Density Matters: Improved Core-Set for Active Domain Adaptive Segmentation

Shizhan Liu[1*], Zhengkai Jiang[2*], Yuxi Li[2*], Jinlong Peng[2], Yabiao Wang[2†], Weiyao Lin[1†]

[1]Shanghai Jiao Tong University
[2]Tencent Youtu Lab
shanluzuode@sjtu.edu.cn, zhengkjiang@tencent.com, cookieyxli@gmail.com,
jeromepeng@tencent.com, caseywang@tencent.com, wylin@sjtu.edu.cn

Abstract

Active domain adaptation has emerged as a solution to balance the expensive annotation cost and the performance of trained models in semantic segmentation. However, existing works usually ignore the correlation between selected samples and its local context in feature space, which leads to inferior usage of annotation budgets. In this work, we revisit the theoretical bound of the classical Core-set method and identify that the performance is closely related to the local sample distribution around selected samples. To estimate the density of local samples efficiently, we introduce a local proxy estimator with Dynamic Masked Convolution and develop a Density-aware Greedy algorithm to optimize the bound. Extensive experiments demonstrate the superiority of our approach. Moreover, with very few labels, our scheme achieves comparable performance to the fully supervised counterpart.

Introduction

Semantic segmentation has become increasingly crucial in various applications, such as autonomous driving (Teichmann et al. 2018), virtual try-on (Ayush et al. 2019), and smart healthcare (Shi et al. 2020). In recent years, remarkable progress has been made in this field (Chen et al. 2017; He et al. 2017; Chen et al. 2018; Cheng, Schwing, and Kirillov 2021). However, annotating each pixel in an image is extremely costly. Thus, domain adaptation techniques have been introduced to overcome the high cost of annotation (Chang et al. 2019; Cheng et al. 2021; Zhang et al. 2021; Liu et al. 2021). Unfortunately, the performance of unsupervised domain adaptation methods still lags far behind that of fully supervised training on the target domain. Therefore, active domain adaptation has emerged as a solution to balance the expensive annotation cost and the performance of the trained segmentation model, which involves labeling a few additional samples from the target domain to help transfer knowledge from the source domain to the target domain.

Existing active domain adaptation methods for semantic segmentation primarily rely on either modeling uncertainty (Prabhu et al. 2021; Shin et al. 2021; Xie et al. 2022a) or data diversity (Ning et al. 2021; Xie et al. 2022a; Wu et al. 2022) as metrics to select samples for annotation. They select regions with either higher uncertainty or larger differences from the source domain. However, these methods usually ignore the fact that selected samples can be highly correlated with their local context (e.g. local sample density and adjacent structures) in the feature space, which leads to inevitable redundancy within annotation budget.

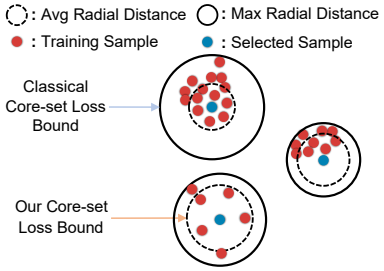

Figure 1: Compared with the classical Core-set loss bound (Sener and Savarese 2017), we find the 'Average Radial Distance' in feature space is a tighter bound, which is highly correlated to local sample distribution.

One potential solution to the aforementioned problem is the Core-set approach, which uses a small set to approximate a large set (Sener and Savarese 2017; Kim and Shin 2022). However, applying the Core-set approach to semantic segmentation with large amount of pixel-level candidates confronts two issues. *Firstly*, the classical solution (Wolf 2011) to Core-set problems is essentially minimizing the radius required to cover extreme points in neighborhood (as shown in Fig. 1). When the number of candidate points becomes larger, extreme points can not effectively reflect the local properties in feature space. *Secondly*, in the classical greedy solution of Core-set, the discrepancy between data points is evaluated equally, regardless of the local context of each candidate. In contrast, the representative ability of samples may vary across different positions in the feature space, hence the discrepancy measurement should also depend on different data points.

To address these challenges, we first revisit the theoretical bound of classical Core-set loss (Sener and Savarese 2017) and derive a new upper bound from the perspective of expectation. This new bound indicates the Core-set per-

[*]These authors contributed equally.
[†]Corresponding author.

formance is closely related to the conditional distribution of samples covered by its nearest selected data. More concretely, through intuitive analysis and empirical observation, we have discovered that the Core-set loss bound is scaled by the average distance from a sample to its closest selected points (as shown in Fig. 1), which indicates that our selection strategy should allocate more label budget to samples with larger and more diverse coverage area. Consequently, we propose a Density-aware Core-set Selection method for domain adaptive semantic segmentation, which takes local sample distribution of candidates into account for domain adaptive semantic segmentation.

To estimate the average distance from a training sample to its nearest selected points, we draw inspiration from VAE (Kingma and Welling 2013) and the context model of learned image compression (Minnen, Ballé, and Toderici 2018), and design a fast local proxy estimator equipped with Dynamic Masked Convolution. This estimator estimates a statistic we call 'coverage density' based on each candidate's neighbors, which is negatively related to the average distance. Additionally, we also develop a Density-aware Greedy algorithm that considers both data discrepancy and estimated coverage density when selecting samples, and minimizes the average radial distance within a defined coverage area of selected samples.

Overall, our contributions can be summarized as follows:

- We derive a new upper bound for the Core-set approach based on an expectation perspective. Our analysis reveals that the average distances within a defined coverage area of selected points are crucial for the performance, which is highly correlated to the conditional probability density within the area.
- We propose a Density-aware Core-set Selection method to optimize the derived upper bound. We applies a proxy estimator equipped with Dynamic Mask Convolution for fast estimation of local density and average distance, the estimated density is then utilized in our proposed Density-aware Greedy algorithm for data sampling.
- We conduct experiments on two representative domain adaptation benchmarks, namely GTAV $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes, achieving performance that surpasses current state-of-the-art methods.

Related Works

Domain Adaptation

Domain adaptation aims to transfer knowledge from a source domain with sufficient labels to a related target domain with little or no labeled data. Depending on the availability of labels in the target domain, domain adaptation can be classified into unsupervised domain adaptation (UDA) and weakly-supervised domain adaptation (WDA). Lately, UDA methods including HRDA (Hoyer, Dai, and Van Gool 2022) and MIC (Hoyer et al. 2023) have achieved promising performance by designing efficient self-training strategies (Wang, Peng, and Zhang 2021; Zheng and Yang 2021; Jiang et al. 2022). As a complement to these works, this paper focuses on the WDA setting, which balances model performance and annotation cost through the use of weak labels such as image-level annotations (Paul et al. 2020) or a limited number of pixel-wise labels (Chen et al. 2021; Guan and Yuan 2023).

Active Domain Adaptation Segmentation

Active learning is a powerful technique that improves model performance with a fixed labeling budget by selecting valuable samples for labeling in multiple rounds. Many active learning works focus on image classification (Wang et al. 2016; Xie et al. 2022b; Kirsch, Van Amersfoort, and Gal 2019). However, these methods are often not applicable to semantic segmentation with numerous candidate samples. For example, BADGE (Ash et al. 2019) and BAIT (Ash et al. 2021) incorporates last-layer gradients or Fisher matrices for active selection, but computing them for each pixel or image patch is computationally and storage-intensive.

Active domain adaptation methods have been developed for semantic segmentation to address the high cost of annotation and the performance of trained models. They typically employ unsupervised domain adaptation to initialize a model and subsequently choose samples in the target domain using indicators such as uncertainty (Shin et al. 2021) or diversity (Ning et al. 2021). Some recent works have integrated uncertainty and diversity measures. RIPU (Xie et al. 2022a) uses the entropy of the model output to quantify uncertainty and measures diversity based on the number of classes predicted by the model within a fixed neighborhood. D2ADA (Wu et al. 2022) utilizes the KL divergence of the feature distribution between the target and source domains to measure diversity. Nevertheless, these methods do not consider that selecting highly similar samples may result in a wasted labeling budget.

Core-set Approach

The Core-set selects a subset of data that approximates the entire dataset by choosing samples that cover the entire training set with the smallest possible radius, thereby enhancing the diversity of selected samples. Sener et al. (Sener and Savarese 2017) extended this approach to convolutional neural networks and developed a Robust k-Center algorithm to improve its optimality. Kim et al. (Kim and Shin 2022) first clustered samples in the training set according to the estimated nearest neighbor distance and then performed active selection in each cluster using the K-Center Greedy algorithm for image classification. However, this approach still relies on extreme data points in selection, ignoring the relationship between data discrepancy and local context.

In this paper, we derived a tighter upper bound for the Core-set loss and optimized it by assigning varying coverage radii to different samples in the selected set. To the best of our knowledge, this is the first time that the Core-set approach has been applied to semantic segmentation.

Problem Definition with a Tighter Core-set Upper Bound

First, we provide a brief overview of the optimization target of the Core-set approach and introduce some notation.

The data space is denoted as $\mathcal{X}$ and the label space is denoted as $\mathcal{Y} = \{1, ..., C\}$, where C represents the number of classes for semantic segmentation. The training set comprises independently and identically distributed (i.i.d.) samples from the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. It is represented as $\{\mathbf{x}_t, y_t\}_{t\in[n]} \sim p_{\mathcal{Z}}$, where n is the size of the training set and $[n]$ is the set of subscripts $\{1, 2, \cdots, n\}$. The Core-set approach aims to select a small subset of labeled samples $\mathbf{s} \subset [n]$ that minimizes the Core-set loss:

$$\min_{\mathbf{s}} \mathcal{L}(\mathbf{s}) = |\frac{1}{n}\sum_{t=1}^{n} l(\mathbf{x}_t, y_t; A_{\mathbf{s}}) - \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} l(\mathbf{x}_k, y_k; A_{\mathbf{s}})|, \quad \textbf{s.t.} \quad |\mathbf{s}| = b, \tag{1}$$

where b is the labeling budget, $A_{\mathbf{s}}$ represents learning algorithm which fits the class-wise distribution $\eta_c(\mathbf{x}) = p(y = c|\mathbf{x})$ for each class given labeled subset $\mathbf{s}$, and $l(\cdot, \cdot, A_s)$ denotes a bounded non-negative loss function. Intuitively, Eq. 1 aims at finding the subset $\mathbf{s}$ such that the performance of the model on $\mathbf{s}$ is close to that on all sampled data.

Previous research (Sener and Savarese 2017) demonstrated that, under appropriate assumptions, the objective of Eq. 1 is bounded by the radius of balls determined by extreme points in data space.

Theorem 1. *Classical Core-set loss bound (Sener and Savarese 2017): Given selected set s, if $l(\cdot, y, A_{\mathbf{s}})$ is λ^l-Lipschitz continuous and bounded by L, $\eta_c(\mathbf{x})$ is λ^η-Lipschitz continuous, and $l(\mathbf{x}_k, y_k, A_{\mathbf{s}}) = 0, \forall k \in \mathbf{s}$, then there exists a radius $\delta = \max_{\mathbf{x}\in\mathcal{X}} \min_{k\in\mathbf{s}} |\mathbf{x} - \mathbf{x}_k|$ such that, with probability no less than $1 - \gamma$,*

$$\mathcal{L}(\mathbf{s}) \leq \delta(\lambda^l + \lambda^\eta LC) + \sqrt{\frac{L^2 log(1/\gamma)}{2n}}. \tag{2}$$

According to Eq. 2, minimizing the upper bound of the Core-set loss requires finding a subset $\mathbf{s}$ with the minimum required radius to cover all other samples, which in turn minimizes δ. One approach is to use the k-Center Greedy algorithm (Wolf 2011) to obtain a sub-optimal solution. Alternatively, the radius δ can be reformulated as follows:

$$\delta = \max_{k} \max_{\mathbf{x}\in\mathbb{N}_{\mathbf{s}}(k)} |\mathbf{x} - \mathbf{x}_k|,$$
$$\mathbb{N}_{\mathbf{s}}(k) = \{\mathbf{x}|\mathbf{x} \in \mathcal{X} \wedge \arg\min_{m\in\mathbf{s}} |\mathbf{x} - \mathbf{x}_m| = k\}, \tag{3}$$

where we define $\mathbb{N}_{\mathbf{s}}(k)$ as the "Coverage Area" of points $\mathbf{x}_k, k \in \mathbf{s}$. As shown in Eq. 3, the upper bound defined by (Sener and Savarese 2017) only considers the furthest points (the inner max operation of Eq. 3) covered by each sample from $\mathbf{s}$, making it a relatively loose upper bound. In contrast, we claim that the Core-set loss can be bounded via a new formulation from the expectation view.

Theorem 2. *With the same assumption as Theorem 1 and definition of Coverage Area in Eq. 3, with probability at least $1 - \gamma$, the Core-set loss is bounded by*

$$\mathcal{L}(\mathbf{s}) \leq \max_{k\in\mathbf{s}} \delta_{\mathbf{s}}(k) \cdot (\lambda^l + \lambda^\eta LC) + \sqrt{\frac{L^2 log(1/\gamma)}{2n}},$$
$$\delta_{\mathbf{s}}(k) = \mathbb{E}_{\mathbf{x}\sim\mathbb{N}_{\mathbf{s}}(k)}[|\mathbf{x} - \mathbf{x}_k|]. \tag{4}$$

Furthermore, it can be easily verified that the upper bound given by Theorem 2 is a tighter approximation compared with the one in Theorem 1.

Theorem 3. *With the same assumption as Theorem 1 and definition of Coverage Area in Eq. 3, the upper-bound of Theorem 2 is smaller than the classical Core-set bound (Sener and Savarese 2017), i.e. $\max_{k\in\mathbf{s}} \delta_{\mathbf{s}}(k) \leq \delta$.*

Theorem 2 shows that the bound is dependent on the maximum value of $\delta_{\mathbf{s}}(k)$, which represents the expected distance between $\mathbf{x}_k$ and other training samples lying in its coverage area $\mathbb{N}_{\mathbf{s}}(k)$. We define this factor $\delta_{\mathbf{s}}(k)$ as the "Average Radial Distance" of $\mathbf{x}_k, k \in \mathbf{s}$. Additionally, Theorem 2 demonstrates that minimizing the maximum average radial distance of labeled set $\mathbf{s}$ is increasingly crucial for reducing the Core-set loss.

Density-aware Core-set with a Proxy Estimator

In this section we introduce how to minimize the upper-bound drawn from Theorem 2 in the context of semantic segmentation. From the definition of average radial distance:

$$\delta_{\mathbf{s}}(k) = \int_{\mathbf{x}} |\mathbf{x} - \mathbf{x}_k| p(\mathbf{x}|\pi(\mathbf{x}) = k) d\mathbf{x}, \tag{5}$$

where $\pi(\mathbf{x}) = \arg\min_{k\in\mathbf{s}} |\mathbf{x} - \mathbf{x}_k|$.

We can observe that $\delta_{\mathbf{s}}(k)$ relies on the conditional probability distribution $p(\mathbf{x}|\pi(\mathbf{x}) = k)$ of samples within the coverage area $\mathbb{N}(\mathbf{x}_k)$ of the chosen sample $\mathbf{x}_k$. We term this distribution as 'coverage sample distribution' of $\mathbf{x}_k$. When samples are uniformly distributed in the feature space, the distributions of all samples are equal. As a consequence, the k-Center Greedy algorithm employed in (Sener and Savarese 2017) effectively optimizes both the upper bounds in Eq. 2 and Eq. 4. Nonetheless, samples in the feature space is seldom evenly distributed (One visual demonstration can be found in the supplementary). Consequently, a selected sample with more eccentric coverage sample distribution exhibit larger average radial distance, thereby it can exert greater influence on the bound of Eq. 4. Therefore, we drew inspiration to devise an optimization algorithm that takes the coverage sample distribution into consideration.

A naive idea is to modify the k-Center Greedy algorithm, i.e. at each step, instead of furthest point sampling, greedily select a sample that minimizes the maximum average radial distance of the newly selected set. However, such solution results in the time complexity of $\mathcal{O}(n^2 b)$, which is almost unacceptable in the large candidate number scenario of semantic segmentation. Therefore, we propose a fast local proxy estimator to estimate a 'coverage density' that reflects the average radial distance for each pixel. Subsequently, this coverage density is utilized in the Density-aware Greedy algorithm (shown in Fig. 2).

Baseline Training and Candidate Selection

In our framework, the images from the source and target domains are separately sent to the Backbone to extract features. The prediction $P \in R^{C\times H_I \times W_I}$ is obtained through a main

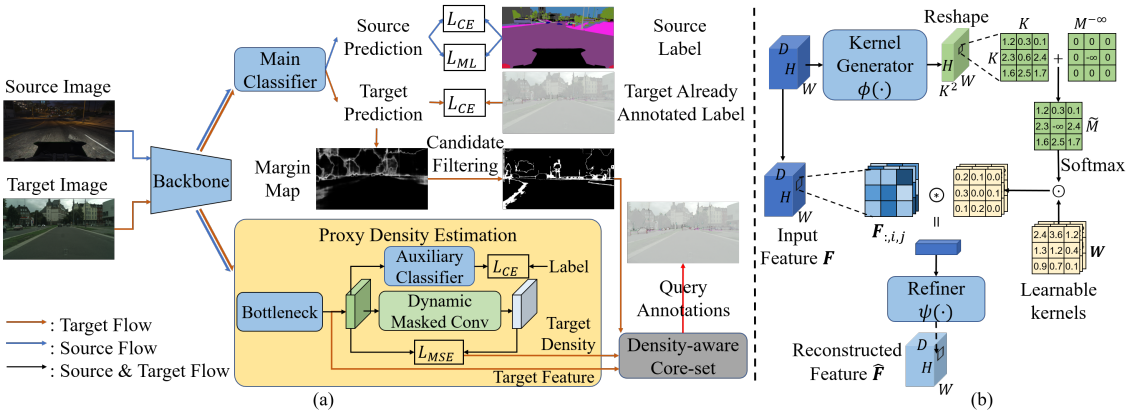

Figure 2: (a) The overview of the proposed method. At each round of active selection, we first select $\alpha b (\alpha > 1)$ pixels closest to the classification boundary as candidate samples based source domain knowledge. We then introduce a density estimation branch to estimate the coverage densities of candidate samples. The features and densities of candidate samples are then fed into our proposed Density-aware Greedy algorithm to perform active selection. Finally, the network will be retrained using all available labels. (b) Structure of our proposed Dynamic Masked Convolution.

classifier, where $H_I \times W_I$ represents the resolution of original image. All labeled data is used to train the backbone and main classifier via cross-entropy.

Source Aided Candidate Filtering. In active semantic segmentation, data is selected at the pixel-level, resulting in extremely large search space for Core-set selection. Therefore, we preselect informative candidate points from the target data. In order to fully utilize the source domain knowledge to filter out informative pixels in the target domain that are distinct from the source domain, we applied the categorical-wise margin loss proposed in (Xie et al. 2022b) to the source data, as shown in Eq. 6.

$$\mathcal{L}_{ML} = \sum_{i} \sum_{j} \sum_{c \neq y} [m - P_{y,i,j} + P_{c,i,j}]_+ \tag{6}$$

where m is set to 1 to control the margin width, y corresponds to the channel index of the ground-truth, and $[x]_+$ denotes $max(0, x)$. Subsequently, we can select informative target samples based on the source knowledge by utilizing $I_{i,j} = 1 - P_{1*,i,j} + P_{2*,i,j}$, where $P_{1*,i,j}$ and $P_{2*,i,j}$ are respectively the maximum and second maximum values within $P_{:,i,j}$. For annotation budget b, we select the top αb $(\alpha > 1)$ pixels with the highest $I_{i,j}$ as candidate samples to reduce complexity. These selected candidates are then resampled through the Density-aware Greedy algorithm based on their estimated density.

Proxy Density Estimator with Dynamic Masked Convolution

Due to the large size and high dimensionality of semantic segmentation candidates, classic distribution estimation methods such as GMM (Reynolds et al. 2009) cannot estimate the coverage sample distribution accurately in an computationally efficient way. Therefore, we propose a proxy estimator for fast estimation. We employ the concept of the Monte Carlo method to estimate a statistic reflecting the local sample distribution of each pixel, termed the 'coverage density'. Given the continuity of images, spatially adjacent pixels tend to be also adjacent in the feature space (Qian et al. 2022). Consequently, the features of neighboring pixels can be considered as samples drawn from the central pixel's coverage sample distribution $p(\mathbf{x}|\pi(\mathbf{x}) = k)$. The process of calculating distances between the central pixel and its neighboring pixels and averaging them is equivalent to applying a Monte Carlo method to approximate Eq. 5. However, such naive estimation can results bias since the local spatial window cannot contain enough samples. To improve this estimation under limited neighboring pixels, we introduce the Dynamic Masked Convolution module (DMC), which aggregates neighboring pixel features to reconstruct the central pixel's feature. Since maximizing the likelihood of a Gaussian distribution is equivalent to minimizing the MSE (Kingma and Welling 2013), our DMC aligns with the use of masked convolution in learned image compression (Minnen, Ballé, and Toderici 2018) to estimate the pixel-wise local conditional distribution.

To enable faster estimation, we first introduce a convolution layer $g(\cdot)$ to convert original feature from backbone to a low-dimensional representation $\mathbf{F} \in R^{D \times H \times W}$ with channel D and spatial size $H \times W$. Next we stack several convolutions $h(\cdot)$ and a softmax operation $\sigma(\cdot)$ as dynamic mask generator $\phi(\cdot)$ to obtain the spatial modulation map $\mathcal{M} = \phi(\mathbf{F}) = \sigma\left(h(\mathbf{F}) + \mathcal{M}^{-\infty}\right) \in R^{K^2 \times H \times W}$, where K is the kernel size for dynamic convolution. Before the softmax σ, we add a mask tensor $\mathcal{M}^{-\infty} \in \{0, -\infty\}^{K^2 \times H \times W}$, where the $\lceil \frac{K^2}{2} \rceil$-th channel is set $-\infty$ to mask out the weight of center coordinates in a $K \times K$ kernel. We then reshape $\mathcal{M}$ as $\tilde{\mathcal{M}} \in R^{K \times K \times H \times W}$, and the reconstructed representation $\tilde{\mathbf{F}}$ is expressed as:

$$\tilde{\mathbf{F}}_{o,i,j} = \sum_{d=1}^{D} \sum_{(u,v) \in \Delta K} \tilde{\mathcal{M}}_{u+\lfloor \frac{K}{2} \rfloor, v+\lfloor \frac{K}{2} \rfloor, i, j} \cdot \mathbf{F}_{d,i+u,j+v} \cdot \mathbf{W}_{o,d,u+\lfloor \frac{K}{2} \rfloor, v+\lfloor \frac{K}{2} \rfloor} \tag{7}$$

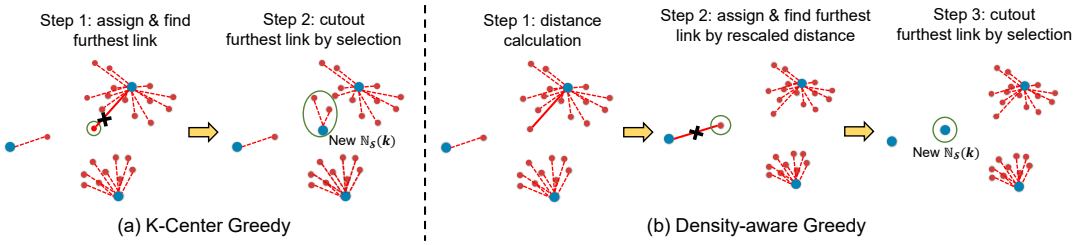

Figure 3: Comparison between k-Center Greedy and Density-aware Greedy. The red points represent candidates to be selected, and the blue points denote selected samples in $\mathbf{s}$. (a) The two-step "Find & Cut" view of k-Center Greedy selection, which cuts the link of furthest distance. (b) Our Density-aware Greedy algorithm rescales the distances so that selected samples with higher density are closer to other candidates and low-density ones are pushed away from linked candidates.

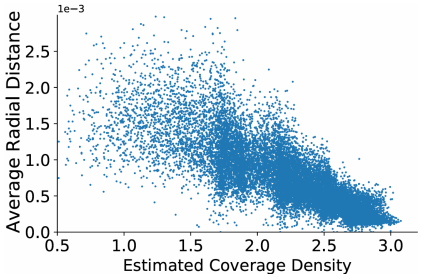

Figure 4: Experimental observation of the relation between estimated density and average radial distance. The distribution of average radial distance tends to approach zero w.r.t increasing coverage density.

where ΔK represents $[-\lfloor \frac{K}{2} \rfloor, \lfloor \frac{K}{2} \rfloor] \times [-\lfloor \frac{K}{2} \rfloor, \lfloor \frac{K}{2} \rfloor]$, and $\mathbf{W} \in R^{D\times D\times K\times K}$ is the learnable parameters of dynamic convolution. Further, additional convolution layers $\psi(\cdot)$ are used to refine the reconstructed results $\hat{\mathbf{F}} = \psi(\tilde{\mathbf{F}})$, where the convolution kernel in $\psi(\cdot)$ is 1×1 to avoid spatial information leakage. Finally, we exploit the reconstruction error to approximate the coverage density at location (i, j) as $\mathbf{D}_{i,j} = \beta \exp\left(-||\hat{\mathbf{F}}_{:,i,j} - \mathbf{F}_{:,i,j}||_2^2/\tau\right)$, where both β, τ are hyperparameters, and we regard the inner reconstruction error $||\hat{\mathbf{F}}_{:,i,j} - \mathbf{F}_{:,i,j}||_2^2$ as estimated average radial distance. We train the estimator to extimate the coverage density for both labeled and unlabeled data:

$$\max_{g,\phi,\psi} \log \prod_{(i,j)} \mathbf{D}_{i,j} \tag{8}$$

Meanwhile, to prevent the optimization in Eq. 8 from having a trivial solution, we append another auxiliary classifier on $\mathbf{F}$ and supervise it with cross-entropy loss using labeled data (as shown in Fig. 2). Before feeding each candidate pixel into the following Density-aware Greedy algorithm, we extract its feature vector $\mathbf{f}$ and density $\mathbf{d}$ from $\mathbf{F}$ and $\mathbf{D}$ at corresponding spatial location respectively.

From Eq. 8, we can observe that the estimated coverage density is negative correlated with the average radial distance. To demonstrate this, in Fig. 4, we analyze the correlation between the estimated density and the average radial distance given the selected set $\mathbf{s}$ from the naive Core-set (Sener and Savarese 2017). The average radial distance of labeled samples with low density is more likely to be larger, while as the density increases, the distribution of the average radial distance tends to approach zero.

Algorithm 1: Density-aware Greedy

Input: candidates $\mathbf{x}_t$, feature $\mathbf{f}_t$ and densities $\mathbf{d}_t$, $t \in [n]$, existing labeled set $\mathbf{s}^0$ and budget b
$\mathbf{s} = \mathbf{s}^0$
$r_t = \min_{k\in\mathbf{s}} ||\mathbf{f}_t - \mathbf{f}_k||_2^2/\mathbf{d}_k \quad \forall t \in [n]$
repeat
  $u = \arg\max_{t\in[n]\setminus\mathbf{s}} r_t$
  $\mathbf{s} = \mathbf{s} \cup \{u\}$
  $r_t = \min(r_t, ||\mathbf{f}_t - \mathbf{f}_u||_2^2/\mathbf{d}_u) \quad \forall t \in [n]\setminus\mathbf{s}$
until $|\mathbf{s}| = b + |\mathbf{s}^0|$
return $\mathbf{s}$

Density-aware Greedy Algorithm

With the estimated density, we propose a density-aware modification of k-Center Greedy algorithm (Wolf 2011) to minimize the Core-set upper bound. To make analogy, we first breakdown the k-Center algorithm into a two-step "Find & Cut" manner as Fig. 3: **STEP1:** link $\mathbf{x}_t$ to its nearest point in $\mathbf{s}$ thus to obtain estimated $\mathbb{N}_\mathbf{s}(k)$ for $\mathbf{x}_k, \forall k \in \mathbf{s}$. **STEP2**: To shrink coverage of $\mathbb{N}_\mathbf{s}(k)$, find and cut the link with longest distance by appending corresponding candidate into $\mathbf{s}$. In contrast, our goal is to downgrade $\delta_\mathbf{s}(k)$ instead of shrinking $\mathbb{N}_\mathbf{s}(k)$, therefore we insert a step between **STEP1** and **STEP2** to **rescale the distance of links in each** $\mathbb{N}_\mathbf{s}(k)$ **by density** $\mathbf{d}_k$ **of corresponding labeled data**. When the coverage density estimated by DMC exhibits a strong correlation with the average radial distance, this inserted step transforms the cut link from furthest link in the feature space into the link associated with the maximum $\delta_\mathbf{s}$. As a result, selected samples with larger $\delta_\mathbf{s}$ are assigned smaller coverage areas and the maximum average radial distance is optimized. In practice, to ensure a robust correlation between the estimated coverage density and the average radial distance, we tune β and τ over the training dataset until they reduce the bound in Eq. 4. Following (Xie et al. 2022a; Wu et al. 2022), we perform multiple rounds of active selection. Algorithm 1 depicts one round of our proposed method.

Method	...	SN	VN	TN	SY	PN	RR	CR	TK	BS	TN	MB	BE	mIOU
Source Only	...	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
DPL-Dual	...	34.0	85.8	41.3	86.0	63.2	34.2	87.2	39.3	44.5	18.7	42.6	43.1	53.3
BAPA-Net	...	55.3	87.8	46.1	89.4	68.8	40.0	90.2	60.4	59.0	0.0	45.1	54.2	57.4
ProDA	...	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
WeakDA (point)	...	51.0	86.1	43.4	87.7	66.4	36.5	87.9	44.1	58.8	23.2	35.6	55.9	56.4
LabOR (V2, 40 pixels)	...	60.6	**89.4**	55.1	**91.4**	70.8	44.7	90.6	56.7	47.9	39.1	47.3	62.7	63.5
RIPU (V2, 40 pixels)	...	55.7	88.5	**55.3**	90.2	69.2	46.1	91.2	70.7	**73.0**	**58.2**	50.1	65.9	65.5
Ours (V2, 40 pixels)	...	**60.7**	89.3	54.9	91.2	**71.0**	**48.7**	**91.6**	**71.9**	71.8	53.9	**55.3**	**68.3**	**66.9**
LabOR (V2, 2.2%)	...	63.5	89.5	57.8	91.6	72.0	47.3	91.7	62.1	61.9	48.9	47.9	65.3	66.6
RIPU (V2, 2.2%)	...	62.2	90.0	57.6	92.6	73.0	53.0	92.8	73.8	78.5	62.0	55.6	70.0	69.6
D2ADA (V2, 5%)	...	64.7	89.3	53.9	92.3	73.9	52.9	91.8	69.7	**78.9**	62.7	**57.7**	**71.1**	69.7
Ours (V2, 2.2%)	...	**67.2**	**90.3**	**58.5**	**92.9**	**74.2**	**55.0**	**92.8**	**75.8**	75.0	**65.3**	54.5	70.4	**71.1**
Fully Supervised (V2)	...	68.0	90.5	58.1	93.1	75.1	53.9	92.7	72.0	80.2	65.0	58.1	71.1	71.3
MADA (V3+, 5%)	...	59.2	89.1	46.7	91.5	73.9	50.1	91.2	60.6	56.9	48.4	51.6	68.7	64.9
RIPU (V3+, 5%)	...	64.1	90.2	59.2	93.2	75.0	54.8	92.7	**73.0**	**79.7**	68.9	55.5	70.3	71.2
D2ADA (V3+, 5%)	...	65.8	90.4	58.9	92.1	75.7	54.4	92.3	69.0	78.0	68.5	**59.1**	**72.3**	71.3
Ours (V3+, 5%)	...	**69.0**	**91.1**	**62.5**	**93.4**	**75.9**	**54.8**	**92.9**	72.5	76.5	**71.3**	54.2	71.2	**72.2**
Fully Supervised (V3+)	...	69.1	91.2	60.5	94.4	76.7	55.6	93.3	75.8	79.9	72.9	57.7	72.2	73.2

Table 1: Comparison[1] with various domain adaptation methods on GTAV $\rightarrow$ Cityscapes.

Experiments

Experimental Setup

Datasets. We evaluate our approach using two popular domain adaptive semantic segmentation benchmarks: GTAV $\rightarrow$ Cityscapes and Synthia $\rightarrow$ Cityscapes. GTAV and SYNTHIA are both synthetic datasets. GTAV shares 19 semantic categories with Cityscapes while SYNTHIA shares 16 semantic categories with Cityscapes.

Implementation Details. To fairly compare with other methods, our training settings and active protocal are aligned with (Xie et al. 2022a). In active selection, β is set to $e^{2.4}$ and τ is set to 0.25, normalizing the reconstruction error to 0-1 before calculating the density. The candidate features $\mathbf{f}$ are also normalized before being fed into Density-aware Greedy. We set α in the candidate filtering to 20 for label budgets lower than 2.2%, and 10 for 5% label budget.

Comparison with State-of-the-Art Methods

We compare our method with various domain adaptation methods[1], as shown in Table 1 and 2. Among them, (Cheng et al. 2021; Liu et al. 2021; Zhang et al. 2021) are UDA methods, (Paul et al. 2020) is a WDA method, while (Shin et al. 2021; Ning et al. 2021; Xie et al. 2022a; Wu et al. 2022) are ADA methods.

It can be observed that: (1) Compared with UDA and WDA methods, our method can achieve more than 10 mIOU improvement. (2) Our method significantly outperforms other active domain adaptation methods. In comparison to RIPU and D2ADA with DeeplabV2, our method showcases a remarkable improvement of over 1.4 mIOU. As the segmentation head improves and the labeling budget increases, the performance gains brought by active strategy somewhat saturates. But our method still achieves better results under the setting of DeeplabV3+ and 5% labeling budget. (3) Our method achieves close performance to fully supervised counterpart with very few annotations. Across various segmentation heads, our method attains 98% of the performance achievable under fully supervised conditions, utilizing just 2.2% to 5% annotations.

[1]Results for each category can be found in the arXiv version.

Ablation Studies

Effect of Dynamic Masked Convolution and Density-aware Greedy. we conduct ablation experiments on the GTAV $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes tasks with a 2.2% label budget to explore the impact of our proposed Dynamic Masked Convolution (DMC) and Core-set selection, as shown in Table 3. The 'Baseline' method uses only the entropy of the model output for active selection. The 'K-Center Greedy' first filters candidate samples using entropy and then applies K-Center Greedy algorithm described in (Sener and Savarese 2017). When combined with DMC, the K-Center Greedy only utilizes DMC as a feature regularization technique. For our Density-aware Greedy, removing DMC involves using the context model from (Minnen, Ballé, and Toderici 2018) to estimate density.

From the results in Table 3, the following observations can be made: (1) Our proposed Density-aware Greedy algorithm demonstrates robust performance improvements compared to K-Center Greedy. Even when DMC is replaced with a 5×5 masked convolution, we still achieve remarkable mIOU improvements. This highlights the significance of coverage density. (2) The proposed DMC is more suitable for estimating density compared to the context model in learned image

Method	...	PE*	LT	SN	VN	SY	PN	RR	CR	BS	MB	BE	mIOU	mIOU*
Source Only	...	31.4	7.0	27.7	63.1	67.6	42.2	19.9	73.1	15.3	10.5	38.9	34.9	40.3
DPL-Dual	...	33.2	22.0	20.1	83.1	86.0	56.6	21.9	83.1	40.3	29.8	45.7	47.0	54.2
BAPA-Net	...	34.9	30.5	42.8	86.6	88.2	66.0	34.1	86.6	51.3	29.4	50.5	53.3	61.2
ProDA	...	44.0	54.6	37.0	88.1	84.4	74.2	24.3	88.2	51.1	40.5	45.6	55.5	62.0
WeakDA (point)	...	34.9	37.3	50.8	84.4	88.2	60.6	36.3	86.4	43.2	36.5	61.3	57.2	63.7
RIPU (V2, 40 pixels)	...	38.3	47.1	54.2	89.2	90.8	69.9	**48.5**	91.4	71.5	**52.2**	67.2	66.1	72.1
Ours (V2, 40 pixels)	...	**41.9**	**50.6**	**60.9**	**89.9**	**91.8**	**71.7**	46.5	**92.1**	**76.2**	47.1	**68.0**	**67.9**	**73.5**
RIPU (V2, 2.2%)	...	45.0	53.0	62.5	90.6	92.7	73.0	52.9	93.1	80.5	52.4	70.1	70.1	75.7
D2ADA (V2, 5%)	...	**48.0**	55.6	66.5	89.5	91.7	**75.1**	**55.2**	91.4	77.0	**58.0**	**71.8**	70.6	76.3
Ours (V2, 2.2%)	...	47.7	**56.5**	**68.0**	**91.2**	**93.0**	74.8	52.2	**93.4**	**83.5**	54.6	70.7	**72.1**	**77.3**
Fully Supervised (V2)	...	49.0	57.5	68.0	90.5	93.1	75.1	53.9	92.7	80.2	58.1	71.1	72.5	77.5
MADA (V3+, 5%)	...	46.7	52.4	60.5	89.7	92.2	74.1	51.2	90.9	60.3	52.4	69.4	68.1	73.3
RIPU (V3+, 5%)	...	48.5	55.2	63.9	91.1	93.0	74.4	54.1	92.9	79.9	55.3	71.0	71.4	76.7
D2ADA (V3+, 5%)	...	54.2	58.3	68.0	90.4	93.4	**77.4**	**56.4**	92.5	77.5	**58.9**	**73.3**	72.7	77.7
Ours (V3+, 5%)	...	**55.1**	**59.1**	**70.0**	**91.9**	**93.8**	77.3	54.4	**93.9**	**80.3**	56.4	71.9	**73.2**	**78.5**
Fully Supervised (V3+)	...	53.8	59.6	69.1	91.2	94.4	76.7	55.6	93.3	79.9	57.7	72.2	73.8	78.4

Table 2: Comparison with various domain adaptation methods on SYNTHIA $\rightarrow$ Cityscapes.

Sampling	DMC	GTAV	SYNTHIA
Baseline	✗	66.2	68.2
K-Center Greedy	✗	69.4	70.7
	✓	70.1	71.1
Density-aware Greedy	✗	70.4	71.4
	✓	**71.1**	**72.1**

Table 3: Ablation Studies of Each Component.

Algorithm	δ	$max_{k\in\mathbf{s}}\delta_{\mathbf{s}}(k)$	**Core-set Loss**
K-Center	**0.132**	0.364	0.646
Density-aware	0.176	**0.124**	**0.550**

Table 4: Bound and Core-set Loss comparison.

Method	GTAV	SYNTHIA
RAND	63.8	65.6
ReDAL	66.2	67.2
BADGE	66.1	67.1
ENT	66.2	68.2
SCONF	66.5	68.4
MARGIN	66.1	68.0
Ours (w/o source data)	69.1	70.0
Ours	**71.1**	**72.1**

Table 5: Comparison with Active Baselines.

compression. (3) Introducing DMC solely as a feature regularization technique also improves model performance.

Moreover, we conducted numerical experiments on the Cityscapes training set to validate that the proposed Density-aware Greedy algorithm reduces the new bound introduced in Eq. 4. We compared the bounds δ from Eq. 3 and $max_{k\in\mathbf{s}}\delta_{\mathbf{s}}(k)$ from Eq. 4 for models trained with annotations selected using K-center Greedy and Density-aware Greedy. The results are shown in Table 4. It is observed that the K-center Greedy algorithm results in a smaller δ, yet a larger $max_{k\in\mathbf{s}}\delta_{\mathbf{s}}(k)$. Conversely, our proposed Density-aware Greedy algorithm results in a larger δ, while yielding a smaller $max_{k\in\mathbf{s}}\delta_{\mathbf{s}}(k)$. The actual Core-set Loss of the Density-aware Greedy algorithm is also smaller.

Comparison with Common Active Learning Baselines. We also compared our method with other active learning methods, including random (RAND), uncertainty-based methods ENT (Shen et al. 2017), SCONF (Culotta and McCallum 2005) and MARGIN (Wang and Shang 2014), and hybrid methods ReDAL (Wu et al. 2021) and BADGE (Ash et al. 2019). The label budget is set to 2.2%. It can be seen from Table 5 that even without any source domain data, our method still outperforms commonly used active learning strategies by a margin over 2 mIOU, demonstrating the competitiveness of our method as an active learning approach.

Conclusion

In this paper, we propose a Density-aware Core-set Selection method for active domain adaptive segmentation. We derive a tighter upper bound for the classical Core-set and identify that the model performance is closely related to the coverage sample distribution of selected samples. Further, we introduce a Proxy Density Estimator and develop a Density-aware Greedy algorithm to optimize the newly derived bound. Experiments demonstrate that the proposed method outperforms existing active learning and domain adaptation methods.

Acknowledgements

The paper is supported in part by the National Natural Science Foundation of China (No. 62325109, U21B2013, 61971277).

References

Ash, J.; Goel, S.; Krishnamurthy, A.; and Kakade, S. 2021. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems*, 34: 8927–8939.

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Ayush, K.; Jandial, S.; Chopra, A.; and Krishnamurthy, B. 2019. Powering virtual try-on via auxiliary human segmentation learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0–0.

Chang, W.-L.; Wang, H.-P.; Peng, W.-H.; and Chiu, W.-C. 2019. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1900–1909.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818.

Chen, S.; Jia, X.; He, J.; Shi, Y.; and Liu, J. 2021. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11018–11027.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875.

Cheng, Y.; Wei, F.; Bao, J.; Chen, D.; Wen, F.; and Zhang, W. 2021. Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 9082–9091.

Culotta, A.; and McCallum, A. 2005. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, 746–751.

Guan, L.; and Yuan, X. 2023. Iterative Loop Method Combining Active and Semi-Supervised Learning for Domain Adaptive Semantic Segmentation. *arXiv*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *European Conference on Computer Vision*, 372–391. Springer.

Hoyer, L.; Dai, D.; Wang, H.; and Van Gool, L. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11721–11732.

Jiang, Z.; Li, Y.; Yang, C.; Gao, P.; Wang, Y.; Tai, Y.; and Wang, C. 2022. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, 36–54. Springer.

Kim, Y.; and Shin, B. 2022. In Defense of Core-set: A Density-aware Core-set Selection for Active Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 804–812.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in Neural Information Processing Systems*, 32.

Liu, Y.; Deng, J.; Gao, X.; Li, W.; and Duan, L. 2021. Bapanet: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 8801–8811.

Minnen, D.; Ballé, J.; and Toderici, G. D. 2018. Joint autoregressive and hierarchical priors for learned image compression. *Advances in Neural Information Processing Systems*, 31.

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 International Conference on Computer Vision*, 9112–9122.

Paul, S.; Tsai, Y.-H.; Schulter, S.; Roy-Chowdhury, A. K.; and Chandraker, M. 2020. Domain adaptive semantic segmentation using weak labels. In *Proceedings of the European Conference on Computer Vision*, 571–587. Springer.

Prabhu, V.; Chandrasekaran, A.; Saenko, K.; and Hoffman, J. 2021. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, 8505–8514.

Qian, Y.; Lin, M.; Sun, X.; Tan, Z.; and Jin, R. 2022. Entroformer: A transformer-based entropy model for learned image compression. *arXiv preprint arXiv:2202.05492*.

Reynolds, D. A.; et al. 2009. Gaussian mixture models. *Encyclopedia of Biometrics*, 741(659-663).

Sener, O.; and Savarese, S. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Shen, Y.; Yun, H.; Lipton, Z. C.; Kronrod, Y.; and Anandkumar, A. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Shi, F.; Wang, J.; Shi, J.; Wu, Z.; Wang, Q.; Tang, Z.; He, K.; Shi, Y.; and Shen, D. 2020. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Reviews in Biomedical Engineering*, 14: 4–15.

Shin, I.; Kim, D.-J.; Cho, J. W.; Woo, S.; Park, K.; and Kweon, I. S. 2021. Labor: Labeling only if required for domain adaptive semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 8588–8598.

Teichmann, M.; Weber, M.; Zoellner, M.; Cipolla, R.; and Urtasun, R. 2018. Multinet: Real-time joint semantic reasoning for autonomous driving. In *IEEE Intelligent Vehicles Symposium*, 1013–1020. IEEE.

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

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin, L. 2016. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12): 2591–2600.

Wang, Y.; Peng, J.; and Zhang, Z. 2021. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 9092–9101.

Wolf, G. W. 2011. Facility location: concepts, models, algorithms and case studies.

Wu, T.-H.; Liou, Y.-S.; Yuan, S.-J.; Lee, H.-Y.; Chen, T.-I.; Huang, K.-C.; and Hsu, W. H. 2022. D 2 ADA: Dynamic Density-Aware Active Domain Adaptation for Semantic Segmentation. In *Proceedings of the European Conference on Computer Vision*, 449–467. Springer.

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*, 15510–15519.

Xie, B.; Yuan, L.; Li, S.; Liu, C. H.; and Cheng, X. 2022a. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8068–8078.

Xie, M.; Li, Y.; Wang, Y.; Luo, Z.; Gan, Z.; Sun, Z.; Chi, M.; Wang, C.; and Wang, P. 2022b. Learning distinctive margin toward active domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7993–8002.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12414–12424.

Zheng, Z.; and Yang, Y. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4): 1106–1120.