptation on LiDAR-Based 3D Object Detection. In *ICLR*.

Choy, C. B.; Gwak, J.; and Savarese, S. 2019. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *CVPR*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *ICML*.

Gong, M.; Chen, C.; Sun, Q.; Wang, Y.; and Huang, H. 2025. Out-of-Distribution Detection with Prototypical Outlier Proxy. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI*.

Goyal, S.; Sun, M.; Raghunathan, A.; and Kolter, J. Z. 2022. Test Time Adaptation via Conjugate Pseudo-labels. In *Neurips*.

Gui, S.; Li, X.; and Ji, S. 2024. Active Test-Time Adaptation: Theoretical Analyses and An Algorithm. In *ICLR*.

Hu, X.; Wang, Y.; Fan, L.; Fan, J.; Peng, J.; Lei, Z.; Li, Q.; and Zhang, Z. 2024. Semantic Anything in 3D Gaussians. *arXiv preprint arXiv:2401.17857*.

Jiang, J.; Zhou, Q.; Li, Y.; Zhao, X.; Wang, M.; Ma, L.; Chang, J.; Zhang, J. J.; and Lu, X. 2024. PCoTTA: Continual Test-Time Adaptation for Multi-Task Point Cloud Understanding. In *NeurIPS*.

Joshi, A. J.; Porikli, F.; and Papanikolopoulos, N. 2009. Multi-class active learning for image classification. In *CVPR*.

Karamcheti, S.; Krishna, R.; Fei-Fei, L.; and Manning, C. D. 2021. Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering. In *ACL*.

Li, Y.; Ma, L.; Zhong, Z.; Liu, F.; Chapman, M. A.; Cao, D.; and Li, J. 2021. Deep Learning for LiDAR Point Clouds in Autonomous Driving: A Review. *IEEE Trans. Neural Networks Learn. Syst.*, 32(8): 3412–3432.

Li, Y.; Su, Y.; Yang, X.; Jia, K.; and Xu, X. 2024. Exploring Human-in-the-Loop Test-Time Adaptation by Synergizing Active Learning and Model Selection. *Transactions on Machine Learning Research*.

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 *ICML*.

Luo, Y.; Chen, Z.; Wang, Z.; Yu, X.; Huang, Z.; and Baktashmotlagh, M. 2023. Exploring Active 3D Object Detection from a Generalization Perspective. In *ICLR*.

Ma, X.; Gao, J.; and Xu, C. 2021. Active universal domain adaptation. In *ICCV*.

mathelin, A. D.; Deheeger, F.; MOUGEOT, M.; and Vayatis, N. 2022. Discrepancy-Based Active Learning for Domain Adaptation. In *ICLR*.

Mittal, S.; Tatarchenko, M.; Çiçek, Ö.; and Brox, T. 2019. Parting with illusions about deep active learning. *arXiv preprint arXiv:1912.05361*.

Munjal, P.; Hayat, N.; Hayat, M.; Sourati, J.; and Khan, S. 2022. Towards Robust and Reproducible Active Learning using Neural Networks. In *CVPR*.

Prabhu, V.; Chandrasekaran, A.; Saenko, K.; and Hoffman, J. 2021. Active domain adaptation via clustering uncertainty-weighted embeddings. In *ICCV*.

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 computing surveys (CSUR)*, 54(9): 1–40.

Saltori, C.; Krivosheev, E.; Lathuilière, S.; Sebe, N.; Galasso, F.; Fiameni, G.; Ricci, E.; and Poiesi, F. 2022. GIPSO: Geometrically Informed Propagation for Online Adaptation in 3D LiDAR Segmentation. In *ECCV*.

Shin, I.; Tsai, Y.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K. 2022. MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation. In *CVPR*.

Sivaprasad, P. T.; and Fleuret, F. 2021. Uncertainty Reduction for Model Adaptation in Semantic Segmentation. In *CVPR*, 9613–9623.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A. A.; and Hardt, M. 2020. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *ICML*.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. In *IJCNN*.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B. A.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*.

Wang, G.; and Ding, C. 2025. Effortless Active Labeling for Long-Term Test-Time Adaptation. In *CVPR*.

Wang, P.; Wang, Y.; Li, S.; Zhang, Z.; Lei, Z.; and Zhang, L. 2024. Open Vocabulary 3D Scene Understanding via Geometry Guided Self-Distillation. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV*.

Wang, Q.; Fink, O.; Gool, L. V.; and Dai, D. 2022. Continual Test-Time Domain Adaptation. In *CVPR*.

Wang, Y.; Liang, J.; and Zhang, Z. 2024. A Curriculum-Style Self-Training Approach for Source-Free Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12): 9890–9907.

Wang, Y.; Peng, J.; and Zhang, Z. 2021. Uncertainty-aware Pseudo Label Refinery for Domain Adaptive Semantic Segmentation. In *ICCV*.

Xiao, A.; Huang, J.; Guan, D.; Zhan, F.; and Lu, S. 2022. Transfer Learning from Synthetic to Real LiDAR Point Cloud for Semantic Segmentation. In *AAAI*.

Xie, B.; Li, S.; Guo, Q.; Liu, C. H.; and Cheng, X. 2023. Annotator: A Generic Active Learning Baseline for LiDAR Semantic Segmentation. In *NeurIPS*.

Xie, B.; Yuan, L.; Li, S.; Liu, C. H.; Cheng, X.; and Wang, G. 2022. Active Learning for Domain Adaptation: An Energy-Based Approach. In *AAAI*.

Xu, Z.; Yuan, B.; Zhao, S.; Zhang, Q.; and Gao, X. 2023. Hierarchical Point-based Active Learning for Semi-supervised Point Cloud Semantic Segmentation. In *ICCV*.

Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; and Wen, F. 2021. Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation. In *CVPR*.

Zou, T.; Qu, S.; Li, Z.; Knoll, A.; He, L.; Chen, G.; and Jiang, C. 2024. HGL: Hierarchical Geometry Learning for Test-Time Adaptation in 3D Point Cloud Segmentation. In *ECCV*.